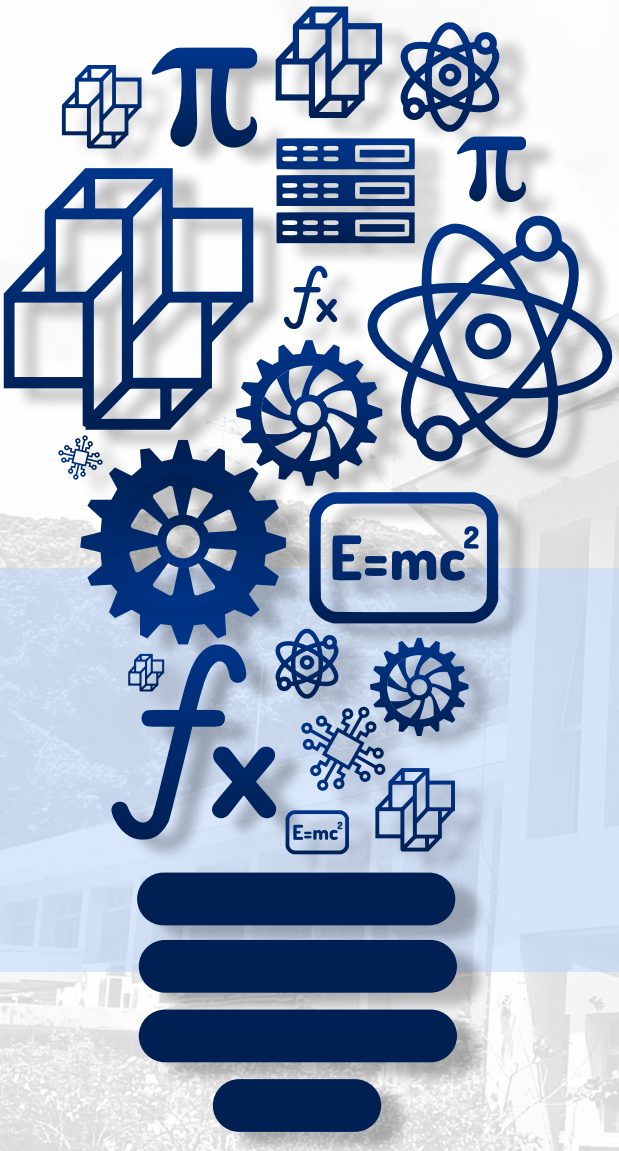
Jornada de Iniciação Científica e Tecnológica

2025



PIBIC · PIBITI

LNCC/MCTI

28 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, 28 de agosto de 2025.

Laboratório Nacional de Computação Científica – LNCC

Diretor
Fabio Borges de Oliveira

Coordenação de Gestão e Administração - COGEA
Paulo César de Freitas Honorato

Coordenação de Pós-Graduação e Aperfeiçoamento - COPGA
Antônio André Novotny

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

Leandro Tavares da Silva – CEFET
Priscila Vanessa Zabala Capriles Goliatt UFJF

Apresentação

O LNCC realiza este ano a XXII 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 2024 a agosto de 2025, o PIBIC e o 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, neste ano presencialmente, os trabalhos são apresentados pelos bolsistas em forma de pôster e avaliados por um comitê científico externo.

Nesta XXII Edição da Jornada, o Comitê Externo de Avaliação do PIBIC/PIBITI tem a seguinte composição:

Prof.ª Priscila Vanessa Zabala Capriles Goliatt - UFJF
Prof. Leandro Tavares da Silva - CEFET

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 de 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 da dedicação de todos os participantes.

Agradecimentos

Agradecemos ao CNPq pelas bolsas concedidas, à Direção e ao SECIN do LNCC pelo apoio, ao Programa de Pós-Graduação do LNCC e à Comissão Interna do PIBIC e PIBITI do 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 e, em particular, à Sra. Roberta Machado.

José Karam Filho
Coordenador do PIBIC/PIBITI - LNCC

Índice

Bolsistas PIBITI ativos

Bolsistas PIBIC ativos

Estudo e implementação de sistema de banco de dados para análise em Saúde Coletiva.
Bolsista: Fernanda Xabudé Moreira Bomfilioli.
Orientadores: Paulo Cabral e José Karam

Estudo e implementação de sistemas de banco de dados para análise me Saúde Coletiva
Bolsista: Gabriel Eduardo Pontes Amaral
Orientadores: José Karam Filho e Paulo Cabral

Uso do padrão de paralelismo de linguagem em benchmarks Científicos.
Bolsista: Gabriel Thomaz do Nascimento
Orientadores: Roberto Pinto Souto e Eduardo Lucio Mendes Garcia

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 e Eduardo Lucio Mendes Garcia

Predição de Eficiência Energética e Desempenho em Aplicações HPC com Aprendizado de Máquina.
Bolsista: Isabella da Silva Muniz
Orientadores: Carla Osthoff Ferreira de Barros e Micaella Coelho Valente de Paula

Análise da Colaboração Científica entre pesquisadores em uma instituição de educação profissional e tecnológica.
Bolsista: Jefferson Pablo Nunes Santos.
Orientadores: José Karam Filho e José Damião de Melo.

Simulação de eventos climáticos na plataforma de Petróleo P40.
Bolsista: Jonatas Halliday Sant Anna Nascimento
Orientador: Jauvane C. de Oliveira

Análise de Desempenho do Programa PA-Star2 no Santos Dumont e sua Aplicação em um Workflow Científico para Alinhamento Múltiplo de Sequências Selecionadas.
Bolsista: Kelen Brito Souza
Orientadores: Kary Ann del Carmen Ocaña Gautherot e Micaella Coelho Valente de Paula

Reestruturação e Otimização de Código Científico para Simulação de Escoamento em Reservatórios.
Bolsista: Luiza Augusto Tavares
Orientadores: Carla Osthoff Ferreira de Barros e Thiago Daniel Quimas Simões Teixeira.

Desenvolvimento de uma ARP para monitoramento de áreas: estação solo, veículo com piloto automático, transmissão de dados e telemetria.
Bolsista: Marcos Paulo de Souza Campanha
Orientadores: Jauvane Cavalcante de Oliveira e Paulo Fernando Ferreira Rosa

Análise de Desempenho com Intel VTune Profiler.
Bolsista: Mariana Aguiar Ribeiro
Orientadores: Carla Osthoff e Thiago Teixeira

Desenho de Substâncias Ativas Usando Abordagens de novo e Ancoramento Molecular.
Bolsista: Pedro Lucas Ferreira Cruz da Mota Mendes
Orientadores: Laurent E. Dardenne e Isabella Alvim Guedes e Matheus M. P. da Silva

HeMoLab1D Web v1.0.
Bolsista: Peter Zeidler
Orientador: Pablo Javier Blanco e Paulo Ziemer

Comparação do Desempenho Computacional de Workflows Científicos de Transcriptômica em Arquiteturas HPC.
Bolsista: Reiglan Soares Di Lourenço
Orientadores: Kary Ann del Carmen Ocaña Gautherot e Carla Osthoff Ferreira de Barros

Introdução à Programação com Scratch e Tutoriais
Bolsista: Ricardo dos Santos de Lima
Orientadora: Regina Célia Cerqueira de Almeida

Construção de um Conjunto de Dados Envolvendo Modelos Tridimensionais e Dados Experimentais de Afinidade Proteína-Ligante.
Bolsista: Shirlei Militão da Silva
Orientadores: Laurent E. Dardenne e José Renato Duarte Fajardo

Título do projeto: MASA-OpenMP e sua aplicação em um *Workflow* Científico para Alinhamento Múltiplo de Sequências Biológicas

Nome do bolsista: Alessandra Almeida de Souza Lima

Nome do orientador: Carla Osthoff

Nome do Coorientador: Micaella Coelho

Tipo de bolsa e período do relatório: PIBIT

Objetivos

Este relatório tem como objetivo apresentar sumariamente o trabalho realizado com o software MASA (Multiple Algorithms for Sequence Alignment), com foco na avaliação de seu desempenho computacional em diferentes configurações de execução. Além disso, destaca-se que o MASA será destinado à futura integração em um *workflow* científico ainda em desenvolvimento, o qual visa a realização eficiente de alinhamento múltiplo de sequências em ambientes de Computação de Alto Desempenho (CAD).

1. Introdução

A bioinformática moderna depende fortemente de ferramentas computacionais para processar e analisar grandes volumes de dados biológicos. Dentre as diversas aplicações existentes, as análises evolutivas, como a reconstrução filogenética, a detecção de seleção positiva e a epidemiologia molecular, desempenham papel importante na compreensão de mecanismos como progressão de doenças, escape imunológico e resistência a fármacos [1]. Para tais análises, o Alinhamento de Sequências (AS) constitui uma etapa crítica [2].

A execução eficiente de algoritmos de AS é especialmente desafiadora frente ao crescimento exponencial dos dados genômicos, impulsionado por tecnologias de sequenciamento de alto desempenho. Nesses cenários, torna-se fundamental explorar o paralelismo e utilizar ambientes de Computação de Alto Desempenho (CAD) para garantir a viabilidade computacional das análises.

Este relatório tem como foco a ferramenta MASA (Multiple Algorithms for Sequence Alignment) [3], que implementa algoritmos exatos para alinhamento par-a-par, com suporte a paralelismo via OpenMP e técnicas de otimização como o Block Pruning. Avaliamos seu desempenho computacional em diferentes configurações de execução, tanto em experimentos controlados com arquivos do banco BAliBASE quanto em um cenário aplicado com dados reais do vírus da dengue (DENV).

2. Fundamentação Teórica

2.1 Aplicação MASA

O MASA (Multiple Algorithms for Sequence Alignment) é uma ferramenta que implementa algoritmos exatos para o alinhamento par-a-par de sequências biológicas. Especificamente, a ferramenta incorpora os algoritmos de Needleman-Wunsch, voltado para o alinhamento global, e Smith-Waterman, voltado para o alinhamento local. Ambos os algoritmos são baseados em programação dinâmica (DP).

Uma das principais características do MASA é sua implementação da técnica de Block Pruning. Essa técnica otimiza o uso da memória durante o processo de alinhamento, permitindo que a ferramenta reconstrua o alinhamento com complexidade de espaço linear em relação ao tamanho das sequências, reduzindo assim significativamente o consumo de memória. O Block Pruning funciona dividindo a matriz de DP em blocos regulares (por exemplo, blocos de 1024x1024 células). Cada bloco é então avaliado com base em um limite inferior de pontuação possível. Se este limite indicar que nenhum alinhamento ótimo pode passar pelo bloco, ele é descartado, evitando cálculos desnecessários e melhorando a eficiência do processo.

Além disso, o MASA suporta paralelismo por meio do OpenMP, permitindo executar alinhamentos em paralelo através de múltiplas threads. Cada thread é responsável por computar um ou mais blocos viáveis, maximizando a utilização dos recursos computacionais disponíveis e acelerando significativamente o tempo de execução dos alinhamentos.

3. Metodologia

3.1 Descrição dos Experimentos e Dados de Entrada

Esta pesquisa está dividida em dois experimentos complementares, ambos com foco na avaliação do desempenho computacional da ferramenta MASA, em diferentes contextos de aplicação.

O Experimento 1 foi conduzido com o objetivo de analisar o comportamento do MASA de forma isolada e controlada, utilizando arquivos do banco de dados BAliBASE. Foram selecionados quatro arquivos de alinhamento de proteínas com menor complexidade computacional, aqui denominados arquivos iniciais: glg, 1sbp, 1aboA e 1ac5. Cada um contém 5 sequências de aminoácidos em formato FASTA. Esses dados permitiram observar a variação do desempenho do MASA conforme o número de threads e avaliar seu comportamento sob diferentes graus de paralelismo.

O Experimento 2 teve como foco avaliar o MASA, aplicado ao caso de estudo com sequências genéticas do vírus da dengue (DENV). Foi utilizado um conjunto de arquivos denominado dataset D1, contendo genes do vírus DENV em quatro versões, com 9, 18, 28 e 38 sequências.

3.2 Ambiente computacional

Os testes foram realizados no supercomputador Santos Dumont, especificamente nos nós computacionais do tipo Sequana X1120. Cada nó é equipado com 2 CPUs Intel Xeon Cascade Lake Gold 6252, totalizando 48 núcleos físicos, 384 GB de memória RAM, sistema de arquivos distribuído Lustre e interconexão InfiniBand EDR de 100 Gb/s.

3.3 Avaliação de Desempenho Computacional do MASA

No Experimento 1, cada um dos quatro arquivos do BAliBASE foi processado com o MASA, variando-se o número de threads utilizadas: 1, 12, 24 e 48. Essa variação permitiu avaliar a escalabilidade e o impacto do paralelismo no tempo de execução do MASA, em um cenário controlado.

No Experimento 2, o dataset D1 foi processado com o MASA utilizando dois níveis de paralelismo: 1 e 48 threads. Foram consideradas as quatro versões do D1 (com 9, 18, 28 e 38 sequências) para verificar o comportamento da ferramenta em volumes de dados crescentes e mais próximos de um cenário real.

Para ambas as etapas, cada configuração foi executada cinco vezes de forma independente, com o intuito de mitigar variações pontuais no sistema e obter estimativas mais confiáveis. O tempo de execução foi medido utilizando a função omp_get_wtime() da API OpenMP.

4. Resultados e Discussão

4.1 Análise do desempenho do MASA: Experimento 1

A Tabela 1 apresenta a média do tempo de execução do MASA para os quatro arquivos selecionados do banco BAliBASE. Como cada arquivo contém apenas cinco sequências, o volume de dados é relativamente baixo, resultando em tempos de execução curtos, entre menos de 1 segundo e até 8 segundos, dependendo da configuração. As execuções apresentaram baixa variabilidade, compatível com a simplicidade das sequências analisadas.

Observa-se que, para o arquivo glg.fasta, o tempo de execução com uma única thread foi inferior ao tempo obtido com múltiplas threads, sugerindo que o *overhead* de paralelização pode superar os ganhos em cenários com volume de dados reduzido. Para os demais arquivos, os ganhos de desempenho com paralelismo foram pequenos ou inexistentes, reforçando que, em alinhamentos de pequena escala, o uso de múltiplas threads pode não trazer benefícios significativos.

	glg.fasta	1sbp.fasta	1aboA.fasta	1ac5.fasta
Threads	**Tempo (s)**			
48	2.39	4.60	3.01	4.00
24	1.34	1.60	3.00	4.00
16	1.00	2.53	6.00	8.00
12	1.89	2.36	4.00	3.00
1	0.98	4.61	4.00	4.00

Tabela 1. Tempo de execução (em segundos) com diferentes números de threads para arquivos do BAliBASE

4.2 Análise do desempenho do MASA: Experimento 2

A Tabela 2 apresenta os resultados do desempenho do MASA no processamento do *dataset* D1, composto por quatro versões com diferentes quantidades de sequências (9, 18, 28 e 38). Em cada versão, todos os pares possíveis de alinhamentos foram processados tanto em modo sequencial quanto em paralelo, utilizando 1 e 48 threads, respectivamente.

Versão do Dataset D1	Quantidade de Pares	Tempo Médio (s) (um par)		Tempo Total (s) (todos os pares)	
		Sequencial	Paralelo	Sequencial	Paralelo
9 seqs	36	0,07	0,23	2,55	8,25
18 seqs	153	0,06	0,09	9,13	14,44
28 seqs	378	0,06	0,11	21,26	40,97
38 seqs	703	0,07	0,11	45,78	74,00

Tabela 2. Desempenho do MASA em execução Sequencial e paralela para diferentes Versões do *dataset* D1

Na primeira linha da tabela, observa-se que, ao utilizar um conjunto com 9 sequências, foram gerados 36 pares para alinhamento. No modo sequencial, o tempo médio por alinhamento foi de 0,07 segundos, resultando em um tempo total de 2,55 segundos. Já no modo paralelo, o tempo médio por par foi de 0,23 segundos, com tempo total de 8,25 segundos, valor superior ao sequencial, indicando que, para volumes pequenos de dados, o *overhead* do paralelismo pode prejudicar o desempenho.

À medida que o número de sequências aumenta, a quantidade de pares cresce substancialmente, chegando a 703 pares na versão com 38 sequências. Nessa versão, o tempo médio sequencial por par se manteve em 0,07 segundos, enquanto o paralelo foi de 0,11 segundos. Apesar de o tempo médio por par continuar relativamente constante, o tempo total de execução aumenta proporcionalmente à quantidade de pares: 45,78 segundos no modo sequencial e 74,00 segundos no paralelo.

Esses resultados indicam que, embora o paralelismo não tenha trazido benefícios de desempenho em *datasets* pequenos, ele se torna competitivo à medida que o volume de dados cresce, especialmente quando há maior paralelismo de pares simultâneos. Esse cenário é compatível com a estratégia do *workflow* científico em desenvolvimento, que integrará o MASA como etapa de pré-processamento e permitirá a execução simultânea de múltiplos alinhamentos par-a-par, potencializando os ganhos de desempenho.

5. Conclusões

Os experimentos realizados demonstraram que a ferramenta MASA apresenta desempenho eficiente para alinhamento par-a-par de sequências biológicas, especialmente em contextos com menor complexidade computacional, como observado nos arquivos do banco BAliBASE. A avaliação em cenários com maior volume de dados, como no caso do *dataset* D1 com sequências do vírus da dengue, indicou que o tempo médio por par se mantém estável, mas o tempo total cresce proporcionalmente à quantidade de pares.

A análise evidenciou que o paralelismo via OpenMP pode trazer benefícios, especialmente quando há possibilidade de execução simultânea de múltiplos alinhamentos. No entanto, em conjuntos pequenos, o *overhead* da paralelização pode superar seus ganhos, tornando a versão sequencial mais eficiente. Tais observações reforçam a importância de ajustar a granularidade do paralelismo de acordo com o volume de dados.

Embora este estudo tenha se concentrado na avaliação do MASA de forma isolada, os resultados obtidos são essenciais para subsidiar sua integração em um *workflow* científico automatizado atualmente em desenvolvimento. Esse *workflow* permitirá a execução paralela de múltiplos alinhamentos par-a-par, aproveitando de forma mais eficaz os recursos de CAD.

Como trabalho futuro, pretende-se expandir os experimentos com conjuntos maiores e mais complexos de dados biológicos, além de validar o desempenho do MASA quando integrado ao *workflow* completo, avaliando seu impacto em escalabilidade, eficiência e reprodutibilidade em ambientes CAD.

6. Referências

[1] Pollett, S., Melendrez, M., Maljkovic Berry, I., Duchene, S., Salje, H., Cummings, D., and Jarman, R. (2018). Understanding dengue virus evolution to support epidemic surveillance and counter-measure development. Infection, Genetics and Evolution, 62:279–295.

[2] Schabauer, H., Valle, M., Pacher, C., Stockinger, H., Stamatakis, A., Robinson-Rechavi, M., Yang, Z., and Salamin, N. (2012). SlimCodeML: An Optimized Version of CodeML for the Branch-Site Model. In 2012 IEEE 26th International

Parallel and Distributed Processing Symposium Workshops & PhD Forum, pages 706–714. IEEE.

[3] De O. Sandes, E. F., Miranda, G., Martorell, X., Ayguade, E., Teodoro, G., and De Melo, A. C. (2016). Masa: A multiplatform architecture for sequence aligners with block pruning. ACM Transactions on Parallel Computing (TOPC), 2(4):1–31.

Prevenção de Incidentes de Segurança no Ciberespaço, com a utilização de técnica de rastreamento, através de Fluxos de Pacotes

Aluno: Gabriel Martins Ramos - FAETERJ

Orientadora: Marita Maestrelli

coorientador: Antonio Tadeu

Objetivos:

Este trabalho tem como objetivo monitorar, analisar e diagnosticar o tráfego de rede utilizando as ferramentas **Netflow**, **Nfdump** e **NfSen**. Busca-se obter visibilidade detalhada dos fluxos de dados, identificar anomalias, melhorar a segurança da rede e auxiliar no planejamento da capacidade da infraestrutura.

Introdução:

Netflow é uma tecnologia criada pela Cisco que permite coletar estatísticas detalhadas sobre tráfego IP. Junto com ferramentas como **Nfdump** (análise em linha de comando) e **NfSen** (interface gráfica), é possível obter uma visão completa do comportamento da rede, identificar gargalos e detectar atividades suspeitas.

Materiais e Métodos:

Infraestrutura → servidor Linux com Nfdump/NfSen e ambiente real de rede para captura de tráfego.

Configurações:

- Habilitação do Netflow nas interfaces de roteadores/switches.
- Instalação e integração do Nfdump e NfSen.
- Coleta de dados via nfcapd e análise com NfSen.
- Análise periódica dos dados para identificação de padrões e anomalias.

Resultados e Discussão:

Visibilidade → Relatórios detalhados sobre IPs, protocolos, volume e horários de tráfego.

Segurança → Detecção de comportamentos anômalos e possíveis ataques de rede.

Diagnóstico → Identificação de pontos de congestionamento e quedas de desempenho.

Planejamento → Base sólida para dimensionamento da rede e melhorias preventivas.

A integração entre Netflow, Nfdump e NfSen se mostrou eficaz, oferecendo recursos visuais, filtragem avançada, geração de gráficos e suporte a plugins personalizados.

Automatizações Desenvolvidas:

Com o ambiente de monitoramento plenamente funcional, surgiu a necessidade de automatizar processos que até então dependiam de intervenção manual. Pensando nisso, foram desenvolvidas soluções em Python e perl que aumentaram a eficiência operacional, reduziram falhas e tornaram o monitoramento mais confiável e inteligente.

nfsen_telegram.py:

Este script identifica o arquivo nfcapd mais recente, gera relatórios personalizados com o comando nfdump, formata os dados em texto estruturado e converte tudo em imagem. Em seguida, o relatório é enviado automaticamente para um grupo no Telegram.

Além disso, o bot aceita comandos via chat como /status, /relatorio e /help, tornando a ferramenta interativa e acessível remotamente. Tudo isso ocorre de forma assíncrona e em segundo plano, sem impactar a coleta de dados.

***Impacto**:* facilita o monitoramento em tempo real, acessível mesmo por celular, com resposta rápida e automatizada.

backupnfsen.py:

Esse script percorre os diretórios dos perfis do NfSen, identifica os arquivos antigos e os compacta em **.tar.gz** , organizando por perfil e por data. Os arquivos são então movidos para um diretório de backup, e os originais são excluídos, liberando espaço no sistema.

***Impacto**:* garante segurança histórica dos dados, reduz riscos de perda e mantém a infraestrutura enxuta e organizada.

auto_nfsen.py:

Foram identificadas quedas intermitentes no serviço NfSen. Para resolver isso, este script permite agendar reinicializações periódicas, usando uma interface gráfica leve e intuitiva. É especialmente útil em ambientes onde a manutenção precisa ser constante, mas automatizada.

***Impacto**:* evita interrupções na coleta de dados e melhora a confiabilidade do sistema.

plugin_alerta.pl:

Além dos scripts autônomos, foi desenvolvido um **plugin personalizado para o NfSen**, integrado diretamente à sua estrutura de plugins. Este recurso permite:

- Enviar relatórios **imediatos** ao Telegram quando **eventos críticos** ocorrem (como picos de tráfego, ataques suspeitos ou falhas).
- Realizar **análises sob demanda** de fluxos específicos conforme parâmetros configurados pelo operador.
- Reportar **falhas internas**, erros de execução ou indisponibilidade de dados diretamente ao responsável técnico, sem a necessidade de supervisão constante.

O plugin foi projetado para ser facilmente adaptado às necessidades da rede e pode ser ampliado para novas regras de detecção no futuro.

***Impacto**:* monitoramento inteligente e proativo, notificações sob demanda e resposta imediata a eventos anormais.

Planos Futuros:

O próximo passo será o desenvolvimento de um módulo com Inteligência Artificial (IA) para:

- Detectar padrões avançados e anomalias em tempo real.
- Antecipar incidentes de rede.
- Automatizar respostas a eventos críticos.

Essa abordagem tornará o sistema proativo, inteligente e capaz de se adaptar às mudanças na rede, elevando o nível de segurança e performance da infraestrutura.

Referências:

1. Cisco Netflow – Site Cisco Netflow
2. Nfdump – Site Nfdump
3. NfSen – Site NfSen

Desenvolvimento de Ferramentas Computacionais Educativas de Química Computacional Aplicada ao Planejamento de Fármacos

Bolsista: Patrick Ferreira da Silva Costa
Supervisores: Isabella Alvim Guedes, Laurent Emmanuel Dardenne
Período da bolsa: 10/03/2025 a 31/08/2025

Introdução

A sub-representação de mulheres nos campos de STEM (Ciência, Tecnologia, Engenharia e Matemática) é um desafio persistente que afeta a diversidade e a inovação em diversas áreas científicas e tecnológicas. Apesar dos avanços recentes, as mulheres continuam enfrentando barreiras culturais e sociais que limitam sua participação nessas disciplinas. Diante desse cenário, é crucial desenvolver estratégias que incentivem e apoiem a participação feminina em STEM, especialmente durante a educação básica, quando os interesses profissionais começam a se formar (Tonini; Araújo, 2019). Este projeto propõe a criação de um jogo educativo voltado para meninas, com o objetivo de auxiliar no aprendizado de conceitos de bioquímica e biologia molecular. O jogo "Caminho dos Átomos" é uma ferramenta educativa desenvolvida no Scratch, voltada para meninas do ensino médio, com o objetivo de ensinar conceitos de bioquímica e biologia molecular de forma interativa. O protagonista é um alienígena que precisa percorrer diversos cenários para retornar ao seu planeta natal. Jogos educacionais desempenham um papel fundamental no estímulo ao interesse por STEM, pois combinam elementos lúdicos e interativos com conhecimento científico, tornando a aprendizagem mais acessível e atraente (LOPES;, 2023). A plataforma Scratch, desenvolvida pelo MIT, foi escolhida para a criação do jogo devido à sua abordagem visual e intuitiva, ideal para iniciantes, especialmente crianças e jovens. Além disso, o projeto utiliza bibliotecas de química computacional, como BioPython e RDKit, para desenvolver aplicações didáticas que simplificam conceitos complexos, tornando-os mais acessíveis a estudantes do ensino fundamental e médio.

Objetivos

Este trabalho tem como propósito central o desenvolvimento de um jogo educacional voltado para o ensino de modelagem molecular aplicada ao planejamento de fármacos, com o objetivo de democratizar o acesso a esse conhecimento na educação básica e estimular o interesse de meninas por carreiras em STEM. Para concretizar essa proposta, o projeto se estrutura em três objetivos principais:

1. Implementar aplicações didáticas que façam uso de bibliotecas especializadas em química computacional, como BioPython e RDKit, permitindo a tradução de conceitos complexos em linguagem acessível e a geração de questões científicas.
2. Desenvolver o jogo educativo "Caminho dos Átomos" na plataforma Scratch, uma iniciativa que combina elementos lúdicos e interativos com conteúdos de bioquímica e biologia molecular, criando uma experiência de aprendizagem interessante para as estudantes. Este recurso pedagógico foi especialmente concebido para despertar o interesse pelo tema de uma forma lúdica e interativa.

3. Aplicação dos conceitos e do jogo desenvolvidos no projeto em iniciativas como Petrópolis Tech Hub - Meninas em STEM e Futuras Cientistas, visando ensinar conceitos de programação e de química computacional de forma lúdica.

Metodologia

O desenvolvimento do jogo "Caminho dos Átomos" foi realizado utilizando a linguagem Scratch, escolhida por sua abordagem visual baseada em blocos que simplifica o aprendizado de programação para iniciantes. O jogo foi projetado para apresentar perguntas de diferentes temas (sendo o foco deste projeto a química computacional), criando uma experiência educativa interativa e lúdica. Visando aproximar as alunas também à programação em Python e a bibliotecas de química computacional, integramos ferramentas de modelagem molecular, como a biblioteca BioPython para manipulação de sequências de DNA e proteínas, e o RDKit, para gerar estruturas moleculares em 2D e calcular descritores físico-químicos simples.

O processo de validação do jogo envolverá testes abrangentes com dois grupos distintos: membros do Grupo de Modelagem Molecular em Sistemas Biológicos (GMMSB), que avaliarão aspectos técnicos e funcionais do jogo e alunas participantes dos programas Petrópolis Tech Hub - Meninas STEM e Futuras Cientistas, que testarão a adequação pedagógica e o apelo lúdico da ferramenta. Essa fase de testes vai ajudar a verificar se as mecânicas do jogo estão funcionando corretamente. Também será possível avaliar se as perguntas estão com um nível de dificuldade adequado, se a parte educativa está sendo eficiente e se a narrativa e os elementos interativos realmente prendem a atenção dos jogadores. Com isso, será possível garantir que o jogo cumpra o objetivo de divulgar a ciência e incentivar o interesse de alunas de escolas públicas pelas áreas de STEM.

Ferramentas de modelagem molecular

BioPython

O BioPython é uma biblioteca gratuita em Python projetada para facilitar a análise de macromoléculas biológicas, como proteínas e ácidos nucleicos (DNA e RNA). Ela oferece funcionalidades como acesso a bancos de dados científicos, análise de estruturas de proteínas e manipulação de sequências de DNA e RNA. No contexto deste projeto, o BioPython foi utilizado para automatizar a criação de perguntas sobre biologia computacional, auxiliando na elaboração de perguntas científicas cientificamente válidas (Chapman; Chang, 2000).

RDKit

O RDKit é uma ferramenta de código aberto para química computacional, amplamente utilizada em áreas como a descoberta de fármacos. Suas funcionalidades incluem a manipulação de estruturas moleculares, o cálculo de propriedades físico-químicas e a geração de representações 2D e 3D de moléculas (The RDKit Documentation — The RDKit 2025.03.3 documentation, [s. d.]). No projeto, o RDKit foi empregado para criar perguntas sobre estrutura molecular, grupos funcionais e propriedades de substâncias orgânicas, contribuindo com o jogo através de conteúdo visual (estruturas 1D, 2D e 3D de pequenas moléculas).

Implementação do jogo educacional “Caminho dos Átomos”

Criando o jogo com o Scratch

Para começar, é necessário criar um cenário inicial e adicionar o personagem principal (o alienígena) como um sprite. Em seguida, as instruções de programação são implementadas através de “blocos de movimento”, que indicam o que o personagem principal deve fazer de acordo com a ação recebida da jogadora. Utilizando blocos de movimento, como "mova __ passos" e "aponte para a direção __", o alienígena pode se deslocar entre os cenários. Cada fase do jogo é representada por um fundo diferente, que pode ser trocado com o bloco "mude o cenário para __" quando a jogadora acerta uma pergunta. As perguntas são exibidas usando blocos de "pergunte __ e espere", armazenando a resposta em uma variável. Se a resposta estiver correta, o jogo avança para o próximo cenário com o bloco "próximo cenário"; caso contrário, uma mensagem de retorno é exibida usando "diga __ por __ segundos". Para tornar o jogo mais dinâmico, blocos de controle como "se __ então" e "repita até __" são usados para verificar respostas e gerenciar a progressão. Blocos de "toque o som __" podem ser adicionados para indicar acertos ou erros. Além disso, variáveis como "pontuação" e "vidas" são criadas usando blocos de dados, permitindo acompanhar o desempenho da jogadora. O bloco "adicione __ a [pontuação]" incrementa a pontuação a cada acerto, enquanto "mude [vidas] por __" controla tentativas restantes. O jogo termina quando o personagem principal chega ao portal dimensional que o leva para seu planeta de origem, exibindo uma mensagem de parabéns com "diga __" e a pontuação total. Para reiniciar, basta usar "quando a bandeira for clicada" para redefinir variáveis e cenários. Com esses blocos, o "Caminho dos Átomos" se torna uma ferramenta educativa eficaz, combinando programação criativa e conteúdo científico.

Utilização do Biopython no projeto

No contexto deste projeto, a utilização do Biopython possibilitou a automação na criação de perguntas sobre bioquímica, garantindo precisão científica e relevância pedagógica. Essa abordagem foi fundamental para elaborar questões tecnicamente corretas e alinhadas com os objetivos educacionais do jogo. Para isso, diversos módulos do Biopython foram empregados, cada um contribuindo com funcionalidades específicas. Os módulos Bio.Seq e Bio.SeqUtils foram essenciais para manipular sequências de DNA, RNA e proteínas, permitindo a criação de perguntas sobre processos biológicos como transcrição e tradução, além de questões relacionadas à estrutura primária de biomoléculas e suas propriedades físico-químicas. Por exemplo, esses módulos possibilitaram gerar perguntas como "Qual é a sequência de RNA transcrita a partir deste segmento de DNA?". Já o módulo Bio.PDB facilitou a análise de estruturas tridimensionais de proteínas, viabilizando a elaboração de questões mais complexas sobre ligações peptídicas, domínios estruturais e interações moleculares. Com ele, foi possível criar perguntas relacionadas à estrutura tridimensional de peptídeos e proteínas, como "Identifique os domínios estruturais nesta proteína" ou "Quantas ligações peptídicas existem nesta cadeia polipeptídica?". Por fim, os módulos Bio.KEGG e Bio.Entrez foram utilizados para acessar bancos de dados científicos, como o KEGG (que contém informações sobre vias metabólicas). Esses módulos permitiram extrair dados atualizados para formular perguntas sobre vias bioquímicas e processos metabólicos, como "Qual enzima catalisa esta reação na via glicolítica?". Dessa forma, a integração desses módulos do Biopython não apenas automatizou a geração de perguntas, mas também enriqueceu o jogo com conteúdo científico em diferentes

níveis, demonstrando na prática como a programação pode ser aplicada ao ensino de química computacional.

Utilização do RDKit no projeto

RDKit é um conjunto de ferramentas de código aberto para química computacional, escrito principalmente em C++ com uma camada de interface em Python. Ele oferece uma ampla gama de funcionalidades para manipulação, análise e modelagem de moléculas. É frequentemente utilizado em diversas áreas, desde o desenho de novas moléculas promissoras como fármacos até a ciência dos materiais, devido à sua flexibilidade e eficiência (The RDKit Documentation — The RDKit 2025.03.3 documentation, [s. d.]). Entre as principais funcionalidades do RDKit, destacam-se a capacidade de ler e escrever diversos formatos de arquivos de pequenas moléculas (como SMILES, MOL2, SDF), gerar descritores moleculares, realizar buscas por subestrutura, calcular propriedades físico-químicas e até mesmo gerar representações 2D e 3D das estruturas moleculares. Por ser uma ferramenta de código aberto e contar com uma comunidade ativa de desenvolvedores, o RDKit está em constante evolução, incorporando novas funcionalidades e melhorias de desempenho. Sua interface em Python facilita a integração com outras bibliotecas populares, como NumPy, SciPy e scikit-learn, tornando-o uma escolha poderosa para a comunidade científica (The RDKit Documentation — The RDKit 2025.03.3 documentation, [s. d.]).

No contexto deste projeto, o RDKit foi usado para a geração automatizada de perguntas sobre química computacional, particularmente em tópicos relacionados à estrutura e propriedades de moléculas orgânicas, fármacos e biomoléculas. Através de seu módulo básico, foi possível visualizar e manipular estruturas químicas em diversos formatos como SMILES, MOL2 ou SDF, permitindo criar questões como "Qual destas estruturas representa a molécula de glicose?", apresentando alternativas tanto em formato SMILES quanto em imagens 2D geradas pelo próprio RDKit. Além disso, a biblioteca possui capacidades avançadas para cálculo de descritores moleculares, incluindo propriedades como massa molar e contagem de átomos de hidrogênio, o que viabilizou elaborar perguntas como "Qual destes compostos possui mais átomos de hidrogênio?". Para aspectos mais avançados da Química, o RDKit oferece recursos de visualização 3D. Essas funcionalidades permitem criar perguntas que envolvem o reconhecimento de propriedades espaciais das moléculas, como "Quantos centros quirais existem nesta molécula?". É importante destacar que a visualização de moléculas em 3D é imprescindível para a compreensão de conceitos básicos de Química Orgnânica, como geometria molecular. Entretanto, muitas vezes estes conteúdos são ensinados nas escolas sem recursos gráficos adequados. Neste contexto, inserir a visualização molecular em 2D e 3D como parte do jogo é importante para auxiliar na compreensão destes conceitos por parte das alunas.

Validação e Testes do Jogo “Caminho dos Átomos”

Como resultado deste projeto, foi elaborado um tutorial educativo visando realizar oficinas para ensinar os usuários a implementar o jogo. O processo de validação do jogo educativo foi estruturado em diferentes fases para assegurar tanto sua qualidade técnica quanto seu valor pedagógico. A primeira etapa consistirá em uma validação interna conduzida pelo GMMSB/LNCC, onde pesquisadores e alunos realizarão testes para verificar o pleno funcionamento da proposta de elaboração do jogo, a compatibilidade com diversos dispositivos e a ausência de erros técnicos. Esta fase será fundamental para garantir a qualidade do jogo

antes de sua exposição ao público-alvo. Em seguida, o jogo será submetido à avaliação das alunas participantes dos programas Meninas STEM e Futuras Cientistas. Durante esta fase, as alunas participarão de uma oficina em que será ensinada a programação do jogo com Scratch, sendo avaliados aspectos como o nível de dificuldade das questões, o grau de engajamento proporcionado pelo ambiente gráfico do jogo e a percepção sobre sua utilidade como ferramenta de aprendizagem. Essas impressões permitirão a realização dos ajustes necessários tanto a estética quanto a funcionalidade do jogo. A etapa seguinte do processo de validação será uma validação final pelas coordenadoras pedagógicas dos programas Meninas STEM e Futuras Cientistas, que avaliarão a adequação do jogo aos objetivos educacionais dos programas e ao potencial da ferramenta para despertar o interesse das alunas de escolas públicas pelas áreas de STEM.

Resultados

Durante a execução deste projeto, foi feita a revisão bibliográfica sobre conceitos básicos de modelagem molecular, ferramentas de química computacional (BioPython e RDKit) e linguagem Scratch. O jogo "Caminho dos Átomos" está em fase final de implementação, juntamente com o tutorial que ensina as alunas a implementar o jogo com a linguagem Scratch.

Assim que finalizado, o jogo será testado com seu público-alvo. Os resultados preliminares sugerem que ele pode ser uma ferramenta valiosa para o ensino de química computacional e programação, aliando conteúdo científico a uma abordagem interativa e engajadora. Como próximos passos de aprimoramento do projeto, pretende-se integrar um banco de dados com perguntas e respostas para selecionar aleatoriamente as questões a serem exibidas no jogo. A atualização trará mais variedade e profundidade às questões, melhorando a experiência do usuário. Por fim, pretendemos publicar os resultados do projeto em eventos de divulgação científica e ministrando aulas com conteúdos adaptados para o objetivo do evento. Desta forma, será possível incentivar alunos e professores a adotarem o "Caminho dos Átomos" como recurso pedagógico lúdico e interativo, despertando maior interesse dos alunos.

Conclusão

Este projeto evidencia o potencial das ferramentas computacionais educativas como aliadas no ensino de STEM, com foco especial em alunas do ensimo da rede pública. Ao unir jogos interativos a conteúdos científicos, a iniciativa apresenta uma abordagem atrativa capaz de tornar o aprendizado mais acessível. O projeto não apenas facilita a compreensão de temas científicos, mas também ajuda a despertar o interesse e a confiança das alunas em conceitos de programação. A utilização de bibliotecas como BioPython e RDKit garante que o conteúdo apresentado nos jogos seja preciso e atualizado, reforçando a credibilidade do jogo como recurso educacional. Por fim, este projeto serve como recurso educacional relacionado ao aprendizado de programação e química computacional com aplicação prática imediata, por exemplo, ao serem ministradas oficinas que ensinam a implementar o jogo com a linguagem Scratch ou ser utilizado diretamente em aulas de diversas disciplinas, visto que a elaboração das perguntas e respostas é personalizável..

Referências bibliográficas

CHAPMAN, Brad; CHANG, Jeffrey. Biopython: Python tools for computational biology. **ACM SIGBIO Newsletter**, [*s. l.*], v. 20, n. 2, p. 15–19, ago. 2000. https://doi.org/10.1145/360262.360268.

COCK, Peter J. A.; ANTAO, Tiago; CHANG, Jeffrey T.; CHAPMAN, Brad A.; COX, Cymon J.; DALKE, Andrew; FRIEDBERG, Iddo; HAMELRYCK, Thomas; KAUFF, Frank; WILCZYNSKI, Bartek; DE HOON, Michiel J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, [*s. l.*], v. 25, n. 11, p. 1422–1423, 1 jun. 2009. https://doi.org/10.1093/bioinformatics/btp163.

COCK, Peter J. A.; FIELDS, Christopher J.; GOTO, Naohisa; HEUER, Michael L.; RICE, Peter M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. **Nucleic Acids Research**, [*s. l.*], v. 38, n. 6, p. 1767–1771, abr. 2010. https://doi.org/10.1093/nar/gkp1137.

DE HOON, M.J.L.; IMOTO, S.; NOLAN, J.; MIYANO, S. Open source clustering software. **Bioinformatics**, [*s. l.*], v. 20, n. 9, p. 1453–1454, 12 jun. 2004. https://doi.org/10.1093/bioinformatics/bth078.

LOPES;, LARA M. Conhecendo mulheres cientistas a partir de jogos. **Women in Information Technology (WIT 2023)**, [*s. l.*], 2023. .

PRITCHARD, Leighton; WHITE, Jennifer A.; BIRCH, Paul R.J.; TOTH, Ian K. GenomeDiagram: a python package for the visualization of large-scale genomic data. **Bioinformatics**, [*s. l.*], v. 22, n. 5, p. 616–617, 1 mar. 2006. https://doi.org/10.1093/bioinformatics/btk021.

RASCHKA, Sebastian. BioPandas: Working with molecular structures in pandas DataFrames. **The Journal of Open Source Software**, [*s. l.*], v. 2, n. 14, p. 279, 7 jun. 2017. https://doi.org/10.21105/joss.00279.

SARDELA, Antônio. **Curso Completo de Química: Volume Único.** 3. ed. São Paulo: Ática, 2002. v. Volume Único., .

TALEVICH, Eric; INVERGO, Brandon M; COCK, Peter Ja; CHAPMAN, Brad A. Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. **BMC Bioinformatics**, [*s. l.*], v. 13, n. 1, p. 209, dez. 2012. https://doi.org/10.1186/1471-2105-13-209.

THE RDKIT DOCUMENTATION — THE RDKIT 2025.03.3 DOCUMENTATION. [*s. d.*]. Disponível em: https://www.rdkit.org/docs/index.html. Acesso em: 31 jul. 2025.

TONINI, Adriana Maria; ARAÚJO, Mariana Tonini. A PARTICIPAÇÃO DAS MULHERES NAS ÁREAS DE STEM (SCIENCE, TECHNOLOGY ENGINEERING AND MATHEMATICS). **Revista de Ensino de Engenharia**, [*s. l.*], v. v. 38, n. n. 3, p. 118–125, 2019. .

VERLI, Hugo. **Bioinformática : da biologia à flexibilidade molecular**. [*S. l.*]: Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014. Disponível em: https://lume.ufrgs.br/handle/10183/166105. Acesso em: 22 out. 2024.

LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA – LNCC/MCTI

Petrópolis – RJ – Brasil

Projeto de bolsa de iniciação tecnológica

Maio a julho de 2025

ADAN EM MOVIMENTO:

Simulações hemodinâmicas em posições anatomicamente realistas

Bolsista: Tatiane Casemira Ferreira Pink

Orientador: Pablo Javier Blanco

Petrópolis

2025

OBJETIVOS

O objetivo principal deste projeto é, a partir da configuração anatômica do modelo ADAN, desenvolver uma metodologia que permita criar cenários que reproduzam posições anatomicamente realistas do cotidiano humano.

O trabalho busca, portanto, elaborar modelos computacionais e protocolos de simulação capazes de representar condições fisiológicas e patofisiológicas, contribuindo para a pesquisa e o desenvolvimento de soluções aplicadas a problemas médicos.

De forma específica, destacam-se os seguintes objetivos:

1. Tornar o modelo ADAN executável e adequado para processos de modelagem.
2. Criar diferentes posições no modelo, possibilitando a simulação de posturas comuns do dia a dia.

INTRODUÇÃO

Segundo Blanco (2025), as doenças cardiovasculares (DCVs) constituem a principal causa de mortalidade no mundo, destacando-se a doença arterial coronariana e o acidente vascular cerebral. A Organização Mundial da Saúde estima cerca de 18 milhões de mortes anuais, podendo alcançar 23,6 milhões em 2030. No Brasil, a situação é igualmente preocupante, com aproximadamente uma morte a cada 90 segundos.

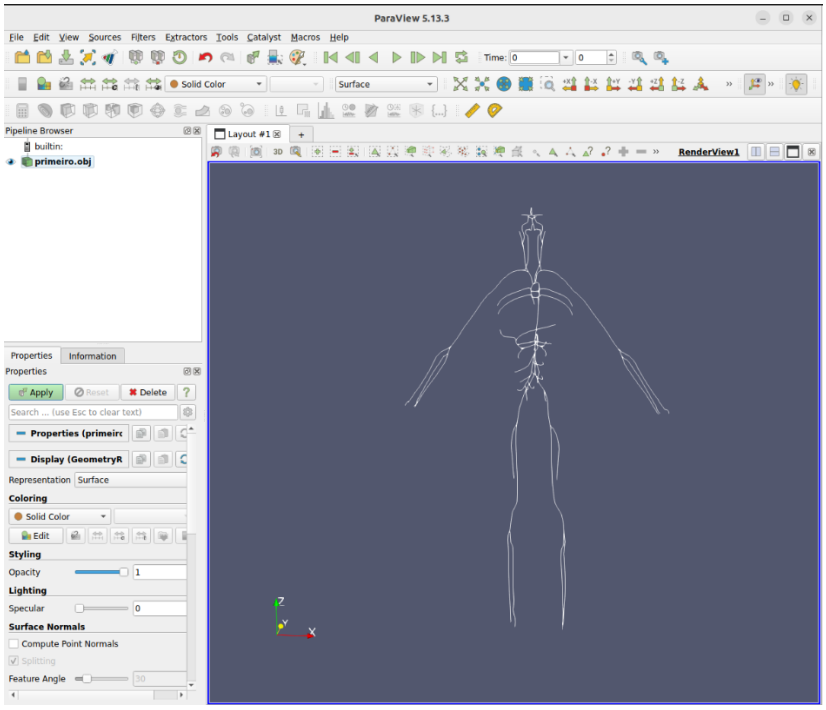
Técnicas de simulação computacional têm contribuído para o entendimento do sistema cardiovascular e introduzido avanços relevantes na prática médica. O aumento do poder computacional permitiu a evolução de modelos que variam desde os tridimensionais, altamente detalhados, mas de custo elevado, até os unidimensionais, que equilibram precisão e eficiência. Esses modelos possibilitam análises em diversos cenários fisiológicos e patológicos, sendo amplamente validados por estudos experimentais e clínicos, o que reforça sua importância para a medicina.

MATERIAL E MÉTODOS OU METODOLOGIA.

Foram utilizadas duas ferramentas principais no desenvolvimento das atividades:

- **Blender**: software utilizado para criar, manipular e animar objetos 3D, além de realizar renderizações. É amplamente aplicado em áreas como animação, jogos, efeitos visuais e design.
- **ParaView**: aplicação multiplataforma de código aberto voltada para a visualização científica interativa.

No presente trabalho, o ParaView foi empregado para a visualização do modelo ADAN e para a conversão do arquivo para o formato (**.**obj), tornando-o compatível com o Blender. Em seguida, o Blender foi utilizado para animar o modelo, atribuindo-lhe movimentos que simulam ações humanas básicas.



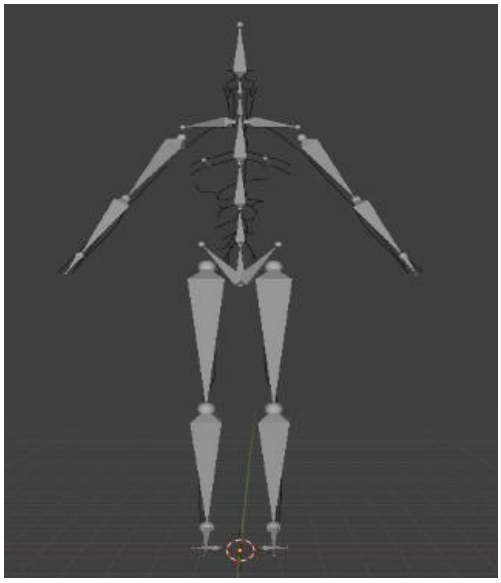
1 - Visualização do modelo original no ParaView.

RESULTADOS E DISCUSSÃO

O modelo anatômico foi inicialmente convertido para um formato compatível com o Blender(.obj), possibilitando sua manipulação e animação. Após a importação do

modelo no Blender, procedeu-se à inclusão de um esqueleto (armature), utilizando o atalho Shift + A > Armature > Human.

Com o esqueleto adicionado, foi necessário ajustar sua escala e posição para coincidir com a malha do modelo ADAN. Essa etapa foi feita no modo de edição (Edit Mode), permitindo o alinhamento preciso entre os ossos da armature e os segmentos anatômicos do modelo.

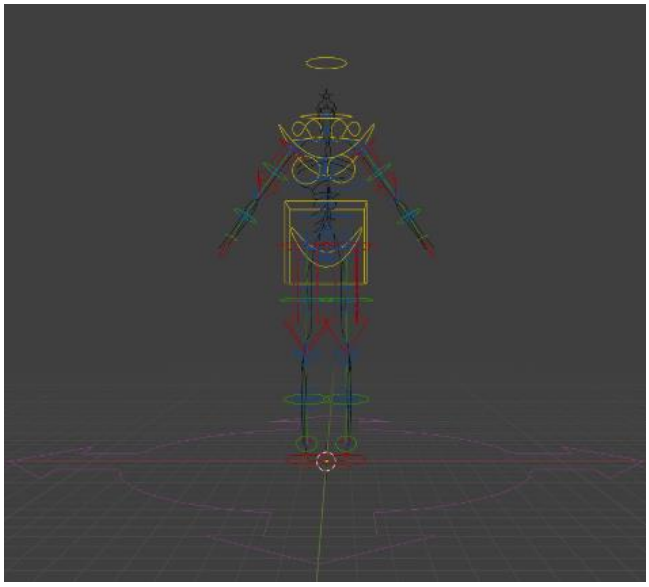


2 - Modelo ajustado

Após os ajustes, utilizou-se o método Ctrl + P > With Automatic Weights para realizar o processo de parenting entre a armature e a malha, possibilitando a deformação da malha com base nos movimentos dos ossos. Com o modelo devidamente associado à armature, foi possível alternar para o Pose Mode e aplicar movimentos às diferentes partes do corpo, usando comandos como “R” (rotação) e “G” (translação). Isso permitiu simular movimentos rígidos simples.

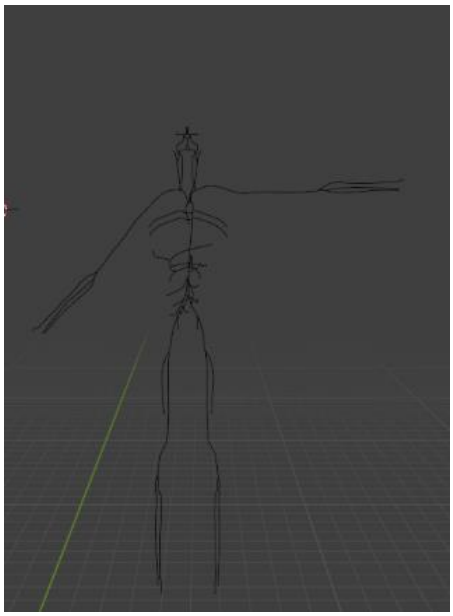
Posteriormente, foi explorada uma segunda abordagem: a criação de poses utilizando o Rigify. Para isso, foi selecionado o armature, acessando o painel Object Data Properties (ícone do boneco verde), e clicou-se em Generate Rig O rig gerado foi ajustado ao tamanho do modelo e o metarig original foi ocultado. Em seguida, aplicou-se novamente o Ctrl + P > With Automatic Weights, desta vez com a malha associada ao rig. No Pose Mode, foi possível movimentar partes específicas do modelo com

maior precisão e controle, permitindo simular posturas humanas do cotidiano com maior realismo, como levantar o braço.



2 - Rigging com o armature humano simples.

Os resultados demonstram que, com o Rigify, foi possível criar poses mais complexas e realistas, simulando posturas cotidianas como levantar o braço. Os resultados demonstraram que a metodologia viabiliza a execução de cenários anatômicos realistas, contribuindo para futuras aplicações em simulações fisiológicas, como a análise da pressão arterial em condições dinâmicas.



4- Pose inicial do modelo animado (um braço levantado).

CONCLUSÕES

O trabalho realizado demonstrou a viabilidade de transformar o modelo anatômico ADAN em um modelo animável, compatível com ferramentas de simulação e visualização como o Blender. A aplicação de rigging, tanto com armature simples quanto com Rigify, possibilitou a criação de poses realistas que reproduzem posturas humanas comuns.

A metodologia adotada mostrou-se eficiente e pode ser expandida para incluir movimentos mais complexos e simulações fisiológicas, tornando-se uma ferramenta promissora para aplicações médicas, educacionais e de pesquisa em biomecânica e fisiologia computacional.

REFERÊNCIAS BIBLIOGRÁFICAS

BLANCO, Pablo. Projeto de pedido de bolsa de iniciação tecnológica. Petrópolis: Laboratório Nacional de Computação Científica (LNCC), 2025. Documento interno.

BLENDER FOUNDATION. Blender 5.0 Reference Manual. Disponível em: <https://docs.blender.org/manual/pt/dev/#>. Acesso em: 29 jul. 2025.

RELATÓRIO DE PROJETO DE INICIAÇÃO CIENTÍFICA

BIOINFORMÁTICA, BANCO DE DADOS E ENGENHARIA DE COMPUTAÇÃO

Título do Projeto Proposto

Análise Computacional de uma Ferramenta de Bioinformática em Arquiteturas de Memória Compartilhada do Santos Dumont

Instituição

Laboratório Nacional de Computação Científica

Nome do Aluno

Albert Siqueira Cosme Emidio

Nome do Orientador

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

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

Tipo de bolsa: PIBIC

Período do relatório: 28/06/2024 - 28/06/2025

1. Objetivo

Este trabalho tem como objetivo analisar o uso dos recursos computacionais, com ênfase na utilização da CPU, durante a execução da atividade Bowtie2 no *workflow* transcriptômico ParslRNA-Seq. A análise ajuda a identificar gargalos de desempenho e avaliar a eficiência da aplicação em um ambiente de computação de alto desempenho (CAD), utilizando o perfilador Intel VTune como ferramenta de monitoramento e diagnóstico.

2. Introdução

Os experimentos em bioinformática abordam uma grande quantidade de dados complexos, desde sequências genômicas até interações moleculares. Devido a essa complexidade, são necessárias soluções eficientes para o tratamento desses dados, o que envolve não apenas recursos computacionais, como a computação de alto desempenho (CAD), mas também a modelagem de *workflows*. [8].

O *workflow* de transcriptômica ParslRNA-Seq foi utilizado para a execução, gerência e análise computacional. ParslRNA-Seq foi modelado na análise de experimentos transcriptômica, na expressão diferencial de genes (EDG), com a linguagem de programação Python e a biblioteca Parsl em um ambiente de computação de alto desempenho (CAD), o que facilitou a integração e automação do *workflow*. A principal atividade executada nesse experimento foi o Bowtie2, escolhido para fazer análise de desempenho devido ao seu elevado custo de CPU e suporte a *multithreading*. As demais atividades do *workflow* incluem: Sort que ordena as leitura dos genes; Split que divide os arquivos de entrada; HTSeq que conta as leituras; Merge, para indexar as contagens; e DESeq, para a análise estatística das EDGs. [7]

Bowtie2 é a primeira atividade do *workflow* e um *software* de alinhamento eficiente utilizado para mapear sequências de DNA curtas em grandes genomas de referência. Ele usa métodos baseados em indexação e alinhamento para garantir eficiência em termos de tempo e memória. O perfilador

Intel VTune é uma ferramenta de análise de desempenho que detalha o uso da CPU, utilizado para otimizar *software* em sistemas Intel. O objetivo da pesquisa é analisar o uso dos recursos computacionais na execução da atividade Bowtie2 já que ele é a atividade mais custosa do *workflow*, assim, observando a utilização da CPU pelo perfilador.

3. Background

3.1 ParslRNA-Seq: *Workflow* Científico de Transcriptômica

O *workflow* ParslRNA-Seq é implementado utilizando o Parsl, uma ferramenta em Python desenvolvida para executar workflows em ambientes de CAD. O Parsl simplifica a criação de workflows permitindo a integração de comandos externos diretamente no código Python por meio de decoradores (Apps), o que facilita a sincronização das atividades.

Esse *workflow* é composto por seis etapas principais e utiliza como entrada o genoma de referência do Mus musculus, arquivos GTF com metadados genômicos e arquivos de sequenciamento em formato FASTQ. Um arquivo no formato CSV associa os arquivos FASTQ às respectivas condições experimentais, sendo três amostras de controle e três da condição Wnt, relacionada à via de sinalização metabólica Wnt.

As aplicações do *workflow* são Bowtie2 que 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 Figura 1 apresenta o modelo conceitual do ParslRNA-Seq. Análises de EDG são essenciais na análise de dados transcriptômicos, permitindo a identificação de genes com expressão significativamente diferente entre condições experimentais.

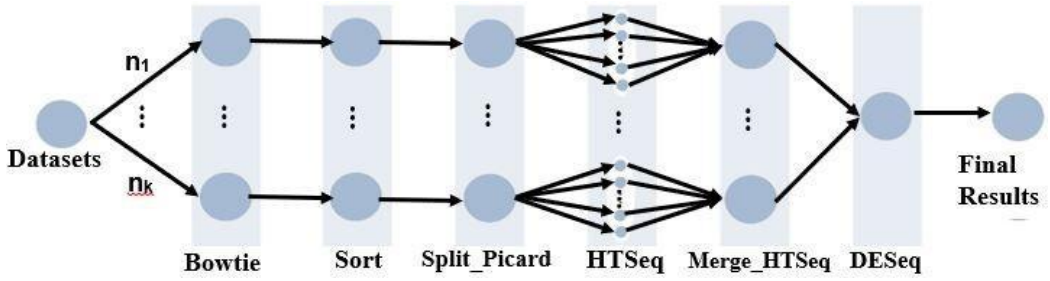


Figura 1. Modelagem Conceitual do *Workflow* Científico ParslRNA-Seq. Adaptada de Silva *et al.* 2021.

3.2 Parsl - Parallel Scripting Library

ParslRNA-Seq é implementado utilizando Parsl, uma ferramenta em Python projetada para execução de *workflows* em ambientes de CAD [10]. Parsl[1] É uma biblioteca de programação paralela desenvolvida em Python que utiliza decoradores para executar funções e software externo como aplicativos Python e Bash. A arquitetura do Parsl é baseada em programação orientada a dados, uma tarefa só é executada quando todas as suas entradas estão disponíveis. Isso permite o gerenciamento dinâmico da execução e o uso eficiente de recursos computacionais, sejam locais, em clusters HPC ou

[1] https://parsl-project.org/

em nuvens. O Parsl oferece suporte nativo para diferentes mecanismos de execução (executors), como ThreadPool, High Throughput Executor (HTEX) e Slurm. O motor de execução do Parsl é flexível, suportando diversos ambientes computacionais e abstraindo as complexidades do *workflow*, tornando mais simples sua implementação e integração com recursos computacionais.

O *workflow* utiliza dois modelos de execução: ThreadPoolExecutor e HTEx. O ThreadPoolExecutor é um modelo que suporta multithreading em recursos locais, gerenciando um pool de threads para executar atividades de forma concorrente. A eficiência desse modelo é essencial para maximizar o desempenho computacional, reduzir erros e melhorar a produtividade. Por outro lado, o modelo HTEx permite a execução do workflow em múltiplos nós computacionais simultaneamente, facilitando o compartilhamento de dados entre as máquinas e proporcionando um controle refinado sobre a alocação de recursos e o paralelismo necessário para uma execução eficiente das tarefas.

4. Metodologia

4.1 Dados do Experimento

Os dados coletados para a análise são de um experimento real de RNA-Seq, extraídos do repositório público Gene Expression Omnibus, divididos em grupo de controle (SRR5445794, SRR5445795, SRR5445796) e grupo de condições das vias metabólica Wnt (SRR5445797, SRR5445798, SRR5445799), sendo o organismo em evidência é o *Mus musculus* e o GEO.ID GSE97763. Os dados de entrada apresentam tamanho entre 1.8GB e 3.0GB, ao todo sendo 13GB, a saída do Bowtie2 também ficou em 13GB . O *workflow* foi executado com os dados de entrada e a atividade Bowtie2, foi submetida ao perfilador para analisar os recursos computacionais utilizados já que ela é a atividade que mais demanda uso da CPU

O SDumont 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. Todos os *software*, algoritmos, dependências de bioinformática (Bowtie, Samtools, Picard, HTSeq e DESeq2) e os componentes do Parsl foram alocados e instalados no ambiente do SDumont.

As execuções foram realizadas em:

- **Cascade Lake:** Equipado com 2 CPUs Intel Xeon Cascade Lake Gold 6252 (48 núcleos no total) e 384 GB de RAM. Utiliza uma arquitetura de clusters com memória distribuída e uma rede Infiniband EDR de 100 Gb/s entre os nós, oferecendo uma solução econômica e escalável horizontalmente. Este processador representa a tecnologia mais recente comparada aos outros no experimento.
- **Ivy Bridge:** Cada nó possui 2 CPUs Intel Xeon E5-2695v2 (24 núcleos no total) e 64 GB de memória RAM DDR3. Utiliza uma arquitetura de *clusters* com memória distribuída, suportando tarefas segmentáveis e distribuídas entre os nós através de uma rede Infiniband EDR de 100 Gb/s. Esta configuração é a mesma utilizada no experimento original descrito nas publicações de Silva et al., 2021, mantendo todos os demais parâmetros inalterados.

4.3 Arquitetura de memória distribuída no SDumont

As arquiteturas Ivy Bridge e Cascade Lake utilizam um modelo de memória distribuída, no qual cada nó possui sua própria memória local, diretamente acessível pelo processador correspondente. A troca de informações entre os nós ocorre por meio de uma rede de interconexão, o que pode introduzir maior

latência devido ao tempo de transferência dos dados. Para manter a coerência entre os caches dos diferentes nós, são empregados protocolos de comunicação como o de passagem de mensagens. Quando um dado é requisitado a partir de outro nó, ele é enviado pela rede e armazenado temporariamente no cache local do nó solicitante. A gestão do cache é feita com políticas locais, como a LRU (Least Recently Used), que controla quais dados permanecem armazenados./

No caso do Cascade Lake, equipado com processadores Intel Xeon dessa geração, destaca-se um cache de último nível (LLC) mais avançado, com uma cache L2 maior do que a encontrada nos processadores Ivy Bridge. Além disso, o LLC do Cascade Lake é não inclusivo, o que contribui para a redução da latência e aumenta a eficiência do cache L2, diminuindo a dependência da LLC. Já na arquitetura Ivy Bridge, baseada em processadores Intel Xeon da geração anterior, as linhas armazenadas no cache L2 de cada núcleo também eram mantidas na LLC compartilhada, que seguia um modelo inclusivo. No Cascade Lake, por outro lado, cada núcleo possui seu próprio cache L2 exclusivo, enquanto a LLC ampliada é compartilhada entre os núcleos.

5. Resultados e Análises

5.1 Análise Computacional do *Workflow* com o Perfilador VTune

A Figura 1 apresenta o histograma das atividades do *workflow* Parsl RNA-Seq, de seis atividades, que foi executado no nó computacional Ivy Bridge. Este experimento envolve a execução de duas atividades em *multithreading* executadas em paralelo (Bowtie e Sort) com 24 *threads*, enquanto as demais atividades são realizadas de forma sequencial, processando um arquivo por núcleo. O gráfico mostra o uso da CPU utilizado ao executar o workflow completo com 6 atividades. Já em contrapartida, as Figuras 2 e 3 apresentam o uso da atividade Bowtie2 em diferentes nós computacionais.

Observa-se um uso intensivo da CPU no nó Ivy Bridge, em que seis tarefas ativam simultaneamente 24 *threads*, maximizando a capacidade de processamento do nó e resultando na menor duração do experimento. As atividades são iniciadas a partir de *tasks* e, ao entrarem em uma etapa que suporta *multithreading*, alocam o número de *threads* permitido pelos núcleos do nó computacional. A seguir iremos apresentar uma análise da atividade bowtie de forma a comprovar que a utilização dos 24 cores da Figura 1 é relativa a aplicação Bowtie2.

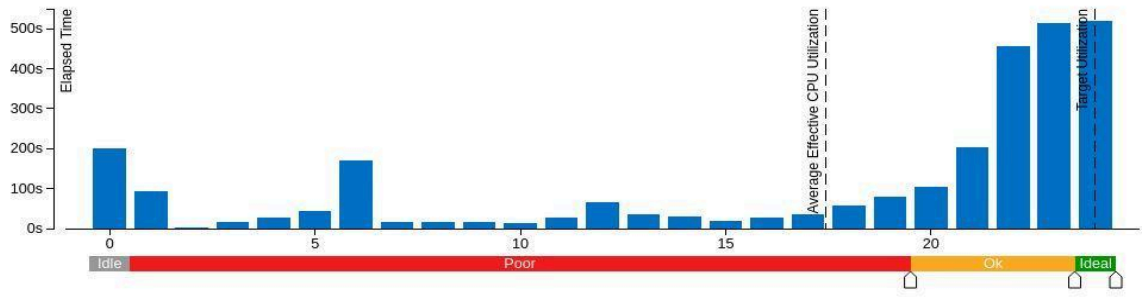


Figura 1. Uso da CPU no Ivy Bridge 24 *cores*

5.2 Análise Computacional da Atividade Bowtie2 com o Perfilador VTune

A distribuição do consumo de CPUs ao longo de processo mostra que ambos o nó computacional Ivy Bridge(Figura 2) e nó computacional Cascate Lake (Figura 3) apresentam uma alta utilização do número máximo de cores do nó computacional durante toda a execução da aplicação. A Figura 2 tem como média o uso da CPU a partir das 20 *threads* e isso quer dizer que todos os trabalhos usaram mais que essa quantidade para executar a atividade. Já na Figura 3 essa média se dá a partir das 35 *threads*.

A Figura 2 utilizou de forma eficiente e distribuída seus 24 *cores,* sem ociosidade, mostrando desempenho superior a partir do uso de 20 *cores*; porém, devido à sua tecnologia de memória mais antiga, apresentou tarefas mais lentas, resultando em tempos de execução maiores. A Figura 3 se destacou pelo uso intensivo de processamento, mostrando desempenho superior a partir do uso de 40 cores, com tempos de execução mais rápidos. A duas arquiteturas se destacam pelo motivo da sua média de uso da CPU estar localizadas em um alto nível de uso, já que todas as threads estão sendo usadas para executar a atividade.

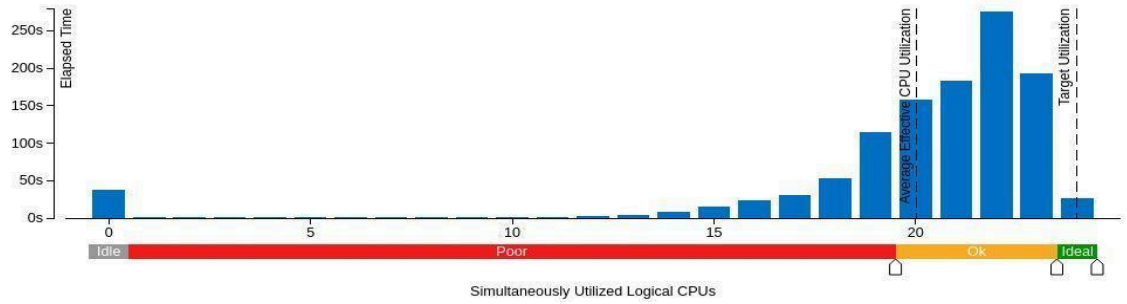


Figura 2. Uso da CPU no Ivy Bridge 24 *cores*

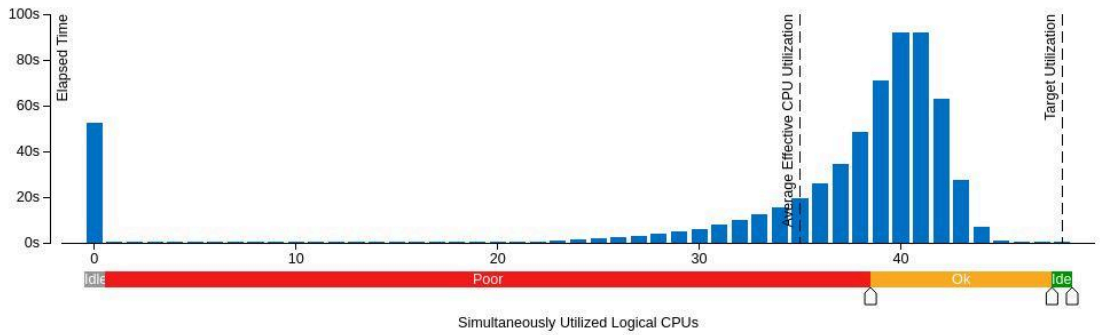


Figura 3. Uso da CPU no Cascade Lake com 48 *cores*

6. Conclusão

A análise realizada com o Intel VTune demonstrou de forma clara e consistente a eficiência da paralelização do código Bowtie2 nas arquiteturas Ivy Bridge e Cascade Lake. A ferramenta evidenciou um aproveitamento elevado e contínuo do número máximo de núcleos disponíveis durante a maior parte do tempo de execução, refletindo um uso intensivo dos recursos computacionais de ambas as plataformas. Ao examinar os histogramas de desempenho, observou-se uma notável semelhança no padrão de comportamento das duas arquiteturas, com uma tendência crescente que indica uma utilização eficiente e balanceada dos recursos de hardware. Essa alta taxa de ocupação dos núcleos de processamento valida não apenas a escalabilidade do Bowtie2, mas também sua robustez e capacidade de adaptação a diferentes ambientes de execução. A maximização da ocupação dos cores está diretamente relacionada à redução do tempo total de processamento, comprovando o bom desempenho do software e a eficácia da sua paralelização em contextos de computação de alto desempenho.

7. Referências bibliográficas

[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.

[10] Babuji, Y. N. et al. (2019). Parsl: Pervasive parallel programming in python. In Weissman, J. B., Butt, A. R., and Smirni, E., editors, HPDC’19, pages 25–36. ACM.

1 Dados gerais

Título do projeto: Desenvolvimento de Jogos para o Ensino-Aprendizagem.

Bolsista: Amanda Vilas Boas Oliveira

Orientadora: Regina Célia Cerqueira de Almeida

Edital / Programa: 2024/PIBIC

Período: setembro de 2024 - agosto de 2025

2 Objetivos

O objetivo desse projeto foi o desenvolvimento de 30 tutoriais e jogos educativos voltados para o ensino-aprendizagem de alunos do ensino fundamental e médio, por meio da plataforma Scratch. Assim, os tutoriais foram estruturados de forma didática com instruções detalhadas e organizados em níveis crescentes de complexidade, visando capacitar os estudantes com noções básicas de programação utilizando o software Scratch. Dessa forma, este trabalho concentrou-se na conclusão dos tutoriais mais avançados, relacionados à criação de jogos didáticos. Além disso, os materiais produzidos foram concatenados em um livro e as imagens originais foram reeditadas por meio do software Gwenview. Por fim, os tutoriais foram aplicados no projeto Meninas STEM - Petrópolis Tec Hub do Laboratório Nacional de Computação Científica - LNCC em Petrópolis/RJ e na disciplina COM364 – Comunicação e Informática, oferecida pela Faculdade de Comunicação da Universidade Federal da Bahia (FACOM/UFBA).

3 Introdução

Desde o final dos anos 1990 a internet transformou a comunicação, a cultura, a economia e principalmente a educação, ampliando o acesso a novos saberes [1]. Apesar disso, muitos estudantes ainda não têm contato com noções básicas de informática e programação durante sua formação inicial, perdendo a oportunidade de desenvolver raciocínio lógico e habilidades para resolução de problemas.

Em grande parte dessa lacuna é resultado da ausência de disciplinas de

computação na grade curricular da maioria das escolas, da falta de infraestrutura tecnológica e de estereótipos que afastam os jovens da área. Desse modo, até mesmo estudantes universitários de áreas não exatas encontram dificuldades para executar sequências lógicas de instruções, competência essencial na programação.

Consoante a isso, o uso de jogos como recurso didático surge como alternativa eficaz para despertar o interesse e facilitar a aprendizagem, transformando um conhecimento que seria maçante em uma atividade lúdica [2]. Dessa forma, este projeto desenvolveu tutoriais na plataforma Scratch para o ensino-aprendizagem dos alunos do ensino fundamental e médio. O Scratch é um software de programação visual criado pelo MIT (Massachussets Institute of Tecnology) e é muito utilizado para introduzir conceitos computacionais de forma interativa e acessível.

Portanto, foram produzidos 30 tutoriais distribuídos em três grupos de complexidade, desde conceitos básicos até a criação de jogos educativos. O material também é voltado ao apoio de professores e busca não apenas ensinar programação, mas estimular o interesse pelas áreas STEM (Ciência, Tecnologia, Engenharia e Matemática) e contribuir para uma educação tecnológica mais inclusiva.

4 Materiais e Métodos

Para elaboração dos tutoriais, principalmente os voltados à criação dos jogos educativos, foi desenvolvido um *template* para criação de uma identidade visual do material, com algumas etapas específicas visando assegurar a didática. Dessa forma, o *template* está padronizado com a identidade visual da plataforma Scratch, utilizando sua paleta de cores como o azul, roxo, laranja e verde. Além disso, todos os tutoriais possuem uma numeração, título condizente com o conteúdo abordado nele, uma explicação do conceito envolvido, apresentação dos blocos que serão trabalhados, uma atividade e um exercício em forma de desafio.

Cada tutorial contém uma introdução, uma breve explicação do conceito envolvido, apresentação dos blocos de programação que serão utilizados e

atividades práticas acompanhadas de desafios. Assim, os tutoriais de jogos exigem uma progressão mais cuidadosa de complexidade, abrangendo jogos como "Par ou Ímpar" e "PacMan", por exemplo. Cada jogo foi projetado com objetivos específicos e desafios que estimulam a interação do usuário, resolução de problemas e pensamento lógico.

Além disso, foi necessário uma reedição das imagens dos tutoriais já produzidos para fins de padronização. Assim, essa etapa foi realizada por meio do software Gwenview, visando aprimorar a clareza visual e facilitar a compreensão das instruções pelos usuários. Foram adicionados sinalizadores indicativos na cor vermelha para destacar áreas específicas em cada etapa da atividade, bem como setas e números de passos a serem seguidos, proporcionando uma orientação visual e intuitiva durante a realização das atividades propostas.

Por fim, o material produzido está sendo compilado em um livro, para fácil acesso e distribuição, além de ter sido aplicado diretamente a estudantes do no projeto Meninas STEM - Petrópolis Tec Hub do Laboratório Nacional de Computação Científica - LNCC em Petrópolis/RJ e da disciplina COM364 – Comunicação e dos Informática, da Faculdade de Comunicação da Universidade Federal da Bahia (FACOM/UFBA), permitindo avaliar sua performance em situações reais de ensino-aprendizagem.

5 Resultados e Discussão

Durante a realização do projeto foi desenvolvido o livro “Introdução à Programação com Tutoriais: Programação de forma lúdica”, em parceria com o projeto Meninas STEM - Petrópolis Tec Hub e reunindo os 30 tutoriais criados. Estes tutoriais foram organizados didaticamente em três grupos de complexidade, abrangendo desde os fundamentos básicos da programação até o desenvolvimento de jogos educativos. Em especial, o terceiro grupo de tutoriais é composto pelos jogos didáticos propostos no projeto, contendo os jogos: "Par ou Ímpar", "Jogo de Perguntas", "Jogo do Labirinto", "Jogo da Memória", "Jogo do Gato e Rato", "Jogo da Cobrinha", "PacMan", "Jogo de Lançamento (Angry Birds)", "Jogo de Corrida" e

"Torre de Hanói".

Paralelamente foi realizada uma aplicação prática desses tutoriais em aulas da disciplina COM364 – Comunicação e Informática, ofertada para estudantes da Faculdade de Comunicação (FACOM) da Universidade Federal da Bahia (UFBA). Durante essas aulas, os bolsistas do projeto participaram como monitores com a supervisão do co-orientador e docente responsável, Crysttian Arantes Paixão. Esse acompanhamento permitiu avaliar o desempenho dos alunos e identificar suas principais dificuldades durante a interação com os tutoriais.

Essa experiência revelou que estudantes provenientes de áreas não exatas tinham pouca familiaridade com conceitos básicos da computação, exigindo algumas adaptações na abordagem didática. Desse modo, optamos por uma comunicação mais acessível, com menos uso de termos técnicos e mais explicações detalhadas sobre conceitos fundamentais da área STEM.

Portanto, os resultados demonstraram o potencial educativo do projeto com a plataforma Scratch. Os estudantes foram bem responsivos, mostrando interesse ao realizar as atividades propostas e enfrentar os desafios dos tutoriais. Ao final da disciplina, entregaram ótimos trabalhos pondo em prática todo o conhecimento adquirido com os tutoriais. Por fim, a qualidade visual dos tutoriais que foi aprimorado através da reedição das imagens com o software Gwenview foi um adicional importante para facilitar o entendimento das atividades.

6 Conclusões

Com esse projeto foi possível comprovar a eficácia dos tutoriais desenvolvidos na plataforma Scratch como um ótimo recurso pedagógico para ensino-aprendizagem de alunos na área de programação, em especial pela abordagem prática e lúdica com jogos educativos. Além disso, a aplicação prática na disciplina COM364 da FACOM/UFBA demonstrou que a adaptação da linguagem e do método didático à realidade dos estudantes e de cada turma em específico é essencial para que o aprendizado seja efetivo, principalmente para aqueles sem

familiaridade prévia com computação.

Outro ponto foi a parceria com o projeto Meninas STEM, que garantiu maior acessibilidade e inclusão ampliando o impacto do material desenvolvido. O envolvimento demonstrado pelos alunos e a qualidade dos trabalhos finais confirmaram o potencial motivacional e educacional da proposta.

Portanto, esses resultados corroboram a continuidade e ampliação da iniciativa para beneficiar ainda mais estudantes, proporcionando-lhes habilidades fundamentais como pensamento lógico, resolução de problemas e criatividade no contexto tecnológico. Dessa forma, fica evidente que despertar desde cedo o interesse pelas áreas STEM contribui diretamente para uma educação tecnológica mais inclusiva, permitindo que jovens de diferentes contextos, especialmente mulheres, tenham oportunidades igualitárias de acesso e sucesso em carreiras científicas e tecnológicas.

7 Referências

SILVEIRA, Sérgio Amadeu. Democracia e Códigos Invisíveis: Como os algoritmos estão modulando comportamentos e escolhas políticas. São Paulo: editora edições Sesc SP, 2019 .

PIAGET, J. Psicologia e Pedagogia. 7ª impressão. Rio de Janeiro: Editora Forense Universitária LTDA, 1985.

VARELA, H. (2017). Scratch: Um jeito divertido de aprender programação. Brasil: Casa do Código.

VALENTE, José Armando et al. O computador na sociedade do conhecimento. Campinas: Unicamp/NIED, p. 11-18, 1999.

DA SILVA, Angela Carrancho. Educação e tecnologia: entre o discurso e a prática. Revista Ensaio: Avaliação e Políticas Públicas em Educação, v. 19, n. 72, p. 527-554, 2011.

NUNES, Felipe Becker et al. Um estudo de caso sobre a importância do uso de objetos de aprendizagem no ensino fundamental como apoio pedagógico. In: Anais do Workshop de Informática na Escola. 2014. p. 542.

LOPES, José Junio et al. A introdução da informática no ambiente escolar. Clube do professor, v. 23, 2004.

SEVERO, C. E. P. (2021). Jogos com Scratch: em projetos práticos com linguagem de blocos. Brasil: Casa do Código.

Salvador, 04 de agosto de 2025.

Amanda Vilas Boas Oliveira
Estudante

Orientador (a)

RELATÓRIO DE ATIVIDADES

Dados Gerais

Título do Projeto: Implementação de Algoritmos Quânticos

Nome da Bolsista: Beatriz Barcelos Cardozo

Nome do Orientador: Renato Portugal

Tipo de Bolsa: Iniciação Científica (PIBIC/CNPq)

Período: 01/09/2024 a 31/08/2025

Objetivos

O presente projeto tem como objetivo principal complementar a formação acadêmica da bolsista na área de Computação Quântica, com ênfase na simulação de algoritmos em computadores clássicos. Os objetivos específicos incluem o desenvolvimento dos fundamentos matemáticos essenciais à teoria quântica, o estudo aprofundado de algoritmos relevantes na área e a sua respectiva implementação computacional.

Além disso, o projeto visa promover a familiarização da bolsista com linguagens, estruturas e ferramentas utilizadas no campo da computação quântica, preparando-a para enfrentar desafios teóricos e práticos que envolvem tecnologias emergentes e de alto impacto científico.

Introdução

A Computação Quântica emergiu como um novo paradigma no início da década de 1980, a partir dos trabalhos pioneiros de Paul Benioff e Richard Feynman. Feynman observou que os computadores clássicos enfrentam limitações fundamentais ao simular sistemas quânticos, uma vez que o espaço de Hilbert, que descreve os estados possíveis de um sistema quântico, cresce exponencialmente com o número de partículas envolvidas.

Para superar essa limitação, Feynman sugeriu a construção de computadores baseados nas próprias leis da Mecânica Quântica. Essa proposta foi formalizada por David Deutsch com a introdução do conceito de máquina de Turing quântica e a generalização do modelo de circuitos lógicos clássicos para o contexto quântico.

Esses marcos teóricos viabilizaram o desenvolvimento de algoritmos com desempenho superior aos algoritmos clássicos, como o de Shor, capaz de fatorar números inteiros em tempo polinomial, com implicações diretas na segurança de sistemas criptográficos modernos. Atualmente, a Computação Quântica apresenta avanços significativos e promete transformar áreas como criptografia, otimização e simulação de sistemas físicos.

Este projeto de iniciação científica insere-se nesse contexto desafiador e inovador, proporcionando à bolsista a oportunidade de estudar conceitos fundamentais, compreender a estrutura de algoritmos quânticos e aplicar esse conhecimento por meio de simulações em computadores clássicos.

Metodologia

A condução do projeto seguiu uma abordagem baseada em estudo dirigido, sob a orientação contínua do pesquisador responsável. A bolsista foi incentivada a revisar e aprofundar conteúdos teóricos por meio de livros, apostilas e artigos previamente selecionados, focando em temas essenciais como álgebra linear, espaços vetoriais, operadores unitários, entrelaçamento quântico, portas lógicas e circuitos quânticos.

A fixação dos conteúdos foi reforçada com a resolução de exercícios, promovendo o raciocínio crítico e a autonomia intelectual da bolsista. A prática computacional incluiu a simulação de algoritmos por meio do simulador de caminhadas quânticas Hiperwalk, desenvolvido em Python. A ferramenta permitiu visualizar e testar o comportamento de algoritmos em um ambiente controlado, proporcionando maior familiaridade com os princípios da computação quântica.

Durante o desenvolvimento do projeto, a bolsista também realizou apresentações de seminários, discutindo temas estudados e compartilhando resultados com o orientador e colegas do grupo de Computação Quântica do LNCC. Reuniões regulares garantiram um acompanhamento próximo das atividades, com ajustes metodológicos sempre que necessário.

Resultados e Discussão

Ao longo do período, a bolsista demonstrou evolução consistente no domínio dos fundamentos da Computação Quântica, com destaque para o entendimento de conceitos complexos como o entrelaçamento quântico, a manipulação de portas lógicas e a modelagem de circuitos. Tais avanços foram apoiados por uma base matemática sólida, construída a partir do estudo de álgebra linear e da análise vetorial.

No aspecto prático, foi realizada a simulação do Algoritmo de Grover utilizando o Hiperwalk. O algoritmo, conhecido por proporcionar vantagem quadrática na busca em listas não ordenadas, foi modelado computacionalmente, permitindo à bolsista compreender sua estrutura e o comportamento probabilístico dos resultados. O uso do simulador serviu como ponte entre a teoria e a aplicação, destacando a importância da abstração matemática para o desenvolvimento de soluções computacionais eficientes.

Os resultados obtidos até o momento foram apresentados na Mini Jornada de Iniciação Científica e Tecnológica, promovida no início do ano, sendo bem recebidos pela banca avaliadora. Durante a execução do projeto, foi identificado que as maiores dificuldades se concentraram na interpretação de notações matemáticas avançadas, que exigem certo grau de maturidade acadêmica. No entanto, essas barreiras foram superadas com o apoio do orientador e a continuidade dos estudos, permitindo à bolsista ganhar mais confiança e fluidez no entendimento dos conteúdos.

Conclusões

Os objetivos traçados para esta etapa do projeto foram, em sua maioria, atingidos. A bolsista desenvolveu uma compreensão aprofundada dos fundamentos da Computação Quântica e demonstrou habilidades sólidas na aplicação prática dos conceitos estudados. A simulação do Algoritmo de Grover representou um marco importante na consolidação desse aprendizado, integrando teoria e prática de forma eficaz.

O projeto proporcionou uma formação robusta em uma área estratégica e em crescimento acelerado, contribuindo significativamente para o preparo acadêmico da bolsista. Como próximo passo, pretende-se explorar novos algoritmos quânticos, aprofundar a modelagem de sistemas mais complexos e, possivelmente, investigar métodos híbridos que combinem técnicas clássicas e quânticas, consolidando ainda mais a trajetória da bolsista na pesquisa científica.

Referências Bibliográficas

[1] P. Benioff. *The Computer as a Physical System: A Microscopic Quantum Mechanical Hamiltonian Model of Computers as Represented by Turing Machines*. Journal of Statistical Physics, 22, 1980.
[2] R. Feynman. *Simulating Physics with Computers*. International Journal of Theoretical Physics, 21:467–488, 1982.
[3] D. Deutsch. *Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer*. Proceedings of the Royal Society of London, A 400:97–117, 1985.
[4] D. Deutsch. *Quantum Computational Networks*. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 425(1868):73–90, 1989.
[5] P. W. Shor. *Polynomial-time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer*. SIAM Journal on Computing, 26(5):1484–1509, 1997.
[6] Portugal, R. et al. *Uma Introdução à Computação Quântica*. SBMAC, São Carlos, 2010.
[7] R. Portugal; F. L. Marquezino. *Introdução à Programação de Computadores Quânticos*. Apostila disponível em: https://github.com/programaquantica.
[8] M. A. Nielsen; I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, UK, 2000.

RELATÓRIO DE PROJETO DE INICIAÇÃO CIENTÍFICA

CIÊNCIAS EXATAS E DA TERRA, CIÊNCIAS DA COMPUTAÇÃO E METODOLOGIA E TÉCNICAS DA COMPUTAÇÃO

1) Dados Gerais

Título do Projeto Proposto: Análise e otimização de E/S paralela em sistemas de processamento de alto desempenho

Instituição: Laboratório Nacional de Computação Científica

Nome do Aluno: Carolina de Oliveira Soares de Menezes

Nome do Orientador:

- D.Sc. Hiago Mayk Gomes de Araújo Rocha (Pesquisador Adjunto - COMAC / LNCC, Orientador)
- D.Sc. Carla Osthoff Ferreira de Barros (Tecnologista Sênior - SEPAD / COTIC / LNCC, Coorientador)

Tipo de bolsa: PIBIC

Período do relatório: 09/05/2025 à 31/08/2025

2) Objetivo

O presente projeto tem como **objetivo geral** avaliar o desempenho da infraestrutura de Entrada e Saída (E/S) paralela do supercomputador Santos Dumont (SDumont), identificando gargalos, padrões de uso ineficientes e oportunidades de otimização no sistema de arquivos Lustre. Contudo, o presente relatório tem o **objetivo específico** de avaliar o número de acessos de leitura e escrita, bem como o desbalanceamento desses acesso nos nós dos SDumont ao longo do mês de Janeiro de 2023.

3) Introdução

Os sistemas de Processamento de Alto Desempenho (PAD) são essenciais para o avanço científico e tecnológico de diversas áreas (e.g., biologia, física, química e engenharia), pois permitem o processamento de grandes volumes de dados graças às suas grandes capacidades de processamento e armazenamento [1]. Contudo, para extrair alta performance das aplicações, não basta apenas ter processadores potentes, é necessário garantir que os dados sejam lidos e armazenados de forma rápida e eficiente. Esse processo, conhecido como entrada e saída (E/S) paralela, é muito usado em ambientes PAD, onde várias operações simultâneas acontecem.

Para atender a grande demanda de E/S, os sistemas de PAD usam uma infraestrutura de E/S paralela que permite operações de leituras e escritas em vários discos simultaneamente [3]. Essa abordagem reduz gargalos como lentidão no acesso aos dados, sobrecarga em componentes como os *Object Storage Targets* (OST), que são discos responsáveis por armazenar os dados, e os *Metadata Servers* (MDS), que gerenciam informações dos arquivos, como nomes, permissões e estrutura de diretórios. Além disso, acessos simultâneos por várias aplicações podem gerar contenção e afetar negativamente o desempenho [3]. Essa infraestrutura de E/S paralela é viabilizada através dos sistema de arquivos paralelo (PFS) que são para gerenciar a distribuição dos dados entre servidores e permitir maior paralelismo nas operações de armazenamento. O Lustre é um dos PFS mais utilizados, presente em cerca de 20% dos sistemas do ranking IO500 [2], tais como nos supercomputadores Frontier (EUA), Lumi (Finlândia), Leonardo (Itália) e o SDumont (Brasil).

Apesar do seu amplo uso, o Lustre enfrenta alguns desafios que afetam a performance das aplicações. Entre os problemas relacionados às próprias **aplicações**, destacam-se os padrões de acesso desbalanceados, o uso intensivo de pequenos arquivos e o comportamento imprevisível de cargas mistas. Já entre os desafios ligados ao **gerenciamento da infraestrutura**, estão a distribuição desigual de dados entre os servidores de armazenamentos (OSTs), a sobrecarga nos servidores de metadata (MDS), e o excesso de acessos ao mesmo tempo. Diante dessas limitações, ***torna-se essencial analisar o comportamento de E/S nos sistemas PAD para identificar gargalos e propor melhorias.***

Dessa forma, o presente trabalho tem como objetivo compreender os gargalos de E/S em sistemas de PAD, por meio da análise do seu comportamento e do impacto causado por diferentes padrões de acesso. Este relatório, contudo, apresenta resultados com foco na seguinte análise: ***avaliar o número de operações de leitura e escrita realizadas no SDumont e investigar o desbalanceamento desses acessos entre os nós de armazenamento ao longo do mês de Janeiro de 2023.***

4) Metodologia

Os dados analisados foram coletados por meio da ferramenta collectl [1], amplamente usada para monitorar o desempenho em sistemas PAD devido a sua baixa intrusão. O collectl oferece várias métricas (e.g., Solicitações de abertura e fechamento de arquivos, Operações de sincronização de dados, etc), contudo, este estudo focou especialmente no volume de dados de leitura e escrita; os campos read_kb e write_kb para entender como o sistema de E/S funcionou ao longo do mês. Foram utilizados dois tipos principais de métricas, derivadas dos dados coletados:

- **Throughput diário**: total de leitura e escrita nos OSTs do Lustre, medido em GiB por dia. Essa métrica foi obtida a partir dos arquivos gerados pelo collectl, somando, a cada dia, tudo o que foi escrito e lido em todos os OSTs.

O volume de dados lidos e escritos estava kilobytes (KiB), mas foi convertido para gibibytes (GiB) para facilitar a visualização.

- **Média e desvio padrão de leitura/escrita por nós:** avaliação da distribuição do uso de E/S entre os nós computacionais. A partir dos volumes de leitura e escrita por nós, também registrados pelo collectl, foram calculados a média e desvio padrão diários. Para representar de forma mais clara o grau de desbalanceamento, foi usado o coeficiente de variação (CV) [1], que chamaremos de *Load Imbalance* (LI) em nossa análise. O CV foi escolhido por ser uma métrica independente de escala, o que permite comparações entre dias com diferentes volumes totais de E/S. Esse valor é calculado dividindo o desvio padrão pela média. Quanto maior esse valor, mais desigual foi a distribuição das operações de E/S entre os nós.

5) Resultados e Discussão

Esta seção mostra os resultados da análise dos dados coletados durante o mês de Janeiro de 2023 no supercomputador SDumont. Inicialmente, será apresentada a análise de vazão diária de leitura e escrita, seguida da análise do desbalanceamento de carga.

5.1 Throughput diário de leitura e escrita

A Figura 1 mostra o volume total de dados lidos (em GiB, no eixo y) e escritos por dia (eixo x) nos OSTs do sistema Lustre ao longo do mês de Janeiro de 2023. Observa-se que o volume de leitura foi consideravelmente maior que o de escrita ao longo do mês, sendo 15,90 na média de todos os dias e até 70,30 no dia 15 de Janeiro. Os dias 08, 09 e 14 de Janeiro mostram os maiores volumes de leitura, ultrapassando 40.000GiB (cerca de 39,06 TiB) em cada um desses dias. Porém, mesmo com picos de atividade, o uso da largura de banda de leitura atingiu cerca de 11,7% da capacidade teórica do sistema, que é de 343 GiB/s, mostrando um aproveitamento abaixo do esperado da infraestrutura disponível.

Essa oscilação no *throughput* entre os dias pode estar associada a fatores como o tipo de aplicações executadas, a quantidade de usuários ativos e o agendamento de tarefas, que pode concentrar cargas intensas em determinados períodos e resultar em dias com pouco uso nos demais. Aplicações com baixo volume de leitura/escrita ou cargas menores exigentes do ponto de vista de E/S também contribuem para a variação vista.

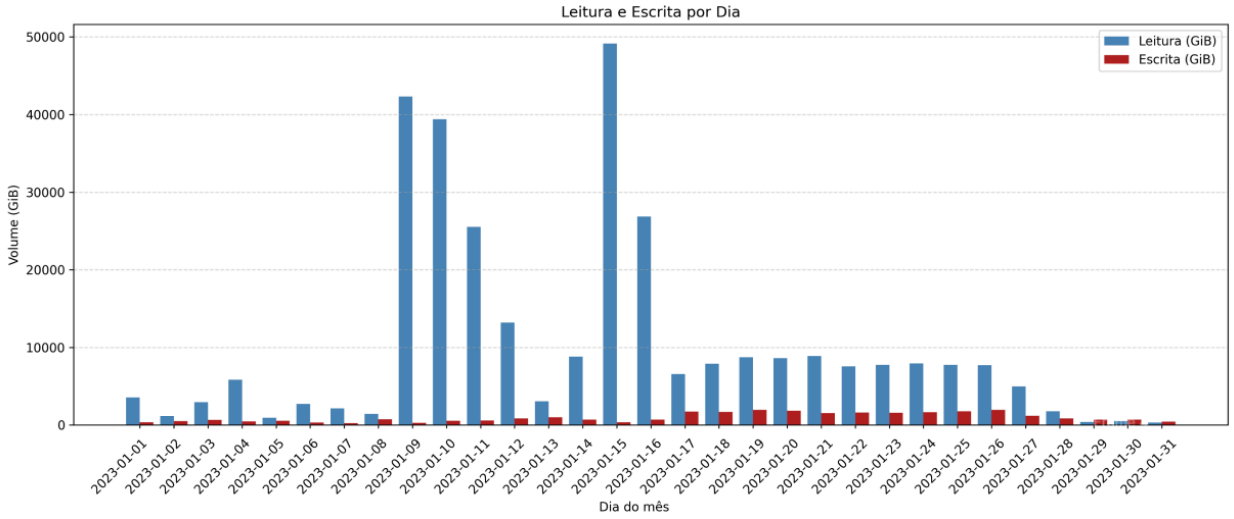


Figura 1: Throughput diário de leitura e escrita

5.2 *Load Imbalance* (LI) diário

Analisando o desbalanceamento no uso dos diferentes nós do SDumont, a Figura 2 mostra o LI (eixo x) e o no decorrer dos dias (eixo y), onde valores de LI maiores que 1 indicam que o desvio padrão ultrapassa a média, representando uma variação superior a 100%. Isso significa que há uma diferença no volume de leitura entre os OSTs, sugerindo que poucos deles estão sendo muito utilizados, enquanto outros estão pouco ou nada nas operações de leitura.

Nossos resultados mostrou picos elevados de LI em diversos dias, como em 06/01 (LI = 24,90), 07/01 (LI= 23,66) e 08/01 (LI= 23,48), além de um período contínuo de alto desbalanceamento entre os dias 16 e 26. Esses valores muito altos indicam que a maior parte das operações de leitura foi concentrada em poucos OSTs, enquanto outros ficam com pouca ou nenhuma atividade. Isso pode causar gargalos, aumentar o tempo de resposta das aplicações e comprometer o desempenho global do sistema. Fatores como a forma como os dados são distribuídos nos arquivos, o modo como as aplicações acessam esses dados e o agendamento das tarefas (*jobs*) podem contribuir para reduzir esse

desbalanceamento.

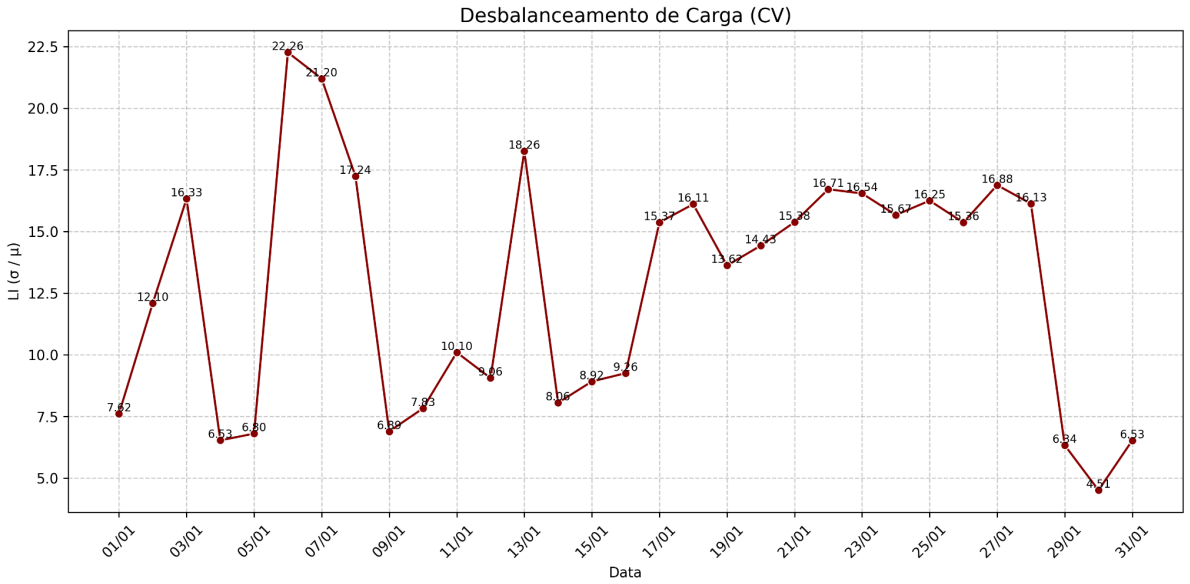


Figura 2: Load Imbalance (LI) diário

6. Conclusão

O presente trabalho avaliou o número de operações de leitura e escrita realizadas no SDumont e investigou o desbalanceamento desses acessos entre os nós de armazenamento ao longo do mês de Janeiro de 2023. Nossos resultados mostraram que o uso do SDumont varia bastante no decorrer dos dias. Além disso, a utilização dos nós é bastante desbalanceada com regiões com alta concentração de acesso de E/S em alguns discos OSTs, o que compromete a eficiência global da infraestrutura de armazenamento. Isso mostra que ainda há potencial para um melhor aproveitamento da infraestrutura de E/S paralela. Para trabalhos futuros, nosso objetivo é ampliar a análise atual com dados de outros períodos e explorando outras métricas, o que pode ajudar a entender melhor onde estão os gargalos e como melhorar ainda mais o desempenho do sistema.

7. Referências bibliográficas

[1] A. R. Carneiro, J. L. Bez, C. Osthoff, L. M. Schnorr, and P. O. A. Navaux, “Uncovering I/O demands on PAD platforms: Peeking under the hood of Santos Dumont,” Journal of Parallel and Distributed Computing, vol. 180, pp. 73–86, 2023.

[2] IO500 Benchmark Committee, “IO500 list – ranked PAD storage systems.” [Online]. Available: https://io500.org.

[3] OpenSFS, Lustre File System Operations Manual, Release 2.15. [Online]. Available: https://docs.lustre.org/lustre_manual.pdf.

Relatório de Atividade – PIBIT/LNCC

1) Dados Gerais

Título do Projeto: *Estudo e implementação de sistema de banco de dados para análise em Saúde Coletiva.*

Nome do Bolsista: Fernanda Xabudé Moreira Bomfilioli.

Orientador: Paulo Cabral Filho.

Coorientador: José Karam Filho.

Tipo de Bolsa: PIBIT

Período do Relatório: 12/2024 a 08/2025.

2) Objetivos

O trabalho teve como objetivos:

- Atualizar a base de dados com os registros de internações hospitalares por câncer referentes ao ano de 2024, integrando ao banco de dados;
- Realizar análises estatísticas com as informações adicionadas;
- Elaborar um plano para avaliação da qualidade dos dados.

3) Introdução

A atividade foi iniciada a partir de um banco de dados relacional (MySQL) existente, contendo registros de internações hospitalares relacionadas ao câncer fornecidos pelo Sistema de Informações Hospitalares (SIH) do DATASUS.

Como parte do trabalho, foram adicionados ao banco novos registros referentes ao ano de 2024, os quais foram obtidos por meio de extração de dados disponibilizados publicamente pelo DATASUS. Para otimizar consultas frequentes, foi criado índices em colunas-chaves utilizando comando MySQL, o que resultou na redução do tempo de resposta em consultas complexas.

4) Materiais, Métodos ou Metodologias

Com o banco atualizado, foram realizadas diversas consultas e análises estatísticas. Para isso, foram utilizados comandos SQL com cláusulas como JOIN, GROUP BY e ORDER BY e funções agregadoras como COUNT, SUM e AVG. As consultas geraram

dados como a descrição de atributos (por exemplo, nome do estado, CID-10, região, entre outros), o número de casos de internações e sua porcentagem em relação ao total. Os

dados obtidos foram organizados e exportados em planilhas, que abordam diferentes estatísticas para análises realizadas dentro de um período de 13 anos, de 2012 a 2024.

Como parte do processo, foi elaborado um plano para analisar a qualidade dos dados utilizados. Como ouvinte da disciplina de Modelagem da pós-graduação do LNCC ministrada pelo Professor Dr. José Karam Filho, foi desenvolvido um modelo analítico de avaliação da qualidade dos dados, que será aplicado futuramente aos dados de internações referentes ao ano de 2025. Esse modelo considera a extração e o meio de armazenamento dos dados, mas o principal foco é a construção de uma matriz, onde será realizado uma avaliação da qualidade dos dados de uma maneira prática e sistemática, uma matriz de pontuação por atributo. Nessa matriz, cada coluna representa um atributo (como sexo, estado, região, etc.), e cada linha corresponde a um registro. A partir do resultado dessa matriz, calcula-se um vetor de pontuação por atributo, obtido pelo somatório dos valores atribuídos a cada coluna. Aplica-se também um processo de normalização para que gere um índice individual de qualidade para cada atributo variando de 0 a 1, onde 1 representa a qualidade máxima e ideal. O conjunto desses índices formam um vetor de qualidade. O qual resume a confiabilidade dos dados em cada atributo. Com os resultados obtidos, também é calculado um índice geral de qualidade dos dados.

Apesar dos avanços na organização e na análise da qualidade dos dados, ainda existem pontos que podem ser aprimorados. Atualmente, a extração dos dados é realizada de maneira manual, o que diminui a eficiência do processo. Por isso, está previsto o desenvolvimento de um extrator automatizado, que será responsável pela coleta dos dados. Essa etapa futura permitirá reduzir significativamente o tempo de coleta de dados e garantir a regularidade na atualização dos dados disponíveis.

5) Resultados e Discussões

Em relação a quantidade de internações por estados, um exemplo, foi feito uma análise comparativa destacando quais estados apresentam os maiores e menores números de casos representado pelo gráfico da figura 1.

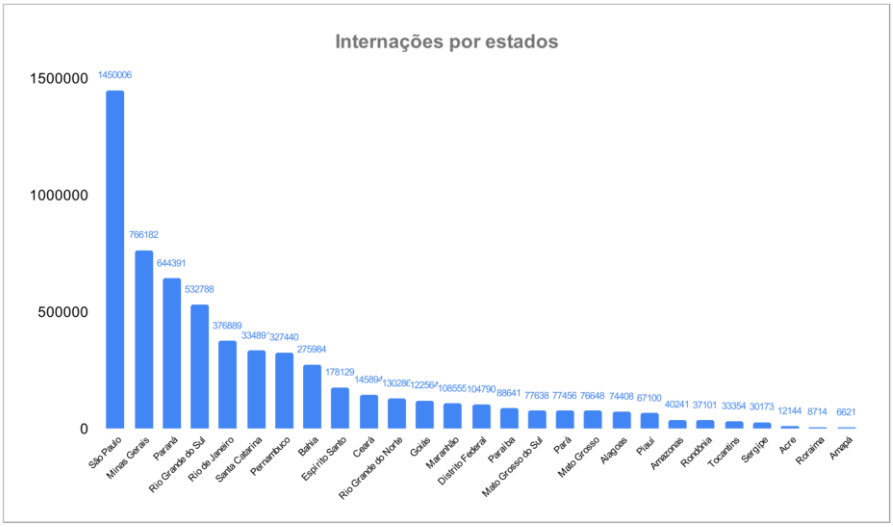


Figura 1 - Internações por Estados

Os dados mostram que o estado de São Paulo lidera com uma quantidade maior de internações hospitalares por câncer, totalizando aproximadamente 1,45 milhões de casos. Em seguida aparecem Minas Gerais com 766 mil casos e Paraná com 644 mil casos, os três estados com o maior volume de internações.

Em comparação, os estados com o menor número de internações são: Amapá com 6 mil casos, Roraima com 8 mil casos e Acre com 12 mil casos. Possivelmente essa diferença pode estar relacionada à população menor desses estados. Além da população, outros fatores podem influenciar essas diferenças, como a distribuição dos centros de tratamento oncológico ou a estrutura hospitalar disponível.

Essa análise por estados é relevante dentro do contexto deste estudo, pois permite identificar padrões regionais e possíveis desequilíbrios nos dados. Diferenças relevantes entre estados podem indicar problemas como inconsistências nos registros. Essas observações ajudam a compreender melhor a estrutura dos dados, porém não são conclusivas. Logo, a motivação para o desenvolvimento de um modelo analítico da avaliação de dados surgiu dessa conclusão.

6) Conclusões

Concluindo, o trabalho realizado se iniciou com a atualização do banco de dados relacional, por meio da inserção dos registros referentes ao ano de 2024. A partir disso, foram aplicadas melhorias de performance diretamente na base de dados com a criação de índices em colunas-chaves, o que otimizou o tempo de execução das consultas. Com o banco atualizado, foi possível realizar novas análises estatísticas envolvendo um período de 13 anos (2012-2024). Além das análises realizadas, o estudo avançou com a elaboração de um modelo analítico para avaliar a qualidade dos dados futuros. Também foi proposto o desenvolvimento de um extrator automatizado para sistematizar o processo de coleta.

7) Referências

CABRAL FILHO, Paulo. Proposta de modelo de acompanhamento dos gastos em saúde: uma análise para o câncer no período 2012 – 2021. 2023. 163 f. Tese (Doutorado em Saúde Coletiva) – Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

KARAM-FILHO, J. Princípios básicos de modelagem. In: Messeder, S. A.; CAMBUI, C. B. C.; MARQUE, M. I. C. (org.). Analista cognitivo: uma profissão interdisciplinar. Salvador: EDUFBA, 2019. p. 23-44.

RELATÓRIO DE ATIVIDADES - BOLSA PIBIC/LNCC

1. Dados Gerais

Título do projeto: Estudo e implementação de sistemas de banco de dados para análise em Saúde Coletiva
Bolsista: Gabriel Eduardo Pontes Amaral
Orientadores: Paulo Cabral Filho, José Karam Filho
Tipo de bolsa: PIBIC
Período do relatório: janeiro - julho 2024
Início da bolsa: janeiro 2024

2. Objetivos

Este projeto de iniciação científica tem como objetivo coletar, tratar, armazenar e analisar dados referentes às internações públicas de oncologia disponibilizados pela plataforma do DATASUS. Após a coleta os dados são ajustados em padrões previamente estabelecidos e inseridos em um banco de dados relacional, permitindo o desenvolvimento e aplicação de análises e modelos diversos.

3. Introdução

A análise de dados na saúde coletiva tem papel importante para a construção de políticas públicas, otimização de recursos e melhoria da qualidade do atendimento. Dessa forma, um sistema de banco de dados estruturado permite integrar, armazenar e processar grandes volumes de informações de forma eficiente, otimizando o processo.
A modelagem matemática consiste em descrever um fenômeno ou sistema por meio de expressões e equações numéricas que possibilitam a previsão do seu comportamento, permitindo assim extrair informações relevantes sobre o fenômeno estudado. Os modelos podem ser classificados de diferentes formas:

- **Qualitativos ou Quantitativos:**
  - Qualitativos: focam na compreensão de relações e características sem redução a números.
  - Quantitativos: expressos em sistemas numéricos e relações matemáticas.

- **Determinísticos ou Probabilísticos:**
  - Determinísticos: utilizam relações exatas e previsíveis, sem variabilidade aleatória.
  - Probabilísticos: incorporam variáveis aleatórias e incertezas, utilizando funções de probabilidade.

- **Transientes ou Estacionários:**
  - Transientes: apresentam variáveis que mudam com o tempo, refletindo comportamento dinâmico.
  - Estacionários: independem do tempo, mantendo características constantes.

- **Contínuos ou Discretos:**
  - Contínuos: baseados em variáveis reais, que podem assumir infinitos valores dentro de um intervalo.
  - Discretos: baseados em variáveis inteiras, que assumem valores separados e distintos.

A partir do conteúdo de modelagem apresentado pelo professor José Karam, foi conduzido um estudo sobre qualidade de dados, motivado pela necessidade de avaliar a confiabilidade das informações utilizadas em análises de saúde coletiva.

4. Materiais e Método

4.1. Materiais

- Os dados utilizados foram obtidos através do portal DATASUS, no conjunto referente às Autorizações de Internação Hospitalar (AIH) para procedimentos oncológicos.
- Sistema Gerenciador de Banco de Dados, MySQL.

4.2. Método

Extração dos dados

- A obtenção dos dados foi realizada através de um programa extrator utilizando o protocolo FTP. Foram selecionados registros referentes ao ano de 2024, abrangendo informações sobre pacientes, procedimentos e custos.
- Após a extração, os dados foram armazenados em conjunto com os dados de 2012 a 2023, em um banco de dados previamente construído.

Análise da qualidade dos dados

- Durante a participação como ouvinte na disciplina de Modelagem Matemática da pós-graduação, foi desenvolvido um método para avaliação de qualidade de dados, que pode ser aplicado através das seguintes etapas:

  1. Estabelecer critérios de qualidade para cada atributo avaliado;
  2. Para cada atributo, construir um vetor atribuindo notas para cada um de seus registros, a partir dos critérios pré-definidos;
  3. A partir dos vetores, é possível construir uma matriz com os vetores obtidos ou um índice de qualidade.

5. Resultados e Discussão

A implementação do banco de dados permite consolidar milhões de registros de internações oncológicas de maneira organizada e acessível. A criação de tabelas específicas para consultas, em conjunto com ferramentas como índices, possibilita maior desempenho na extração de informações.
Utilizando o banco de dados, foi possível criar consultas e desenvolver tabelas contendo estatísticas como a frequência de internações e o custo associado por tipo de câncer, sexo e estado.
Com o método de análise da qualidade dos dados foi possível identificar características de campos associados, onde a qualidade dos dados está relacionada entre si, além de permitir que os dados possam ser comparados de forma temporal a partir de um índice.

6. Conclusão

Este projeto teve êxito na implementação de um sistema de banco de dados eficiente para análise de dados em Saúde Coletiva, focado na área de internações oncológicas do SUS. Também foi possível a criação de um método de avaliação da qualidade dos dados, que pode ser aplicado de forma ampla em outras áreas e conjuntos de informações.
Diante dos resultados obtidos, evidencia-se a importância de utilizar um banco de dados estruturado para apoiar no desenvolvimento de métodos e estatísticas, que facilitam averiguar o funcionamento dos sistemas públicos, além de avaliar a qualidade dos dados disponibilizados.

7. Referências

CABRAL FILHO, Paulo. Proposta de modelo de acompanhamento dos gastos em saúde: uma análise para o câncer no período 2012 – 2021. 2023. 163 f. Tese (Doutorado em Saúde Coletiva) – Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

KARAM-FILHO, J. Princípios básicos de modelagem. In: Messeder, S. A.; CAMBUI, C. B. C.; MARQUE, M. I. C. (org.). Analista cognitivo: uma profissão interdisciplinar. Salvador: EDUFBA, 2019. p. 23-44.

1) Dados Gerais

Título do projeto: Uso do padrão de paralelismo de linguagem em benchmarks científicos
Bolsista: Gabriel Thomaz do Nascimento
Orientadores: Roberto Pinto Souto, Eduardo Lucio Mendes Garcia
Tipo de bolsa: PIBIC/CNPq
Período do Relatório: Setembro 2024 a Julho de 2025
Início do período de bolsa: 01/09/2024

2) Objetivos

Objetivos abordados durante o período da pesquisa:

1. Avaliar o estágio atual de paralelização de código dos *benchmarks* **NPB.**
2. Implementar no código programação paralela utilizando o padrão intrínseco da linguagem Fortran.
3. Avaliar o desempenho paralelo obtido segundo métricas.

3) Introdução

O NAS Parallel Benchmarks (NPB) é um pequeno conjunto de programas destinado a validação de performance paralela de supercomputadores. Os *benchmarks* são derivados de aplicações de dinâmica de fluidos computacional e consistem em 5 *kernels* e 3 pseudo-aplicações. O conjunto de *benchmarks* foi ampliado para incluir novos *benchmarks* para malhas adaptativas não estruturadas, E/S paralela, aplicações *multizone* e *computação em grid.* O tamanho dos problemas no NPB são predefinidos e denotados por diferentes classes. As implementações de referência do NPB estão disponíveis em modelos de programação populares como MPI e OpenMP.

A cláusula DO CONCURRENT, introduzida no padrão Fortran 2008, é uma construção de *loop* projetada para paralelismo explícito e deterministicamente seguro. DO CONCURRENT indica que o *loop* pode ser executado fora de ordem, e também pode ser utilizada para indicar ao compilador que o *loop* pode ser paralelizado, como no auto parallelization do gfortran. A partir da versão 20.11 do NVIDIA HPC SDK, o compilador NVFORTRAN incluído acelera automaticamente rotinas contendo DO CONCURRENT. Através do standard parallelism, o NVFORTRAN permite explorar os recursos da CPU e GPU em paralelismo *manycore* e *multicore* sem o uso de diretivas de linguagem. DO CONCURRENT aceita apenas funções do tipo PURE e oferece suporte a escopo de localidade de variável (shared ou local) e a operações de redução (reduce).

4) Metodologia

Análise dos *benchmarks*

Muitos problemas científicos importantes apresentam vários níveis de paralelismo, e essa propriedade não está refletida no NPB. Para remediar essa deficiência, foram criadas as versões do NPB Multi-Zone (NPB-MZ). (VAN DER WIJNGAART; JIN, 2003)

Os *benchmarks* de pseudo-aplicações Lower-Upper Symmetric Gauss-Seidel (LU), Scalar Penta-diagonal (SP) e Block Tri-diagonal (BT) resolvem versões discretizadas das equações compressíveis e instáveis de Navier Stokes em três dimensões espaciais. Cada um opera em uma malha de discretização estruturada que é um cubo lógico. Em aplicações realistas, no entanto, uma única malha desse tipo geralmente não é suficiente para descrever um domínio complexo, e várias malhas ou zonas são usadas para cobri-lo.

Branches de trabalho foram estabelecidas e podem ser acessadas pelo GitHub: https://github.com/TempoHPC/NPB. Testes foram realizados e pôde-se observar, através do acionamento da variável de ambiente NPB_TIMER_FLAG, que as subrotinas x_solve.f90, y_solve.f90, e z_solve.f90 são responsáveis por ~17-18% do tempo total cada (em 1 *thread*), sendo assim as mais custosas.

Figura 1 - Experimento SP-MZ - 1 *thread*.

```
SECTION   Time (secs)
 total    :   13.276  (100.00%)
 rhsx     :    0.951  (  7.16%)
 rhsy     :    0.861  (  6.49%)
 rhsz     :    0.930  (  7.01%)
 rhs      :    4.782  ( 36.02%)
   --> total  sub-rhs:    2.743  ( 20.66%)
   --> total rest-rhs:    2.040  ( 15.36%)
 xsolve   :    2.272  ( 17.11%)
 ysolve   :    2.376  ( 17.90%)
 zsolve   :    2.367  ( 17.83%)
 txinvr   :    0.269  (  2.03%)
 pinvr    :    0.289  (  2.18%)
 ninvr    :    0.129  (  0.97%)
 tzetar   :    0.417  (  3.14%)
 add      :    0.226  (  1.70%)
 qbc_copy:    0.144  (  1.08%)
 qbc_comm:    0.000  (  0.00%)
   --> total exch_qbc:    0.144  (  1.08%)
```

Fonte: o autor

Adaptações e implementação

Após verificação, escolheu-se a aplicação SP-MZ como potencial para implementação da cláusula DO CONCURRENT, uma vez que esta havia apenas uma chamada de função externa dentro do loop principal dessas sub rotinas. As estruturas destas funções são idênticas, apenas variando as direções (x, y e z).

Figura 2 - Rotina lhsinit, que inicialmente impossibilitaria o uso do DO CONCURRENT.

```
call lhsinit(lhs, lhsp, lhsm, nx-1)


call lhsinit(lhs, lhsp, lhsm, ny-1)


call lhsinit(lhs, lhsp, lhsm, nz-1)
```

Fonte: o autor

Em seguida, foi realizado o *inlining* da rotina lhsinit, já que DO CONCURRENT aceita apenas funções do tipo PURE.

Figura 3 - Rotina lhsinit após realização do *inlining* (x_solve.f90).

```
lhs (:,0) = 0.0d0
lhsp(:,0) = 0.0d0
lhsm(:,0) = 0.0d0
lhs (:,nx-1) = 0.0d0
lhsp(:,nx-1) = 0.0d0
lhsm(:,nx-1) = 0.0d0

lhs (3,0) = 1.0d0
lhsp(3,0) = 1.0d0
lhsm(3,0) = 1.0d0
lhs (3,nx-1) = 1.0d0
lhsp(3,nx-1) = 1.0d0
lhsm(3,nx-1) = 1.0d0
```

Fonte: o autor

De forma análoga à implementação OpenMP nativa da aplicação, buscou-se definir laços com potencial paralelismo e escopos de variáveis.

Figura 4 - Implementação OpenMP do laço paralelo (x_solve.f90).

```
!$omp parallel do default(shared) private(fac2,m,fac1,i2,i1,ru1,i,j,k)  &
!$omp& schedule(static) collapse(2)
       do  k = 1, nz-2
          do  j = 1, ny-2
```

Fonte: o autor

Figura 5 - Estrutura análoga do mesmo laço utilizando DO CONCURRENT (x_solve.f90).

```
do concurrent(k = 1 : nz-2, j = 1 : ny-2) local(fac2, m, fac1, i2, i1, ru1, i, cv,
rhon, lhs, lhsp, lhsm)
```

Fonte: o autor

Definiu-se o escopo de variáveis através da cláusula local, visando garantir *thread-safety*.

5) Resultados e Discussão

Experimentos no Sdumont foram realizados na partição sequana_cpu_dev. As *flags* de compilação: **-O3 -stdpar=multicore -Minfo** foram utilizadas para otimização e paralelismo *multicore*.

Os resultados podem ser observados a seguir:

Tabela 1 - Resultados dos experimentos

Threads	Tempo Total (s)	MOP/s	Speed Up
1	13.28	5556.49	1.00
2	11.23	6570.48	1.18
4	9.46	7795.75	1.40
8	9.00	8193.72	1.48

Fonte: o autor

Tabela 2 - Desempenho por rotina

Threads	Tempo xsolve	Speedup xsolve	Tempo ysolve	Speedup ysolve	Tempo zsolve	Speedup zsolve
1	2.27	1.00	2.37	1.00	2.36	1.00
2	1.37	1.64	1.39	1.70	1.92	1.23
4	0.89	2.54	0.83	2.85	1.00	2.36
8	0.58	3.89	0.48	4.95	0.536	4.41

Fonte: o autor

6) Conclusões

Ao analisar o desempenho por rotina, observa-se um speedup satisfatório, ainda que desbalanceado. Além disso, pôde-se obter êxito em implementar e executar a aplicação utilizando o paralelismo nativo de linguagem. Ademais, as recomendações para trabalhos próximos são:

- **Explorar GPU: Testar -stdpar=gpu para *offload* em GPU's NVIDIA.**

- **Refinar Escopos: Ajustar cláusulas local/shared para minimizar contenção.**
- **Análise Detalhada: Usar ferramentas como nvprof ou Nsight para identificar gargalos de memória.**

7) Referências Bibliográficas

1. **VAN DER WIJNGAART, R.; JIN, H. NAS Parallel Benchmarks, Multi-Zone Versions. [s.l: s.n.].** Disponível em: <https://www.nas.nasa.gov/assets/nas/pdf/techreports/2003/nas-03-010.pdf>.

2. **Using Fortran Standard Parallel Programming for GPU Acceleration | NVIDIA Technical Blog.** Disponível em: <https://developer.nvidia.com/blog/using-fortran-standard-parallel-programming-for-gpu-acceleration/>.

3. **Accelerating Fortran DO CONCURRENT with GPUs and the NVIDIA HPC SDK | NVIDIA Technical Blog.** Disponível em: <https://developer.nvidia.com/blog/accelerating-fortran-do-concurrent-with-gpus-and-the-nvidia-hpc-sdk/>.

4. **POP Centre of Excellence. Performance Assessment. Disponível em:** <https://pop-coe.eu/node/69>.

5. **NAS Parallel Benchmarks.** Disponível em: <https://www.nas.nasa.gov/software/npb.html>.

6. **MARCIANO, Anna V. G.; ANTUNES, Artur dos Santos; SCHEPKE, Claudio. Paralelização do NAS-PB usando Do Concurrent.** In: ESCOLA REGIONAL DE ALTO DESEMPENHO DA REGIÃO SUL (ERAD-RS), 25. , 2025, Foz do Iguaçu/PR. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025 . p. 97-100. ISSN 2595-4164. DOI: https://doi.org/10.5753/eradrs.2025.6787.

7. **TREMARIN, Gabriel Dineck; MARCIANO, Anna Victória Gonçalves; SCHEPKE, Claudio; VOGEL, Adriano. Fortran DO CONCURRENT Evaluation in Multi-core for NAS-PB Conjugate Gradient and a Porous Media Application.** In: SIMPÓSIO EM SISTEMAS COMPUTACIONAIS DE ALTO DESEMPENHO (SSCAD), 25. , 2024, São Carlos/SP. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024 . p. 133-143. DOI: https://doi.org/10.5753/sscad.2024.244796.

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: Agosto de 2024 - Julho 2025
Início da Bolsa: maio 2023

2) Objetivos

Objetivos abordados durante o período da pesquisa:

1. Aprofundamento nos estudos e início de experimentos com modelo MONAN
2. Testes com o modelo MPAS
3. Aprendizagem de conteinerização para as aplicações com docker e singularity

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-**L**a**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 ao desempenho computacional dos modelos, também é de grande importância o entendimento da estratégia de paralelização empregada nos seus códigos fonte. Portanto, este relatório apresenta os estudos iniciais e os primeiros experimentos relacionados ao modelo MONAN, com ênfase na utilização de técnicas de conteinerização. Para isso, foram realizados testes práticos com o MPAS-Model, uma aplicação científica com estrutura e requisitos computacionais semelhantes ao MONAN. Esses testes incluíram a adaptação e criação de ambientes conteinerizados utilizando Docker e Singularity, visando compreender e validar estratégias de instalação, compilação e execução desses modelos em outros ambientes.

4) Material e Métodos

As etapas iniciais das atividades se deram com o conhecimento sobre o modelo MONAN. O MONAN é um modelo atmosférico desenvolvido com base no núcleo dinâmico do **MPAS(Model for Prediction Across Scales)**, que utiliza o método de **discretização por volumes finitos** sobre **grades não estruturadas** (malhas com espaçamento irregular), que permitem **refinamento variável** e

aninhamento de grades, técnicas que otimizam a resolução em áreas de interesse, como regiões costeiras ou ciclônicas, mantendo o desempenho computacional.

O **núcleo dinâmico** do MONAN é:

- **Não hidrostático**: resolve explicitamente os movimentos verticais da atmosfera, permitindo a simulação de fenômenos de pequena escala como tempestades e convecção.
- **Totalmente compressível**: considera variações de densidade do ar com a pressão e a temperatura, fundamentais para simulações em grandes altitudes e em ambientes dinâmicos.
- **Localmente conservador em massa**: preserva a massa de ar em cada célula da malha, assegurando a consistência física das simulações ao longo do tempo.

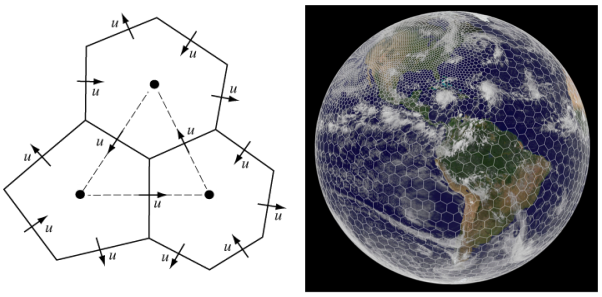


Figura 1 - A - Malha de Voronoi , B- Resolução no globo (Fonte : https://mpas-dev.github.io/)

Após o estudo aprofundado do MONAN, identificou-se a necessidade de utilizar técnicas de **conteinerização** para facilitar a execução e o gerenciamento das aplicações. Para isso, foi realizado um estudo detalhado sobre as ferramentas **Docker** e **Singularity**. A partir disto, foram desenvolvidos containers customizados para o modelo **MPAS**, utilizado como base por sua similaridade com o MONAN, e também iniciada a adaptação com o ambiente. Os trabalhos estão registrados no repositório https://github.com/TempoHPC/MPAS-Model.

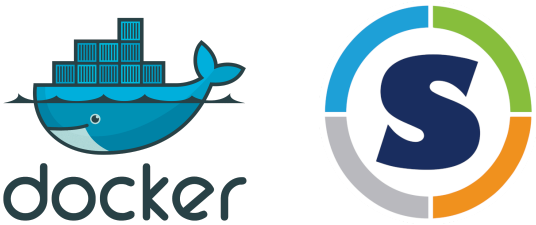


Figura 2 - A - Logo Docker (Fonte : https://pt.wikipedia.org/wiki/Docker_(software)), B - Logo Singularity Appteiner (Fonte: https://docs.sylabs.io/guides/3.5/user-guide/introduction.html)

Docker (Figura 2 - A) é uma plataforma de **conteinerização** que permite empacotar aplicações e suas dependências em ambientes isolados e portáteis chamados containers. Com isso, é possível garantir que o software seja executado da mesma forma em qualquer máquina, independentemente do sistema operacional ou das bibliotecas disponíveis no host. O funcionamento do Docker é configurado por meio de um **arquivo chamado Dockerfile**, que descreve passo a passo como a imagem será construída. Esse arquivo utiliza uma **sintaxe própria baseada em instruções simples como FROM, RUN, COPY, CMD**, entre outras. Não é uma linguagem de programação completa mas o Dockerfile se integra com **scripts em Bash, Python, compiladores de C/C++ e outras ferramentas**, dependendo das necessidades da aplicação.

Singularity (Figura 2 - B) é uma alternativa ao Docker voltada especialmente para ambientes de computação de alto desempenho (HPC), o que o torna mais adequado para supercomputadores e clusters como o SDumont. Sua configuração é feita através de um arquivo de definição **.def**, que também utiliza um formato próprio, com seções como **%post**, **%environment**, **%runscript** e **%files**. Cada seção possui comandos escritos normalmente em **Shell Script (Bash)**, o que facilita a personalização do ambiente.

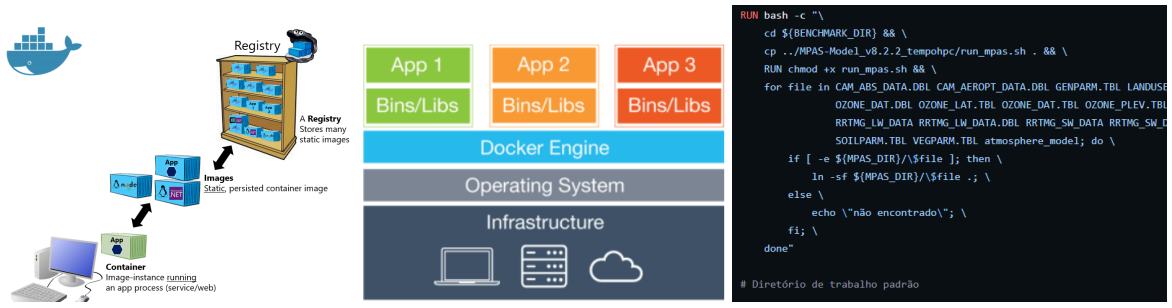


Figura 3 - A - Taxonomia Docker (Fonte: https://learn.microsoft.com/pt-br/dotnet/architecture/microservices/container-docker-introduction/docker-containers-images-registries), B - Estrutura Docker (Fonte: https://docker-unleashed.readthedocs.io/aula1.html), C - Trecho do Dockerfile

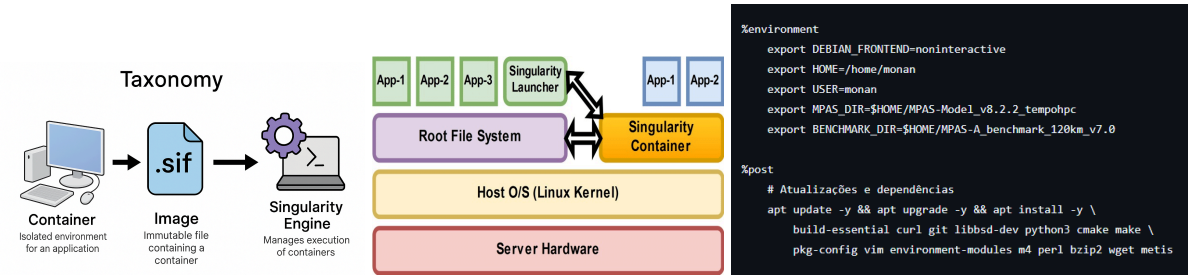


Figura 4 - A - Taxonomia Singularity (Fonte: Esquema gerado por IA), B - Estrutura Singularity (Fonte:https://www.researchgate.net/publication/321637645_A_Review_of_MongoDB_and_Singularity_Container_Security_in_regards_to_HIPAA_Regulations), C - Trecho do script Singularity

5) Resultados e Discussão

Primeiro utilizamos a versão versão adaptada do MONAN para laptops que permite testes realizados em baixa resolução, utilizando o modelo GFS (Global Forecast System, sistema global de previsão numérica do tempo) para dados de entrada presente no fork.

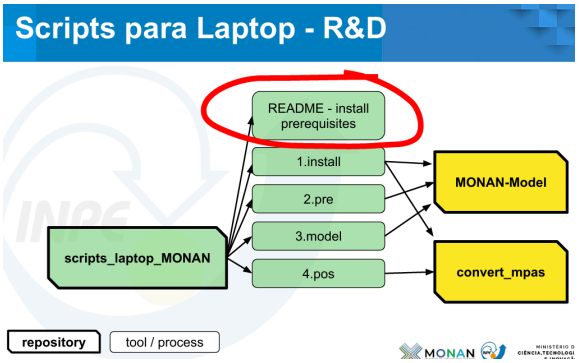


Figura 5 - Esquema do Scripts para Laptop - MONAN (https://github.com/TempoHPC/scripts_laptop_MONAN), um conjunto de scripts voltado para a instalação, configuração e execução do modelo MONAN. (Fonte: 10a_Reuniao_Geral_do_MONAN_GCC)

MONAN-Model - representa o modelo de previsão em si. **convert_mpas** - um processo de conversão, ligando o MONAN ao modelo MPAS, do qual ele deriva certos dados ou estrutura

Abaixo temos alguns resultados com os Scripts MONAN:

1. **1.install_monan.bash** : Este script instala o software e bibliotecas de permissão para o funcionamento do modelo MONAN. Ele lida com pacotes que precisam ser integrados ao sistema e garante que tudo esteja preparado para o pré-processamento.
2. **2.pre_processing.bash** : Prepara as condições iniciais para o modelo, utilizando dados do GFS para gerar uma previsão de 24 horas com baixa resolução, como 480 km.
3. **3.run_model.bash** : Execução do modelo MONAN com as condições iniciais previstas. Ele utiliza a resolução e o período configurado para gerar a previsão. 24 horas com resolução de 480 km

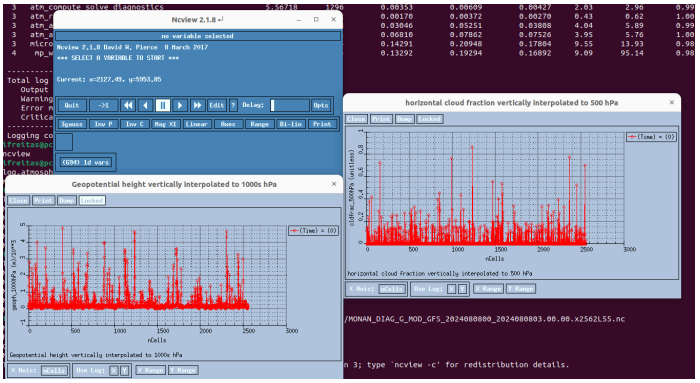


Figura 6 - 24 horas com resolução de 480 km. ***Altura geopotencial interpolada verticalmente para o nível de 1000 hPa, .Fração de nuvem horizontal interpolada para 500 hPa*** *- variação ao longo das células computacionais (nCells) da malha hexagonal*

4. **4.run_post.bash** : O script realiza o pós-processamento, organizando os dados de saída para análise. Os arquivos resultantes são salvos no diretório dataout e podem ser visualizados com ferramentas como ncview ou grads.

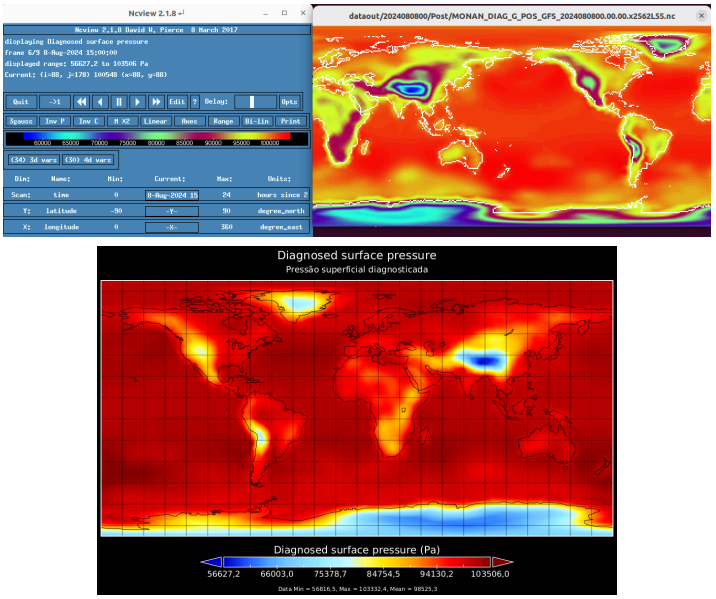


Figura 7 - As imagens correspondem ao instante 8 de agosto de 2024, às 15:00. A - Imagem gerada pelo ncview, B - Imagem gerada pelo Panoply

A variável representada é a **surface pressure** (pressão de superfície). Os valores estão dentro da faixa para simulações atmosféricas, onde regiões de baixa pressão são indicadas em tons de azul e roxo, enquanto áreas de alta pressão aparecem em tons de vermelho e laranja. Na visualização global, observa-se a distribuição da pressão atmosférica, que reflete os padrões atmosféricos e auxilia na análise de sistemas climáticos, como frentes e regiões de alta e baixa pressão.

Temperatura atmosférica interpolada. O intervalo de temperatura varia de 179,216 K a 317,475 K (aproximadamente -93,9 °C a 44,3 °C). Nível de pressão atmosférica de 1000 hPa, representando a camada próxima à superfície.

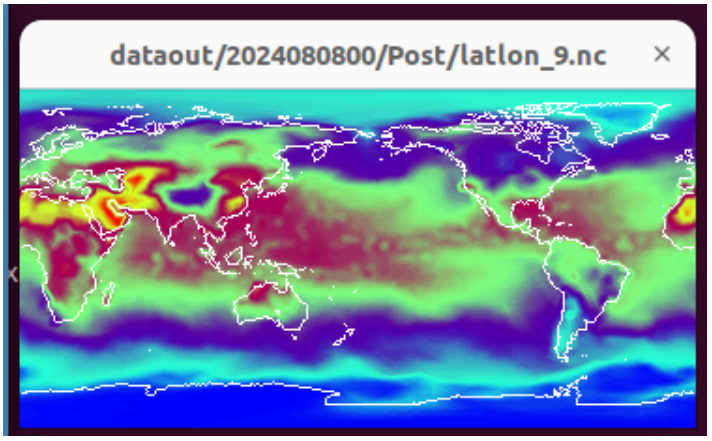


Figura 8 - Data simulada: 8 de agosto de 2024. Imagem gerada pelo ncview

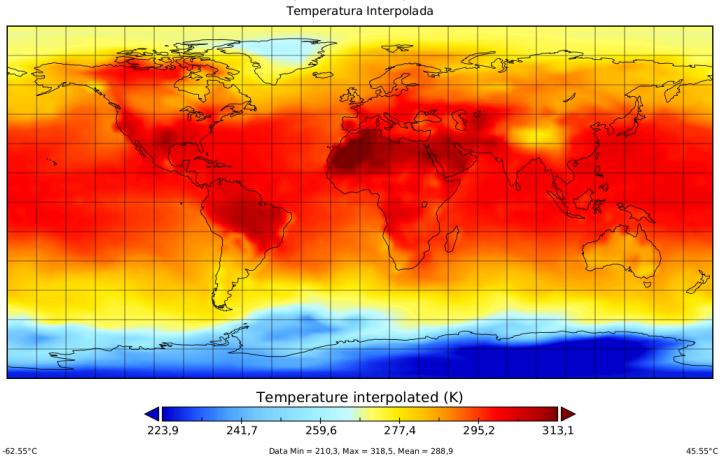


Figura 9 - A imagem corresponde a data simulada 8 de agosto de 2024, às 18:00 UTC. Imagem gerada pelo Panoply,

Escala de Cores: Tons azulados indicam temperaturas muito baixas (regiões polares e altas altitudes). Tons amarelados, alaranjados e vermelhos indicam temperaturas mais elevadas (regiões tropicais e desertos). As regiões polares (em azul) apresentam temperaturas muito baixas, consistentes com suas características climáticas já conhecidas. As regiões tropicais, como a África e a América do Sul, exibem temperaturas elevadas (tons alaranjados/vermelhos), refletindo o aquecimento solar mais intenso próximo ao equador. As áreas oceânicas apresentam variações graduais, destacando o papel dos oceanos como reguladores térmicos.

6) Conclusões

As principais conclusões obtidas no decurso do trabalho realizado destacam:

1. A relevância dos testes e conceitos iniciais com o modelo MPAS como etapa fundamental para a preparação e entendimento do MONAN, modelo principal do projeto.

2. A compreensão aprofundada dos conceitos de linguagem, comandos e do ambiente de trabalho, incluindo o uso de scripts, processos de compilação e execução dos modelos.

3. A importância do uso eficiente das ferramentas de conteinerização, como Docker e Singularity, para garantir portabilidade, reprodutibilidade e facilidade na execução dos modelos.

7) Referências Bibliográficas

CPTEC/INPE. *Treinamento MONAN 2024*. Disponível em: https://www.cptec.inpe.br/treinamento-monan-2024/. Acesso em: 28 jul. 2025.

MONANADMIN. *Documentação do MONAN*. Disponível em: https://monanadmin.github.io/monan_cc_docs/. Acesso em: 28 jul. 2025.

CPTEC/INPE. *MONAN-AI-2024 – Saulo*. Disponível em: https://dataserver.cptec.inpe.br/dataserver_dimnt/monan/meeting_AI_2024_06_19a21/MONAN%20DAY%201/MONAN-AI-2024_Saulo.pdf. Acesso em: 28 jul. 2025.

NASA. *Panoply Data Viewer – Download*. Disponível em: https://www.giss.nasa.gov/tools/panoply/download/. Acesso em: 28 jul. 2025.

MONANADMIN. *Repositório oficial do MONAN*. GitHub. Disponível em: https://github.com/monanadmin. Acesso em: 28 jul. 2025.

MONANADMIN. *MONAN v.0.2.1 – QuickStart Guide*. Disponível em: https://dataserver.cptec.inpe.br/dataserver_dimnt/monan/workshop_2023_10_02e03/MONAN_v.0.2.1_QuickStart.pdf?utm_source=chatgpt.com. Acesso em: 28 jul. 2025.

CPTEC/INPE. *Overview of MONAN Implementation – Treinamento agosto 2024*. Disponível em: https://dataserver.cptec.inpe.br/dataserver_dimnt/monan/trainings_1_MONAN_2024_08_12a16/Overview_of_MONAN_implementation-Treinamento_MONAN_ago_2024.pdf. Acesso em: 28 jul. 2025.

CPTEC/INPE. *MONAN 2024 – 2ª palestra do treinamento*. Disponível em: https://dataserver.cptec.inpe.br/dataserver_dimnt/monan/trainings_1_MONAN_2024_08_12a16/MONAN-2024-training_2nd_talk.pdf. Acesso em: 28 jul. 2025.

TEMPO HPC. *Repositório MPAS-Model*. GitHub. Disponível em: https://github.com/TempoHPC/MPAS-Model. Acesso em: 28 jul. 2025.

CPTEC/INPE. *Produção de dados – Inicialização – Enver*. Disponível em: https://dataserver.cptec.inpe.br/dataserver_dimnt/monan/workshop_2023_10_02e03/Produ%c3%a7%c3%a3o%20de%20dados%20Inicializa%c3%a7%c3%a3o_%20Enver.pdf. Acesso em: 28 jul. 2025.

MPAS DEVELOPERS. *MPAS Atmosphere – Download*. Disponível em: https://mpas-dev.github.io/atmosphere/atmosphere_download.html. Acesso em: 28 jul. 2025.

RELATÓRIO DE PROJETO DE INICIAÇÃO CIENTÍFICA

Título do Projeto

Predição de Eficiência Energética e Desempenho em Aplicações HPC com Aprendizado de Máquina

Instituição: Laboratório Nacional de Computação Científica

Nome do Aluno: Isabella da Silva Muniz

Nome do Orientador: Carla Osthoff Ferreira de Barros e Micaella Coelho Valente de Paula (co orientadora)

Tipo de bolsa: PIBIC

Período do relatório: 10/09/2024 - 28/08/2025

1. Objetivo

Esta pesquisa tem como objetivo geral desenvolver ferramentas para auxiliar o aumento da eficiência energética relacionada com a execução das aplicações no Supercomputador Santos Dumont, e tem como objetivo específico desenvolver e avaliar um modelo preditivo capaz de estimar o consumo de energia da aplicação RAxML[8], que demanda muito recursos computacionais e que está sendo utilizada como estudo de caso, com base em dados reais de execução coletados via SLURM, dando prosseguimento aos estudos desenvolvidos em [3]. A proposta visa apoiar a escolha de configurações mais eficientes para submissão de jobs no supercomputador Santos Dumont, promovendo uma melhor utilização dos recursos computacionais e contribuindo para a eficiência energética em ambientes HPC.

2. Introdução

Em ambientes de computação de alto desempenho (HPC), a crescente demanda por recursos computacionais para execução de aplicações científicas de diversas áreas impacta diretamente no consumo de energia dos supercomputadores. Esse fator afeta tanto os custos operacionais quanto a sustentabilidade desses sistemas. Diante desse cenário, surge a necessidade de desenvolver estratégias para maximizar a eficiência energética sem degradar significativamente o desempenho.

Dentre os fatores que influenciam o consumo de energia e o tempo de execução de aplicações científicas, destacam-se a escolha do número de nós computacionais e de threads por nó para a execução de uma aplicação, configurações inadequadas podem levar à subutilização dos recursos. Nesse contexto, a aplicação de técnicas de aprendizado de máquina surge como uma alternativa eficiente para auxiliar os usuários a selecionar configurações mais eficientes. Modelos preditivos treinados com dados históricos (coletadas durante a submissão e execução de jobs) podem estimar métricas como tempo de execução e consumo de energia a partir de variáveis conhecidas no momento da submissão de

um job, visando o melhor aproveitamento dos recursos disponíveis e fornecendo subsídios para decisões mais eficientes, com foco em desempenho e uso eficiente dos recursos.

Este relatório apresenta as etapas do desenvolvimento e avaliação de um modelo preditivo para estimar o consumo energético da aplicação RAxML, utilizando dados coletados via SLURM (comando sacct). Também foram realizados experimentos com a ferramenta Intel VTune Profiler [7] para aprofundar a análise do comportamento da aplicação e verificar se a aplicação estava utilizando os recursos computacionais de maneira consistente com os resultados observados nas métricas obtidas pelo sacct.

3. Metodologia

Esta seção descreve a construção dos modelos preditivos para estimar o consumo de energia da aplicação RAxML, bem como a análise do seu comportamento durante a execução utilizando o Intel VTune Profiler.

3.1 Construção do Modelo Preditivo

Os experimentos foram conduzidos utilizando a linguagem Python no ambiente *Visual Studio Code*, utilizando o modelo de regressão *Extra Trees Regressor*. Os dados foram coletados previamente pelo grupo a partir de execuções da aplicação RAxML no supercomputador Santos Dumont (nós equipados com duas CPUs Intel Xeon E5-2695v2 (24 cores, 64 GB RAM)), variando o número de nós (1, 5 e 10), threads (2, 4, 8, 12 e 24), bootstraps (10, 100, 1000 e 2000) e arquivos de entrada (15 genomas do vírus da dengue variando de 98.484 a 1.158.854 bytes). A coleta foi realizada via o comando *sacct* do SLURM, permitindo o registro de diversas métricas, utilizadas como variáveis de entrada e saída no modelo.

As variáveis de entrada incluíram informações sobre os recursos computacionais alocados (como AllocCPUS, AllocNodes, AveCPU, AveCPUFreq, NCPUS, NNodes, NTasks, ReqCPUS, ReqNodes, Thread); parâmetros da aplicação (como Bootstrap e Tamanho do arquivo); métricas de uso de CPU (como WaitTime, CPUTimeRAW, ElapsedRaw, SystemCPU, TotalCPU, UserCPU); além de indicadores de uso de memória e disco (AveDiskRead, AveRSS, AveVMSize, MaxDiskRead, MaxDiskWrite, MaxRSS, MaxVMSize, ReqMem). A variável de saída foi o consumo de energia, representado pela métrica ConsumedEnergy.

A avaliação do modelo foi realizada por meio de validação cruzada com 2 folds repetidos 100 vezes, utilizando o erro médio absoluto (MAE) como métrica de desempenho. No primeiro experimento, o modelo foi treinado utilizando todas as variáveis de entrada disponíveis. Em seguida, foi realizada uma análise de importância das variáveis, com base na estrutura do modelo. As variáveis mais relevantes para a predição de ConsumedEnergy foram selecionadas para um segundo experimento, no qual o modelo foi treinado com esse subconjunto reduzido de variáveis, visando verificar o impacto na acurácia das previsões.

3.2 Análise com Intel VTune Profiler

Além da modelagem preditiva, foi realizada uma análise detalhada do comportamento da aplicação utilizando a ferramenta Intel VTune Profiler. Foram realizados dois tipos de análise: (i) Hotspots, para a identificação de funções que mais consomem tempo de CPU; (ii) Threading, para a avaliação do uso e distribuição de threads.

Para aumentar a abrangência da análise, foram variados parâmetros de configuração de execução. Utilizou-se, um arquivo de entrada, denominado DENV_2-colombia (198574 bytes), com 100 e 1000 bootstrap (o aumento do valor de bootstrap acarreta no aumento no tempo de processamento), variando o número de threads (2, 24 e 48) e mantendo a execução em um único nó do SDumont. O objetivo foi verificar a consistência dos resultados e entender os efeitos da paralelização sobre o uso de CPU. As análises indicaram comportamentos semelhantes entre os valores de bootstrap testados, portanto serão apresentados os resultados com 1000 bootstraps.

4. Resultado e Discussões

4.1 Avaliação dos Modelos Preditivos

Com base na análise de importâncias das variáveis de entrada conduzida no primeiro experimento, apresentadas na Figura 1, foram identificadas as variáveis com maior influência na predição do consumo de energia (ConsumedEnergy): AveCPUFreq, AvePages, WaitTime, MaxPages e SystemCPU. A partir desse resultado, o modelo foi treinado novamente utilizando apenas esse subconjunto reduzido de variáveis, com o objetivo de avaliar o impacto na acurácia das previsões. A Figura 2 apresenta as importâncias das variáveis no segundo experimento.

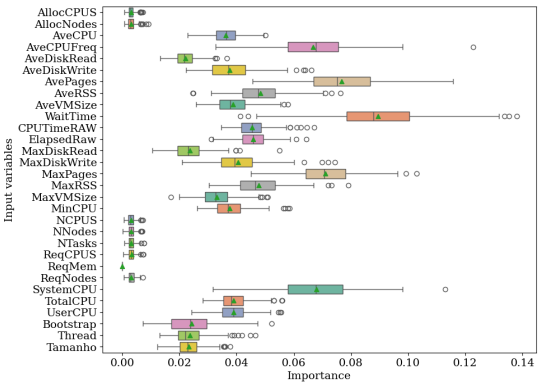


Figura 1: Importância das Variáveis de Entrada do 1º Modelo

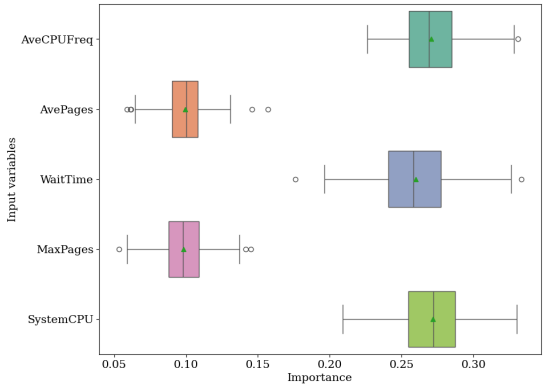


Figura 2: Importância das Variáveis de Entrada do 2º Modelo

A comparação entre o desempenho do modelo inicial (1º experimento) com o modelo ajustado (2º experimento) apresentam o Erro Médio Absoluto (MAE), na Figura 3 o MAE se estabelece entre 0.015 e 0.016, já na Figura 4 o erro se encontra 0.018. Com os resultados observa-se que não houve melhora significativa na

predição ao reduzir as variáveis de entrada. Pelo contrário, o MAE apresentou um leve aumento após a redução, sugerindo que a exclusão de algumas variáveis pode ter comprometido a capacidade preditiva do modelo.

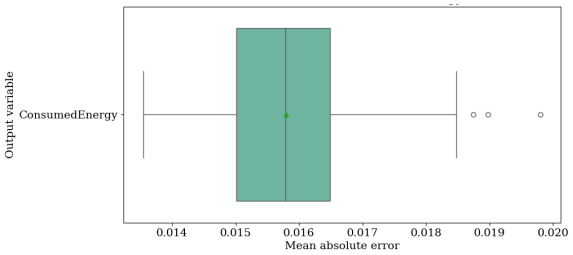


Figura 3: MAE do Modelo Inicial

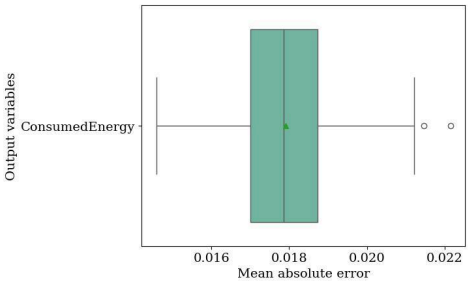


Figura 4: MAE do Modelo Ajustado

4.2 Análise com Intel VTune

Os resultados apresentadas nas Figuras 5, 6 e 7, foram obtidas por meio da execução da análise de threading utilizando o arquivo de entrada DENV_2-colombia com os parâmetros de 1000 bootstraps e variando o número de threads (2, 24 e 48).

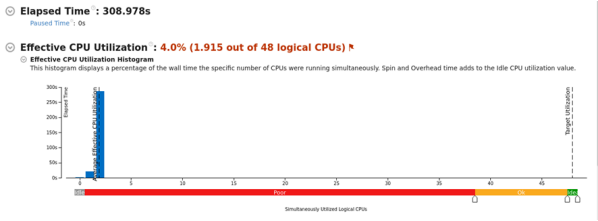


Figura 5: Análise de Threading com 2 nós

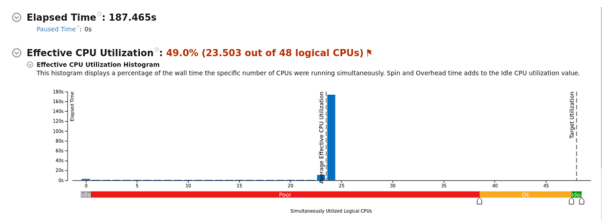


Figura 6: Análise de Threading com 24 nós

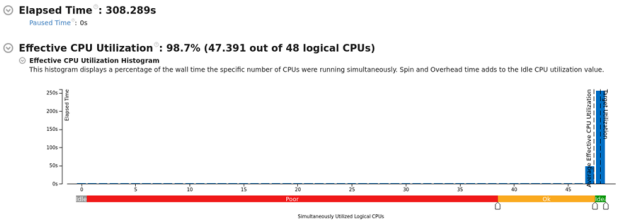


Figura 7: Análise de Threading 48 nós

Na análise de Threading, pode-se observar que com 2 threads, houve baixo aproveitamento dos núcleos (cerca de 4%) e maior tempo de execução, refletindo baixa eficiência na utilização do sistema. Com 24 threads, houve melhor distribuição de carga entre os núcleos (cerca de 49%), resultando em redução do tempo de execução. Já com 48 threads, embora o aproveitamento dos núcleos tenha sido elevado (cerca de 98%), observou-se um aumento no tempo total de execução. Esse comportamento pode estar associado a limitações de escalabilidade da aplicação, uma vez que o RAxML pode não apresentar ganhos proporcionais com o aumento do número de threads para essa carga de trabalho. Além disso, o uso intensivo dos 48 núcleos pode ter introduzido overhead de sincronização entre threads.

Function	Module	CPU Time
newviewGTRGAMMA_AVX	raxmlHPC-HYBRID-AVX	287.432s
coreGTRGAMMA	raxmlHPC-HYBRID-AVX	132.689s
likelihoodThread	raxmlHPC-HYBRID-AVX	49.711s
func@0x77a41	libm.so.6	23.068s
expf64	libm.so.6	22.939s
[Others]	N/A*	78.981s

*N/A is applied to non-summable metrics.

Figura 8: Análise de Hotspots com 2 nós

Function	Module	CPU Time
likelihoodThread	raxmlHPC-HYBRID-AVX	2818.501s
newviewGTRGAMMA_AVX	raxmlHPC-HYBRID-AVX	506.593s
expf64	libm.so.6	377.486s
coreGTRGAMMA	raxmlHPC-HYBRID-AVX	234.075s
masterBarrier	raxmlHPC-HYBRID-AVX	94.072s
[Others]	N/A*	444.963s

*N/A is applied to non-summable metrics.

Figura 9: Análise de Hotspots com 24 nós

Function	Module	CPU Time
likelihoodThread	raxmlHPC-HYBRID-AVX	11509.965s
expf64	libm.so.6	765.689s
newviewGTRGAMMA_AVX	raxmlHPC-HYBRID-AVX	604.543s
__memmove_avx_unaligned_erms	libc.so.6	400.430s
execFunction	raxmlHPC-HYBRID-AVX	378.595s
[Others]	N/A*	1202.049s

*N/A is applied to non-summable metrics.

Figura 10: Análise de Hotspots com 48 nós

A análise de Hotspots, demonstradas nas Figuras 8, 9 e 10, evidenciaram que o RAxML concentra a maior parte do tempo de execução em um conjunto reduzido de funções críticas, responsáveis pelos principais gargalos de processamento. Mesmo com o aumento do número de threads, essas funções permaneceram dominantes no uso da CPU, indicando que melhorias pontuais nelas poderiam gerar ganhos significativos de desempenho.

5. Conclusão

Com base nos testes realizados, foi possível observar que o uso de aprendizado de máquina aplicado aos dados de execução da aplicação RAxML permitiu desenvolver modelos preditivos capazes de estimar o consumo de energia de forma precisa (MAE cerca de 0,018). Apesar de o modelo ajustado não ter melhorado significativamente o MAE em relação ao modelo inicial, os testes evidenciaram a importância de selecionar variáveis relevantes para melhorar a acurácia das previsões.

As análises feitas com o Intel VTune Profiler também foram fundamentais para entender o comportamento da aplicação. A análise de Hotspots mostrou onde o código consome mais tempo de CPU, ajudando a identificar possíveis gargalos. A análise de Threading indicou limitações no paralelismo em certas configurações. Essas informações contribuem para ajustes mais eficientes nos parâmetros de execução da aplicação. Esses resultados reforçam a importância de ajustar os parâmetros de execução e investigar otimizações tanto em nível de código quanto de alocação de recursos computacionais, com vistas à melhoria do desempenho e da eficiência energética da aplicação.

Para trabalhos futuros, pretende-se aplicar os modelos a outras aplicações HPC, ampliar a base de dados, e com o objetivo de aprimorar a qualidade da variável de saída, será coletada medidas diretas de consumo de energia por meio da

ferramenta RAPL [6], a fim de realizar um novo treinamento do modelo com dados mais precisos.

6. Referências

[1] Coelho, Micaella, et al. "Machine learning regression-based prediction for improving performance and energy consumption in HPC platforms." Latin American High Performance Computing Conference. Cham: Springer Nature Switzerland, 2024.

[2] Lorenzon, Arthur F., et al. "Energy-Efficient GPU Allocation and Frequency Management in Exascale Computing Systems." ISC High Performance 2025 Research Paper Proceedings (40th International Conference). Prometeus GmbH, 2025.

[3] Coelho, Micaella, et al. "Desenvolvimento de um Framework de Aprendizado de Máquina no Apoio a Gateways Científicos Verdes, Inteligentes e Eficientes: BioinfoPortal como Caso de Estudo Brasileiro." Simpósio em Sistemas Computacionais de Alto Desempenho (SSCAD). SBC, 2022.

[4] Muralidhar, Rajeev, Renata Borovica-Gajic, and Rajkumar Buyya. "Energy efficient computing systems: Architectures, abstractions and modeling to techniques and standards." ACM Computing Surveys (CSUR) 54.11s (2022): 1-37.

[5] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63 (2006): 3-42.

[6] David, H., Gorbatov, E., Hanebutte, U. R., Khanna, R., & Le, C. (2010). RAPL: Memory power estimation and capping. In Proceedings of the 16th ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), pp. 189–194.

[7] Intel Corporation. "Intel® VTune™ Profiler." Intel Developer Zone. 2025. Disponível em: https://www.intel.com/content/www/us/en/developer/tools/oneapi/vtune-profiler.html. Acesso em: 01 ago. 2025.

[8] https://bio.tools/raxml

1) Dados gerais

Título do projeto: Análise da colaboração científica entre pesquisadores em uma instituição de educação profissional e tecnológica.

Nome do bolsista: Jefferson Pablo Nunes Santos.

Nome do orientador: José Karam Filho.

Co-Orientador: José Damião de Melo.

Tipo de bolsa e período do relatório: Iniciação científica. O período de execução do projeto, de janeiro de 2025 a agosto de 2025.

2) Objetivos

O trabalho a ser realizado no projeto visa alcançar os seguintes objetivos:

Extrair e organizar os dados disponíveis na Plataforma Lattes para a criação de uma base de dados relacional própria e estruturada.

Efetuar análises sobre a cooperação científica entre os pesquisadores pertencentes ao domínio estudado.

Divulgar e publicar os resultados obtidos por meio de relatórios, artigos científicos e apresentações em eventos acadêmicos.

3) Introdução

Este projeto de pesquisa foca na análise da colaboração científica entre pesquisadores de uma instituição de educação profissional e tecnológica. A investigação utiliza como fonte primária de dados a Plataforma Lattes, um repositório central de dados da produção científica brasileira, onde o principal desafio abordado é a extração e estruturação dessas informações, que, apesar de públicas, não estão em um formato ideal para análises complexas.

Para superar essa limitação, o projeto propõe a implementação de um framework metodológico (ETCAP) e a construção de um banco de dados relacional dedicado, visando facilitar a mineração de dados e a gestão do conhecimento para fins de gestão acadêmica.

4) Material e Métodos ou Metodologia

A metodologia foi executada seguindo um fluxo de trabalho sistemático, baseado no framework ETCAP (Extração, Transformação, Carregamento, Análise e Publicação) (Melo e Karam-Filho, 2024).

Coleta e Processamento de Dados: A extração de dados da Plataforma Lattes foi automatizada com o uso de scripts desenvolvidos em Python e da API SoapUI. Os dados coletados passaram por um rigoroso processo de limpeza e padronização para garantir sua consistência e integridade.

Criação da Base de Dados: Está sendo implementado um banco de dados relacional utilizando o sistema PostgreSQL. A modelagem do banco está projetada para armazenar de forma organizada as informações de pesquisadores, publicações, projetos e suas conexões de colaboração, permitindo escalabilidade e consultas avançadas.

Análises e Visualizações: Estão sendo construídas consultas em linguagem SQL para explorar a base de dados, com o objetivo de identificar padrões de colaboração, mapear áreas de pesquisa estratégicas e identificar os pesquisadores mais produtivos. Para a interpretação e comunicação dos resultados, foram desenvolvidas visualizações de dados interativos.

5) Resultados e Discussão

A execução do projeto está resultando em entregas concretas que atendem aos objetivos propostos. O principal resultado esperado é a **criação de um banco de dados relacional funcional e recebendo dados de CV's Lattes**, contendo dados estruturados e limpos. Este banco de dados representa um ativo de informação valioso, pois permite análises que não são diretamente possíveis através da interface pública da Plataforma Lattes.

Outro ponto importante é a possibilidade de criação de amostras temporais, com base de dados específicas, por data de interesse, de forma que as consultas e análise poderão contar com este elemento de qualificação adicional, inclusive para as etapas adicionais do framework ETCAP no que diz respeito à criação e análise das redes de cooperação.

Estão se desenvolvendo e se validando **scripts de extração em Python e um conjunto de consultas, views e instâncias de base de dados**, que constituem ferramentas reutilizáveis para a exploração contínua dos dados. As consultas permitem identificar, por exemplo, quais pesquisadores colaboram com mais frequência, quais são os principais grupos de pesquisa e como as áreas do conhecimento se interconectam, gerando os atributos e valores de interesse para os analistas.

Além disso, esperamos gerar até o final do projeto, **visualizações interativas e redes interativas**. Essas visualizações facilitam a compreensão dos padrões complexos de colaboração, tornando os resultados acessíveis a um público mais amplo, incluindo gestores acadêmicos.

6) Conclusões

Ao final deste período inicial de execução, os objetivos do projeto foram alcançados de acordo com o cronograma proposto. Estamos adaptando e validando o fluxo de trabalho completo e replicável, baseado no framework ETCAP, para extrair, tratar e analisar dados de produção científica da Plataforma Lattes.

A criação de uma instância operacional e funcional do serviço de banco de dados relacional em PostgreSQL está se provando uma solução robusta e eficaz para organizar as informações de colaboração, considerando que poderemos atingir níveis de escalabilidade e disponibilidade de processamento efetivos.

As ferramentas computacionais desenvolvidas e as visualizações interativas demonstraram grande utilidade para a análise de redes de cooperação e para a geração de insights relevantes para a gestão acadêmica.

Ainda tem sido um problema em aberto a consulta e extração de dados diretamente da base de dados Lattes, tanto pela dificuldade de acesso institucional quanto pela alteração da funcionalidade, haja visto o contexto de acesso público da base lattes ter dificuldades relativas para acesso de dados em conjunto, seja pela manutenção do CAPTCHA, seja pela exclusão da possibilidade de acesso ao arquivo do Lattes em formato tratável computacionalmente, via download XML, por exemplo, que não está mais disponível.

O projeto não apenas está produzindo uma análise avançada da colaboração científica, mas também criando uma infraestrutura de dados e um conjunto de métodos que podem ser utilizados para o monitoramento contínuo e a promoção estratégica de dados valiosos e insights que poderão influenciar e colaborar para a melhoria da avaliação das relações de cooperação entre pesquisadores, grupos de pesquisa e programas de pós-graduação.

7) Referências bibliográficas

BALANCIERI, R. al. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. Ciência da Informação, [S.l.], v. 34, n. 1, oct. 2005. ISSN 1518-8353.CNPQ. Sobre a Plataforma Lattes. Disponível

em: < https://memoria.cnpq.br/web/portal-Lattes/sobre-a-plataforma > Acesso em: 12 mar. 2025.

CNPQ Plataforma LATTES. Pesquisadores. Disponível em: < https://Lattes.cnpq.br/ > Acesso em: 12 mar. 2025.CNPQ

GABARDO, A. C. Análise de Redes Sociais: Uma visão computacional. São Paulo. Novatec Editora. 2015.

MATIAS M. S. O. Base referencial para o povoamento de repositórios institucionais: coleta automatizada de metadados da plataforma Lattes. Dissertação de Mestrado. Programa de Pós-Graduação em Gestão de Organizações e Sistemas Públicos da Universidade Federal de São Carlos. São Carlos: UFSCar, 2015.

MELO, J. D.; KARAM FILHO, J.; MESSEDER, S. A. Modelagem científica e aplicação de um framework para análise de redes de cooperação científica. Informação & Informação, [S. l.], v. 29, n. 1, p. 259–282, 2024. DOI: 10.5433/1981-8920.2024v29n1p259. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/48550. Acesso em: 6 ago. 2025.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

LNCC Laboratório Nacional de Computação Científica

Relatório de atividades

Bolsa PIBIC/LNCC
Projeto: Simulação de eventos climáticos na plataforma de Petróleo P40
Bolsista: Jonatas Halliday Sant Anna Nascimento
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 de eventos climáticos na plataforma de petróleo P40 através do PIBIC-LNCC.

- **Objetivos**

O projeto em questão tem por objetivo desenvolver uma simulação 3D de possíveis eventos climáticos para a plataforma de petróleo do tipo P40. Tal simulação aborda diferentes tipos de clima como chuva, ventos e também diferentes horários do dia.

- **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, além de jogos dos mais variados níveis de complexidade. Neste ambiente utiliza-se a linguagem C# adaptada para unity para desenvolver os scripts utilizados no projeto. Usando um modelo adquirido no laboratório labACima de uma plataforma de petróleo do modelo P40. Foi feita a abertura e acoplamento das texturas para as diferentes partes da plataforma. O projeto procura oferecer simulações de diferentes possíveis climas que possam ser controladas por uma pessoa, visando por exemplo estimular efeitos adversos de mudanças climáticas onde essas plataformas costumam ficar.

- **Métodos**

A implementação dos métodos foi feita através do modelo de classes e relacionamentos, ou seja, orientada a objetos na linguagem C#. Foi anexado à cena da plataforma duas câmeras diferentes. O modelo obtido da

plataforma, juntamente com suas texturas foram os primeiros objetos a serem adicionados à cena no *Unity.* Foi incluído também *assets* para adicionar componentes de clima, assim como um outro para as ondas.

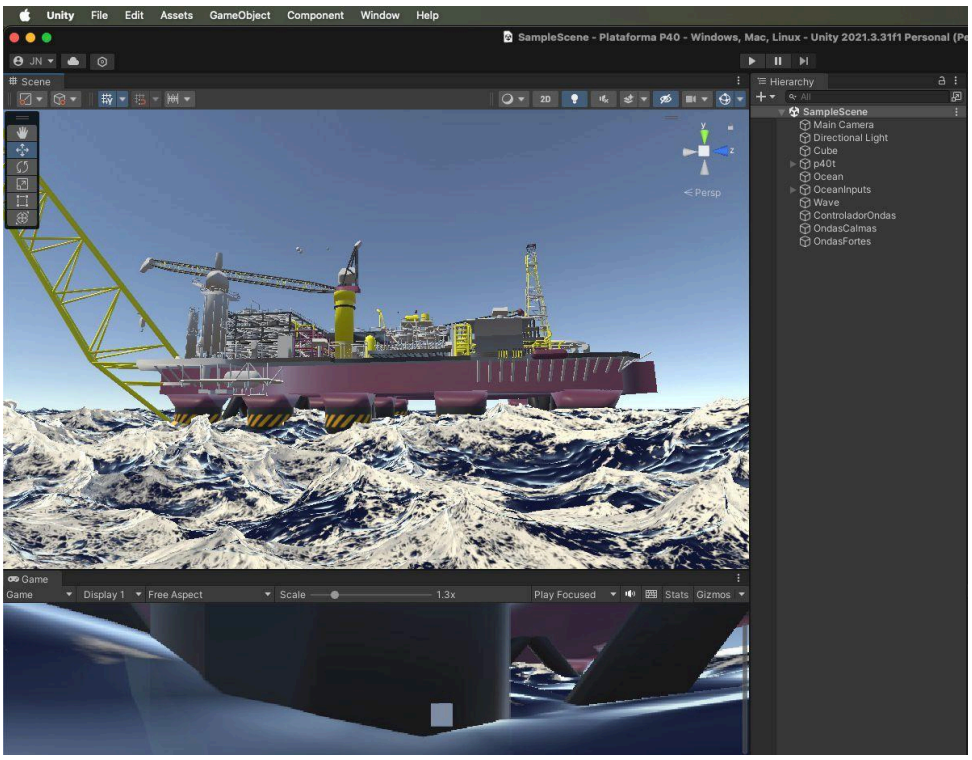


Figura 1: Animação de ondas de maior intensidade

A primeira etapa envolveu a separação de cada componente da plataforma utilizando o software Blender. Em seguida, o modelo completo da plataforma, juntamente com suas texturas, foi importado para o Unity, a fim de garantir a semelhança com a plataforma real. Posteriormente, foram adicionadas à cena simulações de ondas do mar. Para isso, foi desenvolvido um *script* chamado *WaveController.cs*, que permite ao usuário controlar a intensidade e a altitude das ondas, simulando desde um mar calmo até ondas de 20 metros (em unidades de escala da cena). Este *script* verifica o estado das ondas (calmas ou agitadas) através do objeto *Shape Gerstner Batched*, que retorna 0 para mar calmo e 1 para mar agitado, e então ajusta a influência (*weight*) das ondas na animação.

Para variantes climáticas foi adicionado um outro *script* que adiciona à cena a componente de chuva, podendo simular desde uma chuva calma até um temporal severo, como na figura 2. Neste *script* o usuário também é capaz de controlar a intensidade da chuva (que foi colorida artificialmente por motivos visuais).

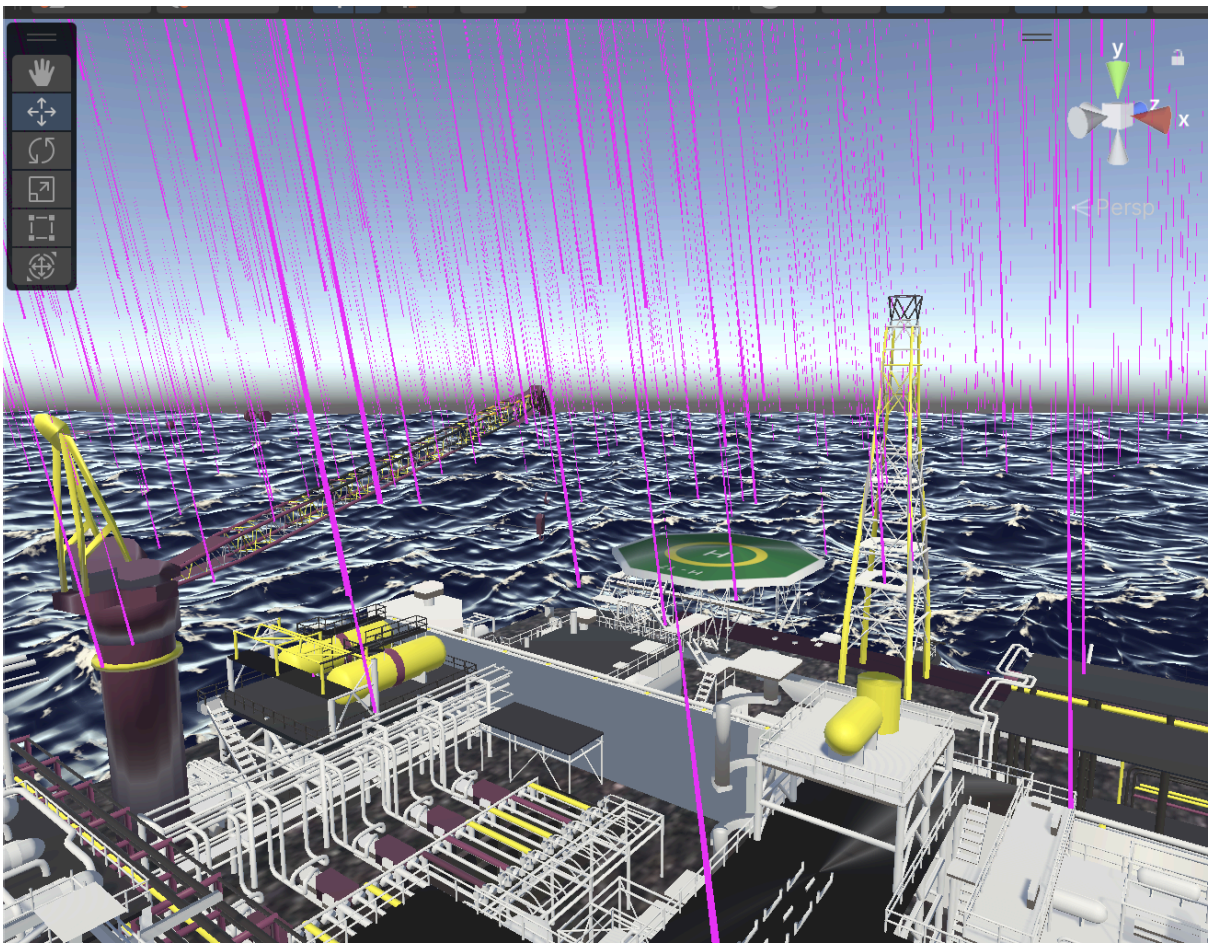


Figura 2: Simulação temporal severo (colorido artificialmente)

Assim que a animação da chuva atinge um nível de intensidade no valor de 3.5, a câmera principal inicia as oscilações que variam de 1 a 5 graus visando uma resposta ao usuário, o objeto que contém a câmera começará a execução de um *script* intitulado *CameraOscilation.cs* que inicia a trepidação da câmera permitindo uma indicação visual que a situação está crítica na animação. Essa oscilação foi criada adicionando uma função no *script* que controla a intensidade da chuva onde esse valor é retornado e é utilizado para aumentar o ângulo máximo e mínimo de oscilação ao longo do eixo **Z.** Quanto maior a intensidade, maior será a oscilação da câmera.

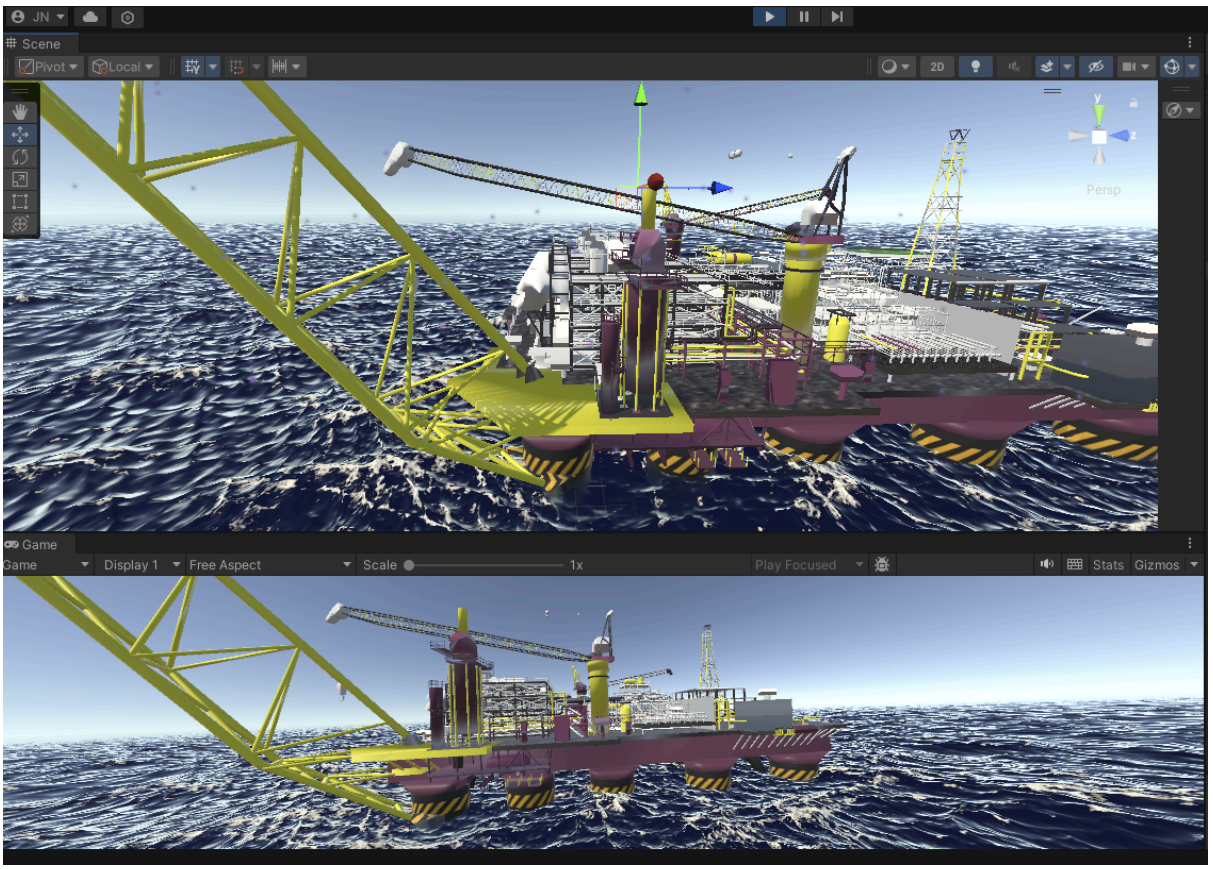


Figura 3: Ambiente de desenvolvimento com oscilação da câmera principal e chuva fraca

- **Resultados e Discussão**

Ao longo do desenvolvimento deste projeto, alguns objetivos foram alcançados, entre eles: A modelagem e integração da plataforma P40 ao ambiente de desenvolvimento, a simulação de ondas do mar permitiu um comportamento controlável e dinâmico, a simulação de condições climáticas para simulação de chuva e por fim a movimentação da câmera proporcionou um feedback visual imersivo para o usuário. Os resultados obtidos demonstram o potencial de escalabilidade do projeto de se tornar uma ferramenta útil para treinamentos e simulações de diferentes condições climáticas podendo inclusive ser escalado para situações de *Virtual Reality*(VR).

- **Conclusões**

O objetivo principal do projeto foi atingido: o de iniciar um desenvolvimento de uma simulação 3D imersiva e interativa de diversas condições climáticas para a plataforma de petróleo P40. A simulação

resultante oferece a capacidade de futuramente poder capacitar equipes e reagir a diferentes cenários climáticos em plataformas *offshore*. O estudo dos impactos de diferentes condições climáticas e seus efeitos nessa plataforma evidencia o potencial de futuras implementações e expansão do projeto, incluindo adição de outros fenômenos climáticos, como neblina, raios e trovões até exploração de tecnologias de realidade virtual para uma experiência ainda mais imersiva.

- **Referências bibliográficas**

Unity Essentials: Unity Essentials Pathway
Representation of 3D vectors and points: Vector3 - Scripting API
Unity Essentials Working with NavMesh Agents:Working with NavMesh Agents - Unity Learn
Unity Documentation Transform RotateAround: Scripting API: Transform.Rotate
Unity Documentation Mathf.Lerp: Scripting API: Mathf.Lerp
Unity Documentation Mathf.Sin: Scripting API: Mathf.Sin
Leading water system: github.com/wave-harmonic/crest

RELATÓRIO DE ATIVIDADES - PROJETO DE INICIAÇÃO CIENTÍFICA BIOINFORMÁTICA, BANCO DE DADOS E ENGENHARIA DE COMPUTAÇÃO

Título do Projeto:
Análise de Desempenho do Programa PA-Star2 no Santos Dumont e sua Aplicação em um Workflow Científico para Alinhamento Múltiplo de Sequências Selecionadas

Instituição:
Laboratório Nacional de Computação Científica

Nome do Aluno:
Kelen Brito Souza

Nome do orientador e coorientador:
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)
M.Sc. Micaella Coelho Valente de Paula (Analista de Sistemas – Petrobras, SEPAD/COTIC//LNCC, Coorientador)

Tipo de bolsa: Iniciação Científica - PIBIC

Período do relatório: 12 de Julho de 2024 a 14 Julho de 2025

1. Objetivo

Esta pesquisa tem como objetivo investigar o desempenho da nova versão do programa de bioinformática PA-Star2 no supercomputador Santos Dumont (SDumont) e no *workflow* científico no qual foi empregado. O PA-Star2, desenvolvido por uma equipe colaboradora da UnB, foi criado para resolver o problema NP-difícil de alinhamento múltiplo de sequências (AMS) utilizando CPU. A nova versão do programa apresenta melhorias de desempenho, com ênfase na otimização da divisão de tarefas em máquinas com processadores assimétricos (*Asymmetric Multicore Processors – AMPs*). Nesta pesquisa, a ferramenta foi integrada a um *workflow* que busca reduzir o custo computacional do AMS e permitir a seleção eficiente de subconjuntos de sequências. Dessa forma realiza-se inicialmente uma filtragem das sequências mais representativas, extraindo sequências específicas de interesse, contornando limitações de memória e de quantidade de sequências a serem trabalhadas. O *workflow* proposto pode beneficiar aplicações que exigem alta acurácia nos alinhamentos, como em estudos filogenômicos.

2. Introdução

A bioinformática desempenha um papel fundamental na biologia, oferecendo ferramentas computacionais essenciais para análises genômicas em larga escala. Entre essas análises, destacam-se os estudos evolutivos, como reconstrução filogenética e epidemiologia molecular, que dependem diretamente do alinhamento de sequências biológicas. Em especial, o alinhamento múltiplo de sequências (AMS)

é uma etapa crítica, pois permite identificar regiões de similaridade entre DNA, RNA ou proteínas, revelando relações funcionais, estruturais e evolutivas. No entanto, o AMS é classificado como um problema NP-difícil, que impõe grande desafio computacional, especialmente diante do crescimento exponencial dos dados.

Nesse contexto, torna-se essencial empregar estratégias eficientes, como as que visam reduzir o espaço de busca para acelerar o processo de alinhamento, com foco em garantir resultados biologicamente relevantes. A Computação de Alto Desempenho (CAD), aliada a Sistemas de Gerenciamento de Workflows Científicos (SWfMS), como o PyCOMPSs (Programming with COMPSs in Python), permite estruturar, automatizar e escalar esses processos, tornando-os mais acessíveis a pesquisadores da área biológica.

Este estudo investiga o desempenho do PA-Star2, uma versão paralela do programa baseado no algoritmo A-star (A*), desenvolvido para resolver o problema de AMS em CPUs. Os testes foram realizados na máquina Sequana, do supercomputador Santos Dumont (SDumont). Ao integrar o PA-Star2 a um *workflow* com PyCOMPSs, o trabalho também busca contribuir com uma solução eficiente e reprodutível para análises evolutivas em larga escala.

3. Metodologia

3.1. Dados de Entrada

Esta pesquisa está dividida em dois estudos complementares: o primeiro avalia o desempenho da ferramenta PA-Star2 de forma isolada, por meio de testes com arquivos de proteínas do banco BAliBASE, comparando a testes feitos com a versão anterior do programa com as mesmas entradas na última Jornada Científica do LNCC. Foram utilizadas sequências de aminoácidos (proteínas) organizadas em arquivos no formato FASTA, extraídas do banco BAliBASE e disponíveis no repositório GitHub astar-msa, por Daniel Sundfeld. Os testes iniciaram com um conjunto de arquivos aqui nomeados como iniciais, sendo eles glg, 1sbp, 1aboA e 1ac5, que apresentam menor complexidade computacional e tempo de alinhamento. Em seguida, um segundo grupo de arquivos multifasta, classificados como intermediários, devido à maior demanda computacional, foi utilizado: gal4, 1gpb, arp, 1sesA, 2myr, 2cba, 1hvA, 2ack e actin. Esses dados permitiram avaliar o comportamento do PA-Star2 frente a diferentes níveis de complexidade e consumo de recursos.

O segundo estudo investiga a aplicação do PA-Star2 integrado a um *workflow* científico, utilizando sequências biológicas do vírus da dengue (DENV) como estudo de caso. Esta segunda parte da pesquisa consistiu na aplicação do *workflow* científico com a ferramenta PA-Star2, utilizando o framework PyCOMPSs para gerenciar a execução paralela das tarefas. Foram utilizados como arquivos de entrada um *dataset* D1 de sequências biológicas, de genes do vírus da dengue (DENV), em diferentes versões: 9, 18, 28 e 38 sequências. No D1, os alinhamentos par-a-par foram realizados com o MASA OpenMP, que filtrou cinco sequências para posterior alinhamento múltiplo no PA-Star2. Os alinhamentos diretos no PA-Star2 foram possíveis apenas com um conjunto reduzido de sequências (9), evidenciando as limitações de uma ferramenta de algoritmo exato.

3.2. O Ambiente Computacional Santos Dumont

O supercomputador Santos Dumont (SDumont), do Laboratório Nacional de Computação Científica (LNCC), é uma das maiores infraestruturas de Computação de Alto Desempenho (CAD) da América Latina. Composto por 376 nós computacionais, totalizando 18.048 núcleos de CPU e 376 GPUs, o sistema é gerenciado pelo escalonador SLURM, que organiza e aloca eficientemente os recursos computacionais. O SDumont conta com um sistema de arquivos de alto desempenho baseado em Lustre (ClusterStor L300), com 1,1 PB de capacidade para armazenamento temporário, e um diretório pessoal fornecido via NFS com 650 TB. Sua arquitetura é interconectada por rede Infiniband EDR de 100 Gb/s, garantindo suporte eficiente a aplicações científicas intensivas. Nesta pesquisa, foram utilizados exclusivamente os nós de CPU do sistema.

3.3. Arquitetura da máquina Sequana

A máquina Sequana, componente principal da expansão do supercomputador Santos Dumont realizada em 2019, é baseada na arquitetura BullSequana X1000 da Atos. Essa infraestrutura de alto desempenho é composta por duas células BullSequana X1000, com capacidade teórica total de 4,0 Pflops (*RPeak*), conectadas por uma rede Infiniband EDR de 100 Gb/s, adequada para cargas computacionais intensivas. Dentro desse ambiente, os testes desta pesquisa foram conduzidos exclusivamente em nós de CPU, mais especificamente no tipo Bull Sequana X1120 (CPU), que conta com 246 nós computacionais, cada um equipado com dois processadores Intel Xeon Cascade Lake Gold 6252, totalizando 48 núcleos de processamento e 384 GB de memória RAM.

Para experimentos com maior demanda de memória, foram utilizados também nós do tipo Bull Sequana X1120 (CPU BIGMEM), que mantêm a mesma configuração de processadores e número de núcleos, porém com 768 GB de memória RAM, distribuídos em 36 nós, sendo adequados para aplicações que exigem elevado consumo de memória. Ambos os tipos de nós integram a partição identificada como Sequana, utilizada nesta pesquisa por meio das filas de CPU disponíveis no sistema de gerenciamento SLURM.

3.4. Desempenho computacional do PA-Star2 na máquina Sequana

A nova fase de testes do PA-Star2, realizada após a atualização do supercomputador Santos Dumont em 2025, permitiu uma análise detalhada do desempenho da versão aprimorada da ferramenta, o PA-Star2, na máquina Sequana, especificamente nos nós do tipo CPU BIGMEM. Essa versão incorporou otimizações significativas, resultando em uma redução expressiva no tempo de alinhamento, que chegou a ser reduzido pela metade, ou mais, em todos os arquivos avaliados, em comparação com os resultados anteriores.

A combinação entre as melhorias da ferramenta PA-Star2 e a capacidade computacional da máquina Sequana tornou possível um ganho real de desempenho, otimizando tanto o tempo de execução quanto o uso dos recursos de hardware disponíveis no SDumont. Esse avanço reforça a aplicabilidade da ferramenta em *workflows* de bioinformática mais complexos, especialmente em ambientes com infraestrutura robusta como a do SDumont.

4. Resultados e Discussões

Os testes realizados com o PA-Star2, em sua versão atualizada, permitiram uma análise do comportamento da ferramenta em diferentes arquiteturas computacionais, cenários de entrada e configurações de paralelismo. A avaliação foi conduzida em três frentes principais: escalabilidade com diferentes números de threads, desempenho frente à complexidade dos dados de entrada, e compatibilidade da ferramenta com o *workflow* bioinformático proposto, especialmente no contexto do alinhamento de sequências do vírus da Dengue (DENV).

Os testes repetidos com os mesmos arquivos de proteínas (aminoácidos) usados previamente revelaram uma redução expressiva no tempo de execução, em alguns casos superior a 50%, o que indica uma evolução clara na eficiência da ferramenta. Segue abaixo na imagem as informações de comparação dos testes.

Arquivos (tipo .fasta)	Tamanho	Nº de Seqs	Tamanho da Menor Sequência	Tamanho da Maior Sequência	Similaridade	Tempo (sequana bigmem 2024)	Tempo (sequana bigmem)	RAM (sequana bigmem 2024)	RAM (sequana bigmem)
glg	2.4K	5	438	486	26.80%	01m:58s	01m:03s	6.25GB	6.26GB
1sbp	1.3K	5	224	263	12.36%	01m:31s	00m:49s	4.28GB	3.31GB
1aboA	335	5	49	80	28.75%	01m:03s	00m:50s	5.08GB	2.90GB
1ac5	1.8K	4	421	483	25.10%	00m:59s	00m:26s	2.99GB	1.00MB
gal4	1.9K	5	335	395	15.80%	03h:43m:59s	01h:10m:18s	297.77GB	249.45GB
1gpb	4.0K	5	796	828	42.60%	01h:01m:12s	22m:29s	132.82GB	114.18GB
arp	2.1K	5	380	418	24.16%	21m:40s	08m:21s	48.78GB	39.65GB
1sesA	2.2K	5	417	442	29.87%	12m:41s	05m:13s	32.27GB	26.03GB
2myr	1.7K	4	340	474	14.94%	05m:17s	02m:02s	19.02GB	14.71MB
2cba	1.3K	5	237	259	22.15%	04m:07s	02m:00s	12.77GB	8.45GB
1hvA	928	5	136	199	14.07%	03m:35s	01m:51s	10.38GB	8.46GB
2ack	2.4K	5	452	482	18.82%	02m:14s	01m:10s	7.90GB	6.47GB
actin	2.0K	5	379	395	40.25%	01m:49s	00m:56s	6.69GB	3.83GB

Figura 1. Comparações de resultados dos testes com aminoácidos no PA-Star2

No contexto do *dataset* de DENV D1, o PA-Star2 foi utilizado como componente final do *workflow*, após filtragem das sequências por meio do MASA-OpenMP. Os resultados mostraram que, embora o PA-Star2 aceite um número limitado de entradas diretas (como verificado na rejeição dos alinhamentos com 18, 28 e 38 sequências), a integração com o *workflow* — que reduz o número de sequências para cinco por conjunto — permitiu a realização dos alinhamentos múltiplos de forma eficiente. Isso confirma que a ferramenta, embora restrita em termos de entrada direta, pode ser adaptada com sucesso para fluxos de trabalho automatizados e otimizados, desde que o volume de dados seja tratado previamente.

Portanto, os resultados indicam que o PA-Star2 representa um avanço relevante em relação à sua versão anterior, principalmente por oferecer melhor escalabilidade e tempo de execução mais eficiente. A combinação do PA-Star2 com o hardware atualizado do SDumont, particularmente os nós da máquina Sequana, representa uma solução promissora para alinhamentos múltiplos de alta demanda e complexidade, desde que as limitações da ferramenta quanto ao número de sequências de entrada sejam tratadas adequadamente por meio de estratégias de pré-processamento. Esses achados confirmam o potencial do PA-Star2 como uma ferramenta viável e eficaz em *workflow*s bioinformáticos de grande quantidade de sequências. A seguir a figura apresenta os resultados dos testes no *workflow*.

Dataset	Tipo	Quantidade de Sequências	Tempo Médio	Desvio Padrão (Tempo)	Gasto de RAM Médio	Desvio Padrão (RAM)
D1	Genes	9 sequências	1m 41s	± 23s	309.49MB	± 46.43MB
D1	Genes	18 sequências	1m 46s	± 61s	392.28MB	± 40.96MB
D1	Genes	28 sequências	1m 20s	± 23s	411.05MB	± 42.48MB
D1	Genes	38 sequências	0m 53s	± 16s	567.08MB	± 28.37MB

Figura 2. Resultados do *Workflow* resultado da integração do PA-Star2 com o MASA-OpenMP

5. Conclusões

A nova versão da ferramenta de alinhamento múltiplo de sequências, PA-Star2, demonstrou avanços significativos em desempenho quando avaliada no SDumont, na máquina Sequana. Os testes evidenciaram reduções expressivas no tempo de execução em comparação à versão anterior da ferramenta, com destaque para arquivos simples, nos quais o tempo foi reduzido pela metade ou mais. Esse ganho de desempenho se deve à atualização do programa que ao desenvolver meios de melhor divisão de trabalho acaba aproveitando melhor todas as CPUs disponíveis no nó para alinhamento.

Nos testes com arquivos de maior complexidade computacional, como aqueles classificados como intermediários, foi possível constatar a importância de ambientes com alta capacidade de memória, como a partição sequana_cpu_bigmem. Nessas condições, o PA-Star2 manteve estabilidade e bom desempenho, mesmo em cenários de elevada exigência computacional. Contudo, para alinhamentos diretos com grandes quantidades de sequências, a ferramenta ainda apresenta limitações, evidenciadas, por exemplo, pela rejeição de execuções com 18, 28 ou 38 sequências no *dataset* D1.

Para contornar essas limitações, a integração do PA-Star2 a um *workflow* automatizado, gerenciado pelo framework PyCOMPSs e utilizando o MASA OpenMP como etapa de filtragem prévia, mostrou-se uma solução eficaz. Esse *workflow* reduziu o número de sequências a serem alinhadas para cinco por conjunto, tornando viável a realização dos alinhamentos múltiplos e garantindo a qualidade biológica dos resultados.

Em síntese, os resultados da pesquisa confirmam que o PA-Star2 é uma ferramenta viável e eficaz para aplicações bioinformáticas com várias sequências, sobretudo quando combinada a estratégias de pré-processamento e execução paralela. Sua melhoria de desempenho, aliada à compatibilidade com infraestruturas modernas de alto desempenho, amplia seu potencial de uso em estudos diversos, como estudos filogenômicos e outras análises computacionalmente intensivas.

6. Referências Bibliográficas

1. SUNDFELD, Daniel; TEODORO, George; MELO, Alba Cristina Magalhães Alves. PA-Star2: Fast Optimal Multiple Sequence Alignment for Asymmetric Multicore Processors. In the 33rd Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP 2025), Torino, Italy, 2025, pp. 146-153.
2. SUNDFELD, Daniel; RAZZOLINI, Caina; TEODORO, George; BOUKERCHE, Azzedine; MELO, Alba Cristina Magalhães Alves. PA-Star: A disk-assisted parallel A-Star strategy with locality-sensitive hash for multiple sequence alignment. Journal of Parallel and Distributed Computing, v. 112, p. 154-165, 2018.
3. LIMA, Daniel Sundfeld. Alinhamento primário e secundário de sequências biológicas em arquiteturas de alto desempenho. 2017. xx, 167 f. Tese (Doutorado em Informática) — Universidade de Brasília, Brasília, 2017.
4. SDUMONT - LNCC. Manual. Disponível em: https://github.com/lncc-sered/manual-sdumont/wiki. Acesso em: 31 jul. 2025.
5. CODECADEMY. Code Documentation. Disponível em: https://www.codecademy.com/resources/docs/ai/search-algorithms/a-star-search. Acesso em: 31 jul. 2025.

PROGRAMA DE BOLSAS PCI/LNCC

Relatório Final de Atividades

Título do Projeto Proposto:

Reestruturação e Otimização de Código Científico para Simulação de Escoamento em Reservatórios

Nome do aluno:

Luiza Augusto Tavares

Orientador:

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

M.Sc. Thiago Daniel Quimas Simões Teixeira (Coorientador)

Nível de bolsa: PIBIC

Período de bolsa: Out de 2024 - Set de 2025

1. Objetivo

O presente trabalho tem como objetivo desenvolver conceitos relacionados com a pesquisa na área de simulação numérica de reservatório de Petróleo, e como objetivo , participar da reescrita de um código originalmente desenvolvido em Fortran, voltado à simulação do escoamento de óleo e gás em reservatórios, para a linguagem Python. A proposta é realizar essa reestruturação de forma progressiva, passando pelas versões unidimensional (1D) e bidimensional (2D), até alcançar a versão tridimensional (3D) final, com escalabilidade tendo sua execução em múltiplos nós e arquiteturas GPU.

Além da reescrita, o projeto visa aplicar técnicas de computação de alto desempenho (HPC), como o uso do MPI, com o objetivo de otimizar o desempenho da aplicação. Nesse contexto, os módulos responsáveis pelos cálculos mais custosos em termos de tempo de execução, como os de transporte e velocidade, serão implementados em Fortran, que é uma linguagem mais eficiente para o HPC. A integração desses módulos com a aplicação principal em Python, por meio de chamadas específicas, tornará o código mais moderno e modular, seguindo as práticas atuais do mercado.

2. Introdução

A simulação numérica de reservatórios de petróleo consiste na elaboração de modelos matemáticos, representativos da física típica de escoamentos em meios porosos, cujas soluções são aproximadas por métodos numéricos apropriados [Correa e Borges 2013]. Seu objetivo é obter um comportamento aproximado da realidade para a realização de previsões do processo de produção. As

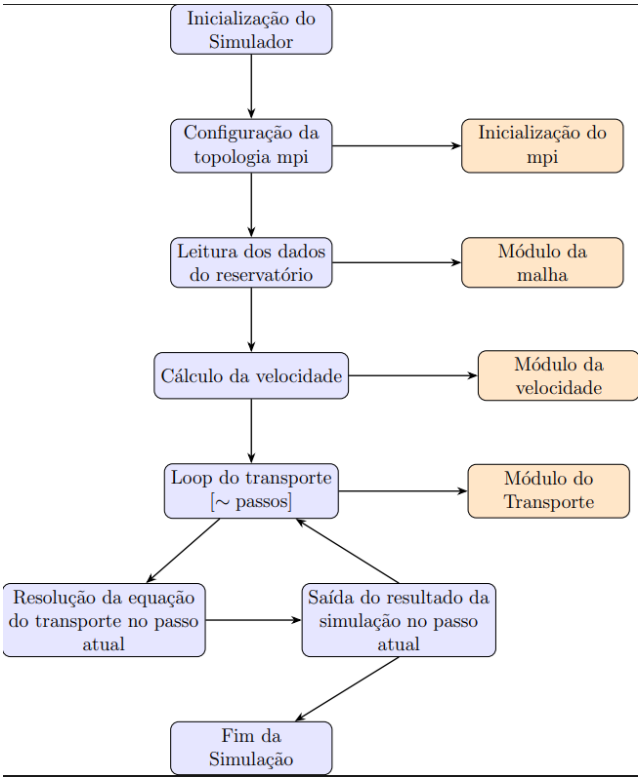
heterogeneidades, presentes nas propriedades das rochas reservatório (porosidade, permeabilidade, módulo de Young, etc.), ocorrem em todas as escalas de comprimento, desde a escala do poro até a escala de campo, que se estende por quilômetros. Tais heterogeneidades exercem marcante efeito sobre o padrão de escoamento. Portanto, em simulações típicas, precisamos discretizar domínios gigantescos (km3) em malhas suficientemente refinadas para representar tais heterogeneidades (m3), o que dá origem a problemas computacionais de grande porte que exigem computação de alto-desempenho para que os mesmos sejam factıveis em tempo razoável.

Este trabalho dá continuidade ao artigo de [Teixeira 2018], no qual foi apresentada uma metodologia de paralelização ao para o método numérico proposto em [Correa and Borges 2013]. Este método numérico envolve grande esforço computacional utilizado para resolver a equação do diferencial parcial do escoamento bifásico (água e óleo) em um reservatório rígido altamente heterogêneo [Carneiro et al. 2020]. O conjunto de equações diferenciais parciais é composto por uma equação hiperbólica não linear para o transporte dos fluidos [Chen 2007] e outra equação do tipo elíptica para o campo de velocidades [LeVeque 2007], e que juntamente com as condições inicial e de contorno permitem encontrar uma solução numérica.

Além da reestruturação, um dos objetivos centrais deste trabalho é adaptar o código para execução eficiente em ambientes de HPC, especialmente no supercomputador SDumont, localizado no LNCC. Atualmente, o SDumont conta com duas importantes arquiteturas de GPU da NVIDIA: a V100 (Volta) e a mais recente H100/GH200 (Hopper/Grace Hopper), além de arquiteturas AMD entre outras. Ambas arquiteturas NVIDIA são voltadas para aplicações científicas de alto desempenho, porém apresentam diferenças marcantes em capacidade computacional, eficiência energética e integração com a memória.

A V100, com memória HBM2 e núcleos especializados para cálculos de precisão mista (Tensor Cores), foi referência durante muitos anos na aceleração de cargas científicas. Já a H100, e em especial sua variação GH200, representa um salto tecnológico, combinando uma GPU Hopper com uma CPU Grace via interconexão NVLink-C2C, proporcionando maior largura de banda, melhor paralelismo e menor consumo energético.

Para que se possa compreender o perfil do código, é necessário antes conhecer sua estrutura básica, com cada um de seus módulos principais, suas funções e a relação entre as chamadas de funções, bem como a complexidade computacional de cada um. Os 3 módulos principais do código são o módulo da malha, o módulo da velocidade e o módulo do transporte. O Módulo da malha tem como objetivo criar a estrutura de dados do código traduzindo os dados do reservatório. O módulo da velocidade calcula a velocidade total dos fluidos a cada uma quantidade escolhida de passos no tempo, e o módulo do transporte é responsável pela resolução do cálculo da equação do transporte a cada passo no tempo.



A principal motivação para essa reescrita é aproveitar os recursos oferecidos pela linguagem Python, que possibilita maior flexibilidade e facilidade para futuras melhorias, como a adição de novos módulos e funcionalidades, além de tornar o código mais acessível e compreensível.

3. O Modelo Matemático

Seja Ω ⊂ R3 um domínio conexo, aberto e limitado e I um tempo de intervalo. Consideramos a lei de conservaçao escalar da seguinte forma: φ∂s∂t + ∇ · f = 0 in Ω × I. (1)

Onde φ : Ω ⇒ (0, φmax] e o coeficiente de armazenamento, s : Ω × I → Im{s} =[smin, smax] e a função escalar, e a funcao vetorial f : Im{s} → R3 é o fluxo da quantidade conservada s.

Particularmente, para os testes considerados neste trabalho: f = sv =s 0 0. (2)

Dados sobre condição inicial e de contorno estão detalhados em [Correa and Borges 2013].

Uma das propostas do novo código é apresentar uma modularidade que possibilite a implementação de demais métodos numéricos de primeira ordem, além do método apresentado acima, assim como permitir a implementação de métodos numéricos de mais alta ordem.

A presença do termo advectivo torna a equação acima hiperbólica. Com isso,métodos centrais podem não ser muito indicados. Uma alternativa bastante explorada é o método upwind, ou método de célula doadora. Neste, a ideia é levar em consideração a direção com a qual as informações se propagam. Sem entrar profundamente nos detalhes matemáticos, os fluxos nas faces são calculados usando os valores da solução na posição atual, ou na posição de fora (na direção da respectiva face), a depender do sinal da velocidade na face. Novamente, a depender da direção da velocidade, um desses fluxos é empregado ou o outro. Na forma implícita do método upwind observa-se convergência incondicionalmente estável, o que é muito desejado.

Contudo, o método também é conhecido por inserir difusividade numérica na resposta. Para superar esse problema, métodos alternativos como o Kurganov-Tadmor (KT), também chamado de upwind centrado, sendo este de segunda ordem, porém mais caro de ser calculado. Atualmente este método não está sendo aplicado, mas é o objetivo final em termos de discretização.

4.Paralelismo em MPI e decomposição de domínio

Conforme descrito em [Teixeira et al. em 2021] a estrutura original do código apresenta um fluxo modular bem definido, iniciando pela leitura dos dados de entrada e geração da malha, e avançando para a resolução dos problemas de pressão e transporte. Dentre esses, destaca-se o módulo de transporte, responsável pela maior parte do custo computacional. A organização da comunicação entre processos é realizada por meio da biblioteca MPI.

Já o artigo de 2020 (submetido ao WSCAD) propõe uma estratégia de decomposição de domínio baseada no fatiamento tridimensional da malha, com ênfase na minimização de *cache miss* por meio da escolha adequada da geometria do particionamento, considerando a ordenação column-major do Fortran e a hierarquia de memória cache em arquiteturas multicore.

4.1 Comunicação entre Processos com MPI

Em ambientes de memória distribuída, a comunicação entre processos é essencial para o funcionamento paralelo de aplicações científicas. O MPI (Message Passing Interface) fornece mecanismos padronizados para que os processos possam trocar informações de forma explícita, mesmo estando localizados em diferentes nós ou máquinas. Essa troca de mensagens é independente da estratégia de paralelização adotada: mesmo aplicações que não fazem uso de decomposição de domínio podem se beneficiar da comunicação via MPI, por exemplo, para sincronização ou envio de resultados parciais.

Nesse experimento, utilizando a biblioteca mpi4py no Python, o processo rank 1 ficou responsável por calcular e enviar um conjunto de variáveis importantes, como constantes físicas, parâmetros de simulação e resultados parciais (por exemplo, K,

dt, Lx, Ly, Lz, c_hist, IC, e os valores iniciais de velocidade em vxsim, vysim e vzsim). O processo rank 0, por sua vez, foi encarregado de receber esses dados usando comm.recv() e armazená-los em variáveis nomeadas para uso posterior.

A comunicação foi feita utilizando as funções comm.send() e comm.recv(), com a atribuição de tags numéricas distintas para cada variável enviada. Esse controle por tag assegura que cada dado recebido seja corretamente identificado, evitando erros de sincronização e colisões entre mensagens. Depois me foi proposto o desafio de verificar uma forma eficiente, que seria enviar somente uma mensagem com todo conteúdo, resolvi este problema com a atribuição de **uma única variável**, com todo o resultado armazenado nela, chamado de: dicionário, deixando o código mais simples e eficaz.

4.2 Decomposição de Domínio

A decomposição de domínio é uma técnica central na computação de alto desempenho, especialmente em simulações científicas que envolvem grandes volumes de dados ou malhas tridimensionais extensas. Essa abordagem consiste em dividir o domínio espacial de um problema, como uma malha que representa um reservatório, um fluido ou uma estrutura física, em subdomínios menores, os quais são atribuídos a diferentes processos. Com isso, cada processo realiza os cálculos relativos à sua região de forma independente, possibilitando a execução paralela e o uso mais eficiente dos recursos computacionais, como memória e núcleos de CPU.

Essa técnica se mostra particularmente útil em aplicações que demandam alto poder computacional, como na modelagem numérica de reservatórios de petróleo e gás. Ao permitir que o código seja executado em arquiteturas distribuídas como clusters e supercomputadores, a decomposição de domínio se torna fundamental para garantir escalabilidade e desempenho.

Compreender os fundamentos e estratégias de decomposição está sendo crucial para a implementação dessa função atual que ainda está em seus estágios iniciais, dividindo a malha em apenas uma dimensão. O estudo da melhor forma de divisão de domínio ainda está em andamento, e estou aprendendo as metodologias para buscar uma maior eficiência e escalabilidade ao aplicar diferentes formas de divisões em diferentes tamanhos de reservatórios.

5. Conclusão e trabalhos futuros

A reescrita do código está progredindo de forma estruturada, respeitando o ritmo necessário para consolidar o conhecimento em Python e MPI. A abordagem gradual tem sido fundamental para a construção de uma base sólida em programação científica e computação de alto desempenho. Ao começar com exemplos simples e bem delimitados, é possível compreender com mais clareza os fundamentos da comunicação entre processos e a lógica envolvida na paralelização de tarefas e conhecimentos essenciais para o desenvolvimento de aplicações mais complexas e eficientes.

Nas próximas etapas, será aprofundado o uso de técnicas de decomposição de domínio, explorando diferentes estratégias e comparando o código antigo com a nova versão. Também serão incorporados métodos de análise e medição de desempenho, utilizando ferramentas apropriadas, como os perfiladores, para identificar gargalos e oportunidades de otimização. O objetivo do estudo é dar continuidade à construção de um código escalável e eficiente, aproveitando os recursos das arquiteturas de HPC, com o objetivo final de um código escalável em múltiplos nós e a execução de seus módulos mais complexos em placas GPU.

6.Referências

[1] MPI4PY Documentation. Disponível em: https://mpi4py.readthedocs.io. Acesso em: 21 jul. 2025.

[2] SOUSA, Fabrício S.; ROCHA, Franciane F. *Métodos de volumes finitos para modelagem computacional de reservatórios de petróleo*. Notas em Matemática Aplicada, nº 96. São Carlos: SBMAC, 2023. Acesso em: 29 jul. 2025.

[3] RIBEIRO, Weber Guilherme Dias. *Estratégias de paralelização para malhas estruturadas: análise da decomposição de domínios e escolhas de geometrias*. 2024. 62 f. Dissertação (Mestrado em Modelagem Computacional) – Laboratório Nacional de Computação Científica. Acesso em: 31 jul. 2025.

[4] TEIXEIRA, Thiago; CABRAL, Frederico L.; COELHO, Micaella; LEITE, Luciano; SURMAS, Rodrigo; BORGES, Márcio; OSTHOFF, Carla. *Estudo de desempenho e de eficiência energética em simulação de dinâmica de fluidos multifásicos nas arquiteturas NVIDIA Volta V100 e Grace Hopper GH200*. LNCC, 2024. (Artigo técnico interno).

[5] HARRISON, Stiw; TEIXEIRA, Thiago; CABRAL, Frederico L.; SOUTO, Roberto P.; BORGES, Márcio R.; OSTHOFF, Carla. *Análise de uma estratégia de decomposição de domínio para a execução no supercomputador SDumont de um método numérico para escoamentos de fluidos*. LNCC, 2020. (Artigo submetido ao WSCAD).

[6] RIBEIRO, Weber; TEIXEIRA, Thiago; CABRAL, Frederico; BORGES, Márcio; OSTHOFF, Carla. *Otimização para ambientes Intel® de um método numérico para o escoamento bifásico de fluidos em meios porosos através da eliminação de barreiras OpenMP*. WIC 2019. LNCC.

[7] HERRERA, Stiw; RIBEIRO, Weber; TEIXEIRA, Thiago; CARNEIRO, André; CABRAL, Frederico; BORGES, Márcio; OSTHOFF, Carla. *Avaliação de desempenho no supercomputador SDumont de uma estratégia de decomposição de domínio usando as funcionalidades de mapeamento topológico do MPI para um método numérico de escoamento de fluidos*. ERAD 2020. LNCC.

[8] CORREA, M.; BORGES, M. *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*, 73(3):205–224, 2013.

[9] CARNEIRO, I.; BORGES, M.; MALTA, S. *Numerical simulation of two-phase flows in heterogeneous porous media*. *TEMA (São Carlos)*, 21(2):339, 2020.

[10] TEIXEIRA, T.; CABRAL, F.; OSTHOFF, C.; BORGES, M.; SURMAS, R.; SOUTO, R. *Analysis of optimization opportunities for Intel Xeon Phi and Intel Xeon Scalable processors environments of a numerical method for the biphasic flow of fluids in porous media*. In: *Anais Estendidos do XX Simpósio em Sistemas Computacionais de Alto Desempenho*, p. 237, 2018.

[11] CHEN, Z. *Reservoir simulation: mathematical techniques in oil recovery*. SIAM, 2007.

[12] LEVEQUE, R. J. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

Relatório de Atividades

Submetido ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC-LNCC) Edital 2024/2025

Título:

Desenvolvimento de uma ARP para monitoramento de áreas: estação solo, veículo com piloto automático, transmissão de dados e telemetria

Bolsista:

Marcos Paulo de Souza Campanha

Orientadores:

Jauvane Cavalcante de Oliveira (LNCC), PhD

Paulo Fernando Ferreira Rosa (IME), PhD

RIO DE JANEIRO
2025

1 Objetivos

Este projeto teve por objetivo estudar e aplicar conhecimentos das áreas de visão computacional, automação e sistemas embarcados no desenvolvimento de Aeronaves Remotamente Pilotadas (ARPs) autônomas, alinhadas aos requisitos técnicos de competições de robótica, como a RoboCup Brasil Flying Robot League[1].

Mais especificamente, buscou-se: (i) utilizar as ARPs já existentes, montadas previamente em laboratório, para implementar e testar sistemas de navegação visual autônoma; e (ii) projetar e construir novas plataformas aéreas com estrutura reforçada e maior capacidade computacional, visando à operação coordenada em enxame e ao transporte de carga leve, consolidando uma base tecnológica robusta no contexto da robótica aérea aplicada.

2 Introdução

Aeronaves Remotamente Pilotadas (ARPs), popularmente conhecidas como drones, constituem plataformas versáteis para aplicações de monitoramento, mapeamento, vigilância e inspeção. Seu uso em ambientes estruturados tem crescido devido à agilidade de implantação, ao baixo custo relativo e à mitigação do risco humano em operações críticas [2].

O presente projeto se insere nesse contexto, com o propósito de desenvolver uma frota de ARPs autônomas capazes de executar missões como reconhecimento de área, localização de alvos e transporte de cargas. A estratégia adotada consistiu em dividir o desenvolvimento em duas fases: a primeira foi dedicada à utilização das ARPs já existentes, com foco em navegação visual baseada em câmera, explorando algoritmos como o ORB-SLAM2 para odometria visual e reconstrução de mapa, e o YOLOv5 para detecção de objetos relevantes à missão, e controle embarcado via protocolo MAVLink; e a segunda concentrou-se na construção, integração e testes de novas ARPs mais robustas e no desenvolvimento de diferentes *scripts* para coordenação e navegação, de modo a torná-las preparadas para operações em enxame e missões cooperativas.

O projeto teve também como objetivo estratégico contribuir para o fortalecimento da Base Industrial de Defesa e para o avanço da linha de pesquisa em robótica aérea do Instituto Militar de Engenharia (IME), por meio da consolidação de uma infraestrutura experimental voltada à formação de recursos humanos qualificados e ao desenvolvimento de soluções aplicadas em contextos reais e simulados.

3 Metodologia

3.1 Fase 1: Utilização das ARPs existentes com foco em navegação visual

Na primeira fase do projeto, foram utilizadas ARPs previamente montadas no laboratório, compostas por armações em fibra de carbono, motores *brushless*, controladores Pixhawk e computadores embarcados (Raspberry Pi). Essas aeronaves serviram como plataforma de testes para a implementação de um sistema de navegação visual totalmente embarcado, integrando percepção visual, estimativa de pose e controle automático de voo.

O pipeline de visão foi desenvolvido com os seguintes componentes principais:

- O algoritmo ORB-SLAM2 foi empregado para realizar odometria visual e reconstrução esparsa do ambiente, permitindo estimar a pose 6D (posição e orientação) da aeronave em tempo real;
- A rede neural YOLOv5 foi utilizada para realizar a detecção de objetos relevantes na cena, como plataformas de pouso;
- Um módulo de fusão geométrica foi responsável por converter as coordenadas da imagem (centroides (u, v) dos objetos detectados) em coordenadas espaciais (x, y, z) no sistema de referência do SLAM, a fim de localizar os alvos no mundo;
- Scripts em Python foram desenvolvidos para enviar comandos de controle ao ArduPilot via protocolo MAVLink, incluindo decolagem, navegação por *waypoints*, aproximação de alvos e pouso.

A comunicação entre os módulos foi realizada via rede Wi-Fi (TCP/IP), com suporte alternativo por enlace serial (115 kbps) para redundância e testes de confiabilidade. A arquitetura geral de comunicação e controle embarcado e o *pipeline* de visão computacional adotados podem ser vistos na Figura 1.

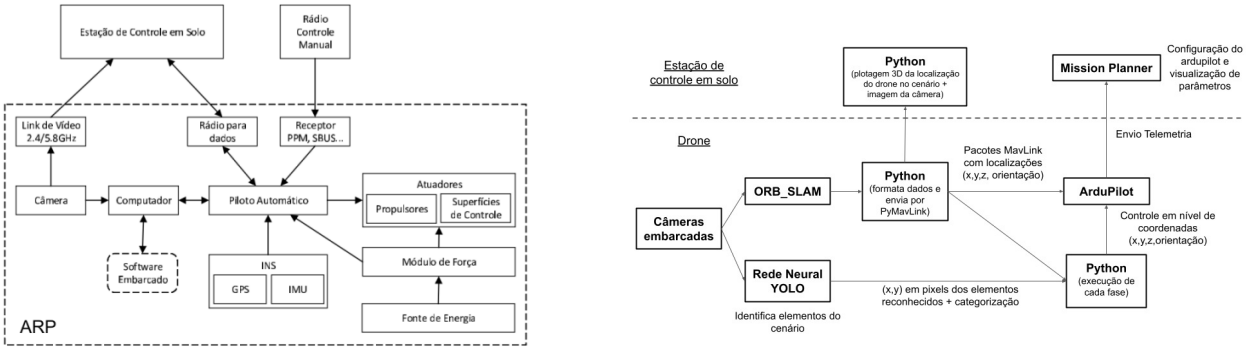
3.2 Fase 2: Construção de novas ARPs e realização de testes

Na segunda fase do projeto, foram projetadas e construídas três novas ARPs, com foco em robustez estrutural, confiabilidade e capacidade de operação em grupo. Os materiais e procedimentos adotados incluíram:

- Estrutura reforçada em fibra de carbono, com protetores de hélice e trens de pouso elevados, visando maior segurança em ambientes internos e externos;
- Instalação de controladores Pixhawk para controle de voo e integração com sensores inerciais;

- Fixação de Raspberry Pi em cada aeronave, destinada ao processamento embarcado de visão computacional em etapas futuras;

A montagem foi realizada integralmente no laboratório, contemplando desde a seleção dos componentes até a soldagem, fixação mecânica, organização dos cabos e configuração de segurança dos firmwares. Após os primeiros testes de bancada, foram realizados voos de validação com execução de comandos autônomos, incluindo decolagem, deslocamento entre *waypoints* e acionamento de atuadores embarcados.



(a) Arquitetura de comunicação e controle embarcado

(b) Arquitetura da visão computacional

Figura 1: Arquiteturas utilizadas: (a) comunicação/controle do drone e (b) pipeline de visão computacional.

4 Resultados e Discussão

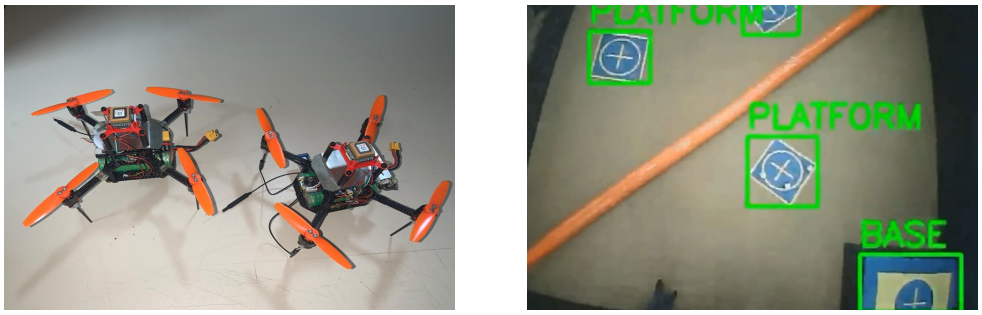
4.1 Fase 1: Validação qualitativa da navegação visual embarcada

Os testes com o drone da primeira fase confirmaram a viabilidade de uma navegação totalmente baseada em visão. O ORB-SLAM2 manteve o mapa e a estimativa de pose estáveis ao longo de trajetórias repetidas no laboratório, mesmo sob variações moderadas de iluminação. Paralelamente, a YOLOv5 identificou de forma consistente as plataformas de pouso e outros marcadores visuais relevantes. A combinação desses módulos permitiu executar pousos autônomos com boa repetibilidade e sem perdas de localização críticas. O encadeamento dos componentes encontra-se ilustrado na Figura 1b, enquanto a plataforma utilizada e exemplos de detecção aparecem na Figura 2.

4.2 Fase 2: Construção e ensaios das novas ARPs

As três aeronaves construídas apresentam estrutura em fibra de carbono, protetores de hélice e trem de pouso elevado. Em voos de ensaio, mantiveram estabilidade durante percursos entre *waypoints* predefinidos e responderam corretamente aos comandos de decolagem, aproximação e pouso. Além disso, foi feita uma demonstração institucional, em que um drone FPV realizou reconhecimento aéreo transmitindo imagens dos alvos, enquanto dois drones de ataque deslocaram-se até as posições designadas e liberaram cargas de forma coordenada. Tais ensaios evidenciam que o sistema é robusto e atende ao requisito de cumprir missões a partir de instruções de alto nível. Vistas das novas ARPs encontram-se na Figura 3.

Os avanços alcançados nas duas fases estabelecem uma base sólida para pesquisas futuras em navegação autônoma, operação em enxame e logística aérea de pequena escala, além de preparar a equipe para participação nas próximas edições da LARC.



(a) Drone utilizado na fase 1 (b) YOLOv5 em operação

Figura 2: Drones utilizados na fase 1 e exemplo de detecção de alvos.

(a) Vista 1 da nova ARP (b) Vista 2 da nova ARP

Figura 3: ARPs construídas na Fase 2, com estrutura reforçada e maior capacidade embarcada.

5 Conclusões

O projeto alcançou avanços significativos no desenvolvimento de uma plataforma de drones robusta e versátil para aplicações de monitoramento, controle de carga e operação em enxame. A montagem das aeronaves, o desenvolvimento de scripts de controle e o planejamento da integração de sensores visuais consolidam a base para participação nas competições nacionais e internacionais de robótica. Espera-se que os resultados obtidos contribuam para o fortalecimento da linha de pesquisa em robótica do IME e de suas aplicações nas Forças Armadas e na sociedade.

Referências

[1] RoboCup Brasil Flying Robot League. *Edital Desafio de Drones 2025.*

[2] LUIS CLAUDIO BATISTA DA SILVA. “SISTEMA DE AERONAVES REMOTAMENTE PILOTADAS COM ALOCAÇÃO DINÂMICA POR CONTROLE BASEADO NA TAREFA PARA COBERTURA DE ÁREAS COM DIFERENTES PRIORIDADES DE INTERESSE”. Diss. de mestr. INSTITUTO MILITAR DE ENGENHARIA, 2018.

Relatório de Atividades

TÍTULO DO PROJETO: Análise de Desempenho com Intel VTune Profiler

Bolsista: Mariana Aguiar Ribeiro

Orientador: Carla Osthoff

Co orientador: Thiago Teixeira

Tipo de bolsa: PIBIC

Período do relatório: Julho de 2024 a Julho 2025

1. Objetivo

Este trabalho tem como objetivo geral analisar como uma típica aplicação da área de "Computação de Alto Desempenho", mais especificamente, da área de "óleo e gás", utiliza recursos computacionais e como o algoritmo desta aplicação pode ser alterado de forma a que o consumo computacional possa ser otimizado. Para atingir este objetivo, o presente relatório tem como objetivo secundário fazer um estudo de perfilagem sobre o comportamento de desempenho de um código desenvolvido em Fortran, voltado à simulação do escoamento de óleo e gás em reservatórios, paralelizado através da biblioteca MPI e implementando uma configuração de decomposição de domínio em um nó computacional com processador multicore. O trabalho utiliza a ferramenta Intel VTune Profiler focando na identificação de gargalos de desempenho, fornecendo recursos para possíveis otimizações futuras no código.

2. Introdução

Simulações da área de óleo e gás precisam de novas técnicas de computação de alto desempenho, para poder lidar com a grande quantidade de alocação de dados e com o alto custo computacional que obtemos do método numérico. Assim, a técnica de decomposição de domínio ou divisão de domínio foi aplicada a um domínio que representa um reservatório de petróleo tridimensional da forma retangular. Existem várias formas de aplicar esta técnica. Neste trabalho estamos analisando um código escrito em FORTRAN que realiza uma decomposição geométrica das dimensões do reservatório e passando essas informações para cada processo MPI (Message Passing Interface) que é executado numa topologia unidimensional, bidimensional ou tridimensional. Essa topologia é definida de acordo ao tipo de particionamento geométrico realizado na fase inicial. Assim, os arquivos que contêm as informações de cada subdivisão do reservatório de petróleo é lida por cada processador e são criadas localmente todas as variáveis de processamento e preenchidas com as funções dos módulos implementados.

O trabalho [Ribeiro] desenvolveu uma análise sobre as partições tridimensionais de divisões de domínio que geram cargas computacionais equivalentes às 48 cores computacionais utilizadas para o nosso experimento. Em ambientes de execução paralela, principalmente aqueles com um grande número de processos, gargalos de desempenho podem atrapalhar a eficiência e o tempo de resposta dos códigos, sendo assim, o uso de ferramentas de análise é essencial. O perfilador "Intel VTune Profiler", é uma dessas ferramentas em que pode-se obter estudos detalhados sobre a utilização dos recursos computacionais do código em sua execução.

A decomposição de domínio surge como uma ferramenta essencial na computação de alto desempenho, especialmente em simulações de óleo e gás. O principal motivo é que ela ajuda a lidar com o grande volume de dados e o alto custo de processamento que essas simulações envolvem.

Basicamente, essa estratégia consiste em pegar um domínio de simulação extenso e dividi-lo em várias partes menores, ou subdomínios. Essa divisão é fundamental para que problemas computacionais complexos possam ser resolvidos em um tempo razoável, aproveitando ao máximo a capacidade de memória e a velocidade dos supercomputadores. Para implementá-la, é comum usar o Message Passing Interface (MPI), que permite organizar a comunicação entre essas partes, garantindo que as informações sejam trocadas apenas nas bordas dos subdomínios, muitas vezes através de uma decomposição em blocos. O trabalho desenvolvido em [Herrera] apresentou uma metodologia para implementar uma decomposição de domínio que gera sub-domínios que apresentam uma mesma carga computacional entre múltiplos processos MPI, comumente chamado de "carga computacional balanceada". O presente trabalho tem como objetivo comprovar que a metodologia desenvolvida em [Herrera] gera subdomínios com carga computacional balanceada. Para isto, este trabalho usa o Vtune para analisar a utilização dos recursos computacionais do código do perfilador para a decomposição de dominio de 48 processos MPI 288x288x288_1x1x48_288x288x6, em um nó computacional de 48 cores.

O perfilador "Intel VTune Profiler", desenvolvido pela Intel, permite obter informações sobre o tempo total decorrido, uso efetivo da CPU, restrição de memória, grau de vetorização. A partir dessas informações o desenvolvedor pode identificar gargalos e fazer modificações no código-fonte ou nas condições de execução. Bem como, traz uma plataforma visual muito clara e comparações entre diferentes execuções, facilitando a análise.

Neste relatório, será apresentada uma visão geral do funcionamento do VTune, os passos necessários para o perfilamento de um código e a interpretação dos principais indicadores fornecidos pela ferramenta. Resultados obtidos com a execução do código em Fortran, serão utilizados como exemplo prático ao longo do estudo.

Foi utilizado uma malha numérica estruturada, cúbica e regular, o que otimiza o desempenho computacional. A divisão dessa malha gera sub-malhas, o que permite o processamento paralelo eficiente entre múltiplos núcleos. Cada sub-malha realiza seus cálculos de forma independente, mas precisa trocar informações nas faces comunicantes para manter a consistência da solução.

3. Metodologia

Inicialmente, foi feito um estudo sobre o perfilador VTune onde pode-se aprender a como manuseá-lo, ganhando conhecimento sobre seu ambiente gráfico e sobre a análise de seus resultados. Foi utilizado então o ambiente computacional do supercomputador Santos Dumont, que possui diversas arquiteturas computacionais, a arquitetura "Sequana Dev" foi escolhida para este estudo por questões de praticidade, dado que as demais arquiteturas ficaram indisponíveis durante a maior parte do desenvolvimento deste estudo.

A arquitetura "Sequana Dev" é voltada ao desenvolvimento e execução de aplicações em processamento paralelo, a mesma é composta por nós computacionais que contém GPUs NVIDIA Tesla V100 com processador Intel Xeon Cascade Lake, sendo muito importantes para trabalhos científicos de grande escala. No contexto deste trabalho, o código analisado não possui versão para GPU, de forma que as análises foram desenvolvidas apenas para o processador Intel Xeon Cascade Lake.

O nó computacional da Sequana Dev possui processadores Intel Xeon Cascade Lake, com 24 núcleos e 48 threads, suportam memória DDR4 de alta velocidade e oferecem frequência turbo de até 3,7GHz. Permitindo análises detalhadas de desempenho com ferramentas como o Intel VTune Profiler.

Após os códigos serem executados, suas saídas foram coletadas e processadas pelo VTune, sendo possível obter informações importantes como o Tempo Total de Execução (Elapsed Time), onde é possível identificar o tempo de execução das funções, tais como função de transporte, velocidade e inicialização. Faz-se necessário sua análise para que seja possível entender os *hotspots*, ou seja, pontos críticos ali presente, também compreender qual função apresenta a pior e melhor performance, para um entendimento mais aprimorado do próprio código.

Em sistemas com múltiplos processadores físicos, como o analisado neste caso, cada "socket" representa um desses processadores. A medição da carga de trabalho por socket permite avaliar se o código está distribuindo a execução de forma equilibrada entre os dois processadores. Uma utilização desequilibrada pode mostrar a necessidade de otimização, como a restrição do uso a apenas um socket para reduzir *overheads* ou a melhoria na distribuição paralela da carga.

A Utilização Efetiva da CPU indica o quanto da capacidade total de processamento está sendo realmente aproveitada, considerando todos os núcleos e threads disponíveis. Sua análise permite identificar se o código executado está conseguindo explorar o potencial da máquina de forma eficiente ou se há limitações que impedem o uso completo do hardware.

Mostrando quantos núcleos estavam ativos durante toda a execução, a Utilização Média da CPU, mostra uma análise complementar à utilização efetiva, apontando, de forma mais geral, se o código está sendo bem escalonado. Valores baixos podem revelar desperdício de recursos computacionais ou limitação no modelo de paralelismo usado.

O *Memory Bound*, ou, Cache Miss, apresenta quanto o desempenho está sendo limitado pelo acesso à memória. Quando um código é "memory bound", significa que a CPU frequentemente espera dados serem carregados da RAM, em vez de processar continuamente. Observar esse comportamento é essencial para melhorar o desempenho, ajustando padrões de acesso ou estruturas de dados.

A vetorização é uma técnica que permite executar múltiplas operações de dados simultaneamente em um mesmo núcleo computacional. Antes mesmo da paralelização entre processos, ela já contribui para a eficiência do código, aproveitando melhor os pipelines internos do processador. Um alto percentual de vetorização indica que o compilador conseguiu traduzir operações matemáticas em instruções otimizadas, o que gera ganhos significativos de desempenho.

4. Resultados e Discussões

Os resultados objetivos demonstram na prática a importância da análise em cada segmento pelo perfilador. A malha utilizada para este estudo é uma malha sintética, possuindo uma dimensão 384x384x384, que a um tamanho típico utilizado em simulações de malhas reais, isso significa que o domínio tridimensional foi dividido em 384 pontos em cada direção: x, y e z. Para possibilitar a execução paralela, essa malha foi particionada em sub-malhas menores, contendo mesma carga computacional, distribuídas entre diversos processos de execução. Nos testes realizados, utilizaram-se configurações com 24 e 48 processos, que correspondem ao número de cores em um e dois sockets do nó de processamento estudado, cada um responsável por calcular uma sub-região da malha. Por fim, cada processo MPI é responsável por uma sub-malha, neste estudo foi utilizada a sub-malha: 288x288x288_1x1x48_288x288x6 (48 processos), para ilustrar cada parte da análise.

A análise da execução da malha demonstrou um tempo de execução total de 17.526 segundos. Esse valor representa o tempo decorrido entre o início e o fim da simulação com a configuração especificada, sendo uma medida direta da performance geral do código naquele cenário.

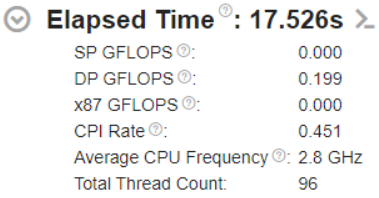


Imagem 1: tempo total de execução.

Observou-se que há dois processadores trabalhando com a porcentagem de utilização equilibradas, o socket 0 estava utilizando 37.5%, isso equivale a 9.005 processos de um total de 24. Enquanto o socket 1 estava utilizando 38.1% de sua capacidade, equivalente a 9.149 processos de um total de 24.

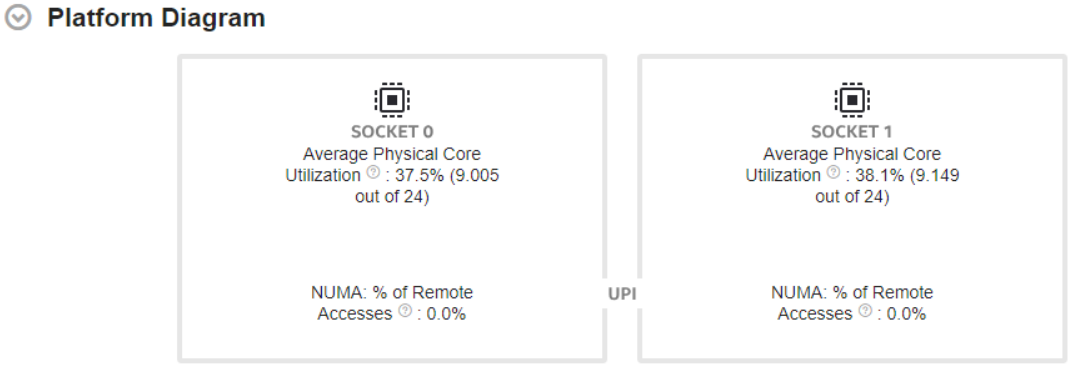


Imagem 2: Processadores: socket 0 e socket 1.

A utilização efetiva da CPU foi de 37.8%, com uma média de utilização de 18.155 de 48. O código não está totalmenteparalelizando de forma eficiente para aproveitar todos os núcleos disponíveis. Pode haver limitações no balanceamento de carga, gargalos de memória, ou mesmo regiões do código que ainda são executadas de forma sequencial. O histograma de utilização da CPU mostra que a principal causa dessa subutilização é o tempo significativo em que poucas ou nenhuma CPU está ativa (um pico em 0 CPUs), sugerindo períodos extensos de ociosidade, possivelmente devido a gargalos.

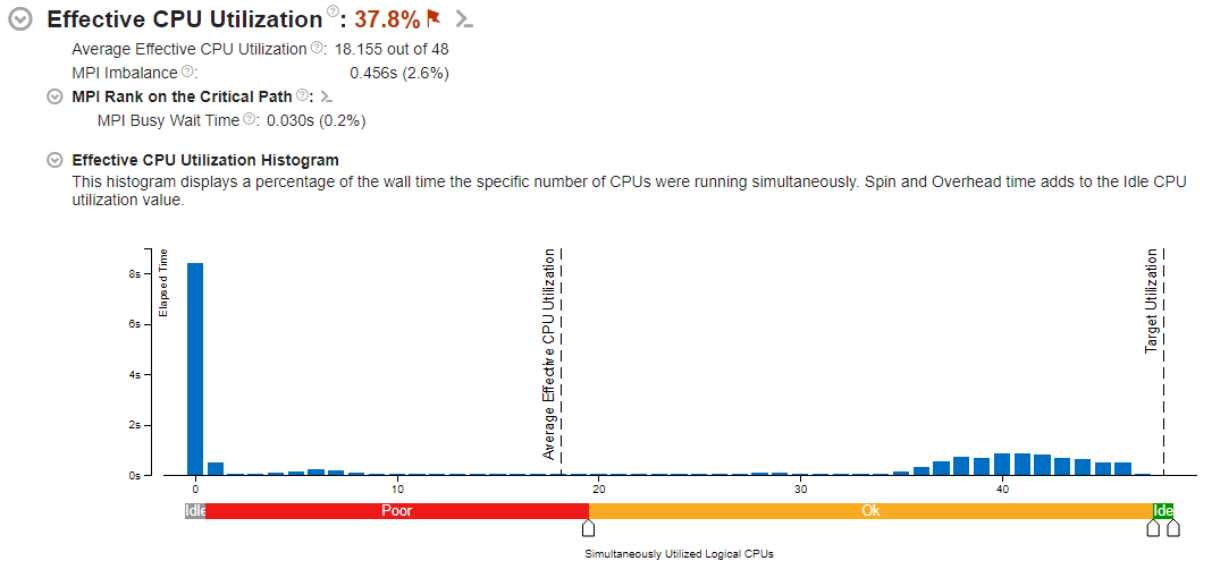


Imagem 3: Utilização Efetiva de CPU.

A análise mostra que 23,4% do tempo de execução foi comprometido com espera por acesso à memória, indicando que o código é memory bound. Dentre esses gargalos, 3,1% dos *clockticks* foram afetados por limitações no uso de cache, enquanto 7,3% foram atribuídos a esperas por acesso à memória DRAM. Esses valores sugerem que o desempenho do código está sendo limitado, em parte, pelo mau gerenciamento da hierarquia de memória.

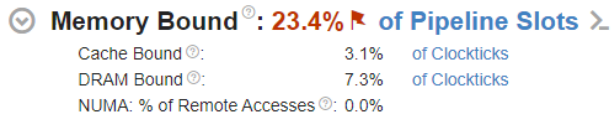


Imagem 4: Memory Bound, ou, Cache Miss.

Em conclusão, 27.3% das operações de ponto flutuante foram realizadas com instruções vetoriais. Embora esse valor representa um certo grau de paralelismo interno, ele ainda está abaixo do ideal para simulações com alto desempenho atrelado à vetorização, como simulações numéricas, sugerindo que há espaço para otimização.

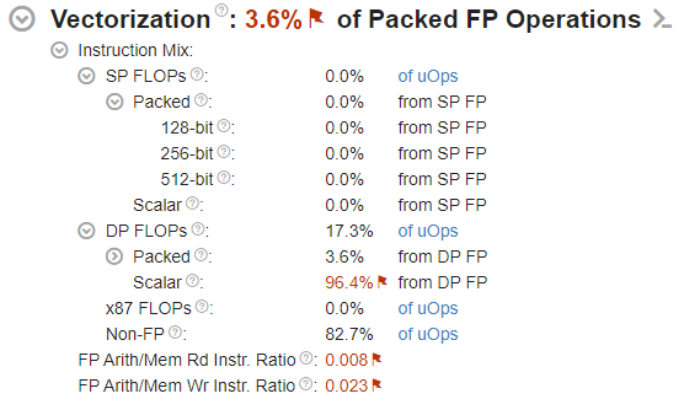


Imagem 5: Vetorização.

5. Conclusões e Estudos Futuros

Diante do exposto, foi possível observar que o uso do Intel VTune Profiler ao longo deste estudo permitiu uma compreensão mais aprofundada do comportamento do código. Permitiu identificar gargalos importantes no desempenho do código em Fortran, mesmo em um ambiente de alto desempenho como o Santos Dumont. A baixa utilização da CPU, a presença de esperas por memória e a vetorização limitada apontam para oportunidades claras de otimização, como por exemplo, implementação de técnicas de "tilling"[referência]. A adaptação ao uso da ferramenta também agregou conhecimento técnico relevante. Para o estudo futuro a análise será repetida nas demais decomposição de domínio para comprovar que a metodologia desenvolvida em [Herrera] apresenta balanceamento de carga. Dessa forma, será possível verificar se a divisão do domínio está resultando em um balanceamento adequado da carga de trabalho e se o tempo dedicado à comunicação entre processos continua significativamente inferior ao tempo gasto com os cálculos, mesmo com o aumento no número de núcleos utilizados.

6. Referências Bibliográficas

- LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA (LNCC). *Supercomputador Santos Dumont*. Disponível em: https://sdumont.lncc.br/. Acesso em: 31 jul. 2025.
- NVIDIA. *NVIDIA Tesla V100 GPU*. Disponível em: https://www.nvidia.com/en-gb/data-center/tesla-v100/. Acesso em: 31 jul. 2025.
- INTEL CORPORATION. *8 ways to maximize applications with VTune Profiler*. 2021. Disponível em: https://www.intel.com/content/www/us/en/developer/articles/technical/8-ways-to-maximize-applications-with-vtune.html. Acesso em: 31 jul. 2025.
- OSTHOFF, Carla et al. A arquitetura do supercomputador Dumont e os desafios da pesquisa brasileira na área de computação de alto desempenho. In: Escola Regional de Alto Desempenho de São Paulo (ERAD-SP). SBC, 2020. p. 1-5.
- RIBEIRO, Weber Guilherme Dias. Estratégias de paralelização para malhas estruturadas: análise de decomposição de domínios e escolhas de geometrias. 2024.
- LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA (LNCC). *General overview of Santos Dumont architecture*. Disponível em: https://sdumont.lncc.br/media/01_General_overview_of_SANTOS_DUMONT_architecture.pdf. Acesso em: 31 jul. 2025.
- HERRERA, Stiw et al. Avaliação de Desempenho no Supercomputador SDumont de uma Estratégia de Decomposição de Domínio usando as Funcionalidades de Mapeamento Topológico do MPI para um Método Numérico de Escoamento de Fluidos. In: Escola Regional de Alto Desempenho do Rio de Janeiro (ERAD-RJ). SBC, 2020. p. 31-35.
- TEIXEIRA, Thiago et al. Estudo de desempenho e de eficiência energética em simulação de dinâmica de fluidos multifásicos nas arquiteturas NVIDIA Volta V100 e Grace Hopper GH200. In: Simpósio em Sistemas Computacionais de Alto Desempenho (SSCAD). SBC, 2024. p. 240-251.

RELATÓRIO DE ATIVIDADES

Título do projeto	Desenho de Substâncias Ativas Usando Abordagens *de novo* e Ancoramento Molecular
Bolsista	Pedro Lucas Ferreira Cruz da Mota Mendes
Orientadores	Laurent E. Dardenne, Isabella Alvim Guedes e Matheus M. P. da Silva
Tipo de Bolsa	PIBIC/LNCC
Período do Relatório	02/2025 - 09/2025

1. INTRODUÇÃO

O desenho de fármacos *de novo* é uma estratégia computacional para criação automatizada de novas moléculas de acordo com um ou mais objetivos desejados. Um dos desafios mais complexos do desenho *de novo* é a otimização multiparamétrica, ou seja, o sucesso de abordagens generativas geralmente depende de sua habilidade em equilibrar uma série de objetivos e restrições de forma simultânea.

2. OBJETIVOS

Este trabalho tem como objetivo descrever a importância das propriedades moleculares no contexto de abordagens de planejamento de fármacos *de novo*. Além disso, este trabalho tem como objetivo a implementação do cálculo de algumas dessas propriedades moleculares através da plataforma RDKit, na linguagem de programação Python, assim como a análise da distribuição dessas propriedades usando um conjunto de fármacos aprovados pela FDA (Food and Drug Administration).

3. METODOLOGIA

A análise realizada neste trabalho foi baseada em uma metodologia computacional, empregando ferramentas de código aberto que são amplamente utilizadas na pesquisa em bioinformática. Foram calculadas 6 propriedades, dentre elas o peso molecular, o LogP, os aceptores de hidrogênio, os doadores de hidrogênio, o TPSA e o Csp3. O cálculo foi realizado com 2582 moléculas, estas que estão presentes no conjunto de moléculas aprovadas até 2023 pelo FDA, que é uma agência federal dos EUA, responsável pela segurança e supervisão de vários elementos relacionados à saúde pública. A linguagem de programação Python foi escolhida devido à sua versatilidade, e uma ampla gama de bibliotecas científicas. A biblioteca usada para o cálculo das propriedades foi a RDKit, uma biblioteca de quimioinformática completa que permite a manipulação de estruturas moleculares e o cálculo de várias propriedades. Para a organização e manipulação dos dados gerados, a biblioteca pandas foi utilizada.

A representação das estruturas moleculares foi realizada através do formato SMILES (*Simplified Molecular-Input Line-Entry System*). Esta notação linear permite codificar a uma

molécula, incluindo sua conectividade atômica, em uma string de texto ASCII. Por exemplo, a molécula de etanol é representada como CCO.

As representações SMILES foram usadas para criar representações moleculares usando a biblioteca RDKit. A partir deste objeto molecular, uma série de propriedades puderam ser calculadas, como por exemplo, o cálculo do peso molecular (MW), que soma as massas de todos os átomos constituintes. O MW, impacta diretamente a taxa de difusão passiva; moléculas grandes, tipicamente acima de 500 Da (LIPINSKI, 1997), enfrentam uma barreira física significativa para atravessar o epitélio intestinal.

O coeficiente de partição (LogP) é uma propriedade chave das moléculas que influencia sua absorção, distribuição, metabolismo e excreção. O comportamento de um fármaco no corpo é inicialmente regido por um equilíbrio entre sua afinidade por ambientes líquidos, como o plasma sanguíneo, e ambientes lipídicos, como as membranas celulares. O LogP é uma medida que quantifica a tendência de um composto se distribuir entre um solvente lipofílico (geralmente octanol) e um meio aquoso, refletindo seu grau de lipofilicidade. Um LogP baixo indica que a molécula é hidrofílica e terá dificuldade em deixar o ambiente líquido para entrar na membrana celular. Em contraponto, um LogP maior que 5 (LIPINSKI, 1997) sugere que a molécula, uma vez dentro da membrana, pode ter dificuldade em sair para o citosol ou pode ser propensa a interações inespecíficas e ao metabolismo hepático acelerado.

Similarmente, a polaridade das moléculas foi estimada pelo cálculo da Área de Superfície Topológica Polar (TPSA). A TPSA é uma medida da área da superfície da molécula que é polar, a polaridade é determinada por principalmente átomos de nitrogênio e oxigênio que estão presentes nas ligações de hidrogênio, e o número de Doadores (HBD) e Aceptores (HBA) de ligação de hidrogênio quantificam essa característica por estarem relacionados . Ou seja, quanto maior o número de grupos funcionais, contendo doadores ou aceptores, maior será sua TPSA. Isso acontece pois cada grupo funcional contribui com uma área de superfície Para que uma molécula atravesse uma membrana lipídica, ela deve primeiro romper as ligações de hidrogênio que forma com as moléculas de água ao redor dela, um processo conhecido como dessolvatação. A energia necessária para essa dessolvatação é proporcional ao número de HBDs e HBAs. Uma TPSA elevada, geralmente > 140 Å² (Veber, 2002) é um forte indicador de baixa permeabilidade intestinal.

A fração Csp3 é a fração de átomos de carbono em um composto orgânico que estão em estado de hibridação sp3. É uma medida que expressa o nível de saturação ou do caráter tridimensional de uma molécula. Moléculas com maior tridimensionalidade tendem a ter melhor solubilidade e a interagir com seus alvos de forma mais específica, resultando em menor toxicidade por interações fora do alvo.

4. RESULTADOS E DISCUSSÃO

4.1 Peso Molecular

É possível constatar na Figura 1 que a maioria das moléculas avaliadas possuem um valor de MW abaixo de 500 Da (Lipinski 1997), concordando com o que está descrito na literatura para fármacos aprovados. Ainda é possível observar algumas moléculas com o peso molecular maior que esse limiar, o que pode ser explicado, por exemplo, por não serem objetos de absorção oral.

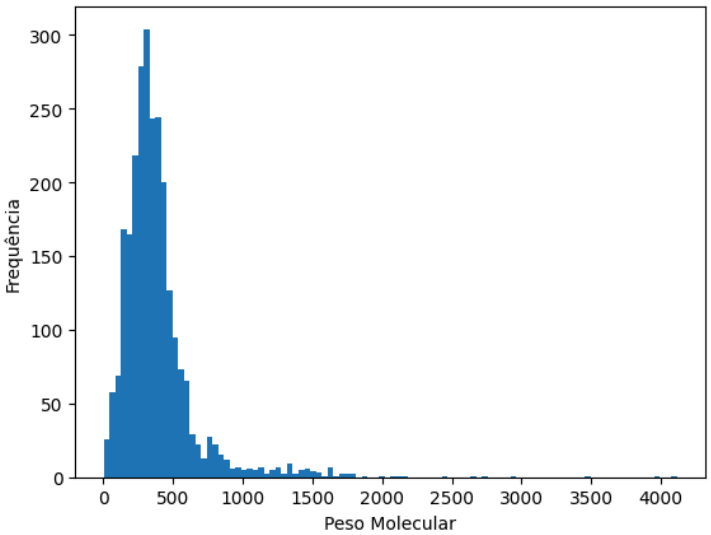


Figura 1: Gráfico de histograma representando a variação do valor de peso molecular entre as moléculas testadas.

4.2 LogP

Na Figura 2, constata-se que a maioria dos compostos possuem o valor de LogP abaixo de 5 (Lipinski, 1997), concordando com os parâmetros descritos na literatura para fármacos aprovados. A importância de não ultrapassar esse valor, é que a molécula terá problemas de absorção e excreção principalmente, no caso de fármacos orais caso ultrapasse.

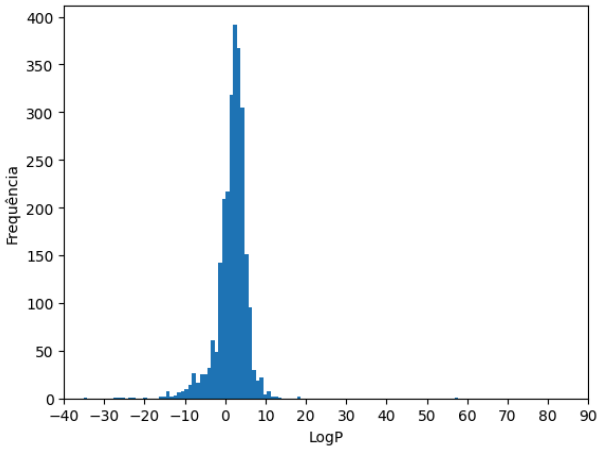


Figura 2: Gráfico de histograma representando a distribuição do valor de LogP para as moléculas testadas.

4.3 Superfície total polar (TPSA)

Analisando a Figura 3, observa-se que a maioria das moléculas possuem um valor de TPSA abaixo de 140 Å² (Veber, 2002), valores maiores podem acarretar em problemas de permeabilidade, principalmente a intestinal.

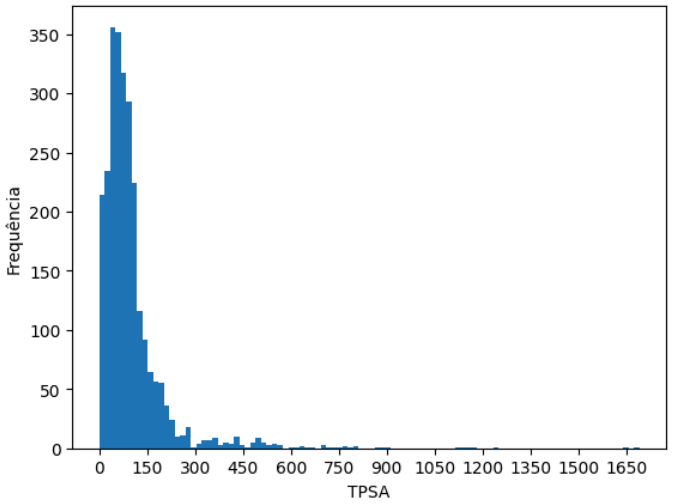


Figura 3: Gráfico de histograma representando a distribuição do valor de TPSA para as moléculas testadas.

4.4 Aceptores de Ligação de Hidrogênio

Com base na Figura 4, observa-se que, novamente, os resultados estão de acordo com a literatura. Para que um composto tenha uma boa absorção ou permeabilidade, é importante que ele não tenha mais de 10 aceptores de hidrogênio (Lipinski 1997), sendo que a maioria dos compostos estão dentro desta faixa.

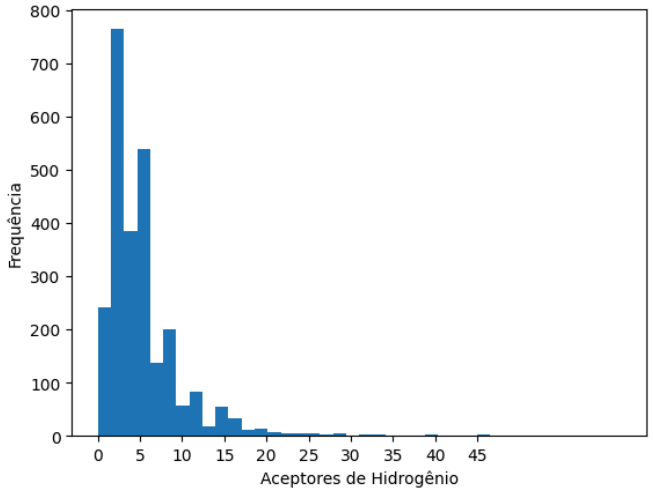


Figura 4: Gráfico de histograma representando a distribuição do valor de aceptores de hidrogênio para as moléculas testadas.

4.5 Doadores de Ligação de Hidrogênio

Na Figura 5, verifica-se que a quantidade de doadores de hidrogênio das moléculas analisadas se encontra, em sua grande maioria, abaixo de 5. Isso está relacionado também com o TPSA do composto, o que é importante para tenhamos uma boa permeabilidade e absorção, através da dessolvatação.

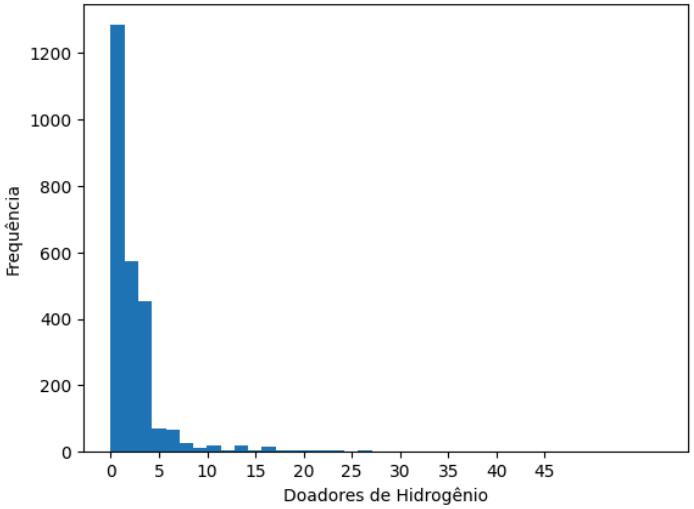


Figura 5: Gráfico de histograma representando a distribuição do valor de aceptores de hidrogênio para as moléculas testadas.

4.6 Fração Csp3

É possível observar na Figura 6 que os valores da fração Csp3 das moléculas são bem variados, indicando que, para as moléculas analisadas, o grau de tridimensionalidade é bastante variável. No caso do valor 0,0 são moléculas que não possuem carbonos, como por exemplo o Fosfato dipotássico, que pode ser utilizado como aditivo alimentar e também um agente tampão o que justifica não necessidade de tridimensionalidade. Podemos confirmar que as moléculas que têm a fração Csp3 mais próximas de 1,0 possuem menos reações indesejadas, por serem mais tridimensionais, assim tendo uma maior afinidade com seu sítio de ligação, gerando assim menos efeitos colaterais indesejados.

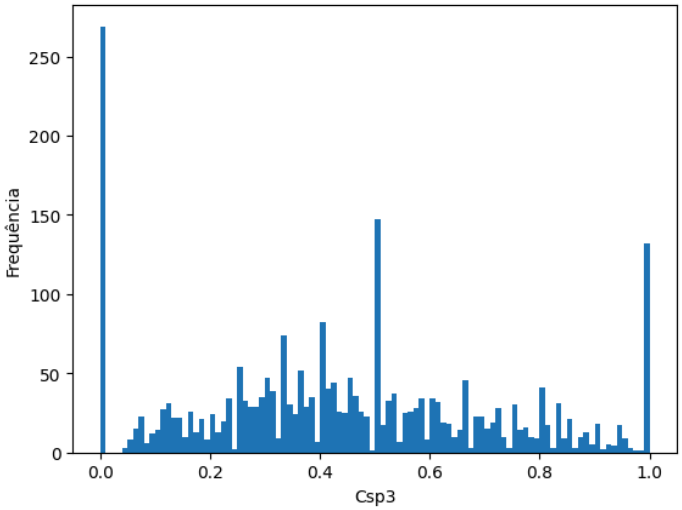


Figura 6: Gráfico de histograma representando a distribuição do valor da fração Csp3 para as moléculas testadas.

5. CONCLUSÕES

A concepção e o desenvolvimento de um novo fármaco através da descoberta de fármacos *de novo*, representam um desafio no meio de diversas propriedades. Neste trabalho, foi detalhado como propriedades físico-químicas importantes para o desenvolvimento de fármacos, como peso molecular, coeficiente de partição, TPSA, HBA, HBD e Csp3, são determinantes fundamentais de seu comportamento *in vivo*. A metodologia computacional empregada, impulsionada por ferramentas acessíveis como a biblioteca RDKit, permite a avaliação desses parâmetros moleculares desde as fases mais iniciais da pesquisa.

Como perspectiva para este trabalho estão a integração de abordagens mais sofisticadas, como a inteligência artificial e o aprendizado de máquina, para prever outras propriedades complexas com maior acurácia, de maneira que possam ser integradas a metodologias *de novo*, visando gerar fármacos melhores, mais seguros, e com um custo de produção reduzido.

REFERÊNCIAS

1. WALTERS, Pat. Practical Cheminformatics Tutorials. [S.l.]: GitHub, 2018.
2. LANDRUM, Greg A. RDKit: Open-Source Cheminformatics. 2006. Disponível em: http://www.rdkit.org.
3. LEACH, Andrew R.; GILLET, Valerie J. An introduction to chemoinformatics. Dordrecht: Springer, 2007.
4. LIPINSKI, Christopher A. et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews, v. 23, n. 1-3, p. 3-25, abr. 1997.
5. RDKIT. Open-Source Cheminformatics Software. [S.l.]: Greg Landrum, [2006-]. Disponível em: http://www.rdkit.org. Acesso em: 2 ago. 2025.
6. SCHNEIDER, Gisbert. Automating drug discovery. Nature Reviews Drug Discovery, v. 17, n. 2, p. 97-113, fev. 2018.
7. CHEN, Hongming et al. Practical considerations for deep learning in de novo drug design. Wiley Interdisciplinary Reviews: Computational Molecular Science, v. 12, n. 4, p. e1584, jul. 2022.
8. FLEMMING, Mette; MORET, Marcelo; JIMÉNEZ-LUNA, José. Multi-and many-objective optimization: present and future in de novo drug design. Journal of Cheminformatics, v. 15, n. 1, p. 111, dez. 2023.
9. GARCÍA-ORTEGA, Armando; HERNÁNDEZ-HERNÁNDEZ, Daniel; MEDINA-FRANCO, José L. Advances in de novo drug design: from conventional to machine learning methods. International Journal of Molecular Sciences, v. 22, n. 4, p. 1676, fev. 2021
10. SANGUINETTI, M. C.; TRISTANI-FIROUZI, M. hERG potassium channels and cardiac arrhythmia. Nature, v. 440, n. 7083, p. 463-469, 2006.
11. VEBER, D. F. et al. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. Journal of Medicinal Chemistry, v. 45, n. 12, p. 2615-2623, 2002.

Relatório de Atividades

Título do Projeto: HeMoLab1D Web v1.0
Bolsista: Peter Zeidler
Orientador: Pablo Javier Blanco, Paulo Ziemer
Tipo de Bolsa: PIBIC
Período do Relatório: 09-04-2025 a 05-08-2025

Objetivos

O projeto cria uma aplicação Web para visualizar simulações 1D do sistema cardiovascular. Ela permitirá manipular dados, configurar modelos e realizar simulações da dinâmica sanguínea em vasos deformáveis, com diversas funcionalidades.

1. Utilização multiusuário, com identificação prévia e suporte a múltiplos usuários simultâneos.
2. Gerenciamento de usuários por administrador.
3. Visualização 3D de modelos via renderização remota, protegendo geometrias no servidor.
4. Visualização de propriedades do modelo em escala global (cores) e por regiões.
5. Alteração de propriedades globais e regionais.
6. Criação de uma base de dados de modelos com metadados (ex.: usuário criador, número de segmentos).
7. Salvamento remoto de modelos na base de dados.
8. Exportação de modelos para a máquina local do usuário.
9. Upload de modelos da máquina do usuário para o banco de dados.
10. Configuração de parâmetros de simulação.
11. Armazenamento de modelos configurados em uma base de dados do usuário.
12. Resolução numérica de modelos usando códigos do HeMoLab.
13. Visualização de resultados de simulação em curvas ou animações.
14. Criação de populações de modelos por variação de parâmetros e geração de dados analíticos.
15. Modos de operação: “Visualização de modelos e resultados” (acesso potencialmente aberto) e “Edição e Simulação” (acesso completo para usuários identificados).
16. Biblioteca de curvas de ejeção, com criação e upload de curvas.
17. Serviço independente de simulações, com consulta de status.

Introdução

As doenças cardiovasculares são a principal causa de morte global, com milhões de óbitos anuais. Modelos computacionais têm avançado para estudar o sistema cardiovascular, mas modelos 3D são complexos e caros. Modelos 1D oferecem precisão eficiente e menor custo, validados por experimentos. O projeto HeMoLab1D Web v1.0 integra essas simulações 1D em uma plataforma Web acessível para apoiar pesquisa e prática médica.

Metodologia

O desenvolvimento do HeMoLab1D Web v1.0 foi estruturado em etapas para implementar as funcionalidades descritas nos objetivos, com os seguintes procedimentos:

1. **Pesquisa inicial**: Estudo das tecnologias Web, com foco no uso de Node.js para o backend e VTK.js para a visualização 3D interativa no frontend.
2. **Desenvolvimento do frontend**: Criação de uma interface web funcional utilizando VTK.js para renderização local dos modelos cardiovasculares em 3D, incluindo suporte à visualização de propriedades regionais (por exemplo, raio dos vasos com gradientes de cores) e seleção de segmentos específicos.
3. **Desenvolvimento do backend**: Implementação de um sistema básico de gerenciamento de usuários via CRUD utilizando Node.js e Express, com integração a um banco de dados relacional para armazenamento dos dados. O sistema completo de autenticação com controle de sessões ainda está em desenvolvimento.
4. **Banco de dados**: Configuração e uso de uma base relacional (PostgreSQL) para armazenamento dos modelos 3D, metadados associados e dados dos usuários.
5. **Testes**: Realização de testes funcionais com upload e visualização dos modelos nos formatos .vtk e .vtp, garantindo o funcionamento básico do sistema.

Resultados

No período de 01-04-2025 a 05-08-2025, o projeto HeMoLab1D Web v1.0 alcançou os seguintes resultados, alinhados aos objetivos propostos:

- **Gerenciamento de usuários (item 1)**: Foi implementado um sistema de cadastro, consulta, edição e exclusão de usuários (CRUD), com persistência em banco de dados utilizando Node.js e Express. A autenticação de usuários (login e controle de sessão) ainda não foi desenvolvida e está prevista para as próximas etapas.
- **Visualização 3D (itens 3 e 4)**: Desenvolvida visualização 3D de modelos cardiovasculares com VTK.js, incluindo renderização local e mapeamento visual com gradiente de cores para propriedades regionais, como o raio dos vasos. Também foi implementada a seleção interativa de segmentos, com exibição de informações associadas em uma interface funcional.
- **Banco de dados (itens 6, 7, 9, 11)**: Configurada uma base relacional para armazenar modelos e seus metadados (ex.: id, id do usuário, nome do modelo, caminho no servidor, tipo de arquivo e data de criação). Funcionalidades de salvamento e upload de modelos (.vtk, .vtp) foram implementadas, incluindo uma página dedicada para upload e testes realizados com sucesso. Por outro lado, o armazenamento de configurações personalizadas dos modelos (alterações de

parâmetros) ainda não foi implementado e está planejado para futuras etapas do projeto.

- **Desenvolvimento do Back-end**: Desenvolvida a estrutura com Node.js para o gerenciamento de usuários (via CRUD), de modelos e da integração com o banco de dados. A autenticação de múltiplos usuários será abordada nas próximas etapas.

Conclusões

O projeto HeMoLab1D Web v1.0 avançou significativamente na criação de uma plataforma Web para simulações cardiovasculares 1D. Foram implementadas funcionalidades essenciais, como o gerenciamento de usuários via CRUD, a visualização 3D com identificação de segmentos, e a estruturação de um banco de dados relacional para armazenamento de modelos e seus metadados.

O sistema permite o upload funcional de modelos nos formatos .vtk e .vtp, com testes realizados com sucesso. Ainda não foram implementadas funcionalidades para alteração e armazenamento de configurações personalizadas dos modelos, que estão previstas para etapas futuras.

Permanecem pendentes a implementação do sistema completo de autenticação com controle de sessões, a exportação de modelos para o usuário e a renderização remota segura. A escalabilidade para múltiplos usuários simultâneos será avaliada após a integração da autenticação.

As próximas etapas incluem finalizar essas funcionalidades pendentes e realizar testes com dados reais, visando a disponibilização da plataforma à comunidade científica.

Referências

[1] Allender, S. et al. European cardiovascular disease statistics. *European Heart Network*, v. 3, p. 11–35, 2008.

[2] Altintas, Z. et al. Cardiovascular disease detection using bio-sensing techniques. *Talanta*, Elsevier, v. 128, p. 177–186, 2014.

[3] Roth, G. A. et al. Demographic and epidemiologic drivers of global cardiovascular mortality. *New England Journal of Medicine*, v. 372, n. 14, p. 1333–1341, 2015.

[4] Ghorpade, A. G. et al. Estimation of the cardiovascular risk using WHO/ISH risk prediction charts in a rural population of South India. *International Journal of Health Policy and Management*, v. 4, n. 8, p. 531, 2015.

[5] Sociedade Brasileira de Cardiologia. Disponível em: https://portal.cardiol.br/br.

[6] Colunga, A. L. et al. Parameter inference in a computational model of haemodynamics in pulmonary hypertension. *Journal of the Royal Society Interface*, v. 20, n. 200, p. 20220735, 2023.

[7] Taylor, C. A. et al. Finite element modeling of three-dimensional pulsatile flow in the abdominal aorta: relevance to atherosclerosis. *Annals of Biomedical Engineering*, Springer, v. 26, p. 975–987, 1998.

[8] Blanco, P. J. et al. Introdução à modelagem e simulação computacional do sistema cardiovascular humano. Petrópolis-RJ, 2009.

[9] Blanco, P. J. et al. Blood flow modeling under load physiology, from global circulation to local hemodynamics. *bioRxiv*, Cold Spring Harbor Laboratory, p. 2022-01, 2022.

[10] Coskun, A. U. et al. Computational modeling of blood flow in arteries. *Journal of Biomechanical Engineering*, v. 128, p. 123–134, 2006.

[11] Friedman, M. H. et al. Arterial geometry and flow modeling. *Journal of Biomechanics*, v. 19, p. 567–575, 1986.

[12] Wong, K. K. et al. Computational medical imaging and hemodynamics framework for functional analysis and assessment of cardiovascular structures. *BioMedical Engineering OnLine*, Springer, v. 16, p. 1–23, 2017.

[13] Avolio, A. P. Multi-branched model of the human arterial system. *Medical and Biological Engineering and Computing*, Springer, v. 18, p. 709–718, 1980.

[14] Blanco, P. J. et al. Computational modeling of blood flow in the cardiovascular system. *International Journal for Numerical Methods in Biomedical Engineering*, v. 31, n. 5, p. 1–25, 2015.

[15] Hughes, T. J. R. et al. A mathematical model of arterial blood flow. *Journal of Applied Mechanics*, v. 40, p. 123–130, 1973.

[16] Liang, F. et al. Modeling of the cardiovascular system with integrated boundary conditions. *Annals of Biomedical Engineering*, v. 39, p. 1234–1246, 2011.

[17] Huberts, W. et al. A pulse wave propagation model for cardiovascular simulations. *Medical Engineering & Physics*, v. 34, p. 123–134, 2012.

[18] Mynard, J. P. et al. A one-dimensional model of blood flow in the arterial system. *Journal of Biomechanics*, v. 48, p. 1234–1245, 2015.

[19] Müller, L. O. et al. Computational modeling of arterial blood flow. *Journal of Computational Physics*, v. 320, p. 123–135, 2016.

[20] Matthys, K. S. et al. In vitro validation of a one-dimensional arterial blood flow model. *Journal of Biomechanics*, v. 40, p. 1234–1245, 2007.

[21] Reymond, P. et al. Validation of a patient-specific one-dimensional model of the systemic arterial tree. *American Journal of Physiology-Heart and Circulatory Physiology*, v. 301, n. 3, p. H1173–H1182, 2011.

[22] Node.js Foundation. Node.js. Disponível em: `https://nodejs.org/`, acesso em: 2025.

[23] PostgreSQL Global Development Group. PostgreSQL. Disponível em: https://www.postgresql.org/, acesso em: 2025.

[24] Kitware Inc. VTK.js: The Visualization Toolkit for JavaScript. Disponível em: https://kitware.github.io/vtk-js/, acesso em: 2025.

RELATÓRIO DE PROJETO DE INICIAÇÃO CIENTÍFICA

BIOINFORMÁTICA, BANCO DE DADOS E ENGENHARIA DE COMPUTAÇÃO

Título do Projeto Proposto

Comparação do Desempenho Computacional de *Workflows* Científicos de Transcriptômica em Arquiteturas HPC

Instituição

Laboratório Nacional de Computação Científica

Nome do Aluno

Reiglan Soares Di Lourenço

Nome do Orientador

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

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

Tipo de bolsa: PIBIC

Período do relatório: 28/06/2024 - 28/06/2025

1. Objetivo

O objetivo deste trabalho é avaliar o desempenho do software multithreading Bowtie2 nas arquiteturas Ivy Bridge (memória distribuída) e MESCA2 (memória compartilhada) do supercomputador Santos Dumont, utilizando o perfilador Intel VTune. A análise tem como foco a eficiência na distribuição das tarefas entre os núcleos de cada nó, considerando as particularidades e diferenças de tamanho entre as duas arquiteturas. Essa avaliação é fundamental para identificar qual configuração apresenta melhor desempenho e pode, assim, orientar sobre a alocação ideal de tarefas que utilizam o multithreading.

2. Introdução

Os experimentos de bioinformática são complexos e exigem alto custo computacional, necessitando de tecnologias especializadas como *workflows* científicos, sistemas de gerência, aprendizado de máquina e computação de alto desempenho (CAD). *Workflows* científicos são abstrações que representam esses experimentos como um fluxo encadeado de atividades, as quais são executadas por aplicações científicas com diversas características. Com o uso de sistemas de gerenciamento ou linguagens de programação, é possível modelar, gerenciar e analisar os *workflows* [Cruz et al. 2020].

O *workflow* de transcriptômica ParslRNA-Seq para análise de expressão diferencial de genes (EDG) foi utilizado como estudo de caso. O ParslRNA-Seq foi modelado e gerenciado com a biblioteca Parsl, desenvolvida em Python, para facilitar a modelagem, integração e automatização do *workflow* em ambientes de CAD. O *workflow* é composto por um conjunto de seis atividades: Bowtie mapeia as leituras; Sort ordena as leituras; Split divide arquivos de entrada; HTSeq conta as leituras geradas no Split; Merge indexa as contagens; e DESeq realiza a análise estatística de EDGs. Este trabalho foca na análise computacional do Bowtie2, que é executado como uma atividade, e é considerado o mais representativo e computacionalmente intensivo no *workflow* [Cruz et al. 2021].

Bowtie2 é um *software* eficiente usado para alinhar leituras de sequenciamento bem longas, como o genoma humano. Ele utiliza uma técnica chamada Índice FM, que ajuda a manter o uso de

memória baixo e suporta a execução em múltiplos processadores para acelerar o alinhamento. A execução do Bowtie foi analisada com o Intel VTune, um *software* de análise de desempenho que detalha o uso dos recursos computacionais da CPU e memória, permitindo identificar gargalos de desempenho e melhorar a eficiência do código.

3. Background

3.1 ParslRNA-Seq: *Workflow* Científico de Transcriptômica

ParslRNA-Seq (Silva et a 2021) é um *workflow* científico de transcriptômica para estudos de expressão diferencial de genes (EDG) modelado com a linguagem de programação Parsl e ser acoplado em ambientes de CAD do supercomputador SDumont. A Figura 1 apresenta o modelo conceitual do ParslRNA-Seq. Análises de EDGs são muito usadas na análise de dados transcriptômicos, permitindo a identificação de genes com expressão significativamente divergente entre condições experimentais. Essa abordagem serve para descobertas de novos transcritos e isoformas de genes, proporcionando *insights* importantes sobre processos biológicos complexos, como diferenciação celular, resposta a estímulos e desenvolvimento de doenças.

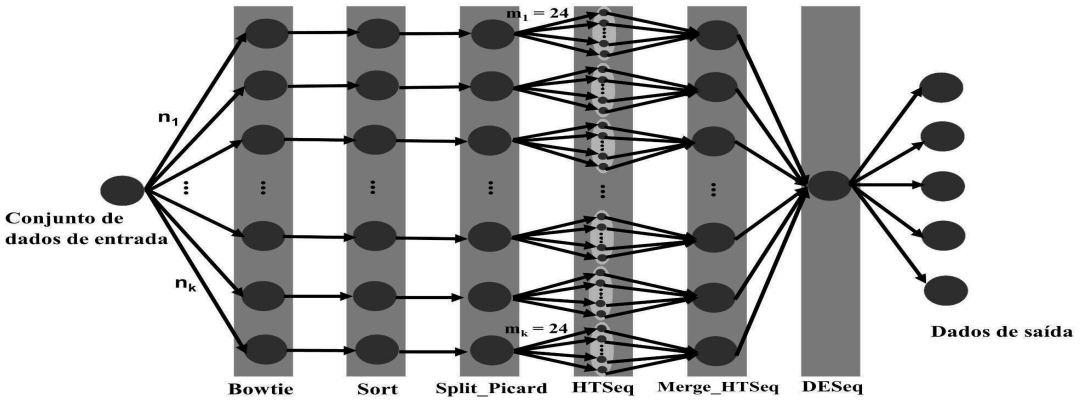


Figura 1. Modelagem Conceitual do *Workflow* Científico ParslRNA-Seq. Adaptada de Silva *et al.* 2021.

3.2 Parsl - Parallel Scripting Library

ParslRNA-Seq é implementado utilizando Parsl, uma ferramenta em Python projetada para execução de *workflows* em ambientes de CAD [Babuji et al., 2019]. Parsl[1] é uma biblioteca de programação paralela desenvolvida em Python que utiliza decoradores para executar funções e software externo como aplicativos Python e Bash. A arquitetura do Parsl é baseada em programação orientada a dados, uma tarefa só é executada quando todas as suas entradas estão disponíveis. Isso permite o gerenciamento dinâmico da execução e o uso eficiente de recursos computacionais, sejam locais, em clusters HPC ou em nuvens. O Parsl oferece suporte nativo para diferentes mecanismos de execução (executors), como ThreadPool, High Throughput Executor (HTEX) e Slurm. O motor de execução do Parsl é flexível, suportando diversos ambientes computacionais e abstraindo as complexidades do *workflow*, tornando mais simples sua implementação e integração com recursos computacionais.

O *workflow* utiliza dois modelos de execução: ThreadPoolExecutor e HTEx. ThreadPoolExecutor[2] é um modelo de execução que suporta *multithreading* em recursos locais, gerenciando um *pool* de *threads* para executar atividades de forma concorrente. A eficiência desse modelo é crucial para maximizar o desempenho computacional, reduzir erros e melhorar a

[1] https://parsl-project.org/
[2] https://docs.python.org/3/library/concurrent.futures.html.

produtividade [Silva e Yokoyama, 2011]. Por outro lado, o modelo HTEx permite a execução do *workflow* em múltiplos nós computacionais simultaneamente, facilitando o compartilhamento de dados entre as máquinas e proporcionando um controle refinado sobre a alocação de recursos e paralelismo necessário para uma execução eficiente das tarefas.

4. Metodologia

4.1 Dados do Experimento

Os dados deste estudo são de um experimento real de RNA-Seq, extraídos do repositório Gene Expression Omnibus (GEO), com o ID GSE97763. Foram divididos em grupo de controle (SRR5445794, SRR5445795, SRR5445796) e grupo de condições das vias metabólicas Wnt (SRR5445797, SRR5445798, SRR5445799), alinhadas ao genoma de referência do *Mus musculus* (UCSC versão mm9) usando o Bowtie2. O conjunto de dados contém arquivos variando entre 1,8GB e 3,0GB, totalizando 13GB. O perfilador Intel VTune foi executado juntamente com o software Parsl na atividade Bowtie2.

Após o alinhamento, o número de leituras mapeadas por gene foi contado com o HTSeq-count. Em seguida, a análise de expressão diferencial de genes (EDG) foi realizada com o DESeq2, a partir das matrizes geradas com base nas contagens dos alinhamentos. Essas matrizes (arquivo GTF) contém o número de leituras que foram alinhadas de forma única (colunas) com os exames de cada gene nas amostras (colunas). Os seis arquivos de entrada totalizaram uma saída de aproximadamente 74 GB.

4.2 O Ambiente Computacional Santos Dumont

O SDumont[3] possui capacidade instalada de processamento na ordem de 20 Petaflop/s (20 x 1015 *float-point operations per second*), apresentando uma configuração híbrida de nós computacionais.

Todos os *softwares*, algoritmos, dependências de bioinformática (Bowtie, Samtools, Picard, HTSeq e DESeq2) e os componentes do Parsl foram alocados e instalados no ambiente do SDumont.

As execuções foram realizadas em:

Nó Base (Ivy Bridge**)**: Composto por 2 CPUs Intel Xeon E5-2695v2 (12 núcleos a 2.4GHz cada), 24 núcleos de 64 GB de memória RAM, utiliza uma arquitetura de memória distribuída que possui sistema de *clusters*, suportando tarefas segmentáveis e distribuídas entre nós através de rede Infiniband EDR de 100 Gb/s, sendo econômico e escalável horizontalmente.

Nó MESCA2: Composto por 16 CPUs Intel Xeon Ivy Bridge (240 núcleos no total) e 6 TB de memória RAM, utiliza uma arquitetura de memória compartilhada onde todos os 240 núcleos têm acesso à mesma memória RAM. Com esse modelo, a comunicação é direta e rápida entre os núcleos.

4.3 Arquitetura de memória compartilhada e distribuída no SDumont

Optamos por comparar dois tipos distintos de memória para avaliar sua eficiência e desempenho na execução do *workflow*. A memória compartilhada do MESCA2 oferece 240 núcleos e 6 TB, permitindo acesso direto e rápido aos dados entre os processadores, ideal para tarefas intensivas que demandam comunicação rápida e sincronização sem complexidade externa. Por outro lado, o Ivy Bridge possui 64 GB de RAM por nó, configurado para tarefas segmentadas que não exigem compartilhamento extensivo de dados entre nós. Cada nó utiliza sua própria memória local, com comunicação entre nós via rede, o que pode resultar em latências maiores comparado ao acesso direto da memória compartilhada.

[3] https://sdumont.lncc.br/

O processador Intel Ivy Bridge, com clock de 2,4 GHz, possui uma arquitetura x86-64 de 64 bits, 15 núcleos por CPU, fabricado com tecnologia de 22 nm, e suporta Hyper-Threading, permitindo 2 *threads* por núcleo. Ele possui até 30 MB de cache L3. O nó de computação MESCA2, com memória compartilhada de grande capacidade (fat node), permite que múltiplos núcleos acessem a mesma região de memória, utiliza cache em três níveis (L1, L2 e L3), e possui suporte à NUMA para otimizar a latência de acesso à memória. A interconexão de alta velocidade e técnicas avançadas de gerenciamento de memória, como prefetching e cache coherence, garantem eficiência e desempenho elevado, além de suportar grandes quantidades de RAM para cargas de trabalho intensivas em dados. O prefetching antecipa a necessidade de dados, carregando-os para o cache antes de serem requisitados, enquanto a cache coherence assegura que todas as cópias dos dados em diferentes caches sejam consistentes, evitando conflitos e garantindo a integridade dos dados durante o processamento paralelo.

Em sistemas de memória distribuída, cada nó possui sua própria memória e cache local, acessíveis diretamente pelo processador do nó. A coerência de cache local é mantida dentro de cada nó, garantindo que as operações de leitura e escrita sejam consistentes para os dados armazenados localmente. No entanto, para manter a consistência entre os caches de diferentes nós, é necessário o uso de protocolos explícitos de comunicação, como a passagem de mensagens. Quando um nó precisa acessar dados que residem em outro nó, esses dados são transferidos pela rede e armazenados temporariamente no cache local do nó solicitante, o que ajuda a reduzir a latência de acesso. As políticas de substituição de cache, como o LRU (Least Recently Used), são aplicadas localmente em cada nó para gerenciar o conteúdo do cache, priorizando a remoção dos dados menos recentemente utilizados quando necessário liberar espaço para novos dados ou atualizações.

5. Resultados e Análises

5.1 Análise Computacional do *Workflow* com o Perfilador VTune

A Figura 1 apresenta o histograma de uso dos núcleos dos dois processadores Intel Xeon Ivy Bridge durante a execução do *workflow* ParslRNA-Seq, composto por 6 atividades, o histograma abaixo mostra principalmente as atividades Bowtie2 e Sort. Essas duas atividades utilizam *multithreading* para executar simultaneamente com 24 *threads*, enquanto as demais são processadas sequencialmente, com um arquivo por núcleo.

Observa-se um uso intensivo dos 24 núcleos durante a execução *multithreading* do Bowtie2, otimizando o desempenho do *workflow* e reduzindo o tempo total de execução. As demais atividades, que envolvem 6 *tasks*, utilizam apenas 6 núcleos, com cada *task* representando um arquivo. Ao entrarem em etapas com suporte a *multithreading*, as atividades alocam o número máximo de *threads* permitido pelo nó computacional. Nas análises a seguir, o foco será direcionado exclusivamente à atividade Bowtie2, visando comprovar a utilização intensiva dos núcleos durante sua execução.

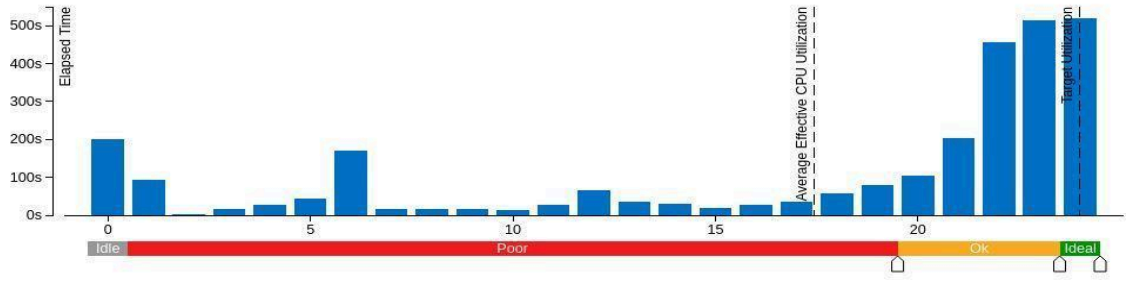


Figura 1. Uso da CPU no Ivy Bridge 24 *cores*

5.2 Análise Computacional da Atividade Bowtie2 com o Perfilador VTune

Nas execuções realizadas com o Ivy Bridge (Figura 2) e o MESCA2 (Figura 3), observou-se um crescimento gradual da carga de trabalho até atingir o pico de paralelismo: cerca de 20 núcleos no Ivy Bridge e entre 150 e 200 núcleos no MESCA2. Também podemos observar na (Figura 3) uma maior utilização dos *cores* múltiplos de 12, devido à arquitetura do MESCA2 ser composta por 16 CPUs de 12 *cores* cada. O eixo horizontal representa o número de *threads*, enquanto o eixo vertical exibe o tempo máximo de execução por *thread* (*wall-clock*) em segundos. A utilização de CPUs lógicas pelas atividades do *workflow* é representada por núcleos: a cor cinza refere-se à fase de inicialização do sistema e atividades; a cor vermelha indica uma utilização deficiente das CPUs; a cor amarela refere-se a uma utilização média das CPUs; e, por fim, a cor verde representa o uso ideal das CPUs, mostrando que os recursos estão sendo usados de forma eficiente. O VTune também mostrou que a eficiência no Ivy Bridge foi de 83,5%, com um tempo de *wall-clock* de 250 segundos. Já no MESCA2, a eficiência foi de 66,3%, com um tempo de *wall-clock* de apenas 100 segundos para acesso à memória.

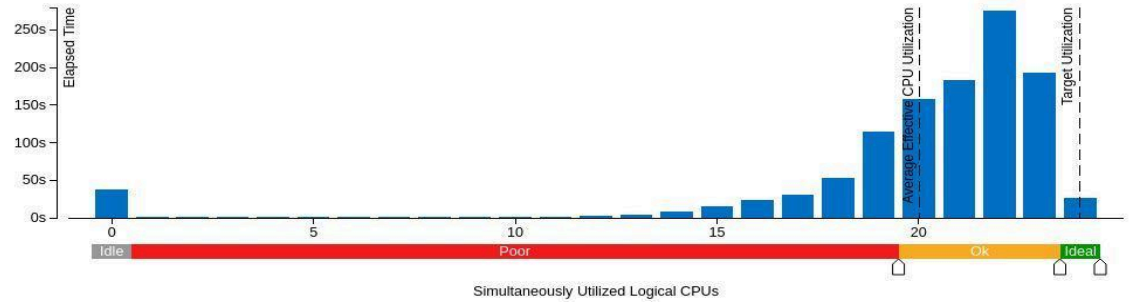


Figura 2. Uso da CPU no Ivy Bridge 24 *cores*

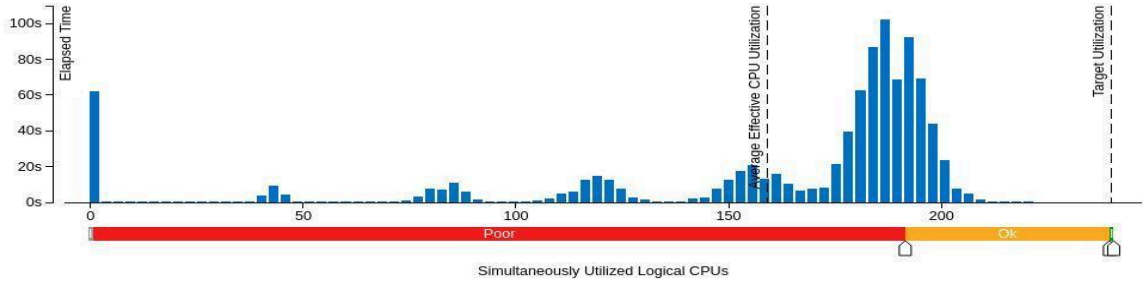


Figura 3. Uso da CPU no Ivy Bridge (MESCA2) com 240 *cores*

6. Conclusão

Neste trabalho, analisamos a distribuição de trabalho entre os núcleos das arquiteturas Ivy Bridge e MESCA2 do supercomputador Santos Dumont, com foco na eficiência de execução de tarefas multithread utilizando o Bowtie2. Os resultados indicam que, embora ambas as arquiteturas apresentam comportamento semelhante quanto à progressiva ocupação dos núcleos ao longo do tempo de execução, o Ivy Bridge demonstrou desempenho superior, com uma eficiência 17,2% maior em relação ao MESCA2. Essa vantagem está associada à melhor utilização e balanceamento dos núcleos, refletindo uma maior estabilidade durante a execução. Assim, para este cenário específico de mapeamento com Bowtie2, a arquitetura Ivy Bridge se mostrou mais adequada, sendo recomendada para alocação de tarefas que demandam paralelismo eficiente.

7. Referências bibliográficas

[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.

1. DADOS GERAIS

Título do projeto: Introdução à Programação com Scratch e Tutoriais

Bolsista: Ricardo dos Santos de Lima

Orientadora: Regina Célia Cerqueira de Almeida

Edital/Programa: 2024/PIBIC

Período: setembro de 2024 - agosto de 2025

2. OBJETIVOS

O projeto tem como objetivo principal a elaboração de tutoriais, utilizando o software Scratch[1], para ensinar noções básicas de programação para estudantes de Ensino Fundamental e Médio. Ao todo, foram elaborados trinta tutoriais abordando diferentes tópicos, envolvendo noções de programação e desenvolvimento de jogos. Com uma metodologia que combina instruções teóricas e atividades práticas, os tutoriais percorrem tópicos como lógica de programação, estruturas de dados e modelagem computacional.

A abordagem do projeto incluiu o desenvolvimento de recursos pedagógicos para facilitar a compreensão dos conceitos fundamentais de programação, bem como a criação de exercícios práticos para fixação dos conceitos apresentados. A ordem dos conteúdos apresentados nos tutoriais propiciam o aprendizado gradual, iniciando pela apresentação dos recursos disponíveis no software Scratch, associados aos recursos de programação. Ao finalizar os tutoriais, o estudante será apresentado às noções de desenvolvimento de jogos, sendo a abordagem distribuída nos tutoriais. Ao término dos tutoriais, espera-se que os estudantes possam estar aptos a iniciar o desenvolvimento de seus próprios projetos, desenvolvendo habilidades de pensamento lógico e resolução de problemas. Para além, durante a execução dos tutoriais, outras atividades podem ser associadas promovendo atividades em grupo e compartilhamento de ideias entre os alunos.

3. INTRODUÇÃO

A compreensão dos fundamentos de programação tornou-se uma habilidade essencial e importante na formação dos estudantes, especialmente diante da crescente evolução digital em diversas áreas de conhecimento. No contexto educacional brasileiro, a utilização de ferramentas didáticas que sejam acessíveis e eficazes é um desafio constante, principalmente no processo de formação cidadã do aluno durante o Ensino Fundamental e Médio. Nesse cenário, a programação deixa de ser reconhecida apenas como uma habilidade técnica, mas com uma forma de desenvolvimento do pensamento lógico e computacional,

[1] https://scratch.mit.edu/

estimulando a resolução de problemas e a criatividade, que são competências importantes para a evolução do estudante.

O Scratch aparece como uma ferramenta pedagógica capaz de introduzir aos alunos, conceitos básicos e fundamentais de uma forma lúdica e visual. Uma ferramenta desenvolvida pelo MIT, sendo uma linguagem de programação visual que baseia-se em blocos, permitindo os estudantes criarem histórias, animações e jogos. Essa abordagem lúdica favorece o engajamento dos estudantes e facilita a introdução dos conceitos fundamentais da computação, essencial para quem nunca teve contato ou familiaridade prévia com a tecnologia.

Ao disponibilizar os materiais educativos de forma acessível e que são de fácil compreensão para os estudantes, o projeto assume um papel importante no combate à desigualdade de acesso à informação e no combate à exclusão digital, ampliando a oportunidade de contato com o mundo da programação. A utilização de uma ferramenta gratuita como o Scratch, aliada a uma linguagem didática e inclusiva, permite que os estudantes tenham acesso ao conhecimento tecnológico, proporcionando maior equidade no processo de ensino e contribuindo na formação de pessoas mais preparadas para os desafios do século atual.

4. MATERIAIS E MÉTODOS

Nesta etapa do projeto, o foco principal foi a finalização dos trinta tutoriais planejados, abordando os conteúdos fundamentais de lógica de programação, estrutura de dados e modelagem computacional através do Scratch. Cada tutorial foi elaborado com o objetivo, atividade e exercício, priorizando a clareza e coerência didática e o alinhamento com os objetivos educacionais do projeto. A estruturação dos tutoriais buscou manter uma progressão gradual dos conteúdos, respeitando o ritmo de aprendizagem. Para além disso, este período focou na construção de tutoriais mais longos e complexidade elevada, criando jogos como PacMan, AngryBirds, Torre de Hanoi, entre outros. As referências bibliográficas utilizadas na primeira etapa como “Scratch: Um jeito divertido de aprender programação” de H. Varela (2017) e “Oficina básica de Scratch” de Vieira, M. F., continuaram servindo como suporte, além dos conteúdos disponíveis no YouTube, tal como projetos dentro do próprio Scratch.

Com todos os tutoriais finalizados, foram realizadas etapas de revisões completas de cada um dos materiais produzidos. O objeto das revisões eram correções e adequação da linguagem dos tutoriais, buscando torná-los mais claros e comunicativos, acessíveis e assertivos tanto para os estudantes, quanto para os professores que utilizarão o conteúdo. Também foi realizada a atualização das imagens dos tutoriais, de modo que, a ilustração dos passos descritos ficassem mais objetivos, buscando facilitar o entendimento, além de torná-lo mais atrativo visualmente. Essa etapa garantiu que os conteúdos estivessem não apenas corretos tecnicamente, mas também com uma abordagem pedagógica mais eficiente.

Para assegurar que as atividades propostas nos tutoriais estivessem alinhadas com os objetivos educacionais e fossem adequadas para o público-alvo. Durante as reuniões foram discutidas as melhores abordagens pedagógicas. Foram realizadas as revisões dos conteúdos selecionados e também o ajuste das atividades conforme atendesse a necessidade dos estudantes. As reuniões foram importantes para discutir as abordagens adotadas e realizar ajustes na estrutura dos tutoriais para garantir que os materiais fossem eficazes para os professores utilizarem em sala de aula. Essa colaboração permitiu a criação de um plano estruturado e coeso.

5. RESULTADOS E DISCUSSÃO

Ao final do período foram elaborados e revisados os trinta tutoriais que possuem como estrutura definida a apresentação do tópico abordado, descrição das atividades a serem elaboradas de forma sequencial, permitindo a reprodução em um segundo momento pelo estudante e a proposta de um desafio ao término. Os tutoriais elaborados durante o período abordaram conteúdos como: tipos de variáveis, operadores aritméticos, estruturas condicionais, laços de repetição, operadores lógicos, manipulação de variáveis, posição de atores, troca de cenários, utilização de sons, animações, etc. Os dez tutoriais finais abordaram a criação de jogos mais complexos, explorando a criatividade e capacidade de raciocínio lógico dos estudantes, além de testar os conhecimentos obtidos ao longo do restante dos tutoriais. Além disso, foram elaborados os gabaritos dos exercícios para que o professor, ao utilizar o tutorial, conseguisse ter uma visão do resultado final dos algoritmos que se pede a cada tutorial.

O principal resultado obtido ao término da elaboração e revisões dos tutoriais, foi a criação do livro “Introdução à Programação com Tutoriais: Programação de forma lúdica”, no qual reúne os 30 tutoriais criados ao longo do projeto, que são organizados de modo a explorar uma didática mais eficiente e um ensino mais assertivo.

6. CONCLUSÕES

O projeto de desenvolvimento de tutoriais em Scratch para introdução à programação é uma iniciativa importante para a educação contemporânea. Ao focar em alunos do Ensino Fundamental e Médio, proporciona uma base sólida em programação desde cedo, propiciando a formação de conhecimentos que poderá ser utilizado no cotidiano. A escolha do Scratch como plataforma é importante, pois sua interface visual é intuitiva e facilita a compreensão de conceitos complexos da programação. A metodologia proposta, que combina teoria e prática, é eficaz para o engajamento dos alunos e promoção da aprendizagem ativa. Ao aprenderem conceitos de lógica de programação, estrutura de dados e modelagem computacional, os alunos desenvolverão habilidades cruciais para o século atual, como pensamento crítico, resolução de problemas e criatividade.

A preparação de materiais para publicação online amplia o alcance do projeto, que permite que os recursos estejam acessíveis a um público maior e diverso. No mais, é importante considerar desafios como a necessidade de dispositivos adequados para todos os

alunos, bem como a capacitação de professores para orientação e apoio ao aprendizado. A avaliação contínua do progresso dos alunos e a adaptação dos tutoriais conforme necessidades serão fundamentais.

7. REFERÊNCIAS BIBLIOGRÁFICAS

1. **VALENTE, José Armando et al.** O computador na sociedade do conhecimento. Campinas: Unicamp/NIED, p. 11-18, 1999.

2. **DA SILVA, Angela Carrancho.** Educação e tecnologia: entre o discurso e a prática. Revista Ensaio: Avaliação e Políticas Públicas em Educação, v. 19, n. 72, p. 527-554, 2011.

3. **NUNES, Felipe Becker et al.** Um estudo de caso sobre a importância do uso de objetos de aprendizagem no ensino fundamental como apoio pedagógico. In: Anais do Workshop de Informática na Escola. 2014. p. 542.

4. **LOPES, José Junio et al.** A introdução da informática no ambiente escolar. Clube do professor, v. 23, 2004.

5. **SEVERO, C. E. P. (2021).** Jogos com Scratch: em projetos práticos com linguagem de blocos. Brasil: Casa do Código.

6. **VARELA, H. (2017).** Scratch: Um jeito divertido de aprender programação. Brasil: Casa do Código.

7. **VIEIRA, Marilene Ferreira.** Oficina básica de Scratch. TUCURUÍ: NTE Tucuruí, 2018.

8. **GUANABARA, Gustavo.** Como criar personagens, cenários e animações no Scratch 3.0 - Tutorial Completo. **2019.** Disponível em: https://www.youtube.com/watch?v=GrPkuk1ezyo&t=765s. Acesso em: 10 jul. 2024.

RELATÓRIO DE ATIVIDADES

Título do projeto	Construção de um Conjunto de Dados Envolvendo Modelos Tridimensionais e Dados Experimentais de Afinidade Proteína-Ligante
Bolsista	Shirlei Militão da Silva
Orientadores	Laurent E. Dardenne, José Renato Duarte Fajardo
Tipo de Bolsa	PIBIC/LNCC
Período do Relatório	12/2024 - 09/2025

1. OBJETIVOS

Construir um conjunto de dados através da elucidação da estrutura tridimensional de complexos proteína-ligante, associando estes a dados de doenças e aos valores de afinidade de ligação obtidos experimentalmente.

2. INTRODUÇÃO

A modelagem da interação entre proteínas e pequenas moléculas ligantes desempenha um papel crucial no desenvolvimento de novos fármacos e terapias direcionadas, sendo essencial para compreender os mecanismos moleculares que regem essas interações e suas implicações clínicas (Li et al, 2024; Patil et al, 2024; Guedes, et al, 2014). A construção de um banco de dados que relacione essas estruturas com informações sobre doenças associadas e dados de afinidade por ligantes é, portanto, um passo importante nesse contexto (Zou et al., 2015; Libório e Resende, 2021).

O conjunto de dados a ser construído neste projeto envolve dados que possuem dados experimentais de afinidade, mas não possuem uma estrutura tridimensional determinada experimentalmente. Esse conjunto não apenas ajudará na identificação de potenciais fármacos, mas também permitirá desenvolver modelos de aprendizagem de máquina que possam prever a eficácia dos mesmos (da Silva et al., 2023).

O uso de técnicas, como modelagem comparativa e ancoramento molecular, será fundamental para o desenvolvimento e validação desse conjunto de dados. A modelagem comparativa permitirá a previsão das estruturas de proteínas sem estrutura experimental (Patel e Mani, 2024; Li et al, 2024; Fiser, 2010), enquanto o ancoramento molecular permitirá obter modelos tridimensionais descrevendo a interação proteína-ligante (Paggi, Pandit e Dror, 2024; Hollingsworth e Dror, 2018). Outro ponto importante a ser perseguido neste projeto é garantir que a metodologia de construção desses modelos tridimensionais seja validada, para garantir a confiabilidade dos modelos construídos. Isso é importante para garantir a qualidade do desenvolvimento de modelos de aprendizagem de máquina. (Turzo, Hantz, Lindert, 2022)

3. METODOLOGIA

O primeiro passo deste projeto consistirá na definição e validação de métodos para elucidação da estrutura tridimensional de complexos receptor-ligante. Estes métodos serão baseados na aplicação da técnica de ancoramento molecular, utilizando estruturas previamente determinadas de complexos relacionados para modelar a interação com novos ligantes que possuam dados experimentais de afinidade.

A metodologia será estruturada em três etapas principais: extração de dados, integração de dados e elucidação das estruturas tridimensionais. As duas primeiras etapas fornecem suporte à terceira, que constitui o núcleo metodológico do projeto. Assim, no momento inicial, o projeto será centrado na validação de métodos para elucidação das estruturas dos complexos receptor-ligante, avaliando sua robustez, aplicabilidade e acurácia com base em casos de referência com estruturas conhecidas.

3.1. ANCORAMENTO MOLECULAR PROTEÍNA-LIGANTE

Com as estruturas tridimensionais das proteínas-alvo definidas, partindo de complexos do mesmo alvo com outros ligantes ou do alvo isolado, a técnica de ancoramento molecular é aplicada para prever como os ligantes interagem com as proteínas. Essa técnica avalia as interações entre molécula e receptor. Segundo Guedes, Magalhães e Dardenne (2014), o método é amplamente utilizado no design de fármacos, tanto para otimizar compostos conhecidos quanto para triagem virtual de novas moléculas biologicamente ativas. Utilizando o DockThor (Guedes et al, 2024) para realizar simulações de docking, onde ligantes candidatos terão sua conformação predita no sítio ativo da proteína.

3.1.1. Avaliação de Afinidade

As energias de ligação dos complexos proteína-ligante também serão preditas utilizando modelos de aprendizagem de máquina desenvolvidos pelo Grupo de Modelagem Molecular de Sistemas Biológicos do LNCC (Guedes et al, 2021), e comparadas com os valores experimentais conhecidos.

3.1.2. Validação

A etapa de validação visa garantir a confiabilidade dos resultados obtidos nas simulações de ancoramento molecular. Para isso, será realizada a reanálise do protocolo de docking utilizando ligantes co-cristalizados, previamente descritos na literatura para as proteínas alvo estudadas. A acurácia do método será avaliada por meio do cálculo de Root Mean Square Deviation (RMSD) entre a conformação experimental do ligante no sítio ativo da proteína. Valores de RMSD $\leq$ 2,0 serão considerados indicativos de reprodutibilidade do método. Além disso, será feita a análise qualitativa das interações entre o ligante e os resíduos do sítio ativo, utilizando softwares de visualização molecular como PyMOL.

Por fim, será realizada a comparação entre as energias de ligação previstas pelos modelos de aprendizagem de máquina, citados no tópico anterior, utilizando as estruturas dos complexos gerados por docking e as energias obtidas por meio de cálculos de FEP (Free Energy Perturbation) extraídos do artigo de Wang et al. (2015), contra os valores de energia experimental. Essa abordagem permitirá avaliar simultaneamente a capacidade

preditiva das metodologias de construção de modelos tridimensionais e a predição da afinidade do complexo.

3.2. BANCOS DE DADOS BIOLÓGICOS

A coleta e a correlação de dados relevantes dependerão da integração estratégica com bancos de dados biológicos validados. Para isso, serão exploradas diferentes fontes, como o Therapeutic Target Database (TTD), o Protein Data Bank (PDB), o BindingDB e o ChEMBL, a fim de reunir informações sobre doenças, alvos terapêuticos, estruturas tridimensionais de complexos proteína-ligante e dados experimentais de afinidade de ligação. Essa abordagem visa estabelecer conexões entre aspectos clínicos, funcionais e estruturais, fortalecendo a interpretação dos resultados obtidos.

O Therapeutic Target Database (TTD) fornece informações sobre alvos terapêuticos, doenças e substâncias ativas, com aproximadamente 4 mil entradas relativas a alvos e cerca de 40 mil referentes a fármacos (Zhou et al., 2024). O Protein Data Bank (PDB) disponibiliza estruturas tridimensionais de proteínas e complexos macromoleculares, totalizando cerca de 225 mil entradas (Berman et al., 2000). O BindingDB concentra dados experimentais de afinidade entre proteínas e ligantes, reunindo aproximadamente 2,9 milhões de entradas (Liu et al., 2007). Por fim, o ChEMBL apresenta informações bioativas de compostos, incluindo estruturas químicas, alvos e dados de ensaios, abrangendo 2,4 milhões de entradas sobre compostos, 1,6 milhão sobre ensaios e cerca de 15 mil relacionadas a alvos e fármacos (Mendez et al., 2019).

3.2.1. Extração de Dados

Nesta fase, será realizada uma busca sistemática em bancos de dados públicos (TTD, PDB, BindingDB, ChEMBL) para coletar informações sobre proteínas, ligantes, afinidades de ligação e doenças associadas. A prioridade será dada a complexos com valores experimentais de afinidade com alguma estrutura tridimensional elucidada experimentalmente para o alvo, seja em um complexo com outro ligante ou o alvo isolado.

3.2.2. Integração de Dados

A última etapa consiste na integração dos dados tratados em um sistema de banco de dados. Será desenvolvida uma interface de acesso com filtros por proteína, ligante, doença e afinidade, utilizando tecnologias de gerenciamento como SQL.

4. RESULTADOS E DISCUSSÃO

Para a fase de validação da metodologia de ancoramento molecular, foram selecionados oito sistemas proteicos com dados experimentais de afinidade disponíveis e estrutura tridimensional resolvida por cristalografia, conforme descrito por Wang et al. (2015). Os sistemas extraídos compreendem alvos farmacológicos relevantes e apresentam séries de ligantes com variações estruturais limitadas em torno de um núcleo químico comum a cada série.

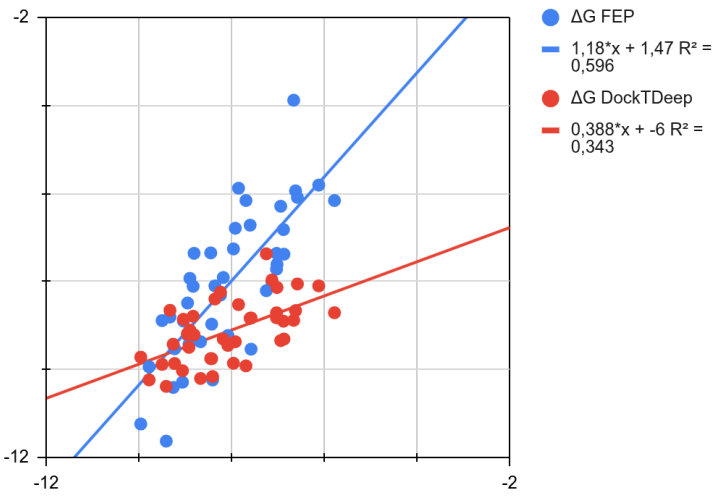
Os sistemas incluídos foram: β-secretase (BACE), utilizando a estrutura cristalográfica 4DJW e uma série de 36 ligantes; Cyclin-Dependent Kinase 2 (CDK2), com

base na estrutura 1H1Q e 16 ligantes; Stress-activated protein kinase 1c (JNK1), a partir da entrada 2GMX e 21 ligantes; Myeloid Cell Leukemia 1 (MCL1), com a estrutura 4HW3 e 42 ligantes; Stress-activated protein kinase 2a (p38), estrutura 3FLY com 34 ligantes; Protein Tyrosine Phosphatase 1B (PTP1B), estrutura 2QBS com 23 ligantes; Thrombin, com a estrutura 2ZFF e 11 ligantes; e Tyrosine Kinase 2 (Tyk2), com a entrada 4GIH e 16 ligantes.

Todos os ligantes utilizados nesses sistemas tiveram suas estruturas tridimensionais previamente disponibilizadas no artigo original. Tais conformações experimentais serviram como referência para o cálculo do Root Mean Square Deviation (RMSD), comparando as poses obtidas por meio do DockThor com as conformações utilizadas para o cálculo de FEP no artigo original.

Até o momento, foram concluídas apenas as simulações correspondentes ao sistema MCL1, o qual apresenta o maior número de ligantes entre os conjuntos analisados. As simulações para os demais sistemas-alvo estão em andamento, seguindo os mesmos critérios metodológicos estabelecidos. Ressalta-se que a finalização dessas etapas será abordada na conclusão do projeto, uma vez que ainda estão em fase de execução.

Figura 1 - Correlação entre energias livres de ligação preditas por FEP e DockTDeep em relação aos valores experimentais para o sistema MCL1



Para o sistema MCL1, foi conduzida uma análise comparativa entre os erros de predição de afinidade gerados pelos métodos DockTDeep e FEP, tomando como referência os valores experimentais. A análise considerou a estrutura cristalográfica 4HW3 e um conjunto de 42 ligantes. Observou-se que, em 18 desses casos, os resultados indicaram que o DockTDeep apresentou energias preditas mais próximas dos valores experimentais, com um erro absoluto mínimo de 0,004 em comparação ao FEP, que teve um erro absoluto mínimo de 0,060. No entanto, a média dos erros foi ligeiramente superior no DockTDeep (1,111) em relação ao FEP (0,843), na mesma linha, o desvio padrão no FEP foi menor, com valor de 0,622, contra 0,795 do DockTDeep. A Figura 1 apresenta a correlação linear entre os valores experimentais e os valores preditos por ambos os métodos. O método FEP demonstrou um coeficiente de determinação igual a 0,596, enquanto o DockTDeep apresentou uma correlação de 0,343. Esses resultados indicam que, embora o desempenho dos métodos varia conforme o ligante analisado, o DockTDeep demonstra potencial competitivo em termos de acurácia e estabilidade.

A análise dos valores de RMSD para as poses geradas mostrou que, na maioria dos casos do sistema MCL1, as cinco melhores conformações apresentaram RMSD

inferiores a 2 Å, o que indica boa precisão na predição das posições dos ligantes. Contudo, algumas exceções foram observadas, com destaque para os ligantes 29 (RMSD = 2,600 Å), 38 (RMSD = 4,253 Å), 40 (RMSD = 6,442 Å) e 41 (RMSD = 4,159 Å), cujos valores ficaram acima do limite de 2 Å, sinalizando predições menos confiáveis. Essas discrepâncias podem estar relacionadas a características específicas desses ligantes, como a presença de determinados grupos funcionais cuja topologia pode não estar adequadamente descrita pelo campo de força empregado, porém, esses quatro casos representam apenas 9,5% do total de 42 ligantes analisados, caracterizando-se, portanto, como uma minoria. Ainda assim, os resultados gerais foram considerados favoráveis, com a maior parte dos ligantes apresentando poses com boa concordância com as estruturas de referência. Essa questão requer investigação adicional, que será realizada considerando também os resultados dos demais sistemas estudados.

5. CONCLUSÕES

Este projeto estabelece um protocolo integrado para a elucidação tridimensional de complexos proteína-ligante, combinando técnicas de ancoramento molecular e aprendizado de máquina, com o objetivo de contribuir para a predição estrutural e funcional em estudos de desenvolvimento racional de fármacos.

O teste de validação utilizando séries químicas com pequenas variações estruturais permitiu avaliar a capacidade do protocolo de docking em reproduzir a orientação e o posicionamento dos ligantes no sítio ativo, considerando modificações sutis em substituintes periféricos. Essa abordagem foi empregada no contexto das etapas de otimização de Leads.

Nos experimentos realizados com o sistema MCL1, 90,5% dos ligantes apresentaram poses com valores de RMSD inferiores a 2 Å em relação às coordenadas de referência. A análise dos dados indicou que, embora o modelo DockTDeep não tenha superado o desempenho do método FEP na predição de afinidades, ele apresentou resultados consistentes e com boa precisão em múltiplos ligantes. Destaca-se, sobretudo, pelo custo-benefício, já que demanda significativamente menos recursos computacionais e tempo de processamento, configurando-se como uma alternativa viável e eficiente para aplicações que exigem maior escalabilidade e rapidez nas análises.

Foram identificadas discrepâncias pontuais em determinados ligantes, o que sugere limitações nos parâmetros atualmente empregados no protocolo de docking, reforçando a necessidade de uma revisão metodológica para garantir a robustez e a confiabilidade do processo de validação.

A continuidade da validação em sistemas adicionais constitui uma etapa essencial para a consolidação do protocolo proposto, direcionado ao desenvolvimento racional de fármacos e à construção de modelos preditivos baseados em aprendizado de máquina. Esse processo envolve o ajuste dos parâmetros como as configurações de docking e critérios de avaliação das predições. Após a obtenção de resultados consistentes e reprodutíveis, que comprovem a robustez e a precisão do protocolo, será possível avançar para a integração dos bancos de dados. Essa integração viabilizará a aplicação escalável do método em múltiplos sistemas moleculares, consolidando um conjunto de dados ampliado com dados experimentais, acrescido de dados sintéticos de alta confiança.

REFERÊNCIAS

1. Li, H.; et al. “Computational drug development for membrane protein targets.” *Nature Biotechnology* 42 (2024): 229-242.

2. Patil, V. M.; et al“Experimental and computational models to understand protein-ligand, metal-ligand and metal-DNA interactions pertinent to targeted cancer and other therapies.” *European Journal of Medicinal Chemistry Reports* 10 (2024).

3. Guedes, I. A.; de Magalhães, Camila S.; Dardenne, Laurent E. "Receptor–ligand molecular docking." *Biophysical reviews* 6 (2014): 75-87.

4. Zou, D.; et al. "Biological databases for human research." *Genomics, proteomics and bioinformatics* 13.1 (2015): 55-63.

5. Libório, L.; Resende, V. H. "Introdução aos bancos de dados biológicos." *Bioinfo–Revista Brasileira de Bioinformática e Biologia Computacional* 1 (2021).

6. Zhou, Y.; et al. (2024). TTD: Therapeutic Target Database describing target druggability information. *Nucleic acids research*, *52*(D1), D1465-D1477.

7. da Silva, M. M. P.; et al. "Deep Learning Strategies for Enhanced Molecular Docking and Virtual Screening." (2023).

8. Patel, K.; Mani, A. “Structural Bioinformatics and Protein Structure Prediction” *Springer Nature Singapore* (2024).

9. Li, X.; et al. “A High-Quality Data Set of Protein–Ligand Binding Interactions Via Comparative Complex Structure Modeling” *Journal of Chemical Information and Modeling* 64 (2024): 2454-2466.

10. Paggi, J. M.; Pandit, A.; Dror, R. O. “The Art and Science of Molecular Docking” *Annual Review of Biochemistry* 93 (2024): 389-410.

11. Hollingsworth, S. A.; Dror, R. O. "Molecular dynamics simulation for all." *Neuron* 99.6 (2018): 1129-1143.

12. Turzo, S. B. A.; Hantz, E. R.; Lindert, S. “Applications of machine learning in computer-aided drug discovery.” *QRB Discovery* 3 (2022).

13. Berman, H. M.; et al. “The Protein Data Bank” *Nucleic Acids Research* 28 (2000): 235-242.

14. Liu, T.; et al. “BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.” *Nucleic Acids Research* (2007).

15. Mendez, D.; et al. “ChEMBL: towards direct deposition of bioassay data.” *Nucleic Acids Research* 47 (2019).

16. Guedes, I. A.; et al. “New machine learning and physics-based scoring functions for drug discovery” *Scientific Reports* 11 (2021).

17. Wang et al. “Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field” *Journal of the American Chemical Society* 137 (2015): 2695–2703