Modelo Multi-Escala Simplificado para Previsão e Disseminação de Variantes Virais

Projeto de Sistemas Inteligentes para a Bioinformática

2025/2026

Contexto e Motivação

A vigilância epidemiológica de vírus respiratórios, como o da gripe (Influenza), representa um desafio constante. A capacidade do vírus de sofrer mutações pode levar ao surgimento de novas variantes com maior transmissibilidade, escape imune ou resistência a antivirais. A previsão destas ameaças exige uma abordagem que conecte as alterações a nível molecular com o seu impacto a nível populacional. Este projeto desafia os alunos a construir um modelo computacional multi-escala, integrando técnicas avançadas de Inteligência Artificial e Bioinformática. O objetivo é criar um sistema que utilize dados moleculares para prever a "aptidão" de variantes virais e, em seguida, use essa previsão para simular a sua disseminação, fornecendo uma ferramenta de apoio à decisão mais robusta e mecanística.

Objetivo Geral

Desenvolver um modelo computacional que:

1. A nível molecular, preveja o impacto funcional de mutações genéticas no vírus (e.g., Influenza H5N1).
2. A nível populacional, utilize essas previsões para simular a disseminação de diferentes variantes.

O sistema final será usado para avaliar o impacto potencial de novas variantes e testar a eficácia de diferentes políticas públicas de intervenção.

Metodologia e Requisitos Fundamentais

O projeto está estruturado em dois módulos interdependentes, fundamentados por pesquisa bibliográfica.

Módulo 1: Previsão de Aptidão Viral a Nível Molecular

Neste módulo, o grupo irá desenvolver um modelo de machine learning para prever a aptidão (*fitness*) de uma variante viral. Este modelo deve integrar inputs de pelo menos dois tipos de dados biológicos distintos, explorando problemas como:

- **Previsão de Resistência a Fármacos**: Modelar a interação entre um composto antiviral (representado por descritores químicos) e a sua proteína-alvo viral (representada pela sua sequência). O objetivo é prever se uma mutação confere resistência, prevendo, por exemplo, uma alteração no IC50.
- **Análise de Dados Multi-Ómicos**: Integrar dados genómicos do vírus com dados do hospedeiro (e.g., expressão genética de linhas celulares, dados clínicos de suscetibilidade) para modelar a virulência.

Fontes de Dados Sugeridas: GISAID, NCBI Virus, Protein Data Bank (PDB), bases de dados de fármacos (e.g., DrugBank, PubChem), e repositórios de dados de expressão (e.g., GEO).

Módulo 2: Simulação Epidemiológica da População

Neste módulo, o grupo irá desenvolver o simulador populacional que utiliza os outputs do Módulo 1 para:

- **Integração Multi-Escala**: O parâmetro chave de aptidão (*fitness*) de cada variante viral, previsto pelo Módulo 1, será usado como input dinâmico no Módulo 2. Este valor irá governar a probabilidade de transmissão, o período de infeção ou a severidade da doença no modelo de simulação.
- **Modelo de Simulação**: O output do modelo preditivo deverá ser integrado num modelo de simulação (e.g., Modelo Baseado em Agentes - ABM) para projetar a disseminação espacial e temporal do vírus na população.

Atividades Esperadas

- **Extração e Tratamento de Dados**: Os alunos devem identificar, recolher e processar dados de, no mínimo, duas fontes públicas distintas:
  - **Dados Ômicos**: Sequências genéticas de variantes virais (Ex: GISAID, NCBI Virus).
  - **Dados Epidemiológicos**: Casos, hospitalizações, taxas de vacinação, etc. (ex: FluNet da OMS, dados de agências de saúde nacionais).
- **Modelagem Preditiva de Evolução**: Utilizando técnicas de machine learning, o grupo deverá desenvolver um modelo que analise características ômicas para prever a "aptidão" (fitness) ou o potencial de disseminação de diferentes variantes virais.
- **Modelo de Simulação**: O output do modelo preditivo deverá ser integrado num modelo de simulação para projetar a disseminação espacial e temporal do vírus na população.
- **Análise de Cenários**: O modelo final deve ser usado para simular o impacto de pelo menos dois cenários de intervenção de saúde pública (ex: implementação de medidas de distanciamento, campanhas de uso de máscara).

Requisito de Inovação

É obrigatória a inclusão, aplicação e justificação de, pelo menos, duas abordagens ou técnicas avançadas que não foram diretamente lecionadas no conteúdo programático da unidade curricular. Sugestões incluem (mas não se limitam a):

- Modelos Baseados em Agentes (Agent-Based Models) para a simulação epidemiológica.
- Modelos de Deep Learning para análise de sequências genómicas (e.g., Transformers, LSTMs).
- Análise Topológica de Dados (TDA) para a identificação de padrões em dados genéticos complexos.
- Aprendizagem por Reforço (Reinforcement Learning) para otimizar estratégias de intervenção simuladas.

Obs.:

- Uma das abordagens pode ser substituida pela validação empírica da capacidade do modelo ao comparar os resultados de uma simulação de um período passado com os dados observados nesse mesmo período.
- Uma das abordagens pode ser substituida por um estudo comparativo, isto é, analisar diferenças na performance a partir da comparação com o resultado de outros estudos semelhantes, presentes na literatura.

Avaliação e Arguição (Total: 20 Valores)

A avaliação dos trabalhos será feita com base nos seguintes componentes:

- Apresentação Intermédia (5 de janeiro de 2026): Peso de 40%. Nesta apresentação, os grupos deverão mostrar o progresso do trabalho, incluindo a revisão de literatura, a seleção e tratamento de dados, e a arquitetura inicial dos modelos.
- Notebook Final (submissão a 25 de janeiro de 2026): Peso de 60%. O notebook deve ser um documento completo e reprodutível, detalhando toda a revisão de literatura, metodologia, implementação, resultados e conclusões. Sugere-se o R notebook.

A. Apresentação Intermédia (7 Valores)

Avalia o progresso, a fundamentação e o planeamento do projeto.

Item de Aprendizagem	Descrição	Pontuação
Revisão da Literatura e Fundamentação	Profundidade da pesquisa sobre o estado da arte e clareza na definição do problema específico a abordar.	**3,0**
Seleção e Análise Exploratória dos Dados	Justificação da escolha dos datasets; qualidade da análise exploratória inicial para compreender os dados.	**3,0**
Clareza da Apresentação e Planeamento	Qualidade da comunicação oral; apresentação de um plano de trabalho detalhado e realista para a fase final.	**1,0**

B. Notebook Final e Arguição (13 Valores)

Avalia o trabalho final entregue e a sua defesa oral.

Item de Aprendizagem	Descrição	Pontuação
Módulo 1: Qualidade da Modelagem Molecular	Rigor na implementação do modelo de ML, incluindo pré-processamento, feature engineering, treino e avaliação.	**4,0**
Módulo 2: Qualidade da Simulação Epidemiológica	Robustez do modelo de simulação e da integração multi-escala com os outputs do Módulo 1.	**4,0**
Aplicação do Requisito de Inovação	Profundidade e justificação da aplicação das duas técnicas avançadas não lecionadas.	**2,0**
Validação, Análise de Cenários e Conclusões	Qualidade da validação empírica, profundidade da análise dos cenários e pertinência das conclusões.	**2,0**
Qualidade do Código e Reprodutibilidade	Clareza, organização e documentação do código. O notebook deve ser totalmente executável e reprodutível.	**1,0**

Nota Importante: Espera-se que **todos os membros do grupo** demonstrem um domínio profundo sobre a **totalidade do trabalho desenvolvido**. A incapacidade de um membro de justificar uma parte do trabalho poderá impactar negativamente a avaliação de todo o grupo.

C. Sobre a Complexidade do Projeto e Avaliação

Este projeto foi desenhado para ser flexível, permitindo que a sua complexidade seja ajustada pela equipa. Um projeto mais simples, mas robusto, bem justificado e funcional, será sempre mais bem avaliado do que um projeto excessivamente ambicioso que não atinja os seus objetivos.

D. Principais Desafios e Pontos de Atenção

Todo projeto de investigação e desenvolvimento acarreta riscos. Identificá-los antecipadamente é fundamental para o sucesso. Cabe ao grupo planear e gerir estes riscos ativamente ao longo do semestre.

O primeiro grande desafio reside na **disponibilidade e curadoria de dados**. Para mitigar este risco, é crucial adotar uma abordagem orientada pelos dados, estando preparado para ajustar o modelo aos melhores dados que encontrar. Em último caso, desde que bem justificado, pode-se adotar um dataset sintético.

Um segundo risco provém da **complexidade técnica**. A melhor estratégia é focar-se em construir uma versão *end-to-end* funcional o mais rapidamente possível, usando as técnicas que o grupo já domina. Pode-se então melhorar ou substituir incrementalmente os componentes com as técnicas mais avançadas.

Adicionalmente, existe o **risco de desempenho do modelo**. Se o modelo não converigr ou apresentar um performance ruim, o foco do trabalho deve transitar para uma análise crítica das suas limitações, discutindo em detalhe as possíveis razões, como a qualidade dos dados ou pressupostos incorretos.

Finalmente, o projeto está sujeito ao **risco de gestão de escopo** (*scope creep*). Para evitar isso, estabeleçam no início um "Produto Mínimo Viável" (MVP) claro, que deve ser o foco principal até à Apresentação Intermédia. Após esta fase, decidam de forma realista que melhorias são viáveis de implementar.

Uso de Ferramentas de IA Generativa (LLMs)

O uso de assistentes de IA generativa (como ChatGPT, Copilot, etc.) é incentivado como ferramenta de apoio à investigação, ideação e programação. Contudo, a sua utilização pressupõe uma responsabilidade académica acrescida, assim todo o conteúdo ou código gerado deve ser rigorosamente estudado, verificado e comparado com fontes académicas fiáveis (artigos científicos, documentação oficial de bibliotecas de software, etc.). Os alunos devem assimilar completamente o conhecimento, sendo capazes de explicar, justificar e derivar cada conceito, algoritmo e linha de código presente no trabalho.