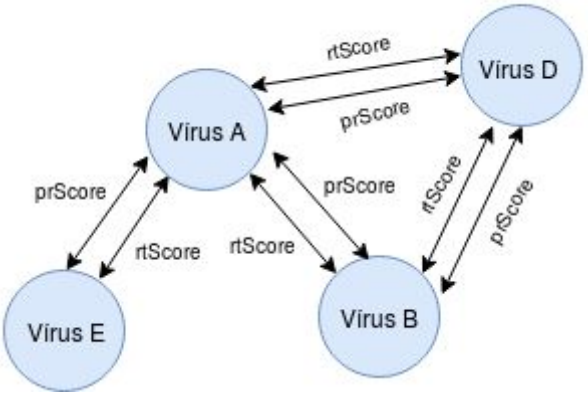
Giordano Bruno Olivetti Mattiello - 173056
Daniela Marques de Morais - 169562

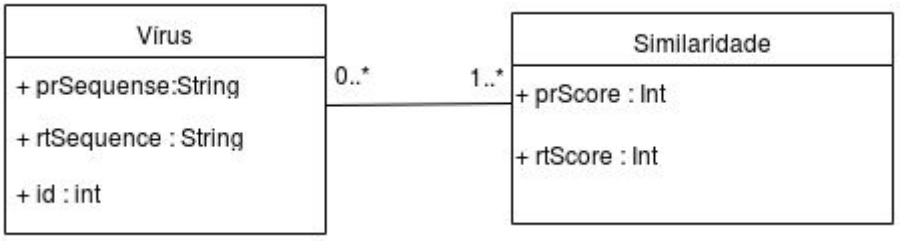
Predição de Efetividade do tratamento de HIV

Nesse projeto, pretendemos estudar a relação da estrutura viral e a efetividade do tratamento de HIV. Para isso, usaremos algoritmos de similaridade genética para estabelecer as similaridades entre os vírus (especificamente a Transcriptase Reversa (RT) e Protease (PR) que compõem seu material genético) e pretendemos utilizar Machine Learning para tentar prever a efetividade do tratamento e suas relações.

Nessa etapa, iremos utilizar o Neo4j para analisar melhor as relações entre os vírus e sua proximidade. Temos inúmeros dados e necessitamos de uma visualização gerada de forma rápida e fácil, note que os vírus possuem relações bidirecionais. Um banco de dados com suporte a grafos é ideal para essa situação, especialmente que nosso problema é facilmente modelado e solucionado como um multi grafo dirigido.

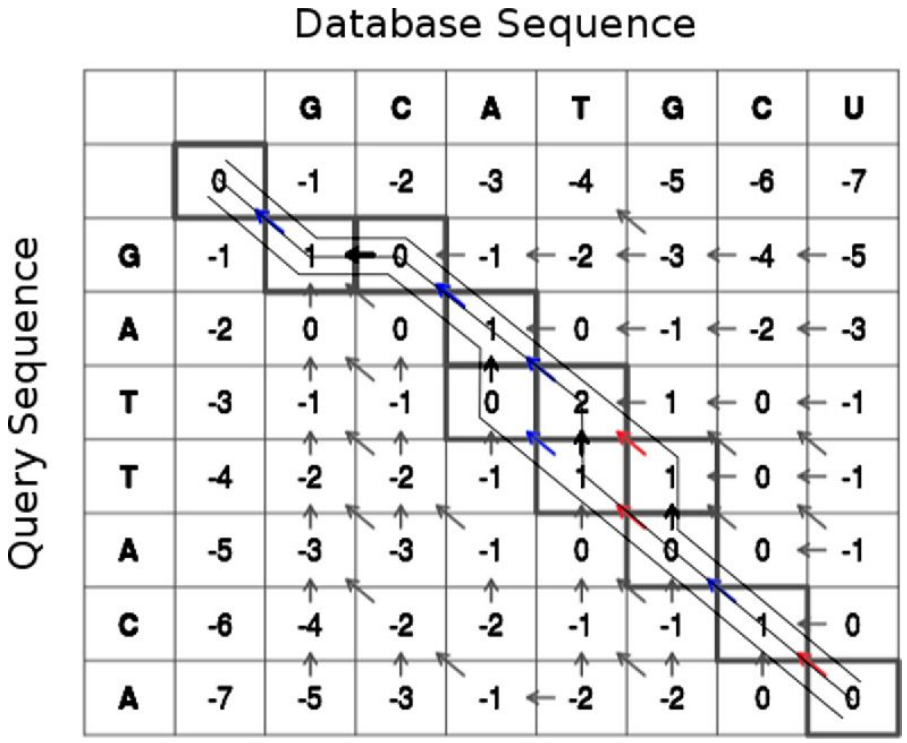


Modelo lógico



Modelo conceitual

Para encontrar as similaridades genéticas, foi utilizado o algoritmo Needleman-Wunsch, o qual propõe uma matriz de scores para determinar o quão próximos são duas sequências.



Exemplo de matriz gerada pelo algoritmo Needleman-Wunsch

Para preparar o CSV e conseguir inserir os nós no Neo4j, utilizamos Python e a biblioteca BioPython para classificar o score de similaridade.

Devido a grande quantidade de dados (um CSV com tamanho de 22MB), tivemos algumas limitações com o Neo4j online e para fins de validação executamos algumas queries em apenas um conjunto pequeno de dados.

Algumas das queries executadas foram:

Buscar os nós que possuem o mais alto score de RT
MATCH (v:Virus)-[s:Similaridade]->(v2:Virus) RETURN v,v2 ORDER BY toInteger(s.rtscore) DESC LIMIT 1

Buscar os nós que possuem o mais alto score de PR
MATCH (v:Virus)-[s:Similaridade]->(v2:Virus) RETURN v,v2 ORDER BY toInteger(s.prscore) DESC LIMIT 1

Buscar os nós que possuem o RT score acima da média
MATCH (v:Virus)-[s:Similaridade]->(v2:Virus) WHERE toInteger(s.rtscore) > 950 RETURN v,v2

Buscar os nós que possuem CD4 num nível crítico (abaixo de 50)
MATCH (v:Virus)-[s:Similaridade]->(v2:Virus) WHERE toInteger(v.cd4) < 50 AND toInteger(v2.cd4) < 50 RETURN v,v2

Buscar a média dos scores de RT e PR
MATCH (v:Virus)-[s:Similaridade]->(v2:Virus) RETURN avg(toInteger(s.rtscore)),avg(toInteger(s.prscore))

Nas próximas etapas estimamos conseguir adicionar todos os dados no Neo4j numa máquina local (fora da Cloud disponibilizada gratuitamente) e poder analisar profundamente os resultados das queries.