UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG
SISTEMAS DE INFORMAÇÃO - DCC
PROPOSTA DE TRABALHO DE CONCLUSÃO DE CURSO II

Mineração e Análise de Dados de Bioinformática Estrutural: Análise de Interações Proteína-RNA e Determinação da Confiabilidade de Modelos de Predição Computacional

Pesquisa Mista

Aluno: João Pedro Braga Ennes - 2021422350
Orientadora: Raquel Cardoso de Melo Minardi

Belo Horizonte - MG
2023

SUMÁRIO

1. INTRODUÇÃO

A relação entre proteínas e RNA é de suma importância para a compreensão dos processos biológicos que regem os organismos vivos. Essa interação desempenha um papel essencial em diversas funções celulares, desde o controle da expressão dos genes até a preservação da integridade do material genético (PIRES et al., 2013). Nesse contexto, o entendimento dessas interações têm implicações significativas para a pesquisa biomédica e o avanço do conhecimento na área biológica. Contudo, apesar de sua relevância, a existência de um banco de dados público abrangente em bioinformática que contenha informações detalhadas sobre as interações proteína-RNA ainda é limitada.

Paralelamente, às investigações no campo das biomoléculas têm sido grandemente influenciadas pelo uso intensivo de métodos computacionais para a organização e análise dos dados (DE ARAÚJO et al., 2008). Dessa forma, um banco de dados dedicado às interações proteína-RNA se torna imprescindível para avançar nossa compreensão das complexas redes regulatórias que governam as células, fornecendo um conjunto robusto de dados para estudos aplicados, incluindo modelos matemáticos.

Tendo isso em vista, na primeira etapa desse projeto, foi construído um banco de dados voltado especificamente para o armazenamento de informações relacionadas às sequências e estruturas envolvidas nas interações proteína-RNA, com o objetivo de reunir e disponibilizar esses dados. Nesse contexto, foram realizadas coletas de informações a partir do Protein Data Bank (PDB) (BERMAN et al., 2002), a fim de agregar dados de estruturas moleculares previamente sequenciadas e construir um recurso valioso para a pesquisa em bioinformática. Este banco de dados se apresenta como um recurso de relevância para a comunidade científica, permitindo o acesso a informações estruturais cruciais para a compreensão das interações proteína-RNA.

A presente proposta de monografia visa, portanto, realizar análises dos dados coletados e armazenados no banco previamente construído, buscando identificar afinidades entre as interações de diferentes proteínas com moléculas de RNA. Além disso, será realizada também, uma comparação entre as estruturas obtidas por meio de técnicas experimentais e estruturas construídas a partir de modelos de predição computacional como o Rosetta (LEAVER-FAY et al., 2011) , com objetivo de avaliar os resultados obtidos e buscar determinar a confiabilidade de tais modelos.

Ao longo deste estudo, será demonstrado que a criação de um banco de dados de interações proteína-RNA não apenas representa um marco importante na compreensão da bioinformática estrutural, mas também se configura como uma ferramenta poderosa para impulsionar descobertas científicas e aplicações clínicas inovadoras. Assim, aprofundar o conhecimento nesse campo tem o potencial de revolucionar a pesquisa biomolecular, fornecendo avanços significativos para estudos acadêmicos e servindo como referência para a prática médica e biomédica ao direcionar novas abordagens para diagnósticos, tratamentos e prevenção de doenças.

2. REFERENCIAL TEÓRICO

A interação entre proteínas e RNA desempenha um papel fundamental nos processos biológicos, afetando a expressão gênica, a estabilidade do RNA e a regulação de diversas funções celulares (PIRES et al., 2013). As proteínas se ligam às moléculas de RNA para executar funções específicas, muitas vezes reconhecendo sequências estruturais ou sequências particulares de RNA. A análise dos dados moleculares provenientes do sequenciamento de DNA e RNA é crucial para entender as características estruturais dos segmentos gênicos e de seus produtos protéicos, além de elucidar suas interações e compreender processos biológicos como tradução, processamento de RNA e regulação pós-transcricional (JÚNIOR et al., 2022).

Apesar de sua importância, a investigação do sequenciamento estrutural das interações proteína-RNA ainda é pouco explorada. No PDB, um banco de dados essencial para a biologia estrutural, que contém estruturas tridimensionais de proteínas, existem mais de duzentas mil estruturas experimentais registradas. No entanto, apenas pouco mais de sete mil estão relacionadas a uma estrutura de RNA após uma rápida filtragem.

Os bancos de dados biológicos, como o PDB, são depósitos de informações biológicas que incluem sequências de proteínas, RNA, estruturas tridimensionais, dados de expressão gênica e interações biomoleculares. Esses bancos são cruciais para a pesquisa biomolecular, fornecendo acesso a informações valiosas para análises e descobertas (MARIANO et al., 2015). Neste projeto, propomos inicialmente a construção de um banco de dados específico para interações proteína-RNA e, em uma segunda etapa, pretendemos utilizá-lo para análises dessas interações e avaliação de ferramentas de modelagem de estruturas que utilizam predição computacional.

Na área de bioinformática, os avanços da predição computacional vêm permitindo a obtenção de informações estruturais com base na sequência de aminoácidos de uma proteína cuja estrutura ainda não foi determinada experimentalmente (SILVA et al., 2021). Anteriormente, essa forma de predição era considerada um desafio, contudo, devido ao progresso dos algoritmos computacionais ao longo dos anos e à disponibilidade crescente de estruturas protéicas conhecidas, tornou-se viável, resultando em previsões plausíveis e razoavelmente precisas em muitos casos. Entretanto, é importante reconhecer que os métodos de predição estrutural computacional apresentam limitações que requerem uma avaliação cuidadosa para determinar a confiabilidade dos modelos gerados. Portanto, é de suma importância realizar uma avaliação desses modelos preditivos, a fim de estimar o nível de confiança a ser atribuído às estruturas preditas.

3. METODOLOGIA

O projeto será executado por meio de um conjunto de etapas principais que abrangem diversas atividades detalhadas e específicas, garantindo uma abordagem completa e eficiente:

1. Atualização dos dados
    - Realizar novamente a pesquisa de estruturas já sequenciadas relativas a interações proteína-RNA, visando obter dados que não haviam sido registrados até a data da última coleta.
    - Exportação dos arquivos em texto para tratamento e construção do banco de dados.
    - Passagem dos dados por todas as etapas de remoção de resíduos e análises, visando a integração desses novos dados ao banco já existente.
2. Análise dos dados
    - Buscar similaridades e identificar padrões a partir do banco construído na primeira etapa do projeto.
3. Análise de modelos estruturais gerados por predição computacional
    - Utilizar nossos dados do banco, para gerar e analisar algumas estruturas construídas por meio de predição computacional, visando coletar informações para comparação futura com estruturas reais
4. Comparação das estruturas nativas e geradas por predição

- Comparar as estruturas geradas por modelos de predição das interações proteína-RNA que coletamos com aquelas nativas geradas por meio de técnicas experimentais.

Com esses passos, pretende-se criar um banco de dados completo e com informações limpas, para que, após sua publicação, seja utilizado para pesquisas e análises de interações moleculares proteína-RNA de forma eficiente.

4. RESULTADOS ESPERADOS

Ao final deste trabalho, espera-se obter informações de semelhanças e padrões notados nos dados reunidos no banco, visando auxiliar pesquisas na área de biologia molecular e bioinformática estrutural, voltadas para o estudo de previsões de interações proteína-RNA. Através dos meios expostos na seção de metodologia, pretende-se analisar os dados previamente coletados sobre as estruturas já sequenciadas por meio de técnicas experimentais, referentes à interação proteína-RNA, e buscar entender os elementos que fazem com que essas moléculas venham a interagir. Com isso, espera-se determinar padrões que possam ser utilizados para previsão destas interações, e identificar possíveis limitações e desafios enfrentados durante o processo.

Outro resultado esperado é determinar a confiabilidade dos modelos atuais de predição computacional, por meio de comparações entre as estruturas nativas e aquelas construídas a partir destes modelos. Dessa forma, temos como objetivo final, tentar determinar qual a acurácia dos modelos computacionais como o Rosetta, que são amplamente utilizados para prever estruturas ainda não conhecidas.

5. ETAPAS E CRONOGRAMA

	01/04 - 07/04	**08/04 - 14/04**	**15/04 - 21/04**	**22/04 - 28/04**	**29/04 - 05/05**	**06/05 - 12/05**	**13/05 - 19/05**	**20/05 - 26/05**	**27/05 - 02/06**	**03/11 - 09/11**	**10/11 - 16/11**	**17/11 - 27/06**
Pesquisa sobre predição computacional	X											
Levantamento bibliográfico	X	X	X	X	X	X	X	X	X	X		
Atualização dos dados		X	X									
Análise dos dados			X	X								
Análise de modelos gerados por predição					X	X	X	X				
Elaboração Apresentação parcial - Pitch						X	X					
Comparação dos dados com estruturas nativas							X	X	X	X		
Elaboração Apresentação final - Pitch											X	X
Elaboração de relatório final											X	X

REFERÊNCIAS BIBLIOGRÁFICAS

BERMAN, Helen M; BATTISTUZ, Tammy; BHAT, Talapady N; *et al*. The Protein Data Bank. **Acta Crystallographica Section D-biological Crystallography**, v. 58, n. 6, p. 899–907, 2002. Disponível em: <https://scripts.iucr.org/cgi-bin/paper?an0594>. Acesso em: 11 set. 2023.

CHENG, Clarence Yu; CHOU, Fang-Chieh; DAS, Rhiju. Modeling complex RNA tertiary folds with Rosetta. In: **Methods in enzymology**. Academic Press, 2015. p. 35-64.

DE ARAÚJO, Nilberto Dias et al. A era da bioinformática: seu potencial e suas implicações para as ciências da saúde. **Estudos de biologia**, v. 30, n. 70/72, 2008. Disponível em: <https://biblat.unam.mx/hevila/Estudosdebiologia/2008/vol30/no70-72/16.pdf>. Acesso em: 12 set. 2023.

HERBERT, Katherine G; JUNILDA SPIROLLARI; WANG, Jianli; et al. Bioinformatic Databases. **Wiley Encyclopedia of Computer Science and Engineering**, 2007. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470050118.ecse561>. Acesso em: 13 set. 2023.

JÚNIOR, Adenivaldo Lima Filgueira et al. OS BANCOS DE DADOS DE INFORMAÇÃO BIOLÓGICA E SUA POTENCIAL APLICABILIDADE ÀS CIÊNCIAS MÉDICAS: UMA REVISÃO. **Visão Acadêmica**, v. 23, n. 1, 2022.

LEAVER-FAY, Andrew et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. In: **Methods in enzymology**. Academic Press, 2011. p. 545-574.

MARIANO, D. C. B.; BARROSO, J. R. P. M. ; CORREIA, T. S. ; de MELO-MINARDI, R. C. . **Introdução à Programação para Bioinformática com Biopython**. 3. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2015. v. 1. 230p .

PIRES, Douglas E V; RAQUEL; CARLOS; *et al*. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics**, v. 29, n. 7, p. 855–861, 2013. Disponível em: <https://academic.oup.com/bioinformatics/article/29/7/855/253252>. Acesso em: 11 set. 2023.

SILVA, Letícia X; BASTOS, Luana L; SANTOS, Lucianna H. Modelagem computacional de proteínas. **BIOINFO-Revista Brasileira de Bioinformática e Biologia Computacional**, v. 1, n. 8, 2021. Disponível em: <https://bioinfo.com.br/wp-content/uploads/2021/07/08_Modelagem-computacional-de-proteinas-_-BIOINFO_compressed.pdf>. Acesso em: 05 jan. 2024.

WU, Cathy H; YEH, Lai-Su L; HUANG, Hongzhan; *et al*. The Protein Information Resource. **Nucleic Acids Research**, v. 31, n. 1, p. 345–347, 2003. Disponível em: <https://academic.oup.com/nar/article/31/1/345/2401247>. Acesso em: 13 set. 2023.