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

Mineração e Análise de Dados de Bioinformática Estrutural: Construção de um Banco de Dados de Interações Proteína-RNA

Pesquisa Mista

Aluno: João Pedro Braga Ennes - 2020425127
Orientadora: Raquel Cardoso de Melo Minardi

Belo Horizonte - MG
2023

SUMÁRIO

1. INTRODUÇÃO

A relação entre proteínas e RNA é fundamental para a compreensão dos processos biológicos que regem a vida. Essa interação desempenha um papel crucial em funções celulares, desde a regulação da expressão gênica até a manutenção da integridade do genoma (PIRES et al., 2013). A compreensão dessas interações tem implicações profundas para a pesquisa biomédica e o avanço do conhecimento biológico. No entanto, apesar de sua importância, a disponibilidade de um banco de dados públicos (BDBs) em bioinformática abrangente e acessível que contenha informações detalhadas sobre interações proteína-RNA ainda é limitada.

Juntamente a isso, as pesquisas na área de biomoléculas têm sido amplamente guiadas pelo uso intensivo de métodos computacionais para a organização e análise das informações (DE ARAÚJO et al., 2008). Isso faz com que um banco de dados dedicado a interações proteína-RNA seja essencial para avançar nossa compreensão das complexas redes regulatórias que governam a vida celular, por proporcionar um conjunto de dados sólido, para utilização em estudos aplicados com utilização de modelos matemáticos.

Esta proposta de monografia busca inicialmente, portanto, construir um banco de dados que se concentre especificamente em armazenar informações sobre sequências e estruturas envolvidas em interações proteína-RNA, onde será possível reunir e disponibilizar estes dados. Neste contexto, serão realizadas coletas de informações a partir do Protein Data Bank (PDB) (BERMAN et al., 2002), a fim de reunir dados de estruturas moleculares já sequenciadas e construir um recurso valioso para a pesquisa na área de bioinformática. Este BDBs será um recurso de relevância para a comunidade de pesquisa, permitindo o acesso a informações estruturais que são fundamentais para a compreensão das interações proteína-RNA. Em seguida, com os dados devidamente tratados e armazenados, pretende-se realizar uma análise, visando buscar afinidades entre as interações das diferentes proteínas com o RNA.

Ao longo deste trabalho, será demonstrado que a criação de um banco de dados de interações proteína-RNA não é apenas um passo importante na compreensão da bioinformática estrutural, mas, também, uma ferramenta poderosa para impulsionar descobertas científicas e aplicações clínicas inovadoras. O aprofundamento do conhecimento nesse nicho tem o potencial de revolucionar a pesquisa na área das biomoléculas e oferecer

avanços para pesquisas acadêmicas, além de ser referência para a prática médica e biomédica, ao direcionar novas abordagens para diagnósticos, tratamentos e prevenção de doenças.

2. REFERENCIAL TEÓRICO

As interações proteína-RNA desempenham um papel vital nos processos biológicos, influenciando a expressão gênica, a estabilidade do RNA, a tradução e a regulação de uma variedade de funções celulares (PIRES et al., 2013). As proteínas interagem com moléculas de RNA para desempenhar funções específicas, muitas vezes reconhecendo sequências estruturais ou sequências específicas de RNA. O estudo dos dados moleculares gerados a partir de análises de sequenciamento das moléculas de DNA e RNA, são essenciais para decifrar características estruturais de segmentos gênicos e de seus produtos proteicos, bem como estabelecer suas interações e compreender processos biológicos, como tradução, processamento de RNA e regulação pós-transcricional (JÚNIOR et al., 2022).

Apesar de possuir tal importância, o sequenciamento estrutural de interações proteína-RNA ainda é uma área pouco explorada. No PDB, banco de dados fundamental nas áreas de biologia estrutural, que armazena estruturas tridimensionais de proteínas, podemos encontrar mais de duzentas mil estruturas experimentais registradas, mas, ao fazer uma filtragem rápida, pode ser notado que apenas sete mil têm alguma relação com uma estrutura da RNA.

Bancos de dados biológicos, como o PDB, são repositórios de informações biológicas que incluem sequências de proteínas, RNA, estruturas tridimensionais, informações de expressão gênica e interações biomoleculares. Esses BDBs são essenciais para a pesquisa de biomoléculas, de forma que permitem o acesso a informações valiosas para análises e descobertas (MARIANO et al., 2015). No contexto deste projeto, o foco será a construção de um banco de dados específico para interações proteína-RNA e utilização para análise e busca de similaridades entre tais interações.

Na área de bioinformática, é possível encontrar três tipos de BDBs, que variam de acordo com o conteúdo neles armazenado. O PDB, citado anteriormente, é um exemplo de banco primário que propõe armazenar os dados originais de sequências de nucleotídeos e proteínas utilizando arquivos de texto simples para armazenar as informações. Além deste, temos também bancos secundários que, por sua vez, são utilizados para armazenar resultados de análises feitas a partir de dados primários, esses já utilizam de Sistemas de Gerenciamento

de Banco de Dados, como MySQL, PostgreSQL, ORACLE, etc. Um exemplo de banco secundário também especializado em informações referentes às proteínas é o Protein Information Resources (PIR) (WU et al., 2003). Por fim, temos os bancos de dados especializados, que se diferem dos outros por atenderem a um interesse particular de pesquisa, ou seja, são especializados para um particular organismo ou tipo de dado (HERBERT et al., 2007).

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. Coleta de dados
    - Pesquisa de estruturas já sequenciadas relativas a interações proteína-RNA.
    - Exportação dos arquivos em texto para tratamento e construção do banco de dados.
2. Análise e tratamento dos dados
    - Análise dos dados obtidos do PDB, identificando a estrutura dos arquivos e características essenciais para tratamento dos dados.
    - Desenvolvimento de um software para tratamento dos dados, visando remoção de resíduos gerados durante o processo de sequenciamento da estrutura, como moléculas de água e estruturas duplicadas.
3. Construção do banco de dados
    - Desenvolvimento de uma estrutura para armazenamento dos dados já prontos, após o tratamento.
    - Alimentação do banco com os dados obtidos.
4. Disponibilização dos dados
    - Desenvolvimento de uma ferramenta para que o banco de dados criado possa ser disponibilizado de forma pública.
    - Publicação da ferramenta desenvolvida.
5. Análise dos dados
    - A partir dos dados coletados, gerar um modelo que procure similaridades entre os dados de interações proteína-RNA

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 primeiro semestre de trabalho, espera-se obter um banco de dados que possa auxiliar pesquisas na área de biologia molecular e bioinformática estrutural, voltadas para o estudo das interações proteína-RNA. Através dos meios expostos na seção de metodologia, pretende-se coletar dados relevantes sobre as estruturas já sequenciadas, referentes à interação molecular proteína-RNA, e disponibilizá-las em um banco unificado que permita um acesso mais fácil e simplificado a estas informações. Além disso, pretende-se realizar um tratamento nos dados coletados, com intuito de remover resíduos gerados durante a estruturação das moléculas, e identificar possíveis limitações e desafios enfrentados durante o processo.

Outro resultado esperado é o desenvolvimento de uma ferramenta para que esse banco possa ser disponibilizado de forma pública e utilizado de maneira fácil por pessoas interessadas na área. Além disso, outro objetivo almejado, é que os dados coletados possam contribuir para o avanço da pesquisa e do desenvolvimento de técnicas e ferramentas que auxiliem no estudo de interações entre proteínas e RNA.

Para a segunda parte do projeto, temos como objetivo final da pesquisa, realizar uma análise a partir dos dados reunidos no banco gerado pela da coleta. Com este passo, pretende-se buscar afinidades entre as interações das diferentes proteínas com o RNA, a fim de investigar a possibilidade de prever a interação entre uma proteína qualquer e uma molécula de RNA.

5. ETAPAS E CRONOGRAMA

	10/09 - 16/09	17/09 - 23/09	24/09 - 30/09	01/10 - 07/10	08/10 - 14/10	15/10 - 21/10	22/10 - 28/10	29/10 - 04/11	05/11 - 11/11	12/11 - 18/11	19/11 - 25/11	26/11 - 02/11
Pesquisa do tema	X											
Levantamento bibliográfico	X	X	X	X	X	X	X	X	X	X		
Coleta de dados		X	X									
Análise e tratamento dos dados			X	X	X	X						
Elaboração Apresentação parcial - Pitch					X							
Construção do banco de dados							X	X				
Disponibiliza-ção dos dados									X	X		
Elaboração Apresentação final - Pitch											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.

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.

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.

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.