CENTRO UNIVERSITÁRIO SAGRADO CORAÇÃO

ROBSON MELCHIADES NETO JUNIOR

REVISÃO E APLICAÇÃO DE TÉCNICAS DE ANÁLISE DE DADOS PARA O PROCESSAMENTO DE DADOS RNA-SEQ

BAURU

2022

ROBSON MELCHIADES NETO JUNIOR

REVISÃO E APLICAÇÃO DE TÉCNICAS DE ANÁLISE DE DADOS PARA O PROCESSAMENTO DE DADOS RNA-SEQ

Pré-projeto de iniciação cientifica sob orientação do prof. Me. Patrick Pedreira Silva apresentada a Pró-reitoria de Pesquisa e Pós-graduação como parte dos Pré-requisitos para aprovação do conselho.

BAURU

2022

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

N469r

Neto Junior, Robson Melchiades

Revisão e aplicação de técnicas de análise de dados para o processamento de dados RNA-SEQ / Robson Melchiades Neto Junior. -- 2022.

18f. : il.

Orientador: Prof. Dr. Patrick Pedreira Silva

Monografia (Iniciação Científica em Ciência da Computação) - Centro Universitário Sagrado Coração - UNISAGRADO - Bauru - SP

1. RNA-Seq. 2. Python. 3. Bioinformática. I. Silva, Patrick Pedreira. II. Título.

Elaborado por Lidyane Silva Lima - CRB-8/9602

RESUMO

A Bioinformática é um campo de estudo de caráter interdisciplinar que atualmente encontra-se em pleno desenvolvimento com aplicações crescentes nas mais variadas áreas das ciências biológicas, tais como a medicina, agricultura, zoologia dentre outras. Por se tratar de um campo interdisciplinar, é necessário ao profissional e pesquisador da área o conhecimento e domínio de diversos tópicos e tecnologias utilizados por esse campo e é nesse aspecto que o presente trabalho busca contribuir. A proposta do presente projeto é o desenvolvimento e a disponibilização pública de um repositório online com aplicações de técnicas computacionais no campo da bioinformática, sobretudo em sua sub-área da genômica. O objetivo é que o referido repositório cumpra o papel como um modelo ao qual iniciantes da área da Bioinformática, advindos das áreas biológicas, possam encontrar exemplos detalhados com explicações claras no idioma Portugues de aplicações de técnicas computacionais requeridas na Bioinformática desenvolvidos em linguagem Python.

Palavras-chave: RNA-Seq; Python; Bioinformática

SUMÁRIO

1. INTRODUÇÃO

O campo de estudo da Bioinformática vem se desenvolvendo de maneira expressiva e apresentando grandes avanços com potencial inovador para muitos aspectos da vida humana, podendo trazer uma maior qualidade de vida para uma grande parcela populacional. Dentre os diversas subáreas desse grande campo destaca-se a genômica que, nas palavras de (ARAÚJO et al, 2008):

> De uma maneira geral, uma aplicação direta da bioinformática nos estudos de genômica refere-se à identificação de possíveis diferenças nas sequências gênicas que possam favorecer o desenvolvimento de ferramentas para melhor diagnóstico de doenças e anomalias.(ARAÚJO et al, 2008)

Esse avanço está intimamente relacionado, ao aumento e aperfeiçoamento das técnicas e recursos computacionais de análise e processamento de grandes conjuntos de dados, o que vem propiciando a extração de informação de uma quantidade cada vez maior de dados gerados nos processos de análises gênicas.

Como exemplo dessa relevância do estudo computacional na área da Bioinformática, segundo VERLI (VERLI, H., 2014) devido a alta quantidade de dados gerados no processo de sequenciamento de RNA pelo método NTS torna-se cada vez mais importante para o profissional da área o domínio e o conhecimento das técnicas de processamento de dados para uma maior eficiência nas etapas do processo.

Contudo, a aplicação de tais técnicas computacionais mostra-se uma tarefa não trivial em muitos casos, exigindo um conhecimento específico do pesquisador no uso de tais tecnologias. Daí a necessidade de se desenvolver materiais de referência com o objetivo de orientar o uso de tais técnicas.

É buscando atender a essa necessidade que o presente projeto se justifica, contribuir para a área da Bioinformática, área essa que exige um conhecimento multidisciplinar, disponibilizando um repositório público para a consulta de aplicações na área utilizando a linguagem de programação Python.

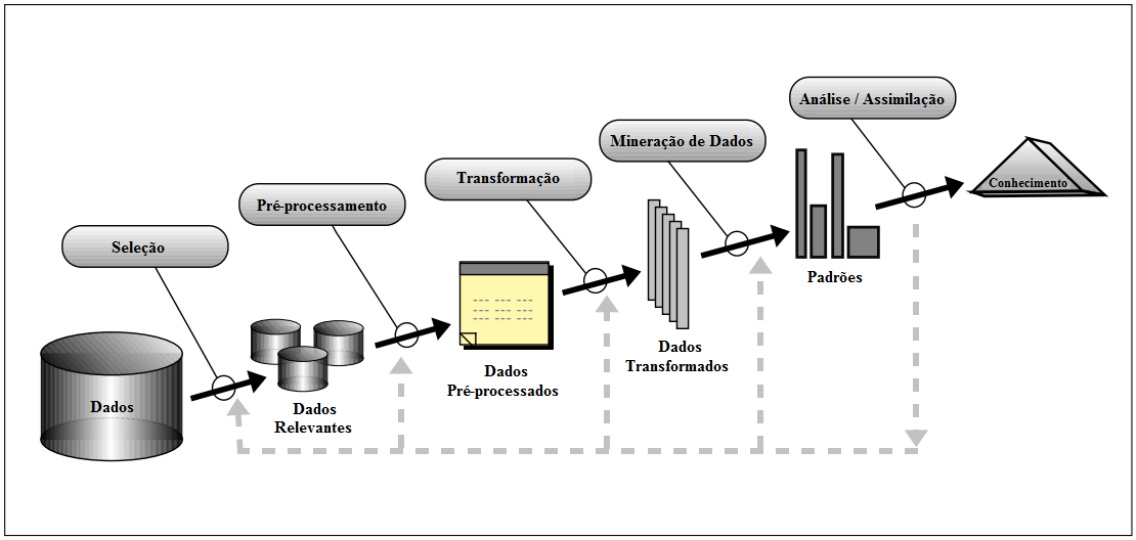
2. REFERENCIAL TEÓRICO

2.1. ANÁLISE DE DADOS

O conceito de análise de dados está relacionado a uma série de técnicas que envolvem a aquisição e tratamento dos dados, extração de informações dos dados por meio do processo de mineração, modelos de Aprendizado de Máquina dentre outros. Tais técnicas são utilizadas em conjunto visando ao entendimento e a extração de *insights* de uma base de dados históricos.

Dentre as técnicas supracitadas, a Mineração de Dados será amplamente utilizada no desenvolvimento do presente projeto. Essa técnica, tem como objetivo a transformação do dado bruto em informações que eventualmente podem se tornar conhecimento por meio da aplicação de uma série de etapas sequenciais, *pipeline* como pode-se ver na Figura 1 a seguir.

Figura 1 - Pipeline Análise de Dados



Fonte: https://www.devmedia.com.br/mineracao-de-texto-analise-comparativa-de-algoritmos-revista-sql-magazine-138/34013

A transformação dos dados brutos em conhecimento KDD (*Knowledge Discovery in Databases*) é descrito por (CORNELIUS JUNIOR, 2015) da seguinte forma:

> KDD (Knowledge Discovery in Databases) é um processo de descoberta de conhecimento em bases de dados que tem como objetivo principal extrair conhecimento a partir de grandes bases de dados. Para isto envolve diversas áreas de conhecimento, tais como: estatística, matemática, bancos de dados, inteligência artificial, visualização de dados e reconhecimento de padrões. São utilizadas técnicas, em seus diversos algoritmos, oriundos dessas áreas. (CORNELIUS JUNIOR, 2015)

Um outro conceito relacionado à análise de dados que será utilizado ao longo do desenvolvimento do presente projeto é o Aprendizado de Máquina. Este conceito, está relacionado com as técnicas que permitem a criação de previsões a partir dos dados históricos. É um conjunto de algoritmos que implementam técnicas estatísticas e matemáticas para a determinação e reprodução de padrões nos dados que em muitos casos são invisíveis em uma análise humana. (ANTONIO DAS NEVES, 2018)

2.2. BIOINFORMÁTICA

A Bioinformática é um campo de estudo multidisciplinar e, como tal, possui diferentes frentes com enfoques específicos. Para o presente projeto, será abordado com maior enfoque os ramos da Genômica e a Transcriptômica.

2.2.1. GENÔMICA

A análise genômica, tem como princípio o estudo dos genes e a sua relação com o ambiente. No estudo dos seres vivos, por exemplo, as aplicações da genômica permitem o desenvolvimento das análises filogenética por meio das quais são criadas as árvores evolutivas dos seres com base na similaridade genética possibilitando a determinação dos graus de parentesco entre as espécies.

Contudo, para o desenvolvimento para que se possa realizar um estudo dessa natureza, é necessário que se siga uma sequência lógica de procedimentos que terão como objetivo a transformação dos dados brutos

provenientes do sequenciamento genético em informações relevantes ao estudo.

2.2.2. TRANSCRIPTOMA

As proteínas são partes constitutivas e funcionais fundamentais para os seres vivos, tendo seu estudo papel fundamental na análise de desenvolvimento, comportamentos, anomalias e doenças que os aflige.

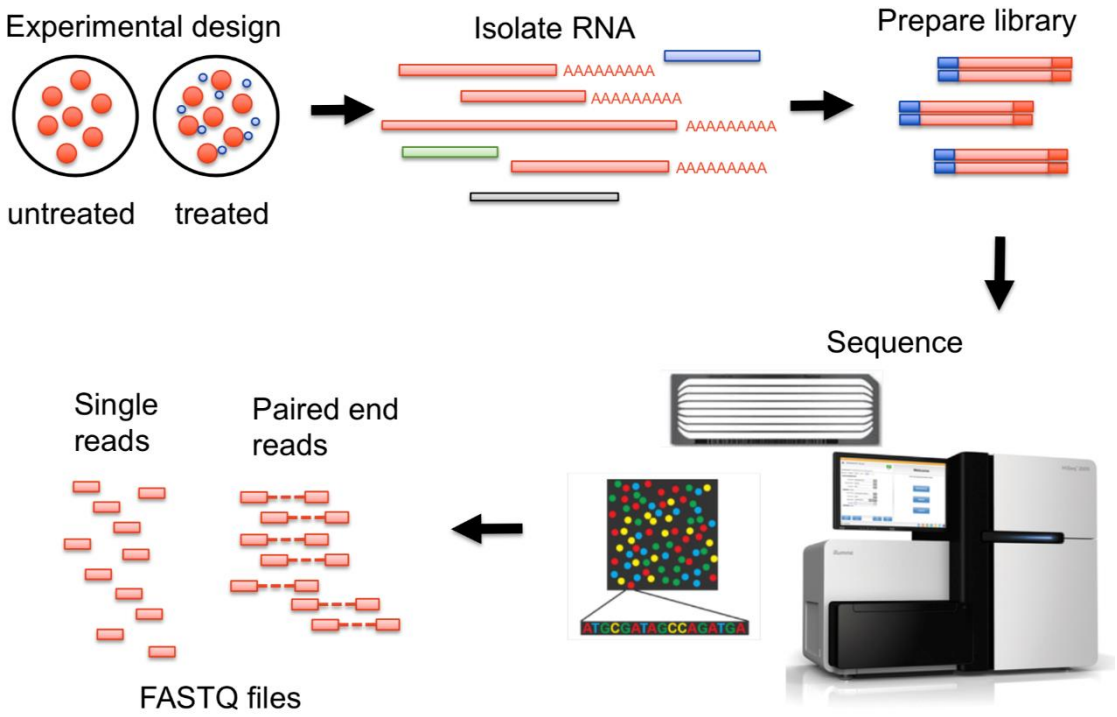
Entretanto, seu estudo é extremamente complexo devido à sua característica morfológica e constitutiva, inviabilizando em muitos casos a extração de conclusões. Por essa razão, torna-se importante a utilização de diferentes técnicas para o estudo do sequenciamento de DNA permitindo as análises comparativas e causais entre os genes. (VERLI, H., 2014).

Uma importante técnica que vem se popularizando, tanto pelo aumento de sua acessibilidade quanto pelos resultados promissores, nos estudos de genômica é o RNA-Seq, sendo os dados provenientes dessa técnica utilizados para o desenvolvimento do presente projeto.

A técnica do RNA-seq, como o nome sugere, tem como foco o sequenciamento dos transcritos de RNAs (transcriptomica), podendo ser utilizada para o sequenciamento dos diversos tipos de RNA que possuem expressões gênicas (tRNA, rRNA, RNAi, etc).

Nessa técnica, o RNA coletado após os processos de análise, separação e limpeza, é fragmentado utilizando-se enzimas específicas e submetidos para o processo de sequenciamento. O processo de sequenciamento é feito por um equipamento que analisa os fragmentos de RNA criando uma biblioteca gênica que é composta pelo sequenciamento de cada fragmento analisado *reads* (E. Korpelainen et al., 2015). Esse processo é resumido de maneira esquemática na Figura 2 abaixo.

Figura 2- Procedimentos RNA-seq



Fonte: https://sydney-informatics-hub.github.io/training-RNAseq-slides/01_IntroductionToRNASeq/01_IntroductionToRNASeq.html#4

As *reads* são um dos principais objetos de estudo desse campo sendo analisados por diversos algoritmos, primeiramente para a realização da limpeza removendo-se as *reads* de baixa qualidade. Posteriormente é realizado o processo de montagem gênica para a formação do sequenciamento completo, seguindo-se das análises computacionais sobre os dados gerados a depender do objeto de estudo.

3. OBJETIVOS

Desenvolvimento de um repositório público online destinado à exemplificação e ao detalhamento no uso de técnicas computacionais usualmente utilizadas na Bioinformática, sobretudo em sua sub-área genômica, utilizando, para tanto, a linguagem de programação Python.

3.1. OBJETIVOS ESPECÍFICOS

A. Levantar na bibliografia correlata às principais técnicas utilizadas em cada etapa do processamento e análise genômica;
B. Obter e selecionar bases de dados para análise nos repositórios públicos;
C. Criar e analisar um *pipeline* para processamento dos dados obtidos no item (b) utilizando as técnicas levantadas no item (a);
D. Desenvolver os *scripts* utilizando a linguagem Python aplicando o *pipeline* criado no item (c);
E. Promover a divulgação da experiência e resultados obtidos em eventos técnicos e científicos, publicações correlatas, e participar do Congresso Anual de Iniciação Científica e Desenvolvimento Tecnológico e Inovação.

4. MATERIAL E MÉTODOS

O método utilizado nesta pesquisa será do tipo descritivo observacional com abordagem qualitativa, na implementação de recurso tecnológico baseado em técnicas de análise de dados (data mining), a fim de refletir a aplicação potencial dessa tecnologia em análise genômica.

Como resultado desta pesquisa, será criado um repositório público online na plataforma de controle de versões e disponibilização de códigos, GitHub. Esse repositório contará, além do código na linguagem Python, com a documentação que fornecerá as informações necessárias para a utilização dos exemplos publicados em outras aplicações análogas.

A seguir será tratado de forma mais detalhada os métodos utilizados para a criação do repositório supracitado.

4.1. ASPECTOS ÉTICOS E PÚBLICO-ALVO

Embora a proposta do presente projeto tenha como objeto de estudo a análise e o processamento de materiais biológicos, não haverá contato direto ou indireto com qualquer paciente ou animal, sendo todos os dados obtidos de repositórios revisados públicos. Exclui-se também, portanto, os problemas relacionados ao sigilo das fontes dos materiais utilizados para a análise uma vez que a camada de coleta dos dados foi abstraída pelos repositórios que disponibilizam os dados de forma pública e on-line na internet.

4.2. TRABALHOS DE REFERÊNCIA

Foram selecionados 6 artigos publicados em periódicos de referência para servirem de base no desenvolvimento dos modelos que serão disponibilizados no repositório online. Buscou-se por meio dessa seleção contemplar temas pertinentes e recorrentemente tratados na área da bioinformática de modo que os modelos desenvolvidos e documentados no repositório possam ser adaptados para outras aplicações.

Os temas selecionados e seus respectivos trabalhos base foram: análise de diferencial de expressão gênica (ABBAS; EL-MANZALAWY, 2020) e (CHEN et al, 2016); análise da interação proteína-proteína (KLEIN et al., 2021) e (YANG et al, 2020); análise da localização das proteínas subcelulares (PANG et al., 2019). Serão, portanto, a princípio, construídos 3 guias que serão incorporados no repositório final.

4.3. OBTENÇÃO DOS DADOS

Os dados utilizados na pesquisa foram coletados de repositórios online onde as bases genéticas das transcrições realizadas são publicadas para o desenvolvimento de pesquisas futuras.

Dentre os diversos repositórios disponíveis dois deles foram utilizado para a coleta dos dados utilizados no desenvolvimento do projeto, o NCBI (“National Center for Biotechnology Information”) e o NIH (“The Cancer Genome Atlas Program - National Cancer Institute”).

4.4. TECNOLOGIA UTILIZADA

Para o desenvolvimento do projeto foi utilizada a linguagem de programação Python. A escolha desta linguagem deu-se pela sua alta popularidade no campo acadêmico bem como no campo da análise de dados. Especificamente com relação à Bioinformática, a linguagem Python conta com diversas bibliotecas desenvolvidas especificamente para a solução de problemas inerentes a essa área (como a biblioteca *biopython*) o que acaba automatizando e padronizando algumas etapas do processo onde são utilizadas técnicas computacionais.

4.5. REDAÇÃO FINAL E APRESENTAÇÃO DA PESQUISA

Integrante da documentação para o resultado metodológico aplicação do uso tecnologias para análise de dados, especialmente a análise genômica proposta por essa pesquisa, a redação final desse projeto deverá conter todo o levantamento bibliográfico utilizado, os materiais e métodos empregados para a elaboração da própria documentação e do produto final, os resultados alcançados, as discussões e considerações sob a aplicação que foi proposta, as referências utilizadas e, ainda, a documentação do usuário e demais anexos indispensáveis para a reprodução e evolução dessa pesquisa.

Por fim, após o término desse projeto de pesquisa, a proposta, os resultados obtidos serão apresentados no Fórum de Iniciação Científica do UNISAGRADO, a fim de compartilhar ao público interessado todos os procedimentos, limitações e singularidades do produto desenvolvido.

5. RESULTADOS

O projeto foi desenvolvido de modo a que se tenha um repositório público com exemplos práticos de aplicações da bioinformática em linguagem Python. Para a elaboração dos exemplos foram utilizadas bibliotecas da linguagem Python como *biopython* e *biotite* como no exemplo abaixo (Figura 3) em que é demonstrado como se realiza a importação de arquivos no formato .fasta.

Figura 3 - Carga Arquivos Fasta

Como carregar um arquivo no formato .fasta em um script Python

```
import biotite.sequence.io.fasta as fasta

# Carregando o Arquivo Fasta
file = fasta.FastaFile()
file.read("ebola_sequence.fasta")
```

Fonte: *Elaborado pelo autor*

Foram implementadas funções responsáveis pela execução de tarefas corriqueiras nos processos de análise em Bioinformática como Contagem de pares de bases (Figura 4), processo de transcrição, entre outros.

Figura 4 - Contagem dos Pares de Bases

Como realizar a contagem de nucleotídeos:

Python

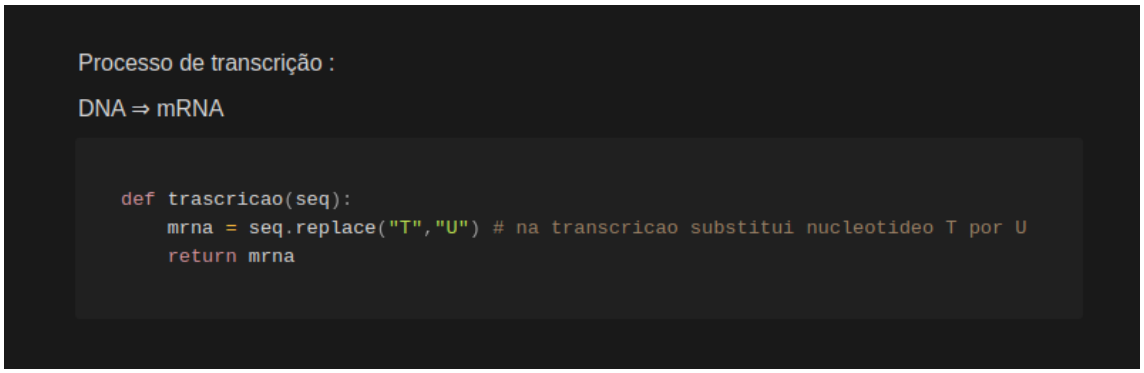
```
def retorna_contagem_nucleotideos(seq):
    base_dict = {"A":0,"T":0,"G":0,"C":0} # inicializa contagem em 0

    for base in seq:
        base_dict[base] +=1

    return base_dict
```

Fonte: *Elaborado pelo autor*

Figura 4 - Processo de transcrição



Fonte: *Elaborado pelo autor*

Além disso, foi elaborado um documento inicial descrevendo etapas gerais para a instalação das dependências utilizadas nos códigos, links úteis de suporte para instalação do Python e informações gerais sobre o projeto e sua estrutura, inclusive informações sobre contribuições.

6. CONSIDERAÇÕES FINAIS

Conforme anteriormente descrito, o campo da Bioinformática trata-se de um campo multidisciplinar exigindo, portanto, um conhecimento de múltiplas áreas do profissional e estudioso da área.

Nesse sentido, o presente projeto pôde contribuir ao fornecer uma base de referência ao qual iniciantes da área advindos de áreas diversas à computação possam ter contato com um material detalhado e de fácil compreensão de exemplos de algumas aplicações corriqueiras nos processos de análise da Bioinformática utilizando a linguagem Python.

Contudo, trata-se de uma base de conhecimento inicial, com aplicações pontuais das aplicações na área. O objetivo é que a base permaneça aberta a contribuições permitindo que se crie um material mais completo para consultas.

REFERÊNCIAS

ABBAS, M.; EL-MANZALAWY, Y. **Machine learning based refined differential gene expression analysis of pediatric sepsis**. BMC Medical Genomics, v. 13, n. 1, 28 ago. 2020.

ANTONIO DAS NEVES, SAMUEL. **Técnicas de Aprendizado de Máquina Aplicadas a Classificação da Qualidade de Pavimentos Asfálticos utilizando Smartphones**. p. 48, 2018. Disponível em: <http://www.monografias.ufop.br/bitstream/35400000/799/1/MONOGRAFIA_TécnicasAprendizadoMáquina.pdf>.

ARAÚJO, Nilberto Dias De e colab. **A Era Da Bioinformática: Seu Potencial E Suas Implicações Para As Ciências Da Saúde**. Estudos de Biologia, v. 30, n. 70/72, p. 143–148, 2008.

Chen Yifei, Yi Li, Rajiv Narayan, Aravind Subramanian, Xiaohui Xie, **Gene expression inference with deep learning**, Bioinformatics, Volume 32, Issue 12, 15 June 2016, Pages 1832–1839, https://doi.org/10.1093/bioinformatics/btw074

CORNELIUS JUNIOR, Romeu. **Uso Da Mineração De Dados Na Identificação De Alunos Com Perfil De Evasão Do Ensino Superior**. p. 138, 2015. Disponível em: <https://repositorio.unisc.br/jspui/bitstream/11624/535/1/Romeu Cornelius Junior - TCC - Final.pdf>.

E. Korpelainen, J. Tuimala, P. Somervuo, M. Huss, G. Wong **RNA-seq Data Analysis: A Practical Approach** CRC Press, Abingdon (2014)

Klein, B., Holmér, L., Smith, K.M. et al. **A computational exploration of resilience and evolvability of protein–protein interaction networks**. Commun Biol 4, 1352 (2021). https://doi.org/10.1038/s42003-021-02867-8

National Center for Biotechnology Information. Disponível em: <https://www.ncbi.nlm.nih.gov/>. Acesso em: 5 abr 2022.

Pang L, Wang J, Zhao L, Wang C, Zhan H. **A Novel Protein Subcellular Localization Method With CNN-XGBoost Model for Alzheimer's Disease**. Front Genet. 2019 Jan 18;9:751. doi: 10.3389/fgene.2018.00751.

The Cancer Genome Atlas Program - National Cancer Institute. Disponível em: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>. Acesso em: 5 abr 2022.

VERLI, H. **Bioinformática: da Biologia à Flexibilidade Molecular.** 1. ed. São Paulo: Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

Yang, F., Fan, K., Song, D. et al. **Graph-based prediction of Protein-protein interactions with attributed signed graph embedding**. BMC Bioinformatics 21, 323 (2020). https://doi.org/10.1186/s12859-020-03646-8