UNIVERSIDADE ESTADUAL DE SANTA CRUZ – UESC
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS - DCET
COLEGIADO DO CURSO DE CIÊNCIA DA COMPUTAÇÃO

Pré-projeto

Sistema para análise de proteínas: otimização na descoberta de funções e Regiões Intrinsecamente Desordenadas (IDRs)

Ilhéus - Bahia
2019

Ícaro Maradei Costa Borges

Sistema para análise de proteínas: otimização na descoberta de funções e Regiões Intrinsecamente Desordenadas (IDRs)

Pré-projeto para o Relatório de Estágio, apresentado à Universidade Estadual de Santa Cruz - UESC, Colegiado de Computação, como parte das exigências para aprovação na disciplina Estágio Supervisionado.

Orientador: Prof. Paulo Eduardo Ambrósio

Ilhéus - Bahia
2019

Sistema para análise de proteínas: otimização na descoberta de funções e Regiões Intrinsecamente Desordenadas (IDRs)

RESUMO

As proteínas são macromoléculas as quais possuem estruturas que, até alguns anos atrás, eram definidas por serem estáticas e possuírem o formato adequado para a interação com outra substância em específico, modelo chamado de chave-fechadura. No entanto, com o passar dos anos, foi descoberto que a maioria das cadeias de aminoácidos possuem as chamadas Regiões Intrinsecamente Desordenadas (IDR, do inglês *Intrinsically Disordered Regions*), que se ligam com outros substratos de forma dinâmica, e podem se moldar de acordo com a situação. Este trabalho visa, então, analisar as cadeias polipeptídicas, facilitando e automatizando seu processamento em várias etapas, observando as Regiões Intrinsecamente Desordenadas, e trazendo como resultado a função da região analisada ou da proteína por completo, porém de forma mais eficiente e mais rápida.

Palavras-chave: Proteínas, Regiões Intrinsecamente Desordenadas, bioinformática.

Sumário

1. INTRODUÇÃO

A proteína é a molécula com a maior variedade de funções existentes, portanto, sempre há uma grande gama de formas de estudo sobre ela, seja sobre sua estrutura, os elementos que a compõe, como cada uma se liga com a outra, entre outros. Para isso, há vários *softwares* disponíveis tanto para *download* quanto para uso através de sistemas *web*, onde os dados são submetidos e processados *in silico*.

Em sua grande maioria, quando representadas computacionalmente, as proteínas geram arquivos de grande volume de informação. Estas macromoléculas são formadas por cadeias de, no mínimo, 50 aminoácidos, e de acordo com Pinheiro (2019), as maiores apenas no corpo humano passam dos 30 mil aminoácidos. Porém, este número pode ser muito maior: o genoma humano, por exemplo, segundo dados do *NCBI*, *National Center for Biotechnology Information* (2019), pode gerar mais de 100 mil proteínas, estas que são compostas por mais de 70 bilhões de aminoácidos.

Além disso, foi descoberto que as proteínas possuem as chamadas Regiões Intrinsecamente Desordenadas (*IDR* – do inglês *Intrinsically Disordered Regions*), trechos estes que possuem uma estrutura dinâmica, e que podem atribuir mais de uma função a uma proteína ou a parte dela.

Sendo assim, é importante que seja desenvolvida uma ferramenta que automatize e facilite a análise das proteínas *in silico*, de forma a deixar o trabalho simples, os dados legíveis para qualquer pessoa com conhecimento em bioinformática, sem ter necessariamente um conhecimento aprofundado em computação, além de trazer resultados mais eficientes com um desempenho melhorado, seja através de menos tempo ou de melhores resultados.

1.1. Justificativa

É necessário que haja uma otimização do processamento de dados de proteínas, pois, a depender do tipo de estudo, pode-se facilmente levar dias para, por exemplo, uma cadeia de aminoácidos ser comparada com informações de um banco de dados. Para tal situação, uma análise prévia seria

fundamental para a melhora do tempo de execução ou até mesmo do desempenho.

1.2. Objetivos

1.2.1. Gerais

Desenvolvimento de um sistema *web*, em que os dados inseridos serão manipulados, de forma a facilitar o processamento de acordo com o que se espera de resultado, sendo na maioria das vezes a análise da função da proteína.

1.2.2. Específicos

a) Obtenção de informação sobre os *softwares* já disponíveis para predição de funções de proteínas, tais como formato dos dados, resultados esperados, área específica de atuação
b) Desenvolvimento de *scripts* para manipulação de arquivos e do sistema *web* com estes programas embutidos
c) Aplicação de testes e comparação com o desempenho sem a utilização da ferramenta

2. REVISÃO DA LITERATURA

Segundo Celi (2019), “Proteína é um tipo de substância formada a partir de um conjunto de aminoácidos ligados entre si. [...] existem apenas 20 tipos de aminoácidos, os quais se ligam de forma variada para originarem diferentes proteínas”. Estas ligações diferentes em conjunto com infinitas combinações de aminoácidos resultam em funções distintas para cada cadeia polipeptídica.

As proteínas podem ser representadas computacionalmente através de arquivos no formato *fasta*, onde possuem um padrão de vários cabeçalhos (iniciados pelo caractere “>” e contendo sua descrição) seguidos por suas sequências descritas anteriormente, estas que são representadas através de um conjunto de aminoácidos representados cada um por uma letra do alfabeto, sendo que são 20 aminoácidos

possíveis (as letras “B”, “J”, “O”, “U”, “X” e “Z” não são utilizadas). Na figura 1, podemos observar um arquivo *fasta*:

Figura 1 – Início do arquivo *fasta* que representa o genoma humano codificado em proteínas

Fonte: Elaborada pelo autor

Como dito anteriormente, os aminoácidos são descritos através de letras, os quais são representados de acordo com a tabela 1 abaixo:

Tabela 1 – Representação dos aminoácidos

Letra representante	Nome do aminoácido
A	Alanina
C	Cisteína
D	Ácido Aspártico
E	Ácido Glutâmico
F	Fenilalanina
G	Glicina

H	Histidina
I	Isoleucina
K	Lisina
L	Leucina
M	Metionina
N	Asparagina
P	Prolina
Q	Glutamina
R	Arginina
S	Serina
T	Treonina
V	Valina
W	Triptofano
Y	Tirosina

Fonte: Elaborada pelo autor

Estas representações em forma de arquivos computacionais demonstram apenas a chamada estrutura primária, que se trata puramente da sequência de aminoácidos que define a cadeia de peptídeos. No entanto, as proteínas possuem também as estruturas secundária, terciária e quaternária, que não podem ser representadas no arquivo *fasta*, pois estão diretamente ligadas às formas em três dimensões que a proteína pode adotar.

No entanto, embora inicialmente fosse concretizado que as proteínas possuíam estruturas físicas estáticas, foi descoberto, segundo Li (2015), em 1978 através de cristalografia de raios X e espectroscopia por ressonância magnética nuclear, que estas substâncias poderiam possuir as chamadas Regiões Intrinsecamente Desordenadas (*IDR* – do inglês *Intrinsically Disordered Regions*).

De acordo com Meng (2017) e Li (2015), as *IDR* são partes da cadeia polipeptídica que falham em formar uma estrutura 3D específica, indo contra o modelo chave-fechadura e possuindo uma certa “adaptabilidade”, permitindo à proteína modificar sua estrutura terciária para se adaptar na interação com outras substâncias. Esta transfiguração permite que uma proteína possa variar nas suas funções inicialmente definidas, e estas podem ser analisadas e preditas *in silico*.

3. MATERIAIS E MÉTODOS

Para o desenvolvimento deste trabalho, será utilizada a linguagem de programação *Python* em conjunto com o *framework Django* para desenvolvimento do sistema *web* em conjunto com seus *scripts*. Além disso, os dados a serem trabalhados, manipulados e analisados virão como resultado do DISOPRED3, InterProScan ou BLAST.

4. REFERÊNCIAS

CELI, R. **Proteínas: o que é, função e fontes de proteína!** 2019. Disponível em: < https://www.stoodi.com.br/blog/2019/03/13/proteinas-o-que-e/ >. Acesso em: 26/07/2019.

LI, J. et al. **An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014**. International Journal of Molecular Sciences, 2015. p. 23447.

MENG, F., UVERSKY, V., & KURGAN, L. **Computational prediction of intrinsic disorder in proteins**. *Current Protocols in Protein Science*, 2017. 88,2.16.1.

PINHEIRO, P. **O QUE SÃO PROTEÍNAS E AMINOÁCIDOS?** 2019. Disponível em: < https://www.mdsaude.com/nutricao/proteinas/ >. Acesso em: 30/07/2019.