INSTITUTO FEDERAL Brasília	**Instituto Federal de Brasília - Campus Taguatinga** **Superior em Computação** **Atividade Prática 1** **Dados em Bioinformática** **1/2020**

Aquisição de dados para a análise de toxinas na aranha Viúva-Negra (*Latrodectus)*

Introdução

"A viúva-negra americana **(*Latrodectus mactans*)** é uma espécie de aranha da família dos teridiídeos, distribuída por toda a América. O nome provém do fato de a fêmea geralmente se alimentar do macho após a cópula. No Brasil, é encontrada atualmente próxima ao mar, sobretudo em praias pouco frequentadas. É amplamente encontrada em torno da Baía da Guanabara. (...) Por ser uma aranha muito conhecida no Brasil e por saber-se que sua peçonha é muito forte, há a alusão de que esta é a aranha mais venenosa que existe. Porém, mesmo no Brasil, existem dois gêneros considerados de maior perigo: a aranha-marrom **(Loxosceles sp.)**, e a armadeira **(Phoneutria sp.),** sendo esta última considerada por muitos a aranha mais peçonhenta do mundo, ambas encontradas em praticamente todo o país."

Wikipedia, Latrodectus, Disponível em:
https://pt.wikipedia.org/wiki/Latrodectus

Descrição da atividade

Imagine que foi encontrada uma aranha com características muito similares à viúva negra na praia de Copacabana, porém com um veneno mais poderoso, similar às encontradas em aranhas-marrom e armadeiras. Para melhor estudar esta aranha, foi proposto um sequenciamento de seu genoma, de forma a identificar genes associados à produção de toxinas.

Como visto em sala de aula, o sequenciamento de um genoma não é capaz de produzir a sequência completa do DNA, mas sim uma imensa quantidade de **fragmentos de DNA (*reads*).** De forma a entender as características do sequenciamentos de outras aranhas "próximas" responda a questão a seguir:

1) **A primeira atividade de um bioinformata é adquirir e analisar a qualidade dessas reads. Os dados de sequenciamento são normalmente armazenados no banco de dados SRA (*Sequence Read archives)* do NCBI.**
    a) Pesquise o termo **Latrodectus** no banco de dados SRA. Quantos dados de sequenciamento foram retornados? Quantos deles são dados de DNA e quantos são de dados de RNA?
    b) Filtre a busca para DNA. Qual tecnologia de sequenciamento foi utilizada para estes dados?
    c) Escolha uma das coleções de *reads* retornado e responda:
        i) Qual seu código de acesso (*acession)?*
        ii) Qual a estratégia de sequenciamento foi usada? O que faz essa estratégia?

iii) Qual layout foi usado, *Paired-end ou single-end?* o que significam estes termos e quais as vantagens do uso de um em relação ao outro?

iv) Quantas *runs* foram realizadas? quais os tamanhos dos arquivos?

d) Volte para a página da pesquisa (com o filtro de DNA), clique no link *send to* e crie um arquivo no formato *summary*. A partir deste arquivo, faça uma tabela apresentando as características, levantadas nas questões anteriores, de cada sequenciamento.

Após o tratamento das reads, foi realizada as etapas de **filtragem**, responsável pelo descarte de *reads* de baixa qualidade, e de **montagem**, responsável por transformar *reads* em "pedaços" maiores de DNA (iremos explorar estas duas fases ao longo do curso).

Depois disso, entramos na fase de anotação, que consiste em dar um significado (função) para as sequências encontradas. Nesta fase, voltamos ao objetivo principal de nossa pesquisa, a de encontrar genes associados à produção de toxinas. Normalmente, a identificação de uma função **putativa** (ainda não validada *in-vivo*) de uma proteína é feita através da comparação das sequências obtidas pelo processo descrito anteriormente com sequências "anotadas" em banco dados biológicas.

Foi-lhe dada a tarefa de construir duas coleções de proteínas, relacionadas com a produção de toxinas, a serem utilizadas para essa comparação, **uma com dados manualmente curados e outra sem validação *in-vivo.*** Essa coleção de dados deve ser armazenada em um arquivo no formato ***fasta****.*

2) **Resolva os passos a seguir para obter a coleção de dados sem validação *in-vivo.***

   a) Entre no site do Uniprot. Busque quantas sequências de toxinas existem no TrEMBL?

   b) Ao lado da barra de pesquisa clique em Advanced. Faça uma pesquisa que filtre a busca por:

      i) Reviewed: no

      ii) All: toxin

      iii) Organism: Spider

      Quantas sequências foram retornadas? Qual informação foi adicionada à barra de pesquisa?

   c) Faça o download de todas as sequências no formato FASTA. (anexe o arquivo junto com o documento de respostas)

   d) Faça também o download do arquivo TEXT. A partir deste arquivo obtenha:

      i) AC - Accession Number de cada proteína

      ii) O nome recomendado que descreve a função de cada proteína (pesquisar o significado das linhas que começam com DE).

Por fim, crie uma tabela que relacione AC com o nome recomendado que descreve a função de cada proteína encontrada.

3) De forma análoga ao exercício anterior, crie um arquivo fasta de uma coleção de dados manualmente curados e uma tabela que descreve AC e a função de cada proteína da coleção.