Neste segundo encontro abordaremos um poucos dos metadados dos bancos de dados biológicos. Essas informações têm relevância tanto na extração do contexto dos dados (Quando? Quem? Como? Onde?), como também nos esforços por padronizá-las para tornar as análises viáveis.

Como estudo prévio, você pode também conferir quais tipo de metadados são disponibilizados pelo banco de dados aleatório que você buscou para o primeiro encontro. O banco de dados MathFiction Homepage, por exemplo, fornece metadados tanto sobre a mídia (livro, filme etc.) quanto sobre a área da matemática que tais histórias abordam (fractais, geometria etc.).

Além dos materiais de apoio de cada encontro, vamos disponibilizar também um Desafio da Semana. Caso você já se sinta confortável com algum dos tópicos da seção "Materiais de Apoio", pode ignorá-lo sem problemas, a única parte obrigatória é a instalação dos programas da parte prática antes do encontro em si.

Segundo encontro (01/10/2020)

TEMA:

Qual a importância dos metadados?

DESAFIO DA SEMANA:

Para esse segundo encontro, seu objetivo será criar um script em Python (ou terminal, se estiver com vontade de se aventurar) para limpar um arquivo de metadados de um projeto do SRA. Nesse arquivo, algumas colunas, apesar de possuírem cabeçalho, não têm dados, limpar o arquivo, portanto, é remover essas colunas do arquivo .csv.

PROGRAMAS DA PARTE PRÁTICA:

Durante a parte prática usaremos as ferramentas do motor de busca Entrez para linha de comando. Dentre as ferramentas disponíveis, usaremos o *esearch* e o *efetch*. No Linux, as ferramentas podem ser instaladas via apt, através do comando:

```
apt install ncbi-entrez-direct
```

Os parâmetros do comando *esearch* e do comando *efetch* podem ser encontrados aqui. Se se sentir confortável para se aprofundar em cada um dos comandos, há as páginas dos manuais de cada um, que possuem informações mais técnicas (esearch e efetch).

Um detalhe importante do *efetch* é a formatação dos dados, que depende do tipo de banco de dados, sendo que uma lista com essas correspondências pode ser encontrada através do comando: `efetch --help`

Usaremos também algumas ferramentas do Biopython, que pode ser baixado aqui ou instalado, tanto no Linux quanto no Windows, através do comando:

```
pip install biopython
```

Do Biopython usaremos as ferramentas do motor de busca Entrez, e alguns detalhes dos parâmetros das mesmas podem ser encontradas aqui.

As diferenças entre o Entrez para linha de comando e o Entrez do Biopython são poucas, porém importantes, isso vale tanto para os parâmetros, “query” no lugar de “term”, quanto para o tipo de dado retornado, dados em tabela no lugar de XML (Valores de saída do efetch para Biopython), respectivamente.

MATERIAIS DE APOIO

1. Introdução aos metadados

   Vídeos introdutórios curtos e em português sobre metadados em livros e em vídeos do youtube.

   Alguns textos introdutórios em inglês podem ser encontrados aqui e aqui, e em português aqui e aqui.

2. Padronização de metadados

   Texto do Instituto Europeu de Bioinformática sobre a importância de usar um vocabulário padronizado para os metadados.

   Artigo do Genomic Standards Consortium (GSC) introduzindo um dos padrões para definir a menor quantidade de metadados necessários para sequências genômicas.

3. Metadados do SRA

   O SRA aceita diversos padrões de metadados na submissão das amostras biológicas (BioSamples), que dependem do tipo de amostra utilizada. Dessa forma, saber com o que se trabalha permite descobrir que tipo de metadados você vai encontrar.

4. Exercícios de prática sobre limpeza de metadados

   Parte de um do workshops de análise de dados em genômica do Data Carpentry, nessa aula seu objetivo é limpar os metadados do arquivo de submissão, uniformizando eles e corrigindo eventuais falhas e erros de digitação.

5. Artigo de exemplo da parte prática

   O seguinte artigo: Small RNA Transcriptomes of Two Types of Exosomes in Human Whole Saliva Determined by Next Generation Sequencing, será utilizado como base para a realização da parte prática do encontro.

   Em específico, os dados do artigo no banco de dados estão disponíveis no seguinte link: <https://www.ncbi.nlm.nih.gov/bioproject/247214>, sendo que os metadados podem ser encontrados nas seções SRA Experiments (informações do sequenciamento em si) e BioSample (informações da amostra sequenciada).