Departamento de Ciências de Computação (SCC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo

SCC02713 - Introdução à Bioinformática
Trabalho Prático: Classificação de Elementos Transponíveis

Descrição do Problema

Elementos Transponíveis (TEs) são sequências de DNA repetitivas dispersas nos genomas, mais prevalentes em organismos eucarióticos e frequentemente impactando a evolução e arquitetura do genoma, principalmente devido à sua redundância e rearranjos que promovem. Por exemplo, TEs podem compreender até 90% em muitos genomas de plantas, por exemplo, milho e trigo. Por outro lado, independentemente do número e quantidade presentes em um determinado genoma, TEs também podem desempenhar papéis importantes moldando a expressão gênica e a estrutura da cromatina [1].

Neste trabalho vocês utilizaram um conjunto de dados de TEs da planta Z. mays (Zea mays) disponível no site do Atlas dos Elementos Transponíveis em plantas (http://apte.cp.utfpr.edu.br/download). Vocês devem fazer o download dos seis arquivos *.gff3* correspondentes às seis classes apresentadas para download: *LTRs, LINEs, SINEs, TIRs, MITEs e Helitrons*. Não façam o download da coluna *All*, pois essa pode conter sequências não classificadas. Vocês devem baixar seis arquivos por meio dos links de *Download*, como destacado na figura abaixo. Todos os exemplos nos arquivos baixados pertencem às suas classes correspondentes de acordo com os nomes das colunas.

Transposable Elements Records

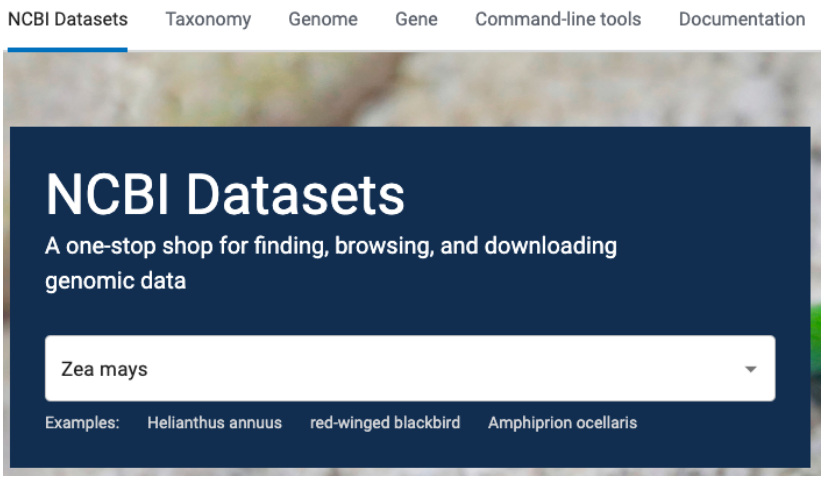
Results of the annotation of Transposable Elements are avaiable for *download* right in the table below.
** All records are formatted (GFF3) output like:*
Chr | Source Annotation | Class/Order/Superfamily | Start | End | Score | Strand | Phase | Attributes

Species	LTRs	LINEs	SINEs	TIRs	MITEs	Helitrons	All
A. chinensis - (974,621 TEs)	Download	Download	Download	Download	Download	Download	Download
A. tauschii - (3,613,042 TEs)	Download	Download	Download	Download	Download	Download	Download
V. vinifera - (522,720 TEs)	Download	Download	Download	Download	Download	Download	Download
Z. mays - (1,105,158 TEs)	Download	Download	Download	Download	Download	Download	Download

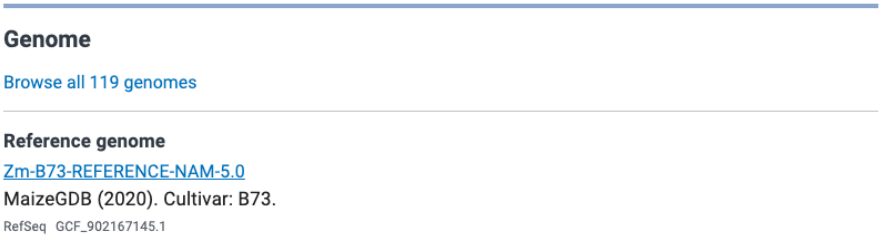
Feito o download dos arquivos *.gff3*, vocês verão que suas colunas fornecem as seguintes informações, nessa ordem, sobre os TEs representados nas linhas: *Chr | Source Annotation | Class/Order/Superfamily | Start | End | Score | Strand | Phase | Attributes*. De cada arquivo, vocês devem recuperar apenas as linhas em que o valor da coluna *Strand* seja positivo (+). Dessas linhas vocês devem recuperar o cromossomo (*Chr*) e os valores das colunas *Start* e *End*. A figura abaixo apresenta um trecho do arquivo de *LINEs* da planta *Z. mays*, destacando na sexta linha os valores das colunas *Chr, Start, End* e *Strand*.

```
4  APTEdb  Class I/LTR/Gypsy 122450121   122450632   .  +  .  TE-Score=0.285;Software=RepeatModeler;Length=511bps
5  APTEdb  Class I/LTR/Gypsy 47788694779015 .  +  .  TE-Score=0.428;Software=RepeatMasker;Length=146bps
4  APTEdb  Class I/LTR/Gypsy 200832603  200842716  .  -  .  TE-Score=0.428;Software=RepeatModeler;Length=10113bps
1  APTEdb  Class I/LTR/Gypsy 295530124  295548111  .  -  .  TE-Score=0.571;Software=RepeatModeler;Length=17987bps
7  APTEdb  Class I/LTR/Gypsy 157114644  157142357  .  -  .  TE-Score=0.571;Software=RepeatModeler;Length=27713bps
2  APTEdb  Class I/LTR/Gypsy 197296548  197297070  .  +  .  TE-Score=0.142;Software=RepeatMasker;Length=522bps
9  APTEdb  Class I/LTR/Gypsy 89034632   89035827   .  +  .  TE-Score=0.428;Software=RepeatModeler;Length=1195bps
9  APTEdb  Class I/LTR/Gypsy 39829863   39830128   .  +  532 TE-Score=0.285;Software=LTRRetriever;Length=265bps
4  APTEdb  Class I/LTR/Gypsy 142884529  142885177  .  -  .  TE-Score=0.285;Software=RepeatModeler;Length=648bps
4  APTEdb  Class I/LTR/Gypsy 213876111  213885917  .  -  .  TE-Score=0.428;Software=RepeatModeler;Length=9806bps
3  APTEdb  Class I/LTR/Gypsy 209986600  209986834  .  -  .  TE-Score=0.142;Software=RepeatModeler;Length=234bps
```

Depois de separar essas informações (linhas e colunas dos arquivos *.gff3*), vocês devem acessar o site do NCBI (https://www.ncbi.nlm.nih.gov/datasets/) e buscar pelo organismo *Zea mays* conforme figura abaixo.



Na página resultante, procurem pelo genoma de referência do *Z. mays*, conforme figura abaixo, e cliquem no link que leva ao genoma de referência.

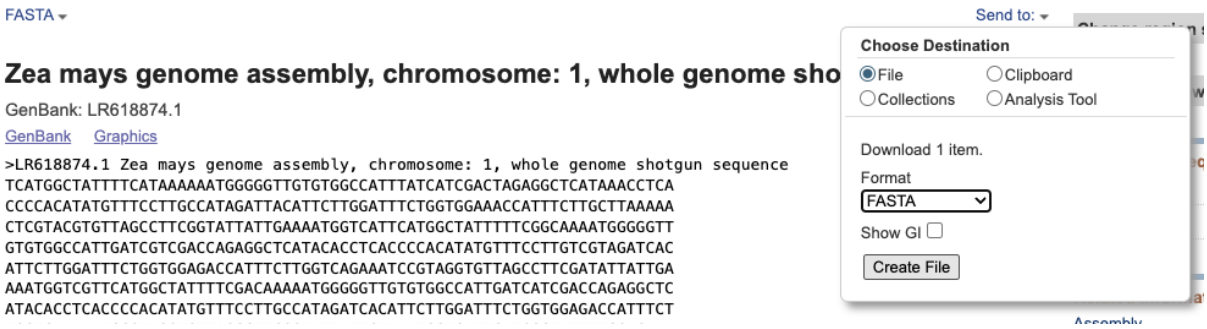


Nessa página vocês podem fazer o download de todos os cromossomos do *Z. mays* clicando no link correspondente do *GenBank*:

Chromosome	GenBank	RefSeq	Size (bp)	GC content (%)	Unlocalized count	Action
1	LR618874.1	NC_050096.1	308.452.471	47	0	⋮
2	LR618875.1	NC_050097.1	243.675.191	47	0	⋮
3	LR618876.1	NC_050098.1	238.017.767	47	0	⋮
4	LR618877.1	NC_050099.1	250.330.460	46,5	0	⋮
5	LR618878.1	NC_050100.1	226.353.449	47	0	⋮
6	LR618879.1	NC_050101.1	181.357.234	47	0	⋮
7	LR618880.1	NC_050102.1	185.808.916	46,5	0	⋮
8	LR618881.1	NC_050103.1	182.411.202	47	0	⋮
9	LR618882.1	NC_050104.1	163.004.744	47	0	⋮
10	LR618883.1	NC_050105.1	152.435.371	47	0	⋮
MT	AY506529.1	NC_007982.1	569.630	44	0	⋮
Pltd	X86563.2	NC_001666.2	140.384	38,5	0	⋮

Na página de resultados, vocês podem acessar a sequência do cromossomo em formato fasta:

A sequência fasta pode então ser salva em um arquivo:



O próximo passo agora é recuperar as sequências de TEs dos cromossomos correspondentes. Em posse dos arquivos de todos os cromossomos e das posições de início (Start) e fim (End) de cada TE, vocês devem fazer um script para recuperar todas as sequências de TEs. Esse script deve criar um único arquivo contendo três colunas: Cromossomo, Sequência de TE, Classe. Cada linha é uma sequência de TE. Esse será o conjunto de dados que vocês utilizarão em uma tarefa de classificação de TEs.

DICA: Tente automatizar todo o processo acima usando BioPython. Pesquisem, pois o BioPython permite o acesso direto ao NCBI e recuperação de várias informações.

Para a tarefa de classificação vocês devem extrair atributos de cada sequência de TE. Para isso, utilizem a ferramenta MathFeature (https://github.com/Bonidia/MathFeature) [2]. Essa ferramenta permite a extração de diversos tipos de características. Vocês devem ler sobre as características e escolher algumas para serem extraídas das sequências de TEs. Vocês podem escolher um único tipo de característica ou um conjunto de tipos. Isso fica a critério de vocês e faz parte do trabalho.

Após a extração das características, cada sequência será então representada por um vetor, em que cada posição contém um valor de característica. Vocês devem então escolher um classificador e fazer um experimento de classificação de TEs. Para isso usem o scikit-learn (https://scikit-learn.org/stable/). Qualquer classificador pode ser utilizado e sua escolha justificada. Isso faz parte do trabalho.

Tarefas

1. **Montagem dos Dados**: Construção do conjunto de dados como descrito acima.
2. **Separação do Conjunto de Dados**: Separar o conjunto de dados em partições de treino (70%) e teste (30%) de forma aleatória.
3. **Pré-processamento**: Aplicar qualquer estratégia de pré-processamento que considerarem necessária, como normalização, padronização ou seleção de atributos.
4. **Construção do Modelo**: Construir um modelo de classificação (classificador) utilizando o scikit-learn. A escolha do classificador fica a critério de vocês, mas deve ser justificada.
5. **Avaliação de Desempenho**: Avaliar o desempenho do modelo nas partições de treino e teste utilizando a área abaixo da curva Precisão-Revocação [3] (https://dl.acm.org/doi/pdf/10.1145/1143844.1143874).

Entregáveis

1. **Notebook Python no Google Colab**: Enviar um notebook (.ipynb) com o código bem documentado e instruções para reprodução;
2. **Relatório de Pesquisa**: Elaborar um relatório de 8 a 10 páginas no formato de artigo científico contendo as seções: Resumo, Introdução, Metodologia, Experimentos, Discussões, Conclusão.

Templates de artigo para Overleaf e .docx. Vocês devem seguir esses templates.

- Overleaf: https://www.overleaf.com/latex/templates/sbc-conferences-template/blbxwjwzdngr
- Docx: https://www.sbc.org.br/wp-content/uploads/2024/07/modelosparapublicaodeartigos.zip

Data de Entrega

Data limite para submissão: **30/11/2024**.

Referências

[1] Pedro DLF, Amorim TS, Varani A, Guyot R, Domingues DS, Paschoal AR. An Atlas of Plant Transposable Elements. F1000Res. 2021 Nov 24;10:1194. PMID: 35035898; PMCID: PMC8729191, https://doi.org/10.12688/f1000research.74524.1.

[2] Robson P Bonidia, Douglas S Domingues, Danilo S Sanches, André C P L F de Carvalho, MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors, Briefings in Bioinformatics, 2021; bbab434, https://doi.org/10.1093/bib/bbab434.

[3] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/1143844.1143874.