UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO INTERUNIDADES EM BIOINFORMÁTICA

GUILHERME MIURA LAVEZZO

Análise de dependência entre posições de bases de sequências motivos de fatores de transcrição aplicada à comparação de modelos baseados em *Position Weight Matrix* e Gramática Regular Estocástica

São Paulo
2021

GUILHERME MIURA LAVEZZO

Análise de dependência entre posições de bases de sequências motivos de fatores de transcrição aplicada à comparação de modelos baseados em *Position Weight Matrix* e Gramática Regular Estocástica

Orientador: Profa. Dra. Ariane Machado Lima

Coorientador: Prof. Dr. Luiz Paulo Moura Andrioli

São Paulo

2021

Dedico este trabalho a todas as pessoas queridas que perdi durante a pandemia, Dulce, Moacyr e Ricardo. Sinto a falta de todos.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Agradeço ao programa CAPES pelo financiamento e incentivo, principalmente em tempos difíceis.

Agradeço meus orientadores Ariane Machado Lima e Luiz Paulo Moura Andrioli pela oportunidade concedida e pela amizade adquirida durante o processo.

Agradeço aos meus pais, que me apoiaram integralmente na minha decisão de realizar minha carreira acadêmica e que me deram todo o suporte possível.

Agradeço aos meus familiares que durante toda minha vida me deram suporte para eu chegar onde estou.

Por fim, agradeço também aos meus amigos, e em particular ao meu amigo e primo Gustavo, pelo carinho e companheirismo em tempos difíceis.

Resumo

LAVEZZO, Guilherme Miura.**Análise de dependência entre posições de bases de sequências motivos de fatores de transcrição aplicada à comparação de modelos baseados em *Position Weight Matrix* e Gramática Regular Estocástica** 2021. 125 f. Dissertação (Mestrado em Bioinformática) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Para elucidar os mecanismos de regulação transcricional, é essencial determinar acuradamente a interação entre DNA e fatores de transcrição (FT). Embora FTs mantenham certa especificidade em reconhecer sequências curtas de DNA, os sítios de ligação de fatores de transcrição (SLFT) são sequências degeneradas. Técnicas experimentais *high throughput in vivo*, como ChIP-seq, ainda que muito utilizadas, identificam regiões de 100-600 pares de bases (pb), enquanto FTs geralmente se ligam a sequências de 6-15 pb. Por outro lado, técnicas experimentais *in vitro*, apesar de avaliarem a interação FT-DNA com maior resolução, não correspondem às condições fisiológicas em que ocorrem a regulação transcricional. O padrão de reconhecimento de DNA mais provável que interage com um FT, ou seja o motivo do FT, precisa ser descoberto a partir de sequências maiores de DNA, obtidas experimentalmente e em grande volume. Existem diversos algoritmos que se encarregam de descobrir motivos, porém esses algoritmos divergem em considerar ou não dependência entre bases, questão essa ainda em aberto na comunidade científica. Com o motivo descoberto, geralmente deseja-se obter representações do mesmo ao longo de genoma ou região genômica de interesse e, para isso, é necessário um modelo preditor de SLFTs. Existem também diversos modelos computacionais que procuram predizer SLFTs de tamanhos exatos. No entanto, devido ao curto tamanho dos sítios, tais modelos tendem a produzir muitos falsos positivos, dificultando uma interpretação biológica acurada do contexto biológico. Além disso, nenhum modelo preditor excede os demais em todos os casos, tornando a escolha de um melhor modelo caso-específica para cada FT de interesse. Considerando as diversas combinações entre o tipo de experimento e o algoritmo de descoberta de motivos, a tarefa de escolher o melhor modelo preditor de SLFTs não é trivial. O modelo mais utilizado para predição de SLFT são PWMs (*Position Weight Matrix*), que assumem independência entre as bases do sítio, o que pode não ser verdadeiro para determinados fatores de transcrição. Gramáticas regulares estocásticas (GRE) são uma alternativa às PWMs, pois são modelos que conseguem capturar uma relação de dependência entre posições de bases. Considerando esse problema, foi possível escolher pelo modelo PWM ou GRE baseando-se apenas no conjunto amostral de SLFTs obtidos e em novas medidas de dependências propostas. Com o cálculo dessas medidas, foi possível criar uma regra de decisão, via árvore de decisão, que opte pelo melhor modelo de maneira que garanta seu desempenho.

Palavras-chaves: Sítios de ligação de fatores de transcrição. *Position Weighted Matrix*. Gramáticas Regulares Estocásticas. Predição *in silico*. ChIP-seq.

Abstract

LAVEZZO, Guilherme Miura. **Inter-position dependency analysis of transcription factor motif sequences applied to compare PWM (Position Weight Matrix)and SRG (Stochastic Regular Grammar)-based models**. 2021. 125 p. Master's Dissertation in Bioinformatics – Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2020

In order to elucidate the mechanisms of transcriptional regulation, it is essential to accurately determine the interaction between DNA and transcription factors (TFs). Although TFs maintain a certain specificity in recognizing short DNA sequences, transcription factor binding sites (TFBS) are degenerate sequences. High throughput in vivo experimental techniques, such as ChIP-seq, although widely used, identify regions of 100-600 base pairs (bp), while TFs generally bind to sequences of 6-15 bp. On the other hand, in vitro experimental techniques , despite assessing FT-DNA interaction with higher resolution, do not correspond to the physiological conditions under which transcriptional regulation occurs. The most likely DNA recognition pattern that interacts with a TF, i.e. the TF motif, needs to be discovered from larger DNA sequences obtained experimentally and in large volume. There are several algorithms that take charge of discovering motifs, but these algorithms differ in whether or not to consider position dependency, which is still an open question in the scientific community. With the motif discovered, it is usually desired to obtain representations of it along the genome or genomic region of interest and, for this, a predictive model of TFBS is required. There are also several computational models that seek to predict TFBSs of exact sizes. However, due to the short site sizes, such models tend to produce many false positives, making an accurate biological interpretation of the biological context difficult. Furthermore, no single predictor model outperforms the others in all cases, making the choice of a best case-specific model for each TF of interest. Considering the various combinations between experiment type and motif discovery algorithm, the task of choosing the best predictive model for TFBSs is not trivial. The most widely used model for TFBS prediction are PWMs (Position Weight Matrix), which assume independence between site bases, which may not be true for certain transcription factors. Stochastic regular grammars (SRGs) are an alternative to PWMs, as they are models that can capture a dependency relationship between base positions. Considering this problem, it was possible to choose between the PWM or SRG model based only on the sample set of TFBSs obtained and novel proposed dependency measures. By calculating these measures, it was possible to create a decision rule, via a decision tree, that opts for the best model in a way that guarantees its performance.

Keywords: Transcription Factor Binding Sites. Position Weight Matrix. Stochastic regular grammars (SRG). Prediction in silico. ChIP-seq.

Lista de figuras

Lista de algoritmos

Lista de tabelas

Lista de abreviaturas e siglas

pb	pares de base
MCR	Módulo *cis*-regulador
FT	fator de transcrição
SLFT	sítio de ligação do fator de transcrição
GRE	Gramática Regular Estocástica
PWM	*Position Weight Matrix*
PBM	*Protein Binding Microarray*
ROC	*receiver operating characteristic*
AUC	*Area Under Curve*
NGS	*next generation sequencing*
AFDEA	Autômato Finito Determinístico Estocástico Acíclico
ChIP-seq	*Chromatin ImmunoPrecipitation sequencing*
AP	*Average Precision*
AUPRC	*Area Under Precision-Recall Curve*

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Uma área de intensa pesquisa atualmente nas ciências biológicas é dedicada ao estudo da formação corporal dos organismos. Como um conjunto de instruções genéticas partindo de uma única célula é capaz de gerar um organismo multicelular composto por milhares de milhões de células integradas realizando funções diferentes no corpo é uma das questões centrais dentro dessa área de estudos.

Um processo fundamental do desenvolvimento é a regulação da transcrição diferencial dos genes. Os estudos pioneiros de Edward B. Lewis, Christiane Nüsslein-Volhard e Eric F. Wieschaus, ganhadores do Prêmio Nobel em 1995, foram os responsáveis por elucidar os mecanismos moleculares por trás do desenvolvimento do embrião de *Drosophila melanogaster*, popularmente conhecida como mosca das frutas (NüSSLEIN-VOLHARD; WIESCHAUS, 1980). A *D. melanogaster* é um organismo modelo para os estudos de regulação a nível transcricional, que é o principal mecanismo de regulação da expressão gênica envolvido na especificação do eixo ântero-posterior do seu corpo. A especificação do eixo ântero-posterior é responsável por orientar e definir onde será a região anterior e posterior do embrião e delimitar os segmentos, que serão as unidades anatômicas do corpo larval e do adulto (GILBERT, 2018).

A partir desses estudos pioneiros verificou-se a existência de genes codificadores para proteínas que atuam como reguladoras da transcrição de outros genes, que por sua vez também podem ser reguladores transcricionais de ainda outros genes, formando uma cascata de regulação. Essas proteínas são fatores de transcrição (FTs), que se ligam em sequências de DNA, chamadas sítios de ligação de fatores de transcrição (SLFTs), localizadas nas regiões reguladoras dos genes. Os SLFTs constituem o alvo principal do estudo dessa dissertação.

Apesar da importância da atuação dos mecanismos pelos quais os FTs regulam a expressão gênica, ainda continua sendo um desafio entender como SLFTs, quando reconhecidas pelos respectivos FTs culminam na regulação da atividade transcricional (LAMBERT *et al.*, 2018). Proteínas FTs podem atuar por diversos mecanismos, seja individualmente, seja cooperativamente, seja pela ligação de um cofator associado ou por obstruir a interação de outro FT no DNA (LAMBERT *et al.*, 2018). A atuação de um FT é difusa: um mesmo FT pode ser responsável pela expressão de genes distintos em

diferentes tipos celulares (GERTZ *et al.*, 2012). Além disso, FTs atuam dinamicamente, mudando quais regiões genômicas são reconhecidas e reguladas, por exemplo, conforme ocorre uma mudança de estados da cromatina (VOSS; HAGER, 2013). Apesar de existir uma distinção entre FTs funcionalmente ativadores e repressores transcricionais, essa noção dicotômica funcional de um FT está atualmente em conflito (LAMBERT *et al.*, 2018). Diversos FTs podem recrutar cofatores com efeitos resultantes opostos como caso do FT Max em *D. melanogaster* (FRIETZE; FARNHAM, 2011; ROSENFELD, 2006; SCHMITGES *et al.*, 2016). Determinar como FTs podem se associar de diferentes formas e como as específicas combinações de SLFTs envolvidas podem estar relacionadas com a regulação transcricional é, portanto, essencial para decodificar propriedades específicas e funcionais no genoma (LAMBERT *et al.*, 2018). Para isso, uma etapa fundamental anterior é necessária, correspondente à determinação da localização o mais precisa possível de SLFTs.

Não obstante aos grandes avanços tecnológicos, métodos experimentais que estudam a interação FT-DNA, como ChIP-seq, ainda possuem limitações. Por exemplo, a precisão da localização de SLFTs ainda é comprometida pelas técnicas de sequenciamento *high throughput*. Outras técnicas experimentais, sejam *in vivo* ou *in vitro*, podem destoar na caracterização do conjunto de SLFTs reconhecidos por um mesmo fator de transcrição, isto é, na sequência motivo. Além disso, os métodos experimentais demandam tempo e recursos financeiros dificultando, por exemplo, a realização de diferentes experimentos em menores intervalos de tempo. Todos esses aspectos justificam, portanto, o uso de modelos computacionais para a obtenção da localização mais precisa possível das regiões genômicas de interesse.

O modelo computacional para identificação de SLFTs mais utilizado é a *Position Weight Matrix* (PWM) (INUKAI; KOCK; BULYK, 2017), sendo o mais difundido em ferramentas de bioinformática e mais fácil de ser implementado. Apesar disso, PWMs desconsideram que haja uma relação de dependência entre bases de posições distintas dentro de um SLFT, questão essa já levantada diversas vezes na literatura (EGGELING, 2018; TOMOVIC; OAKELEY, 2007; WEIRAUCH *et al.*, 2013). Consequentemente, é questionado se a acurácia na predição de SLFTs, usando PWMs, poderia ser afetada.

Por outro lado, ainda é um duro desafio caracterizar as sequências funcionais reconhecidas por um FT (FORNES *et al.*, 2019; WASSERMAN; SANDELIN, 2004; WEIRAUCH *et al.*, 2013), dificultando mensurar a relevância da dependência entre

posições de bases para um modelo classificador de SLFTs. Um modelo alternativo às PWMs capaz de caracterizar tais dependências são as Gramáticas Regulares Estocásticas (GRE) mais conhecidas por seus dispositivos equivalentes chamados autômatos probabilísticos. Porém, por ser um modelo mais complexo e com mais parâmetros a serem estimados, o desempenho torna-se dependente da qualidade e quantidade de SLFTs que sirvam como amostra. Além disso, a etapa de treinamento de GREs (o algoritmo LAPFA) são mais custosas computacionalmente e requerem mais tempo de processamento. Apesar disso, uma vez o modelo treinado, a utilização do modelo preditor é equivalente, em termos de custos computacionais, à PWM.

Considerando esses fatos, há a necessidade de analisar a relevância de dependência entre posições de bases de sequências motivos e o impacto no desempenho de modelos classificadores de SLFTs: PWMs e GREs.

1.1 Hipóteses

Este projeto baseia-se nas seguintes hipóteses.

A primeira hipótese é a de que o modelo preditor (PWM ou GRE) mais adequado para a predição de SLFTs seja caso-específico, de acordo com as características da amostra de treinamento como tamanho (número de sequências conhecidas de sítios daquele FT específico), Conteúdo de Informação e nível de dependência entre bases. Um trabalho anterior já sugeriu que essa hipótese parece ser verdadeira (NETO, 2018), mas faltavam dados suficientes para aplicar estatísticas robustas de comparação entre desempenho médio dos modelos e, além disso, a forma de estimação do nível de dependência das amostras utilizada precisava ser aprimorada.

A segunda hipótese assume que cada motivo descoberto por um algoritmo de descoberta de motivo consiga se aproximar do verdadeiro motivo hipotético de um FT. Cada algoritmo obtém um motivo com algum grau de distinção, para um dado FT. Portanto, para que seja possível comparar modelos entre diversos algoritmos e experimentos, essa hipótese precisa ser assumida.

1.2 Objetivos

O objetivo geral deste trabalho é comparar o desempenho entre PWMs e GREs como preditores de SLFTs e analisar a relevância em se considerar a dependência entre posições de bases para um conjunto de SLFTs obtidos de diversos métodos.

Os objetivos específicos são:

1. propor uma forma de obtenção de amostras positivas para treinar os modelos aproveitando o grande volume de dados em larga escala, como por exemplo, de *Chromatin ImmunoPrecipitation sequencing* (ChIP-seq) e *Protein Binding Microarray* (PBM);
2. propor uma ou mais estratégias para medir o nível de dependência entre posições de bases em uma dada amostra;
3. comparar os desempenhos de PWMs e GREs como modelos de predição de SLFTs e, com base nos resultados, criar uma regra de decisão que, dado o critério que verifica dependência e tamanho da amostra, escolhe entre os modelos preditores PWM ou GRE para cada FT específico;

1.3 Organização deste documento

Além dessa introdução, este documento conta com os seguintes capítulos: o capítulo 2 introduz o leitor aos principais conceitos biológicos e computacionais necessários para o entendimento desta dissertação; o capítulo 3 detalha como o projeto foi desenvolvido, o capítulo 4 mostra os resultados e discussão do trabalho, o capítulo 5 apresenta as contribuições e considerações finais, e, por fim, o capítulo 5.2 comenta sobre trabalhos futuros e perspectivas para novos projetos.

2 Conceitos Fundamentais

Para o melhor entendimento desta dissertação, este capítulo apresenta nas seções 2.1 e 2.2 os principais conceitos biológicos e computacionais, respectivamente, envolvidos neste trabalho.

2.1 Conceitos Biológicos

2.1.1 Regulação Transcricional

A regulação transcricional é essencial para a diferenciação celular e para a adaptação a eventos intra e extracelulares. Existem diversos mecanismos celulares atuando para a transcrição diferencial de genes, tais como mudanças de estado da cromatina (favorecendo ou impedindo a transcrição) e por mecanismos que atuam sobre o complexo de pré-iniciação e alongamento da transcrição (ALBERTS; JOHNSON; LEWIS, 2015). Além disso, RNAs não codificadores (ncRNAs, do inglês *non-coding* RNAs) compreendem uma porção substancial dos genes transcritos do genoma de eucariotos, cujo mecanismo de ação tem sido mostrado como regulador da atividade transcricional (CASAMASSIMI; CICCODICOLA, 2019). FTs, por sua vez permeiam entre os diversos mecanismos de controle transcricionais. A característica fundamental de um FT é a sua capacidade de reconhecer e se ligar diretamente ao DNA e nuclear eventos subsequentes. Por possuírem múltiplos domínios[1], os FTs podem atuar tanto diretamente quanto indiretamente, por exemplo, recrutando proteínas que modificam ou remodelam a cromatina e efetivamente contribuem para alteração do status transcricional (FRIETZE; FARNHAM, 2011; LAMBERT *et al.*, 2018). O foco do presente estudo reside no controle transcricional por meio dos fatores de transcrição.

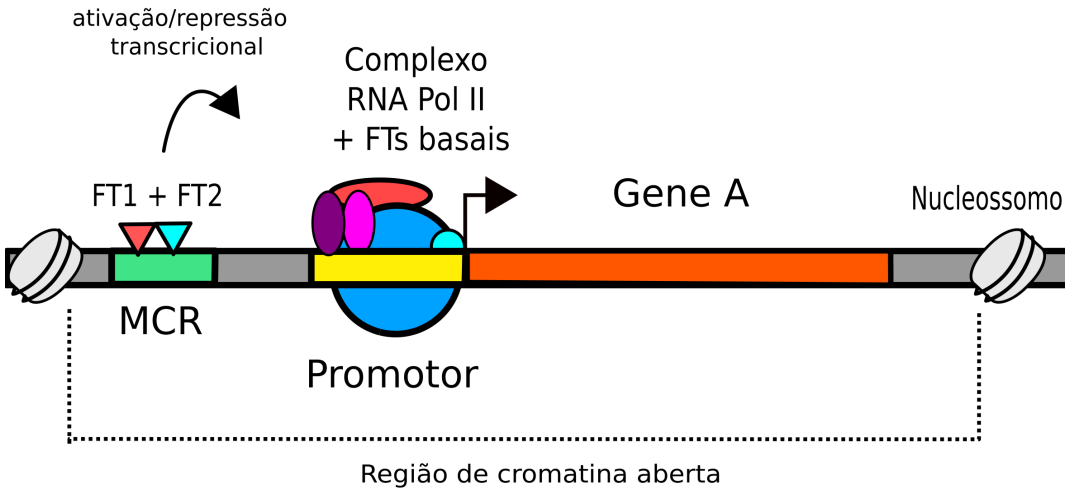
A expressão de qualquer gene depende da ação de síntese de RNA efetuada por uma enzima RNA polimerase. Nos eucariotos existem 3 RNA polimerases, sendo que a RNA polimerase II (RNA Pol II), é responsável pela síntese dos RNA mensageiros, intermediários da síntese de proteínas. A RNA Pol II é recrutada para regiões genômicas, tipicamente ao redor do sítio de início da transcrição[2], denominadas promotores que, por

[1] O domínio de uma proteína é uma região (ou subunidade) tridimensional conservada e independente, podendo estar atrelada a alguma especialização funcional, como uma atividade enzimática ou o reconhecimento de uma sequência de DNA.

[2] Sítio de início da transcrição, do inglês, *Transcription Start Site* (TSS) é o primeiro nucleotídeo transcrito de um gene alvo.

meio dos fatores de transcrição basais, que também estabilizam a interação DNA-RNA Pol II (figura 1). O papel dessa ação conjunta direciona o complexo de RNA Pol II formado para a exata localização do sítio inicial e a direção da transcrição. Apesar do promotor e do complexo RNA Pol II ser suficiente para o início da atividade transcricional, a taxa de transcrição é modulada por regiões genômicas adicionais, sejam estas distais ou proximais ao promotor (ANDERSSON; SANDELIN, 2019; SPITZ; FURLONG, 2012). Essas regiões genômicas, denominadas *enhancers*, são regiões tipicamente de algumas centenas de pares de base (pb) que contém regiões específicas de reconhecimento para determinados FTs. *Enhancers* são exemplos dos chamados módulos *cis*-reguladores (MCR), que atuam na modulação dos genes (figura 1) (SPITZ; FURLONG, 2012; ANDERSSON; SANDELIN, 2019; SCHOENFELDER; FRASER, 2019). MCRs funcionam como um módulo operacional formado por uma sequência de DNA onde estão agrupados sítios de ligação para uma variedade de FTs (ANDERSSON; SANDELIN, 2019). Os fatores de transcrição se ligam aos seus respectivos SLFTs atuando como ativadores ou repressores da transcrição, e o balanço resultante da interação desses FTs determina a transcrição ou não do gene. (ALBERTS; JOHNSON; LEWIS, 2015).

Figura 1 – Esquema simplificado da regulação transcricional. Está representado o complexo formado da RNA Pol II (em azul) com fatores de transcrição (FT) basais ligados à região promotora (em amarelo) para iniciar a transcrição do gene A (em laranja). A flecha sobre o gene A aponta para o sentido ($5' \rightarrow 3'$) de início da transcrição. O módulo *cis* regulatório (MCR) está representado à esquerda do promotor. No MCR, alguns FTs, representados como FT1 (triângulo vermelho) e FT2 (triângulo ciano), se ligam em regiões específicas. As combinações específicas de FTs (representadas genericamente pelo sinal de +) no MCR podem resultar em diferentes efeitos, seja ativação ou repressão gênica. Além disso, a expressão e regulação de um gene também depende da abertura da cromatina (livre de nucleossomos), como mostrado na figura.



Fonte: Guilherme Miura Lavezzo (2020).

Embora FTs sejam específicos, na prática é observado que os SLFTs podem variar na composição de bases em algumas posições. Esse padrão de sequências de DNA reconhecidas por um mesmo FT é denominado motivo, cujo tamanho varia geralmente de 6 a 15 nucleotídeos, mas podendo ser maior(LAMBERT *et al.*, 2018; DABROWSKI *et al.*, 2015). Por motivos serem curtas sequências degeneradas, um típico gene alvo pode conter múltiplos sítios de ligação em potencial para um mesmo FT e também para distintos FTs (LAMBERT *et al.*, 2018). Apesar disso, poucos são os FTs que se ligam em quase todos os potenciais SLFTs, como é o caso do FT CTCF (LAMBERT *et al.*, 2018). Por outro lado, estudos massivos com o genoma mostraram ocupação de FTs em

SLFTs, provavelmente sem significado funcional (DONIGER, 2005; FISHER *et al.*, 2012; SURYAMOHAN; HALFON, 2014).

Enhancers controlam a expressão de genes distais ou proximais podendo alcançar distâncias típicas de 4 a 10 mil bases, e para *D. melanogaster* até casos em que a informação regulatória dista de 70 a 100 mil bases, como por exemplo no caso do gene *cut* (FURLONG; LEVINE, 2018). Além disso, alguns *enhancers* foram mostrados como possuindo atividade promotora (ANDERSSON; SANDELIN, 2019) e silenciadora (inibindo a transcrição) (HALFON, 2020). Portanto *enhancers* não somente aumentam como também diminuem a transcrição, e por isso, o termo *enhancer* (realçador) tem sido substituído pelo termo módulo *cis*-regulador.

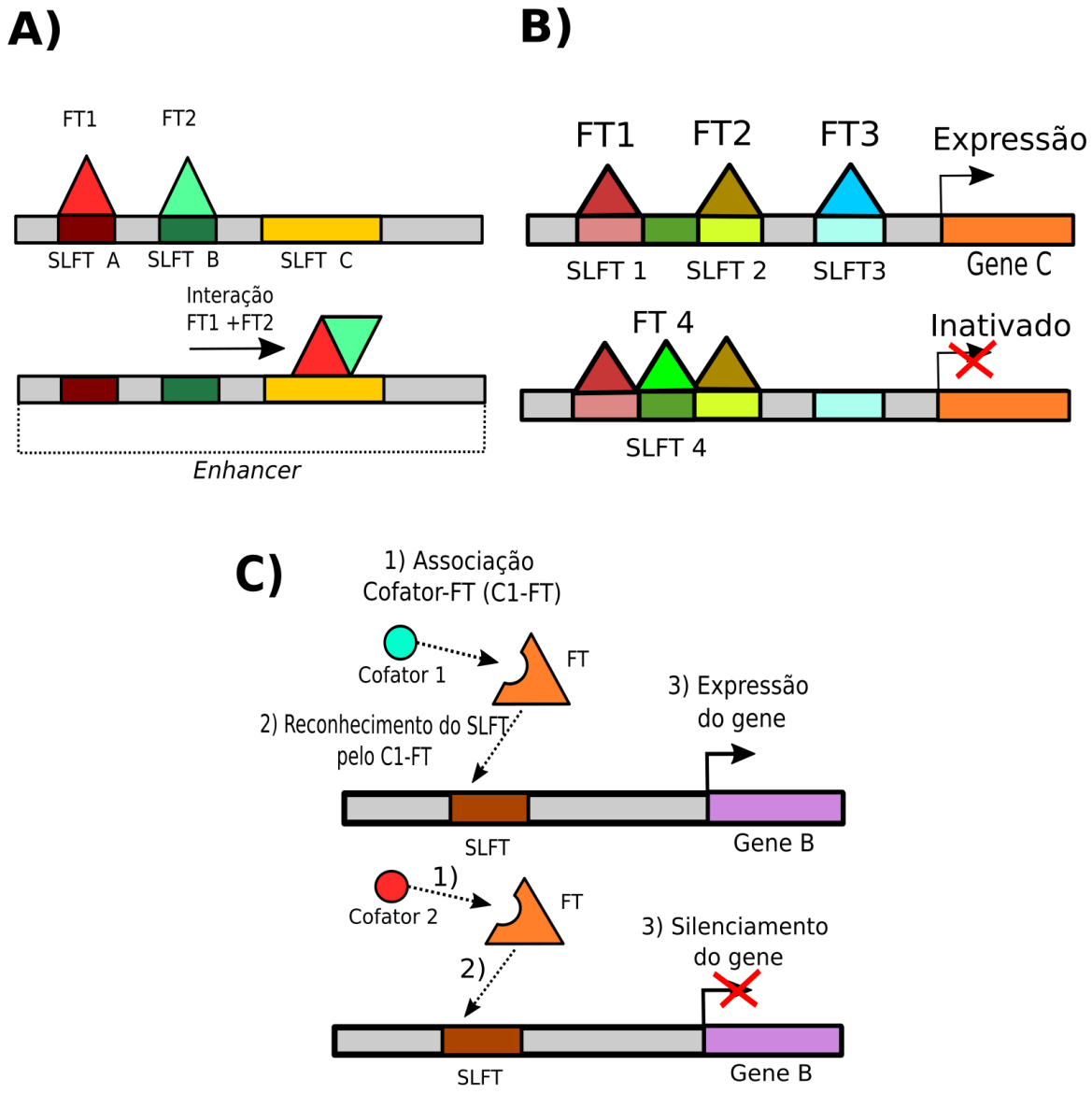
A forma como FTs ligados ao DNA dentro de um mesmo módulo interagem entre si e atuam para a transcrição ainda é uma área de pesquisa com perguntas básicas a serem respondidas. Existem evidências de SLFTs cuja localização se sobrepõem no DNA (EGGELING, 2018; MACARTHUR *et al.*, 2009). Isto sugere que possa haver algum tipo de interação entre proteínas, tanto negativamente (impedimento estérico) como positivamente (formação de um complexo proteico cooperativo) (MACARTHUR *et al.*, 2009). Adicionalmente, existem proteínas chamadas cofatores que regulam a transcrição, se ligando ao fator de transcrição, que por sua vez pode, por exemplo, sofrer uma mudança conformacional e reconhecer motivos de DNA diferentes (ROGERS; BULYK, 2018). Dessa forma, o conhecimento das localizações exatas dos SLFTs podem auxiliar na elucidação sobre esses mecanismos de ação conjunta dos FTs.

Para o caso do embrião de *D. melanogaster*, sabe-se que a repressão transcricional é um dos mecanismos principais responsáveis por regular a expressão gênica na cascata de segmentação(GRAY; SZYMANSKI; LEVINE, 1994; SMALL *et al.*, 1991; ANDRIOLI *et al.*, 2002). A repressão transcricional pode se dar de diferentes formas, mas basicamente o resultado final é a inativação de um gene alvo. Por exemplo, pelo mecanismo de *quenching*, um FT com papel repressor pode se ligar em uma região adjacente a uma região reconhecida por uma proteína ativadora de transcrição, ocorrendo uma interação direta entre as proteínas ou por meio de cofatores. A interação direta impede que a proteína ativadora interaja com o complexo transcricional, e que portanto, desempenhe sua função (GRAY; SZYMANSKI; LEVINE, 1994). Em outro exemplo, o mecanismo de competição por SLFTs parcialmente coincidentes consiste na ligação de um FT sobre uma região promotora ou ativadora da transcrição, obstruindo portanto a ligação do

complexo de início de transcrição (LEVINE; MANLEY, 1989). Isso sugere que existam regiões reguladoras de um gene alvo que contenham algum espaçamento recorrente entre si, sendo um importante mecanismo para ser investigado com ferramentas de predição de SLFT *in silico* (PAPATSENKO; GOLTSEV; LEVINE, 2009).

A interação entre FTs tem sido alvo de estudos para entender o mecanismo de regulação transcricional (MOORMAN *et al.*, 2006). Apesar de alguns FTs possivelmente se ligarem aos sítios de forma individual, uma parte dos FTs atuam cooperativamente, formando dímeros, trímeros ou tetrâmeros (figura 2 A). Uma vez que FTs podem formar um complexo proteico, a conformação estrutural resultante pode reconhecer um motivo diferente dos motivos individualmente reconhecidos pelos FTs (figura 2 A) (MORGUNOVA; TAIPALE, 2017). Diversos mecanismos são sugeridos para a interação cooperativa entre FTs. Dentre esses mecanismos, a cooperação de FTs pode favorecer o reconhecimento de estruturas tridimensionais distintas de DNA, ou o complexo de FTs pode mudar a conformação do DNA para favorecer particulares domínios proteicos de ligação ao DNA (MORGUNOVA; TAIPALE, 2017; KIM *et al.*, 2013). Consequentemente à mudança conformacional do DNA, a ligação de um FT pode, a longa distância, propiciar específicos FTs de se ligarem em regiões mais estabilizadas do DNA e, portanto, podendo formar um padrão de espaçamento entre FTs (KIM *et al.*, 2013). Alguns FTs, poderiam atuar somente em conjunto com uma ou mais proteínas e, cada complexo proteico distinto formado com um FT alvo pode, além de mudar o motivo reconhecido, mudar o mecanismo de regulação de transcrição, funcionando ora como ativador ora como repressor da transcrição (figura 2 B)(MORGUNOVA; TAIPALE, 2017). Nucleossomos podem causar impedimento estérico da ligação de FTs ao DNA. Portanto, é sugerido que um FT pioneiro que liga-se a uma região de DNA possa favorecer indiretamente a ligação de outros FTs (MIRNY, 2010; MORGUNOVA; TAIPALE, 2017).

Figura 2 – Mecanismos gerais da interação entre fator de transcrição (FT) e sítio de ligação do fator de transcrição (SLFT). No item A está mostrado dois FTs (FT1 e FT2) ligados respectivamente aos SLFT A e SLFT B em uma região de *enhancer*. Com a interação física entre os dois FTs, ocorrendo uma mudança conformacional, o SLFT reconhecido passa a ser o C. No item B, a combinação dos FTs 1, 2 e 3 em um *enhancer* resulta na expressão do gene C (em laranja). Por outro lado, a combinação do FT 1, 4 e 2 promove a inativação do gene C. No item C, a associação do cofator 1 (em azul ciano) com o FT (em laranja) promove a expressão do gene B. Porém, ligado ao cofator 2 (em vermelho), o FT promove o silenciamento do gene C.



Fonte: Guilherme Miura Lavezzo (2020)

2.1.2 Experimentos biológicos para estudo das regiões de ligação de fatores de transcrição

Existem diversas abordagens experimentais, tanto *in vitro* quanto *in vivo*, para identificar SLFTs. Dentre elas pode-se citar PBM, ChIP-chip e ChIP-seq.

Protein binding microarray (PBM) é uma técnica *in vitro* que possibilita medir a preferência de oligonucleotídeos por um FT de interesse, em alta resolução. Em um experimento de PBM, o FT de interesse é incubado junto com sequências de DNA de tamanho fixado e a afinidade por uma determinada sequência é aferida através de um sinal de fluorescência produzida por um anticorpo que reconhece o FT alvo. As sequências, *probes*, de DNA inseridas em um *array* ("placa") consistem de uma parte randômica de tamanho 10 e de uma composição fixa que contêm *primers* para formação de sequências dupla-fita e adaptadores que se fixam no *array*. Os oligonucleotídeos de tamanho 10 são produzidos sinteticamente e de maneira a cobrir todos os possíveis arranjos sequenciais das bases. Assim, esse arranjo randômico permite testar todas as possíveis afinidades do DNA com a proteína de interesse. O FT de interesse é modificado para conter um antígeno, *epitope flag*, que será reconhecido pelo anticorpo e emitirá um sinal de fluorescência. Esta técnica está limitado a produzir cerca de 40000 *probes* a serem testadas e um *array* único ou universal.

Apesar da alta resolução e da vantagem de produzirem grandes quantidades de dados, graças à tecnologia *high throughput*, todas as técnicas *in vitro* possuem a desvantagem de medir afinidade DNA-FT fora das condições celulares fisiológicas. É sabido (e como já mencionado anteriormente na seção 2.1), que apesar de haverem inúmeros potenciais SLFTs ao longo de um genoma para um FT, nem todas as sequências são funcionais. O fato de que proteínas formam complexos e possam estar impedidas de interagir com o DNA pode resultar em preferências distintas das medidas por técnicas *in vitro*. Além disso, é de interesse saber não somente a afinidade por sequências de DNA, mas também a coordenada genômica em que houve interação, para fins de estudos mais aprofundados em regulação da atividade transcricional de um gene de interesse. Para esse fim, existem técnicas *in vivo*, que também usufruem da produção em massa de dados e conseguem cobrir todo o genoma de interesse, mas não possuem resolução suficiente para localizar os SLFTs em que houveram interação direta com o FT alvo.

Uma das técnicas *in vivo* mais utilizadas atualmente é o *chromatin immunoprecipitation sequencing* (ChIP-seq). Nesta técnica, diferente de experimentos *in vitro*, sequências genômicas que interagem com um FT alvo são estudadas. No ChIP-seq, interações DNA-proteína das células são estabilizadas com um tratamento de formaldeído e então, o DNA é fragmentado por sonicação em pedaços de cerca de 150 a 600 pares de base. Anticorpos produzidos especificamente para reconhecer a proteína de interesse são utilizados para imunoprecipitar o complexo DNA-proteína. As interações DNA-proteína são desfeitas e os fragmentos de DNA são isolados e sequenciados, de onde vem o termo "seq" para se referir a sequenciadores de larga escala "*next-generation sequencing*" (NGS). Especificamente, apenas porções (*reads*) de aproximadamente 36 a 100 pb no sentido 5' para 3' do DNA são sequenciados. O tamanho da *read* é determinado pelo número de ciclos do sequenciamento, sendo um parâmetro ajustado no desenho experimental. As *reads* são então remapeadas a um genoma de referência do organismo alvo, usando um algoritmo computacional, como o Bowtie2 (LANGMEAD; SALZBERG, 2012; NAKATO; SHIRAHIGE, 2016; PARK, 2009).

O ChIP-seq é a forma mais direta de se estudar a interação do complexo DNA-proteína relativa à localização genômica deste complexo, que é um assunto ainda não completamente desvendado. Aliado ao NGS, o tempo e os custos de se produzir tais experimentos foram reduzidos drasticamente, permitindo a utilização disseminada e aprimorada da técnica na comunidade científica mundial. Além disso, comparado ao seu antecessor ChIP-chip[3], o ChIP-seq ainda possui cobertura total do genoma de interesse e melhor resolução e sensibilidade (FUREY, 2012; LIU; POTT; HUSS, 2010; JAYARAM; USVYAT; MARTIN, 2016)). Porém, ressalta-se que técnicas de imunoprecipitação, apesar de específicas aos FT de interesse, não são necessariamente seletivas ao FT individual. O anticorpo pode se complexar especificamente a um FT, mas este pode estar na forma individual ou complexada com demais proteínas e cofatores (EGGELING, 2018). Apesar de haver experimentos que tenham o objetivo de explorar as interações entre proteínas de um complexo, este trabalho foca apenas em FTs que se ligam diretamente ao DNA e assume que todo experimento coletado contém a informação necessária para inferir o contato direto entre FT-DNA.

[3] ChIP-chip é uma técnica antecessora ao ChIP-seq, que como o primeiro termo "ChIP"menciona, utiliza da técnica de imunoprecipitação para estudar uma proteína alvo. O segundo termo "chip"se refere a etapa de identificação das sequências ligantes (FT-DNA) via DNA *microarray* ("chip"). O ChIP-chip possui uma etapa de hibridização do DNA com a plataforma de *microarray*, tornando a etapa de identificação do DNA ligante ao FT menos precisa em relação ao sequenciamento direto desse DNA ligante (por ChIP-seq).

Existe uma etapa de processamento prévia desses dados: dados de ChIP-seq são as sequências *reads* obtidas diretamente do equipamento sequenciador. As *reads* precisam ser remapeadas ao genoma de referência. Posteriormente, são contadas e normalizadas por sequências obtidas pelo experimento controle, que geralmente se trata de fragmentos de DNA sequenciados sem a imunoprecipitação de um FT alvo.

Tais informações são utilizadas para fazer uma visualização das contagens de *reads*, através de uma densidade estimada projetada ao longo de um genoma de referência. Algoritmos computacionais como o MACS2 realizam uma verificação por picos de contagens que são estatisticamente significativos, o chamado *peak calling* (GASPAR, 2018). Os picos assim obtidos são evidências estatísticas para uma região onde verdadeiramente existe a ligação do FT alvo com o DNA. Ainda assim, as regiões de picos do ChIP-seq e outras abordagens variam em centenas de pb, enquanto estima-se que o SLFT seja, sequências curtas, tipicamente de 6 a 15 bases (JAYARAM; USVYAT; MARTIN, 2016; LAMBERT *et al.*, 2018). Geralmente, as regiões centrais dos picos possuem maior contagem ou maior densidade do sinal e, portanto, mostram com maior probabilidade onde houve ligação do FT. A partir disso, pode-se extrair as regiões centrais dos picos para a descoberta de motivos ou usar as sequências recortadas como amostras de treinamento para os modelos preditores de SLFT (FORNES *et al.*, 2019; LANDT *et al.*, 2012). Mas ainda assim é necessário um processamento posterior para tentar identificar os sítios exatos, como por exemplo usando uma ferramenta de descoberta de motivos (INUKAI; KOCK; BULYK, 2017). Será abordado mais sobre esse assunto no capítulo 3 e na seção seguinte.

Em resumo, visando a estudar os mecanismos de regulação transcricional via fatores de transcrição, existem duas grandes classes de experimentos complementares: *in vivo* e *in vitro*. Na primeira, é possível obter a coordenada genômica onde o FT de interesse se ligou, mas com menor resolução de bases. Por outro lado, as técnicas *in vitro* podem ajudar a aferir com melhor resolução quais sequências interagem preferencialmente com o FT. Independente da técnica experimental, são necessários algoritmos computacionais para: processar os dados em grande volume, extrair possíveis sequências motivos, conciliar informações.

2.1.3 Algoritmos computacionais de descoberta de motivos

Todas as técnicas experimentais, sejam *in vivo* ou *in vitro* passam por um processamento computacional para se obter o conjunto de SLFTs com afinidade ao FT, isto é, a sequência motivo. Mesmo em técnicas de alta resolução, como PBM e HT-SELEX [4], ainda existem pelo menos duas problemáticas: a de separar sequências de alta afinidade das demais, e determinar o tamanho da sequência motivo. No caso de experimentos *in vivo*, existe ainda a problemática adicional de tentar localizar e obter sequências representativas a partir de fragmentos longos de DNA. Entende-se descoberta de motivo como a utilização de um ou mais algoritmos que, partindo de sequências de DNA, consiga extrair um padrão que melhor represente a afinidade de um FT alvo. Durante o processo, cada sequência que representa e constrói o motivo pode ser considerado um SLFT (KULAKOVSKIY; MAKEEV, 2013). Assim, sequência motivo e conjunto de SLFTs, na definição deste trabalho, são duas faces da mesma moeda.

Existem diversos algoritmos computacionais propostos para realizar a descoberta de motivos (EGGELING, 2018; BAILEY, 2020; NGUYEN *et al.*, 2018). Essencialmente, a descoberta de motivos é uma etapa de custos computacionais elevados e, precisa resolver todas as problemáticas listadas acima em milhares de sequências obtidas ao final do experimento. Além disso, necessita fazer uso de um modelo classificador que é calibrado e treinado conforme o número de iterações do algoritmo. A PWM, por ser mais simples de implementar e com menores custos computacionais, costuma fazer parte do cerne desses algoritmos em sua maioria. Além disso, a busca por um motivo, principalmente em dados *in vivo*, precisa ser reiterada múltiplas vezes para buscar o motivo mais representativo. Isso é devido ao fato de que o FT alvo pode interagir com demais FTs, sendo que sequências longas de DNA permitem adquirir representações de SLFTs para diversos FTs. Portanto, múltiplos motivos podem ser adquiridos e uma análise de enriquecimento de motivos (BAILEY; MACHANICK, 2012; BAILEY; GRANT, 2021) deve ser feita para escolher o mais significativo, dada a amostra. Uma vez escolhido, geralmente é necessário converter a sequência motivo no modelo classificador que foi implementado (por exemplo PWMs) e adquirir de maneira retroativa as sequências SLFTs que compõem a sequência

[4] HT-SELEX (*high throughput systematic evolution of ligands by exponential enrichment)* é uma técnica *in vitro* que testa a afinidade do FT com moléculas sintéticas de DNA de tamanho definido. A técnica faz uso de PCR para amplificar as sequências que obtiveram maior afinidade ao FT e, a cada ciclo refinar o processo.

motivo obtida. Provavelmente, uma explicação para isso está nos custos computacionais elevados dos algoritmos de descoberta de motivos, dificultando reter em memória as SLFTs representativas do motivo e sem trazer custos elevados de tempo de processamento ao usuário.

Como ambos PWM e GRE (descritos nas seções 2.2.5 e 2.2.6) são modelos que precisam ser treinados diretamente com SLFTs, existe uma dúvida se a etapa anterior de descoberta de motivos poderia afetar a extração dos respectivos SLFTs e, consequentemente, o treinamento desses modelos, principalmente pelo fato de descobridores de motivos serem em maioria baseados em PWMs. Por isso, a tarefa de comparar PWMs e GREs está em dois níveis, como listado na seção de hipóteses. O primeiro nível de comparação é direto e compara ambos os modelos a partir da amostra de SLFTs adquirida pelo mesmo algoritmo de descoberta de motivos. O segundo nível de comparação avalia os modelos sob diversas técnicas de descoberta de motivos, assumindo que todas convergem em algum grau para uma sequência motivo razoavelmente semelhante para um mesmo FT. Existem também a dúvida se algoritmos de descoberta de motivos baseados em PWMs poderiam fornecer SLFTs em que possa ser medido dependência entre posições de bases. Isso justificado no fato que um modelo independente à posições de bases (PWM), simplesmente ignoraria tal informação na obtenção de SLFTs, não significando que esta informação não poderia estar disponível e ser avaliada.

Como última observação, o resultado final desta etapa é fornecer SLFTs que possam treinar os modelos PWMs e GREs. Mas, o propósito final desses modelos é serem utilizados com maior generalização possível, podendo classificar sequências de DNA adicionais e possivelmente descobrir novas localizações genômicas de SLFTs. Tudo isso, com melhor acurácia possível.

Algoritmo RSAT

O algoritmo de descoberta de motivos RSAT (*peak motifs* é constituído por quatro algoritmos que, separados, possuem a tarefa de descobrir padrões mais representativos no conjunto de sequências. Ao final, o RSAT retorna uma lista com cada motivo descoberto e um escore de significância. Cada subalgoritmo do RSAT consiste de uma abordagem de contar composições de *kmers* (sequências de tamanho k) que estejam mais frequentes

e estatisticamente mais significativos do que outras composições. A seguir, uma breve descrição sobre cada um dos quatro algoritmos independentes:

- *dyad-analysis*: específico para FTs em que há duas subunidades proteicas (*dyad*), geralmente simétricas, reverso complementares entre si e bem conservadas que estão em contato direto com o DNA, mas separadas por um espaçamento fixo. Cada subunidade geralmente reconhece um motivo conservado e curto de cerca de 3 pares de bases, enquanto o espaçamento contém uma composição não relevante ao motivo, pois não está em contato direto com o DNA (por hipótese). O algoritmo conta todos os pares de kmers (com tamanhos fixados) possíveis formados (com restrições dadas pelos usuários) e compara a frequência observada com a frequência esperada (um *dyad* formado ao acaso). Nota-se que cada contagem é de uma composição fixa de nucleotídeos, portanto para formar motivos degenerados, uma medida de distância obtém os *dyads* mais similares entre si e junta num único motivo (HELDEN, 2000);
- *oligo-analysis*: assim como o *dyad-analysis*, este algoritmo realiza uma contagem de todos os kmers (oligos) de um tamanho escolhido pelo usuário (e fixado), mas sem formar pares inter-espaçados. A contagem (frequência observada) é comparada com a frequência esperada, obtida com um modelo nulo, e por fim, um teste estatístico verifica quais sequências observadas são significativas em relação ao número esperado (HELDEN; ANDRÉ; COLLADO-VIDES, 1998);
- *position-analysis*: assume que as sequências estejam alinhadas por alguma referência [5] e procura enriquecimento de kmers que estejam mais representados em uma posição do que randomicamente ao longo das sequências usadas de entrada (exemplo: picos de ChIP-seq);
- *local-word analysis*: esse algoritmo combina as ideias do oligo-analysis com o position-analysis, resultando na busca por kmers diferencialmente representados considerando uma posição no conjunto de sequências de entrada contra um conjunto de sequências controle. Portanto, é uma análise de representação diferencial dos kmers. Por padrão, o conjunto de sequências controle é o embaralhamento das sequências de entrada, mas podem ser definidas pelo usuário (THOMAS-CHOLLIER *et al.*, 2011) [6]

[5] Neste trabalho, as sequências usadas para o algoritmo de descoberta de motivo estão alinhadas pelo início e fim de uma sequência *probe* de PBM ou pelo *peak summit* de um ChIP-seq.

[6] O algoritmo é mencionado na referência, mas definido na própria página do manual, encontrado em (DEFRANCE, 2011).

Todos os motivos encontrados pelo RSAT são kmers únicos, que após um critério com limiar, são unidos para formar um motivo com degenerações (como esperado de um motivo). A união dos kmers encontrados por cada algoritmo é convertida e retornada na forma de uma PWM.

Algoritmo STREME

STREME (Simple, Thorough, Rapid, Enriched Motif Elicitation) (BAILEY, 2020) é um algoritmo que procura por motivos enriquecidos na amostra de entrada em comparação a um grupo controle, este na forma de sequências embaralhadas, assumindo que cada sequência possua zero ou uma ocorrência do motivo a ser encontrado. Segue as etapas sequenciais do algoritmo:

1. Preparação dos dados. Etapa que verifica a integridade dos dados, gera a amostra controle embaralhando, em kmers, a amostra de entrada e cria um modelo de markov de ordem $k-1$ como modelo nulo para os cálculos de log-odd da PWM a ser estimada.
2. Criação de árvore de sufixos a partir da amostra de entrada e de controle. A árvore de sufixos é uma estrutura de dados que contém todos os possíveis sufixos de uma sequência, indexadas por posição. Um sufixo s é uma cadeia (de símbolos) de uma cadeia t, se existe uma cadeia $p = ts$. A árvore é utilizada para agilizar as buscas de instâncias de motivos pelo conjunto de sequências de entrada e de controle, que pode ser um processo demorado e custoso.
3. Avaliação de cada possível cadeia. Para cada possível cadeia, de tamanho definido pelo usuário, contar a presença da cadeia na amostra de entrada e controle usando a árvore de sufixos e depois obter o p-valor de enriquecimento.
4. Refinamento do motivo. STREME converte as quatro melhores cadeias de cada tamanho em PWMs e refina o motivo formado pelas quatro cadeias. O motivo de PWM formado desta maneira é refinado obtendo o score log-odd em janela deslizante de cada sequências de entrada e controle e obtendo melhores representações de motivo a partir do melhor score possível (de um limiar predeterminado). De cada sequência de entrada, se obtém apenas um ou zero ocorrências do motivo, que são realinhados para montar um novo motivo que será utilizado iteradamente até um critério de parada.

5. Significância do motivo. Com o motivo e a respectiva PWM obtida, classificar sequências de teste, reservadas como uma parcela das sequências de entrada e não vistas pelo algoritmo durante a descoberta de motivo, e calcular a significância.
6. Mascaramento de motivos. Para obter novos motivos, o algoritmo mascara instâncias dos motivos (o melhor score de cada sequência de entrada) que já foram coletados pelas etapas anteriores e itera novamente até um critério de parada, especificado pelo usuário. Esse critério geralmente é baseado no tempo de processamento, na quantidade e na significância dos novos motivos que forem coletados.

Algoritmo InMoDe

O algoritmo InMoDe (*intra-motif-dependencies*) (EGGELING, 2018) é um modelo de descoberta de motivos baseado em árvores de contextos parcimoniosas (ou *parsimonious context trees*, PCT), que na sua forma equivale a cadeias de Markov não-homogêneas e parcimoniosas. Em outras palavras, é um modelo que captura dependência entre posições de bases (como uma cadeia de Markov) a partir de um contexto, ou seja, do conjunto de símbolos prévios ao símbolo considerado em uma sequência. O contexto desses modelos é parcimonioso, ou seja, o modelo procura assumir uma forma simples (com redução de parâmetros) sem perder sua capacidade preditiva alta, podendo ser ajustado em composição e número de símbolos [7] de acordo com um critério de aprendizado e de poda da árvore. Em especial, o algoritmo InMoDe consegue aprender contextos de no máximo ordem seis. Alguns parâmetros como a ordem máxima do contexto, o número de iterações e o tamanho do motivo a ser encontrado devem ser ajustados pelo próprio usuário.

O aprendizado do algoritmo assume que as sequências de ChIP-seq de entrada possuem tamanho arbitrário com exatamente uma instância do motivo a ser encontrada em cada sequência (incluindo o reverso complementar), usando o algoritmo *Expectation-Maximization* EM. O algoritmo EM busca maximizar localmente a probabilidade de um modelo, condicionado aos parâmetros selecionados, explicar a posição da instância do motivo em cada sequência de entrada. O algoritmo itera múltiplas vezes para obter tanto os parâmetros do modelo como a melhor distribuição das instâncias de motivo ao longo das sequências (verossimilhança da sequência dada o modelo com os parâmetros estimados) até um critério de parada ou convergência. A estrutura das PCTs tenta reduzir o espaço de

[7] Isto é, a ordem da cadeia de Markov potencialmente muda de acordo com a transição considerada.

parâmetros a serem estimados enquanto estes são reobtidos a cada iteração. Essa estrutura é podada e selecionada de acordo com o critério BIC (*Bayesian Information Criterion*), que em modelos de Markov, são conhecidos por fazer a seleção da ordem da cadeia (ou seja, do contexto), sendo adotado uma abordagem semelhante para escolher os contextos que pertencem a PCT final. Sendo assim, exige um elevado tamanho amostral e de alta qualidade, para garantir que a estrutura do modelo seja aprendida corretamente.

Ao final, o algoritmo retorna um modelo ajustado aos dados com algum grau flexível de dependência entre posições de bases, podendo generalizar até para independência em algumas posições (equivalente a PWMs) em alguns casos particulares. Também retorna as instâncias do motivo equivalente ao número de sequências de entrada, pela hipótese assumida pelo algoritmo.

2.2 Aprendizado Computacional

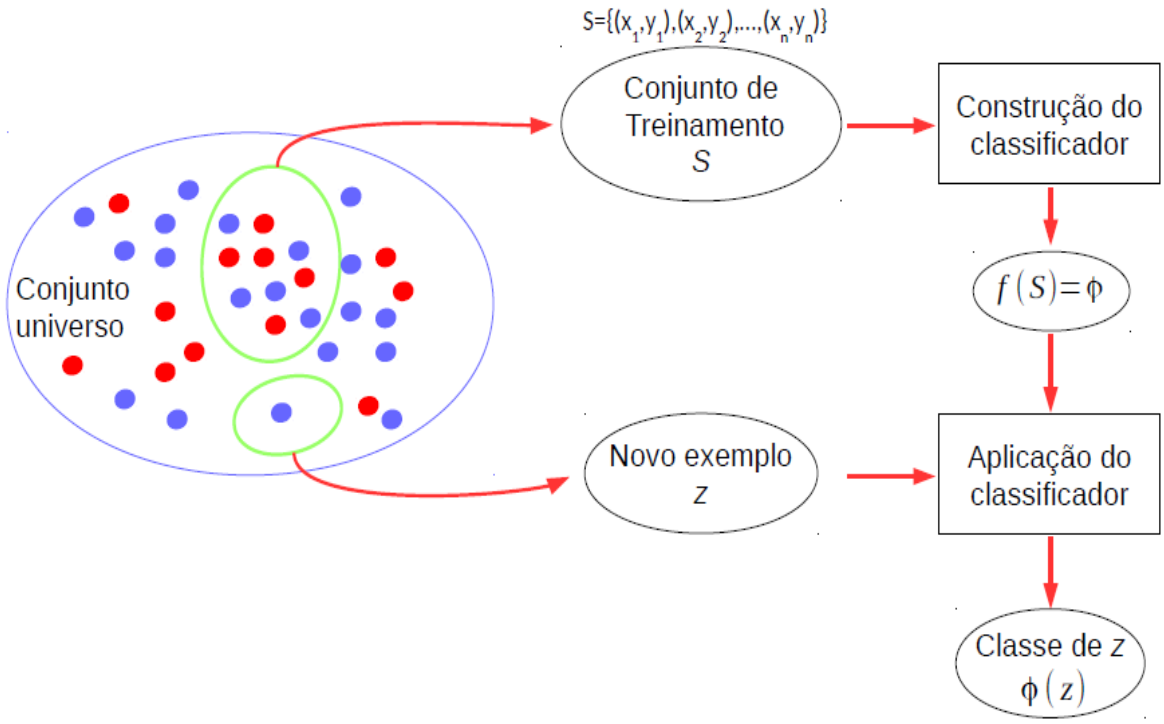
Esta seção aborda os conceitos computacionais e estatísticos envolvidos neste trabalho, tendo sido baseada em grande parte nas obras dos autores James, Hastie e Tibshirani (2013) e Duda, Hart e Stork (2001). Para mais informações sobre o tema Aprendizado Computacional, o leitor poderá se referir a esses autores citados.

Entende-se por aprendizado computacional ou *machine learning* como um conjunto de métodos computacionais, normalmente associados a técnicas estatísticas, que extraem características e identificam padrões em um conjunto de dados com a finalidade de realizar alguma capacidade preditiva (ou estimativa) pré-definida. Apesar do termo "aprender"seja biologicamente associado a um processo cognitivo de adquirir conhecimento baseado em evidências, neste contexto, o termo "aprendizado" possui um sentido conotativo.

Este trabalho utiliza-se de aprendizado supervisionado, sendo esse, portanto, o foco desta seção. Nesse tipo de aprendizado assume-se a existência de uma amostra de treinamento, que é um conjunto de exemplos dos conceitos que se pretende aprender. Mais precisamente, a amostra de treinamento é um conjunto de N pares ordenados (x_i, y_i) no qual x_i é o i-ésimo exemplo e y_i é seu rótulo (classe).

A Figura 3 ilustra um processo típico de aprendizado supervisionado de um classificador e sua posterior utilização. A amostra de treinamento serve de entrada para o algoritmo de aprendizado (ou de indução) que retorna um classificador aprendido. Consi-

Figura 3 – Representação ilustrativa do aprendizado supervisionado. Um conjunto universo de exemplos representa a distribuição real de todos os exemplos, ou seja, a população. O modelo a ser aprendido recebe um conjunto finito S do conjunto universo. Utilizando o conjunto S como treinamento, o algoritmo de aprendizado f constrói um modelo ϕ, que é o modelo classificador aprendido. Esse classificador é utilizado para classificar um novo exemplo z do conjunto universo.



Fonte: Clebiano da Costa Sá (2018).

derando o contexto desse trabalho de modelos de classificação, um algoritmo indutor de classificadores é uma função $f(S) = \phi$, ou seja, uma função que dada uma amostra de treinamento S gera (aprende) um classificador ϕ. Esta função ϕ, por sua vez, recebe como entrada uma nova instância z e retorna como resultado $\phi(z) = y$, sendo y a classe predita.

2.2.1 Amostra de teste e medidas de desempenho de classificadores

Uma propriedade essencial desejável de qualquer aprendizado é que esse tenha a capacidade de generalização. Um modelo de aprendizado computacional, para realmente ser funcional, precisa ter a capacidade de predizer corretamente novas instâncias que forem apresentadas, e não apenas as instâncias da amostra de treinamento. O fenômeno de aprendizado de um classificador que apenas classifica corretamente instâncias da amostra de treinamento é denominado *overfitting*. O *overfitting* é um problema que deve ser evitado a

qualquer custo no processo de aprendizado, pois estaria implicando na não-generalização do modelo de aprendizado, considerando que a amostra de treinamento é apenas uma pequena parcela do conjunto de possíveis instâncias observáveis no conjunto população. Desta forma, para testar se os modelos generalizaram corretamente a amostra de treinamento, um conjunto de amostra de teste é necessária.

Quando o problema a ser tratado envolve apenas duas classes dizemos que trata-se de um problema de classificação binária. Nesses casos é usual chamar uma das classes de "positiva" (por exemplo a classe dos sítios de ligação de um fator de transcrição específico) e a outra de "negativa" (não ser um sítio de ligação de tal fator). Assim, uma amostra de teste é composta de sequências positivas e também negativas. A grosso modo, uma vez treinados os modelos, a amostra de teste é utilizada para avaliar o desempenho do classificador em classificar corretamente tanto o que é verdadeiramente um SLFT como o que não é um SLFT.

Para uma amostra de teste contendo N instâncias, considera-se os seguintes resultados do classificador:

- número de verdadeiros positivos (VP): número de instâncias de teste que são positivas e que foram corretamente classificadas como positivas;
- número de falsos positivos (FP): número de instâncias de teste que são negativas mas que foram erroneamente classificadas como positivas;
- número de verdadeiros negativos (VN): número de instâncias de teste que são negativas e que foram corretamente classificadas como negativas;
- número de falsos negativos (FN): número de instâncias de teste que são positivas mas que foram erroneamente classificadas como negativas.

Com base nessas quatro medidas acima é possível obter métricas mais elaboradas, que são comumente utilizadas para avaliar o desempenho de um classificador. Dentre elas tem-se:

- Revocação $= \frac{VP}{VP+FN}$

  A revocação – também chamada de sensibilidade – é a taxa de Verdadeiros Positivos que um classificador rotula dentre todas as instâncias originalmente positivas;

- Especificidade = $\frac{VN}{VN+FP}$

  A especificade é a taxa de Verdadeiros Negativos que um classificador rotula dentre todas as instâncias originalmente negativas;

- Precisão = $\frac{VP}{VP+FP}$

  A precisão é a taxa de Verdadeiros Positivos dentre todos os classificados como positivos pelo classificador;

- Acurácia = $\frac{VP+VN}{VP+VN+FP+FN}$

  A acurácia é a porcentagem de acertos (totais) em relação à totalidade dos dados.

Dependendo do algoritmo indutor de classificador, o classificador gerado não fornece a classificação mas sim um valor que representa um escore ou uma probabilidade da instância ser, por exemplo, positiva. Nestes casos é necessário estipular um limiar de forma que instâncias com valores acima de tal limiar sejam consideradas positivas. Assim, cada valor de limiar resulta em diferentes valores de VP, VN, FP e FN, e portanto alteram as medidas de desempenho. Por isso deve existir, além das medidas de desempenho descritas anteriormente, alguma forma de visualização de todos os resultados para diferentes limiares. Para isso, existem as curvas *Receiver operating characteristic* (curva ROC) e a curva *Precision-Recall* (curva PR)[8]. Tanto na curva ROC como na PR, cada ponto representa o resultado proveniente de um limiar distinto de classificação. A curva PR possui dois eixos: o eixo horizontal representa a Revocação (ou Sensibilidade), enquanto a coordenada o eixo vertical representa a Precisão. Já a curva ROC possui como eixos horizontal e vertical 1-especificidade e a sensibilidade, respectivamente.

É fato que existem mais sequências não-SLFTs do que SLFTs. Logo, ao avaliar o desempenho de um modelo ao longo de um genoma, haverá um número maior de sequências esperadas como negativas do que positivas. Desta forma, com amostras positivas e negativas desbalanceadas, a curva ROC tende a mascarar os resultados. Isso porque a medida de $1-$Especificidade (ou Taxa de Falsos Positivos) não captura adequadamente o desempenho do modelo sob condições de desbalanceamento dos dados porque, uma vez que o número de sequências negativas é grande, mesmo um número não desprezível de falsos positivos (FP) irá diluir-se com o número de verdadeiros negativos (VN), apresentando uma especificidade alta. Por isso, a interpretação do desempenho dos modelos na curva ROC, em cenário de amostra desbalanceada, fica comprometida.

[8] Traduzida como Precisão-Revocação

Já a medida de Precisão, presente somente na curva PR, varia de acordo com o desbalanceamento das amostras positivas e negativas (SAITO; REHMSMEIER, 2015), e por isso, a curva PR captura melhor o desempenho dos modelos no cenário deste trabalho.

Para a comparação de modelos classificadores, é desejável que exista uma métrica de valor escalar para comparar duas curvas PR distintas. Para tal, existe a área sob a curva PR ou $AUPCR$ [9] que é uma medida que representa o desempenho de um classificador considerando os vários limiares de classificação. Uma medida alternativa de se resumir uma curva e que tende à área sob a curva é a AP (SU; YUAN; ZHU, 2015).

Seja $P(l_i)$ e $R(l_i)$ a precisão e a revocação, respectivamente, no limiar l_i, em que i é o índice ordenado de maneira crescente pela revocação. Ou seja, $R(l_{i-1}) \leq R(l_i)$, $i = 1, 2, 3, .., n\}$

A AP, na prática, pode ser calculada[10] como (KATERENCHUK; ROSENBERG, 2018; PEDREGOSA *et al.*, 2011):

$$AP = \sum_{i=1}^{n} P(l_i)\Delta_R(l_i) \tag{1}$$

em que, $\Delta_R(l_i) = R(l_i) - R(l_{i-1})$.

Embora a AP permita verificar o desempenho de um classificador considerando os vários limiares adotados, é necessário ainda escolher um limiar ótimo a ser reportado como resultado ótimo do classificador.

É possível realizar tal escolha inspecionando a curva e baseando-se no balanço desejável da Precisão e Revocação, ou então utilizar um critério mais geral e automático.

Para isso, a medida F_1 (ou F_1-score) é uma medida que representa a média harmônica entre as medidas Precisão e Revocação, delimitada pelo intervalo entre zero e um. Define-se o F_1-score como :

$$F_1 = 2\left(\frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}\right) \tag{2}$$

Na curva PR, cada ponto representa a medida de Precisão e Revocação para um limiar escolhido. Logo, seria desejável adotar um limiar ótimo cujo valor Precisão e Revocação sejam ambos altos (isto é, o mais próximo de 1). Pode-se então obter como limiar ótimo (l^*) aquele cuja medida F_1 para um modelo seja a maior. Seja L o conjunto de

[9] Na prática, a AP equivale a Área sob Curva Precisão-Revocação(SU; YUAN; ZHU, 2015)

[10] Observe como esse cálculo se assemelha à soma de Riemann para aproximar a AUPRC.

todos os limiares de uma curva PR e F_1-score(l) a medida F_1 do l-ésimo limiar, define-se l^* :

$$l^* = \underset{l \in L}{\arg\max} \left(F_1\text{-score}(l) \right) \tag{3}$$

2.2.2 Validação Cruzada em k-partes

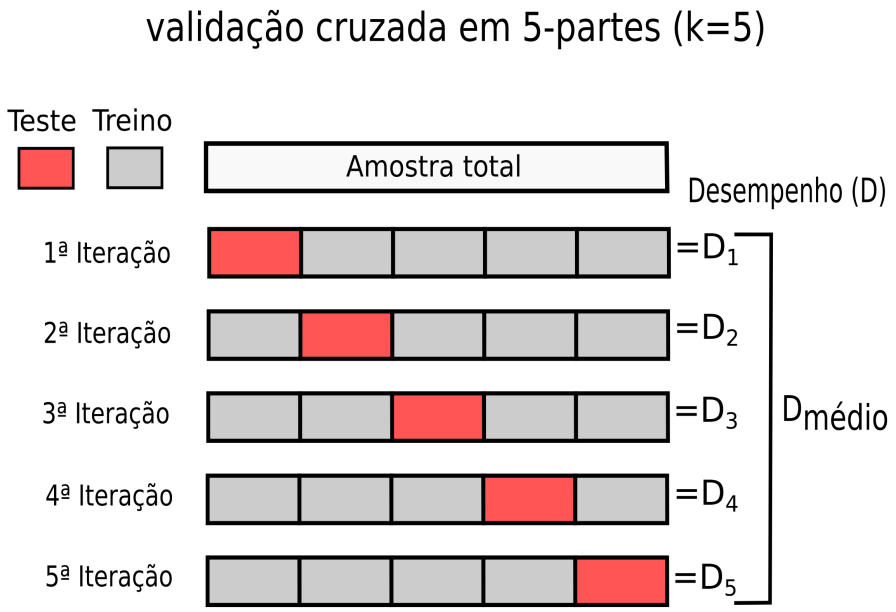
Como visto, o aprendizado do modelo necessita de uma amostra de treinamento e de uma amostra de teste sobre a qual são estimadas as medidas de desempenho do classificador, como as mencionadas na seção 2.2.1. Na prática, apenas um conjunto de amostra é fornecido, então é necessário um critério para dividir a amostra em treinamento e teste. Por um lado, usar toda a amostra como treinamento resulta em uma imprecisão na estimação do erro do modelo. Por outro lado, quanto menos amostra for alocada ao treinamento, menos informação um modelo possui sobre a variabilidade dos dados e, portanto, menor poder preditivo. Além disso, fazer uma única divisão em treinamento e teste normalmente não fornece uma boa estimativa das medidas de desempenho do classificador que será treinado com a amostra toda e aplicado às novas instâncias, pois cada divisão potencialmente fornece resultados diferentes.

Uma solução para o problema ilustrada na Figura 4, é realizar uma validação cruzada em k-partes, (REFAEILZADEH; TANG; LIU, 2009) que consiste em :

- dividir a amostra original em k partes disjuntas de tamanhos iguais, em que k é um valor arbitrário, que geralmente varia entre 5 e 10;
- separar uma das k partes para ser a amostra de teste e utilizar o restante para amostra de treinamento;
- rotacionar a amostra de teste, escolhendo outra parte como teste e todo o restante como amostra de treinamento, repetindo essa rotação até que todas as k partes tenham sido utilizadas unicamente como teste por rodada.

Como em uma validação cruzada em k-partes há k etapas de treinamento e teste, há k valores para cada medida de desempenho sendo estimada. Para cada uma delas, o valor estimado é a média desses k valores (Figura 4).

Figura 4 – Validação cruzada em k-partes. A figura representa uma validação em que k= 5 partes. A amostra total, contendo todo os exemplos que foram fornecidos para o modelo está representado como um retângulo, na qual a largura representa, de maneira didática, o tamanho (fixado) dessa amostra. Dividindo essa amostra em 5 partes iguais e distintas, resulta na divisão mostrada como 5 retângulos adjacentes. O retângulo vermelho representa a parte da amostra total que será usada como teste, enquanto o retângulo cinza representa a parte que é usada como treinamento. Em cada iteração, um dos conjuntos é utilizado como teste, enquanto o restante para treino, até que todas as partes tenham sido utilizadas uma única vez como teste. Cada iteração resulta em uma medida arbitrária de desempenho D. Ao final da validação cruzada, a medida D desejada é a média de todas as k (no caso 5) medidas de cada iteração realizada.



Fonte: Guilherme Miura Lavezzo (2020).

2.2.3 Amostras de sequências para classificadores de SLFT

Nesse contexto de classificadores de SLFTs, uma amostra positiva de sítios de um determinado fator de transcrição se refere a um conjunto de N sequências de tamanho k [11] Essas amostras são positivas pois estão rotuladas corretamente como sendo sequências de DNA que são SLFT. Dessa forma, para o aprendizado de classificadores probabilísticos (seções 2.2.5 e 2.2.6) adequados, é importante possuir uma amostra de treinamento positiva

[11] Note: a letra k neste contexto não equivale ao "k" encontrado na validação cruzada k-fold, que é comum se referirem por "k-fold" como um número arbitrário de folds. Já a notação k do tamanho da sequência foi usada por ser comum na bioinformática para se referir a um k-mer (sequência de tamanho k) e foi a notação usada em (TOMOVIC; OAKELEY, 2007), o qual este trabalho se baseou para aprimorar novas medidas de dependência entre posições de bases.

cujas sequências sejam exatamente os sítios de ligação e possua uma distribuição o mais próxima da realidade possível. Uma forma de criar tal amostra é a partir de sequências obtidas por experimentos que sejam capazes de fornecer a coordenada genômica de onde houve a ligação, assunto abordado na seção 2.1.2.

Em Bioinformática é normalmente difícil dispor de uma amostra negativa, pois isso implicaria em garantir que tais sequências não são positivas nunca. No contexto deste projeto, isso corresponderia a obter um conjunto representativo de sequências às quais se garante que o FT de interesse não se ligue, sob nenhuma hipótese dada pela condição experimental estudada. Assim, é oportuno a utilização de classificadores probabilísticos binários que consideram duas classes: a positiva e a "nula" (detalhes metodológicos em 3.5.1). A classe de sequências nulas corresponderia a qualquer sequência de mesmo tamanho que as sequências positivas, independente de qual seja a sua afinidade pelo FT alvo. Por questões didáticas, denominamos uma amostra negativa como sendo composta de sequências não-positivas, ou nulas.

Além da composição de uma amostra negativa, também é comum o uso de modelos nulos para vias de comparação com os classificadores selecionados. Neste contexto, modelos nulos se referem a modelos que não foram treinados com uma amostra positiva e que portanto não possuem seletividade para classificar SLFTs, como é esperado dos classificadores treinados. Possibilidades de modelagem da classe nula são, por exemplo, considerar uma distribuição uniforme (cada nucleotídeo tem igual probabilidade de ocorrer, independente de sua posição) e a distribuição genômica da espécie sendo estudada (cada nucleotídeo ocorreria com a mesma probabilidade desse nucleotídeo ocorrer no genoma, independente de sua posição). O bom ajuste desse modelo nulo é importante para a diminuição da taxa de falsos positivos (MACHADO-LIMA; KASHIWABARA; DURHAM, 2010). Isso porque, como as sequências dos sítios são curtas, a chance de encontrar uma similar, ao acaso, costuma ser maior do que o desejável. Neste trabalho, o modelo nulo utilizado foi baseado na frequência de bases de cada genoma utilizado.

2.2.4 *Effect size* e D de Cohen

O tamanho do efeito, ou *effect size*, é uma medida complementar ao p-valor de um teste de hipóteses. Seu objetivo é associar uma diferença ou distância ao efeito observado

entre os grupos testados, via a estatística utilizada (SULLIVAN; FEINN, 2012). Sob hipótese nula, o *effect size* assumido é geralmente 0, pois assume-se que não há diferença entre as estatísticas medidas entre os grupos (COHEN, 1977). O p-valor de um teste de hipótese mede a probabilidade da estatística medida ser o valor testado ou valores mais extremos, puramente por chance, ou seja, sob hipótese nula. O p-valor é dependente do tamanho da amostra, da variabilidade dos dados, entre outros fatores que podem acabar distorcendo sua interpretabilidade. Por outro lado, o *effect size* é uma maneira quantitativa de se medir essa diferença observada e pode ajudar intuitivamente com a dimensão do efeito que está sendo testado (DUNKLER *et al.*, 2019).

Uma das medidas de *effect size* é o D de Cohen (COHEN, 1977), que possui valor mínimo em 0, e valor máximo indefinido. Apesar disso, por ser uma medida padronizada, existem referências da intensidade do efeito medido. Valores de D em 0 indicam nenhuma diferença da média entre dois grupos, valores de 0.1 indicam uma fraca diferença, valores próximos a 0.5 indicam uma diferença intermediária, e valores pertos ou acima de 1.0, indicam forte diferenças (SAWILOWSKY, 2009).

O D de Cohen pode ser definido como:

$$D = \frac{\overline{x_1} - \overline{x_2}}{s} \tag{4}$$

sendo $\overline{x_i}$ a média do grupo i e, s, o desvio padrão entre grupos, definido como:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{5}$$

sendo que s_i^2 se refere à variância do grupo i.

2.2.5 PWM como modelos preditores de SLFT

Position Weight Matrices são matrizes $4 \times k$ que matematicamente caracterizam motivos de tamanho k (STADEN, 1984). Cada PWM possui um escore na posição (b, i), sendo b correspondente aos nucleotídeos A, C, G e T e i ($i = 1, 2..., k$) a posição na sequência. Esse escore representa, informalmente, o quão mais provável é a ocorrência de um nucleotídeo b na posição i em relação à ocorrência do mesmo nucleotídeo representada pelo modelo nulo.

A PWM dos sítios de ligação de um FT específico é construída a partir de uma amostra positiva de treinamento da seguinte forma. Para cada nucleotídeo b em uma determinada posição i, o valor $W_{b,i}$ (posição (b, i) da matriz) denota o escore atribuído pela PWM, sendo calculado como :

$$W_{b,i} = \log_2 \left(\frac{p(b,i)}{q(b)} \right), \tag{6}$$

sendo $p(b, i)$ a probabilidade do nucleotídeo b na posição i sob a hipótese dele ser um sítio do FT em questão e $q(b)$ a probabilidade do nucleotídeo b sob um modelo nulo escolhido posição-independente. Geralmente $q(b)$ é escolhido como sendo 0.25 para todas as bases (modelo nulo uniforme) ou obtido a partir da frequência relativa de bases do genoma do organismo de estudo, como mencionado na seção 2.2.3. Já $p(b, i)$ é calculado a partir da amostra positiva de treinamento, utilizando a seguinte equação:

$$p(b,i) = \frac{f_{b,i} + \psi(b)}{n + \sum_{b' \in \{A,C,T,G\}} \psi(b')}, \tag{7}$$

sendo $f_{b,i}$ a contagem de ocorrências da base b na posição i na amostra de treinamento, $\psi(b)$ uma função de pseudocontagem, atribuída arbitrariamente para corrigir problemas cujas probabilidades estimadas sejam zero[12] e n é o número de sequências da amostra de treinamento (WASSERMAN; SANDELIN, 2004).

Neste trabalho, a função pseudo-contadora $\psi(b)$ para uma base $b \in \{A, T, C, G\}$ foi definida como (XIA, 2012) :

$$\psi(b) = \alpha f_b \tag{8}$$

com $\alpha = \frac{1}{10n}$, um valor inversamente proporcional ao tamanho amostral n (já definido anteriormente) e f_b a frequência absoluta da base b.

Seja $S = s_1, s_2, ..., s_L$ uma sequência de tamanho $L \geq k$, com $i = 1, 2, ...L$ e s_i um nucleotídeo do conjunto {A,C,G,T}. Uma PWM $4 \times k$ atribui um escore $W(s')$ para cada sequência s' de tamanho k contida em S. Intuitivamente, uma PWM com k colunas

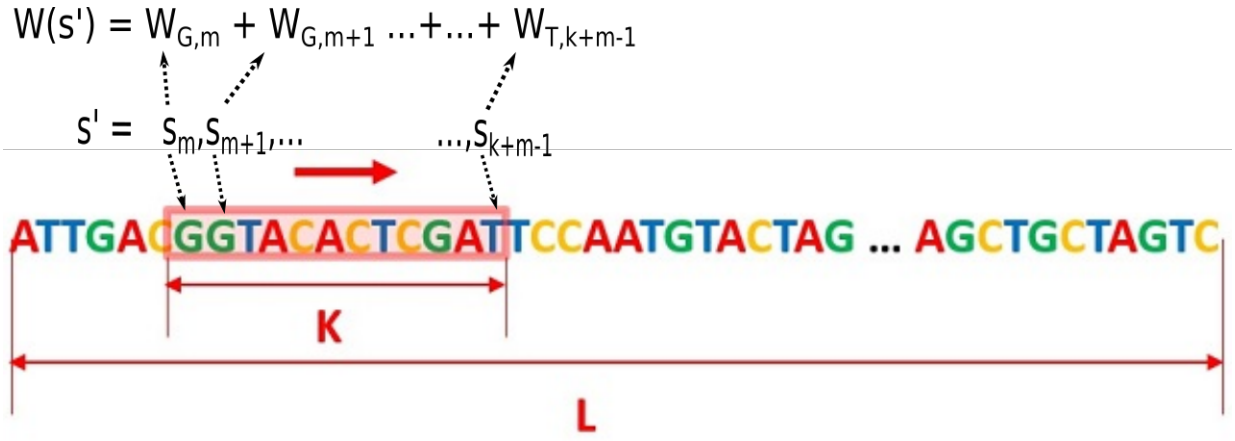
[12] O pseudocontador é atribuído para evitar o problema de baixas contagens da amostra de treinamento e evitar que a conversão log do score seja indefinida (em zero) (WASSERMAN; SANDELIN, 2004). Sendo um valor arbitrário, normalmente se utiliza como sendo $\psi(b') = 1$, mas em alguns casos pode ser variado para representar algum conhecimento *a priori* sobre determinado nucleotídeo.

atribui um escore sobre uma subsequência de tamanho k contida em S como uma janela deslizante de tamanho fixado em k (Figura 5). O escore $W(s')$ é definido como :

$$W(s' = s_m, ..., s_{k+m-1}) = \sum_{i=m}^{k+m-1} W_{b,i} \tag{9}$$

sendo m uma posição arbitrária ($m = 1, 2..., L - k + 1$); sendo $W_{b,i}$ como definido na equação 6, em que b se refere ao nucleotídeo que se encontra na posição i da sequência avaliada. Note que a sequência $s_m, ..., s_{k+m-1}$ tem tamanho k.

Figura 5 – Ilustração de como atua a janela deslizante de uma PWM $4 \times k$ sobre uma sequência de tamanho L. Para cada sequência s' equivalente à janela de tamanho k, um escore $W(s')$ é atribuído. O escore $W(s')$ é a soma dos $W_{b,i}$ individuais: no primeiro elemento (s_m, equivalente ao nucleotídeo G) da sequência s' é atribuído o escore $W_{G,m}$ e assim por diante até o último elemento da sequência (s_{k+m-1}, equivalente ao nucleotídeo T) com escore correspondente de $W_{T,k+m-1}$.



Fonte: Antônio Ferrão Neto (2017), modificada por Guilherme Miura Lavezzo (2020).

Os escores $W(s')$ para uma sequência s' podem ser tanto positivos quanto negativos. A classificação de tal sequência em SLFT ou não SLFT depende de um limiar l, escolhido arbitrariamente (DABROWSKI *et al.*, 2015). Uma PWM classifica uma sequência como sendo SLFT se :

$$W(s') > l \tag{10}$$

Esse limiar poderia, por exemplo, assumir valor $l = 0$ para qualquer sequência dada. Então, a classificação de uma sequência em SLFT seria dada por $W(s') > 0$. A figura 6 representa a atribuição de um escore por uma PWM treinada e o respectivo critério de classificação das sequências em SLFT ou não SLFT, conforme um limiar $l = 0$.

As PWMs são os modelos mais utilizados como preditores de SLFTs (WEIRAUCH *et al.*, 2013; DABROWSKI *et al.*, 2015; EGGELING, 2018; WASSERMAN; SANDELIN,

Figura 6 – Exemplo de uma PWM atribuindo escores para sequências, seguido da respectiva classificação em SLFT ou não SLFT dessas sequências. Caso a sequência satisfaça a condição que o Escore da PWM seja maior que 0 (Escore> 0), a sequência é classificada como SLFT. Caso contrário, a sequência é classificada como não sendo SLFT.

Fonte: Antônio Ferrão Neto (2017).

2004), todavia assumem independência entre as posições dos sítios. Diversos estudos mostraram que, para alguns casos, modelos que consideram dependência entre posições podem melhorar a acurácia de predição de SLFTs (ZHAO; STORMO, 2011; ZHAO *et al.*, 2012; GRAU; NETTLING; KEILWAGEN, 2019; EGGELING, 2018; MATHELIER; WASSERMAN, 2013).

A tarefa de encontrar o melhor modelo preditor para um FT é caso-específico. Em alguns casos, obter um modelo preditor mais complexo que exige maior número de parâmetros a serem estimados pode piorar os resultados de predição (ZHAO; STORMO, 2011; EGGELING, 2018; WEIRAUCH *et al.*, 2013).

2.2.6 Gramáticas Regulares Estocásticas como modelos preditores de SLFT

Gramáticas Regulares Estocásticas (GREs) possuem a vantagem de serem capazes de representar dependências entre bases adjacentes. Como alguns estudos mostraram (EGGELING, 2018; TOMOVIC; OAKELEY, 2007), modelar a dependência entre bases pode ser uma forma de reconhecer SLFTs mais acuradamente do que PWMs, em alguns casos. Informalmente, entende-se dependência entre posições de bases como sendo uma correlação entre as frequências de bases em posições distintas.

Gramáticas Regulares Estocásticas são dispositivos da teoria de linguagens formais que consistem em regras de substituição (ou produções) associadas com probabilidades. A subseção aqui descrita será introdutória aos conceitos básicos de Gramáticas Regulares , baseando-se nos trabalhos de Higuera (2010), Sipser (2013), Ron, Singer e Tishby (1997), Ron, Singer e Tishby (1998), os quais o leitor pode consultar para obtenção de conceitos mais aprofundados sobre o tema.

Definição 1 *Uma Gramática Estocástica é uma quíntupla* $G = \{V, \Sigma, R, S, P\}$*, na qual :*

- V *é um conjunto de símbolos não terminais (ou variáveis);*
- Σ *é um conjunto de símbolos terminais (ou alfabeto);*
- R *é um conjunto de regras de substituição ou produção, com cada regra seguindo a forma :* $\alpha \rightarrow \beta$

  com $\alpha \in (V \cup \Sigma)^* V (V \cup \Sigma)^*$ *e* $\beta \in (V \cup \Sigma^*)$*, sendo* α *chamado lado esquerdo da produção e* β *lado direito. A operação* $*$ *é o operador de Kleene, que define a operação de zero ou mais concatenações sobre os elementos imediatamente à sua esquerda;*
- S *é o símbolo inicial de onde todas as sentenças (ou cadeias) derivam, sendo que* $S \in V$*;*
- P *é uma função* $P : R \rightarrow [0, 1]$*, que associa uma probabilidade p para cada regra de substituição em R, tal que a soma de todas as produções com o mesmo lado esquerdo somem em 1.*

Uma gramática gera uma cadeia utilizando as regras de produção R. Começando com o símbolo inicial S, a produção de uma cadeia termina quando todos os símbolos não terminais (pertencentes a V) derivam em símbolos terminais (pertencentes a Σ).

Definição 2 *O conjunto de todas as cadeias produzidas por uma gramática G é chamado de linguagem $L(G)$.*

Definição 3 *Uma gramática estocástica é regular se todas as suas produções $\alpha \rightarrow \beta$ são regulares à esquerda ou à direita, sendo :*

- *regular à esquerda: com $\alpha \in V$ e ($\beta \in \Sigma$ ou $\beta \in V\Sigma$);*
- *regular à direita: com $\alpha \in V$ e ($\beta \in \Sigma$ ou $\beta \in \Sigma V$)*

A figura 7 mostra à direita as produções de uma gramática regular estocástica que gera a linguagem definida pela conjunto de sequências descrito à esquerda naquela figura. Tal gramática é definida por $V = \{S, E1, E2, ..., E22\}, \Sigma = \{a, c, g, t\}$, S é o símbolo inicial, P é a função que atribui probabilidade às regras em R (descritas ao lado direito de cada produção) e R é o conjunto de produções descritas na figura.

Uma gramática G atribui uma probabilidade $P_G(s)$ para uma sequência s como sendo o produto (de probabilidades) das produções que geram a sequência s. A classificação em SLFT ou não da sequência s, semelhante à classificação das PWMs, depende de um escore e de um limiar C. Para uma sequência s, pode-se definir o escore $H_G(s)$ de uma gramática G como sendo :

$$H_G(s) = \log_2(P_G(s)) - \log_2(P_N(s)) \tag{11}$$

em que $P_N(s)$ é a probabilidade que o modelo nulo N atribui para a sequência s, calculada como sendo o produtório das probabilidades de cada nucleotídeo de s segundo o modelo nulo, sendo este o mesmo modelo nulo utilizado em PWMs.

Portanto, classifica-se uma sequência s em SLFT caso $H_G(s) > C$, sendo C um limiar arbitrário.

O processo de aprendizado de uma gramática, estocástica ou não, a partir de uma amostra de sequências da linguagem alvo é chamado *inferência gramatical.* (SAKAKIBARA, 1995). GREs podem ser obtidas por algoritmos de inferência gramatical, como por exemplo o *Learn-APFA* (RON; SINGER; TISHBY, 1998) e o *Amnesia* (RON; SINGER; TISHBY, 1997). Assim, no contexto deste trabalho, o objetivo da inferência gramatical é utilizar uma amostra de sequências de sítios de um FT específico para aprender uma GRE cuja linguagem gerada represente o conjunto de todas as sequências de sítios daquele FT.

A seção a seguir contém uma breve explicação do processo de inferência gramatical pelo algoritmo LAPFA usado neste trabalho. Para mais informações sobre o assunto, o leitor pode consultar os trabalhos dos autores Ron, Singer e Tishby (1998), Ron, Singer e Tishby (1997) e Higuera (2010). Além disso, no que se refere a Inferência Gramatical, todos os algoritmos utilizados neste trabalho foram implementados no trabalho de Vieira e Durham (2005).

No campo de inferência de gramáticas regulares são mais comuns os algoritmos que inferem, no lugar de uma GRE, um dispositivo equivalente[13] chamado autômato finito determinístico estocástico. Cada algoritmo de aprendizado aqui utilizado pressupõe um tipo específico de AFDE, apresentados a seguir.

Algoritmo Learn-APFA

O algoritmo Learn-APFA (LAPFA), descrito originalmente por Ron, Singer e Tishby (1998), é um algoritmo de modelagem de dados sequenciais baseado em Autômatos Finitos Determinísticos Estocásticos Acíclicos AFDEA ou, do inglês "APFA" *Acyclic Probabilistic Finite Automata*. Dada uma amostra, ou seja um conjunto de sequências de tamanho k, o algoritmo consegue gerar um modelo M aprendido.

A seguir são apresentados a definição de um AFDE acíclico (definição 4), um exemplo de tal AFDE e sua GRE equivalente (Figura 7) e uma breve apresentação do algoritmo LAPFA.

Definição 4 *Define-se um AFDE acíclico (AFDEA) como uma 7-upla* $(Q, q_0, q_f, \Sigma, \zeta, \tau, \gamma)$*, em que:*

- Q *é um conjunto finito de estados;*
- $q_0 \in Q$ *é o estado inicial;*
- $q_f \notin Q$ *é o estado final;*
- Σ *é um alfabeto finito;*
- $\zeta \notin \Sigma$ *é o símbolo final (que representa o final de uma sequência);*
- $\tau : Q \times \Sigma \bigcup \{\zeta\} \rightarrow Q \bigcup \{q_f\}$ *é a função de transição;*
- $\gamma : Q \times \Sigma \bigcup \{\zeta\} \rightarrow [0, 1]$ *é a função probabilística de escolha do próximo símbolo.*

As funções devem satisfazer as seguintes propriedades:

[13] Dois dispositivos são equivalente se eles reconhecem exatamente a mesma linguagem (SIPSER, 2013).

- *para todo* $q \in Q$, $\sum_{\sigma \in \Sigma \bigcup \{\zeta\}} \gamma(q, \sigma) = 1$;
- *a função de transição* τ *só poderá não ser definida para estados* q *e símbolos* σ *em que* $\gamma(q, \sigma) = 0$;
- *para todo* $q \in Q$, $\tau(q, \zeta) = q_f$ *e* $\gamma(q, \zeta) > 0$;
- *o estado* q_f *pode ser alcançado por qualquer estado* $q \in Q$ *que possa ser alcançado a partir de* q_0;
- *a função de transição* τ *só pode ser definida entre dois estados* q *e* r $(r \neq q_f)$ *se* q *for um estado do nível* d *e* r *for um estado do nível* $d+1$, *para* $d = 0, ..., L$, *sendo 0 o nível da raiz (na qual encontra-se o estado* q_0*) e* L *o penúltimo*[14] *nível do autômato (sendo esta última propriedade a que torna o AFDE acíclico).*

Informalmente, o AFDEA é um autômato que começa no estado q_0 e termina no estado q_f, e que neste caminho gera sequências do tipo $\Sigma^*\zeta$. Para gerar um símbolo a partir de um estado, o símbolo previamente necessita ser escolhido pela função probabilística γ, associada a esse estado atual. Ou seja, uma função probabilística que escolhe o próximo símbolo é associada a cada estado. Se o símbolo ζ é escolhido, o estado q_f é atingido e a geração da sequência é encerrada. Isso significa na prática, que para o modelo encerrar, precisa passar pelo estado final q_f, que é determinado pelo tamanho das sequências de entrada. Portanto, sequências maiores ou menores que o conjunto de amostras não são aceitos pelo AFDEA.

Em outras palavras, AFDEAs podem ser vistos como grafos[15] acíclicos[16] direcionados[17], em que se começa com um estado inicial q_0. A parte estocástica do modelo é dada somente pela probabilidade de escolha do próximo símbolo, pela função γ definida para o atual estado. Uma vez definido o próximo símbolo por γ, a transição de um estado para outro passa a ser determinística, definida pela função τ.

Considere um AFDEA M e uma cadeia $s = s_1 s_2 \dots s_l$ em que $s_i \in \Sigma$ para $i = 1, ..., l-1$ e $s_l = \zeta$. Dizemos que s é aceita por M com probabilidade $P^M(s)$ se existe

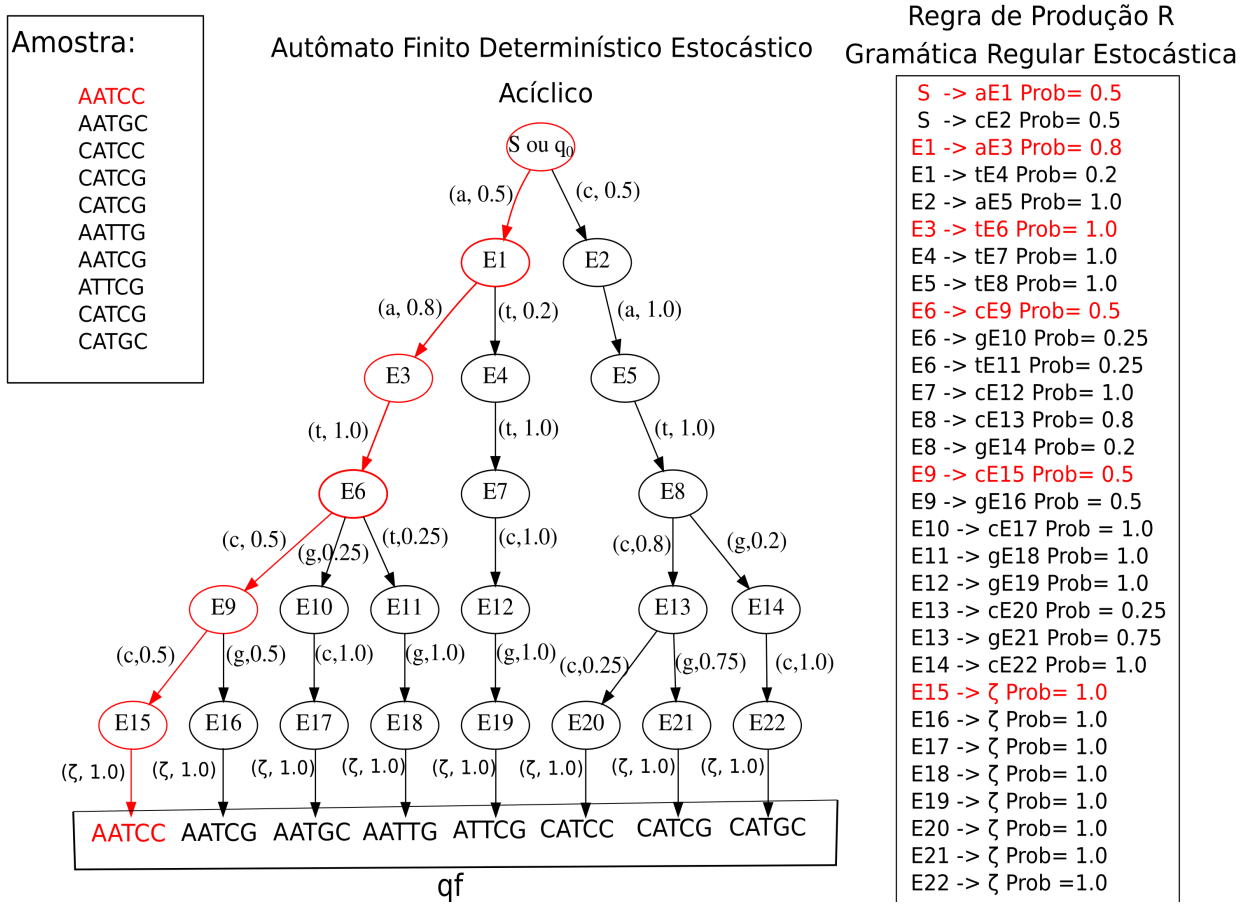
[14] O último nível desse AFDE possui apenas o estado q_f, para o qual não vale a restrição.

[15] Grafos são estruturas compostas de nós e arestas. Uma aresta liga dois nós, se estes possuem alguma medida arbitrária (ao contexto do problema) que os relaciona. Os nós representam estados, ou algum tipo de informação que se queira representar como um diagrama.

[16] Grafos são acíclicos se não existe um circuito fechado de nós interligados por arestas.

[17] Um grafo direcionado possui arestas que representam um caminho com destino e chegada. Ou seja, considerando dois nós e uma aresta relacionando-os, o sentido da aresta (seta) diferencia a relação entre os dois nós.

Figura 7 – Exemplo de uma amostra positiva sendo utilizada para treinar o modelo de Gramática Regular Estocástica(GRE, à direita) e o Autômato Finito Determinístico Estocástico Acíclico (AFDEA, ao centro) equivalente. Em vermelho, uma sequência de exemplo é representada tanto no AFDEA como na GRE. No AFDEA, o primeiro nó, chamado de S ou equivalentemente de q_0 transita para o próximo estado (indicado por uma seta) E_i $(i = 1, 2, ..., 22)$ com uma probabilidade. Os estados (E) representam a posição de uma base na sequência. Cada aresta está sendo rotulada com um nucleotídeo (a,c,g,t) e a probabilidade respectiva deste ser gerado nesta posição, dado que se está num determinado estado E_i. O caminho de um grafo traçado de cima para baixo gera uma cadeia, que cresce da esquerda para a direita. Note que a soma de todas as probabilidades que saem de um nó E_i é 1, pela definição 4. Ao receber o símbolo ζ, o AFDE atinge o estado final q_f, que foi indicado equivalentemente como uma sequência completa. Equivalentemente, a regra de produção R de uma GRE está sendo representada para a mesma amostra. Como pela definição 1, a soma de todas as produções com o mesmo lado esquerdo somam em 1.



Fonte: Guilherme Miura Lavezzo (2020).

uma sequência de estados $q_0, q_1, ..., q_l$ tal que $\tau(q_i, s_{i+1}) = q_{i+1}$ para $i = 0, ...l-1$ e $q_l = q_f$, sendo que:

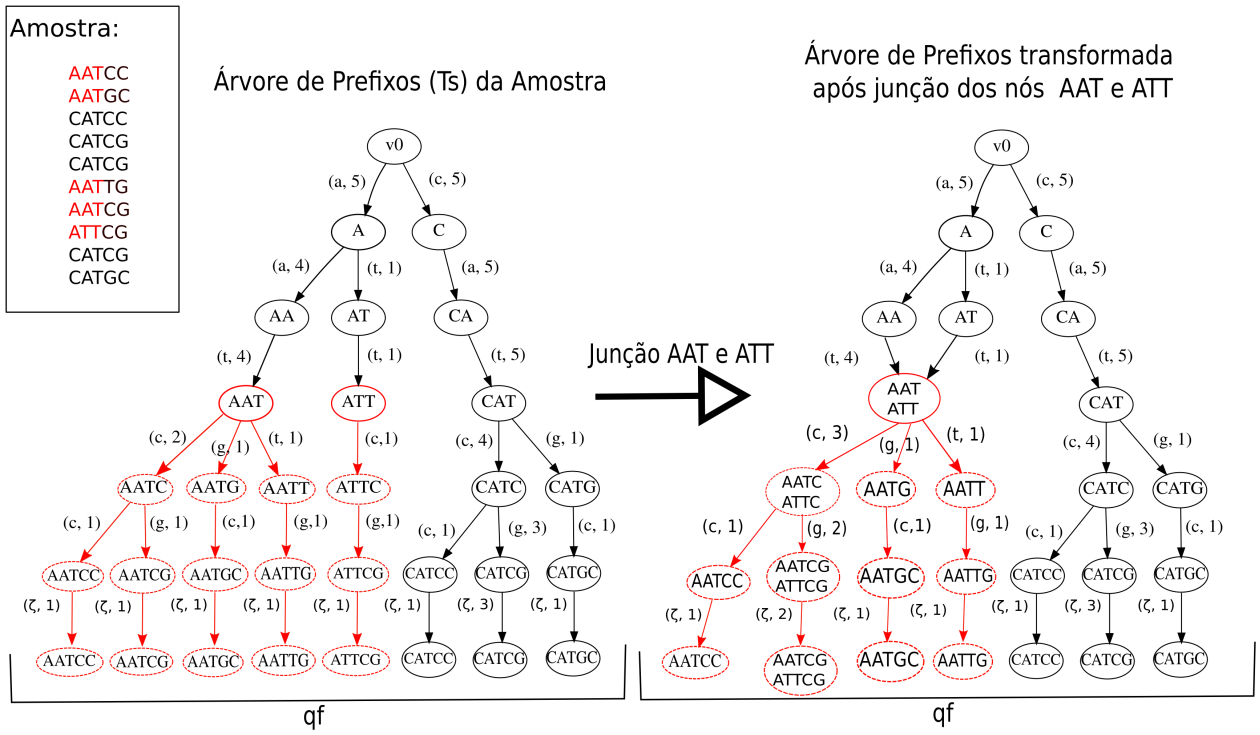
$$P^M(s) \stackrel{def}{=} \prod_{i=0}^{l-1} \gamma(q_i, s_{i+1}) \tag{12}$$

Seja S um conjunto de sequências a ser utilizada como amostra de treinamento. O algoritmo LAPFA inicia construindo uma árvore de prefixos T_S a partir de S e, depois transforma T_S em um AFDEA como definido previamente. A árvore de prefixos é um grafo acíclico com estrutura semelhante ao AFDEA, que recebe todas as sequências de S e constrói um grafo que representa o caminho, símbolo a símbolo, que forma cada sequência da amostra (figura 8, à esquerda). Cada aresta dessa árvore é rotulada por um símbolo $\sigma \in \Sigma \bigcup \zeta$. Cada nó da árvore é rotulada por uma cadeia, que representa um prefixo de alguma sequência contida em S. Uma árvore de prefixos é estruturada por níveis de profundidade. O nível zero de profundidade (ou raiz) contém apenas um nó v_0 que representa a cadeia vazia. Qualquer outro nó v é associada à cadeia prefixo que faz a trajetória começando em v_0 e alcança v. Se existe um nó v no nível i que se conecta no sentido do nó u no nível $i+1$, a aresta $v \rightarrow u$ é rotulada por um símbolo σ, que representa o prefixo v concatenado com σ. A trajetória descrita pode ser denotada como $v \stackrel{\sigma}{\rightarrow} u$ [18]. A profundidade máxima de T_S é dada pela maior sequência contida em S, ou seja, pela maior trajetória a partir de v_0.

Além do rótulo σ, cada aresta contém também uma contagem que representa o número de sequências em S que passam por aquela trajetória. Isto é, a contagem m de um nó v, m_v, representa o número de sequências em S que possuem o prefixo associado ao nó v. De forma semelhante, $m_v(\sigma)$ representa o número de sequências em S que partem de v sob a aresta rotulada por σ. Tem-se então que vale a propriedade : $\sum_{\sigma \in \Sigma} m_v(\sigma) = m_v$. Além disso, $m_v(s)$ representa o número de sequências de S que contêm a subsequência s e que s é gerada por M a partir do nó v.

[18] Pelo conceito de autômato determinístico, uma vez definidos v e σ, u é único

Figura 8 – Árvore de Prefixos e a junção de nós por similaridade. No exemplo à esquerda, é apresentada uma árvore de prefixos para a mesma amostra utilizada na figura 7. Em vermelho, à direita, é representado o processo de junção por similaridade. Mas duas árvores, em vermelho, estão marcados todos os nós que participam do processo de junção correspondente. Observe que em vez de probabilidades, cada aresta mostra a contagem m daquela cadeia. A junção dos nós AAT e ATT é mostrada. Para fins didáticos, considere que esses nós sejam similares, portanto serão juntados. Note que os nós filhos de AAT e ATT também são juntados para aqueles que contenham o mesmo símbolo de transição σ na aresta. A definição de similaridade é recursiva, então se AAT e ATT são similares, os nós filhos também são. Note também que caso um nó seja juntado, as arestas que atingem esse nó são somadas para um mesmo símbolo, por exemplo $m_{AATC/ATTC}(g) = m_{AATC}(g) + m_{ATTC}(g)$.



Fonte: Guilherme Miura Lavezzo (2020)

.

Construída a árvore, o LAPFA verifica a cada iteração se, em um mesmo nível (começando por zero), existem dois pares de nós v e u que são similares para poder uni-los. Dois estados são considerados similares se, para cada cadeia s,

$$|m_v(s)/m_v - m_u(s)/m_u| \leq \mu/2 \tag{13}$$

sendo μ um parâmetro do algoritmo importante para o processo de generalização durante o aprendizado do AFDEA.

Em outras palavras, considere todas as subcadeias s que sejam geradas a partir de v e u até algum estado (não necessariamente o final). Se a frequência relativa de s, a partir de dois nós v e u do mesmo nível de profundidade, diferir por no máximo $\mu/2$, então v e u são similares. Adicionalmente, no algoritmo LAPFA também é checado se $m_v > m_0$ e $m_u > m_0$, em que m_0 é outro parâmetro que define a contagem mínima que os estados u e v precisam possuir para serem candidatos à junção. Caso sejam similares, u e v são juntados como um único nó w, mantendo a relação $m_w(\sigma) = m_v(\sigma) + m_u(\sigma)$. Além disso, todos os nós filhos de v e u são juntados também, ou seja, todos os nós que partem de v e u. No caso, os nós filhos são juntados para um mesmo símbolo de aresta (figura 8). Nota-se que, de acordo com esse critério e a depender dos parâmetros μ e m_0, o modelo pode potencialmente modelar uma PWM, caso as evidências não justifiquem adotar mais parâmetros para diferir a transição de um símbolo em uma dada posição.

Feito esse procedimento até então descrito para todos os níveis de T_S, a árvore é transformada em um AFDEA. Para tal, o nó inicial da árvore passa a ser o estado q_0 do autômato, e todos os nós terminais (ou folhas) da árvore passam a ser o nó único q_f. Para todos os nós u de um mesmo nível d que satisfazem $m_u < m_0$, são juntados em um nó chamado $pequeno(d)$. Nós cujas arestas estão rotuladas com ζ são direcionados para q_f. Se $m_v(\sigma) = 0$ para algum $\sigma \in \Sigma$ e v situado no nível d, então uma aresta é conectada como $v \xrightarrow{\sigma} pequeno(d)$.

Com a estrutura do grafo montada, as probabilidades γ de M são estimadas. Para cada estado q que é associado a um nó u no nível d, e para cada $\sigma \in \Sigma \bigcup \zeta$, define-se:

$$\gamma(q, \sigma) = (m_u(\sigma)/m_u)(1 - (|\Sigma| + 1)\gamma_{min}) + \gamma_{min} \tag{14}$$

em que γ_{min} é o terceiro parâmetro do algoritmo que define a probabilidade mínima de um próximo símbolo em um estado.

Ao estimar as probabilidades γ, a árvore T_s é transformada em um AFDEA (passo 12 do algoritmo 1)

Segue abaixo uma versão simplificada do algoritmo LAPFA em questão, em que $d[i]$ denota o nível i do grafo, ζ representa a cadeia vazia e m_0 é um parâmetro pré-definido para o algoritmo.

Algoritmo 1 Pseudocódigo LAPFA simplificado

1: Inicializa $i \leftarrow 0$, $Ts =$ Árvore de Prefixos montada sobre o conjunto Amostra S, $d[0] \leftarrow 1$, $D \leftarrow$ maior profundidade deT_s
2: **enquanto** $d[i] < D$: **faça**
3: Procure por nós j e j' no nível $d[i]$ que satisfaça
4: **se** $m_j \geq m_0$ e $m_{j'} \geq m_0$ **então**
5: **se** j e j' são similares e, todos os filhos (nós de profundidade maior que a profundidade/nível considerado) de j e j' também são similares (recursivamente) **então**:
6: juntar nós j e j' e todos os filhos similares recursivamente, também faça $i \leftarrow i + 1$, ou seja, olhe a próxima profundidade (maior)
7: **caso contrário**, isto é, nenhum par de nós j e j' é similar, **então** $i \leftarrow i + 1$ (olhe o próximo nível) e retorne à condição de **enquanto**
8: Redirecione todas as arestas finais para um estado final q_f e para $d = 1, ...D - 1$, junte todos os nós j no nível d em questão que satisfazem $m_j < m_0$ em um único estado chamado $pequeno(d)$
9: **para** todo nível d, exceto D e para todo nó j nesse nível e para todo $\sigma \in \Sigma$,
10: **se** $m_j(\sigma) = 0$ então redirecione esta aresta para o estado $pequeno(d)$, isto é $j \xrightarrow{\sigma} pequeno(d)$
11: **se** $m_j(\zeta) = 0$ então redirecione esta aresta para o estado q_f, ou seja $j \xrightarrow{\zeta} q_f$
12: Construa um AFDEA sobre a Árvore Ts resultante

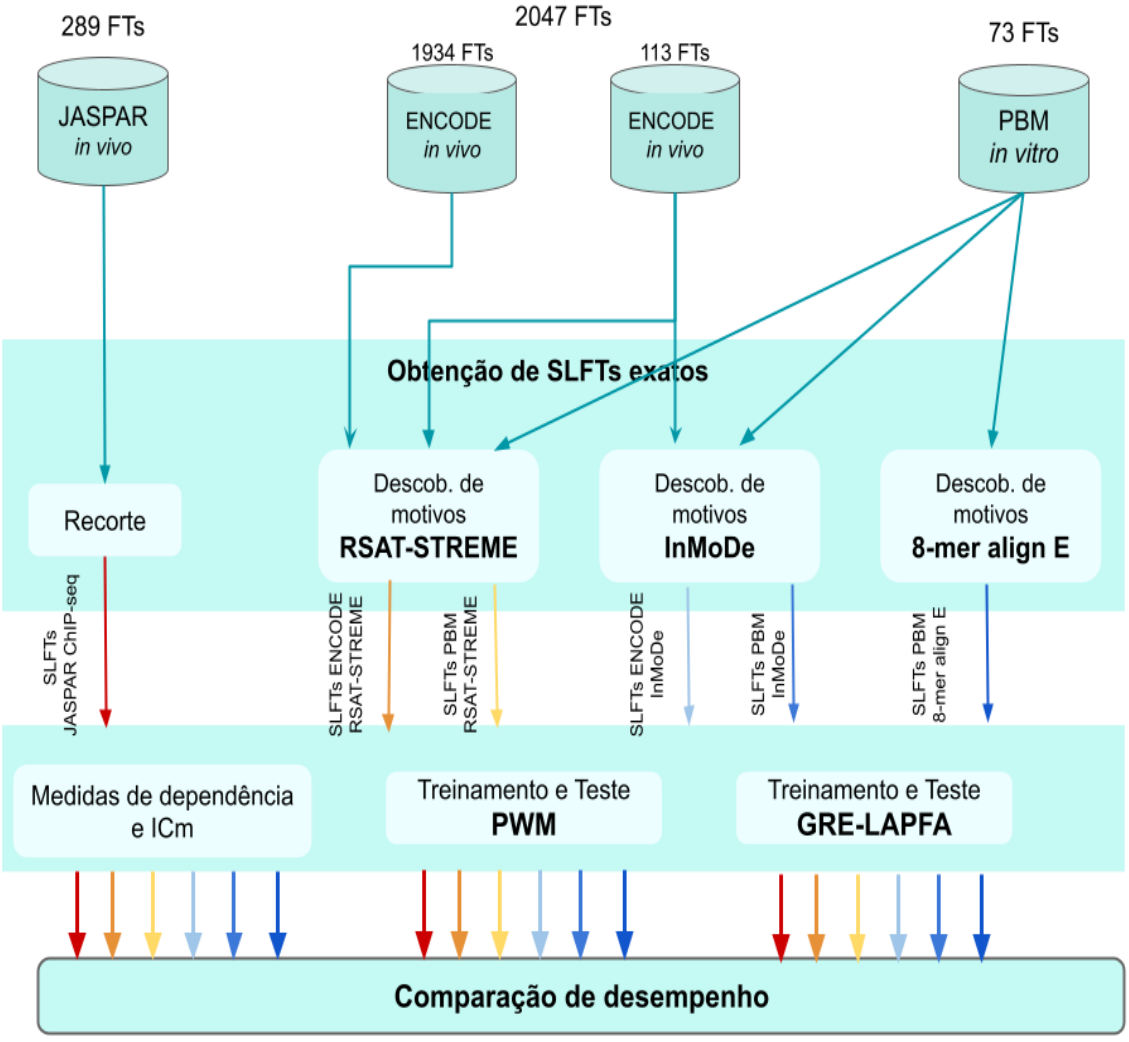
Fonte: Ron, Singer e Tishby (1998), Vieira e Durham (2005), modificado por Guilherme Miura Lavezzo(2020)

3 Materiais e métodos

Este capítulo descreve os materiais e métodos utilizados para atingir o objetivo deste projeto, que como já mencionado anteriormente consiste em comparar o desempenho entre PWMs e GREs como preditores de SLFTs e analisar a relevância em se considerar a dependência entre posições de bases para um conjunto de SLFTs obtidos de diversos métodos.

A figura 9 apresenta um panorama geral dos métodos empregados neste trabalho e descritos nas próximas seções. Inicialmente, foram coletadas sequências de bancos de dados públicos, especificamente do JASPAR, ENCODE e PBM (WEIRAUCH *et al.*, 2013). Esses dados foram processados utilizando três diferentes ferramentas de descoberta de motivos, gerando um total de seis conjuntos de SLFTs de tamanhos exatos, conforme descrito nas seções 3.1, 3.2 e 3.3. Cada um desses seis conjuntos de SLFTs foram analisados quanto a medidas de dependência e Conteúdo de Informação médio (seção 3.4) e utilizados para treinamento e teste dos modelos PWM e GRE (seção 3.5). Por fim, o desempenho desses dois modelos são comparados à luz das características de dependência e Conteúdo de Informação médio, conforme descrito na seção 3.6.

Figura 9 – Panorama geral dos métodos empregados. A figura mostra as etapas, desde a obtenção e processamento de dados públicos de *ChIP-seq* e *PBM* até o processo final de comparar os modelos PWM e GRE-LAPFA. Note que os dados do ENCODE foram separados em dois, uma contendo 1934 FTs, e outra contendo 113 FTs, totalizando 2047 FTs. Esta última parte (113 FTs) foi utilizada como entrada para os algoritmos RSAT-STREME e InMoDe, enquanto a primeira foi utilizada apenas para o RSAT-STREME. Note também que as cores das flechas (vermelho, laranja, amarelo e três tonalidades de azul) representam duas consecutivas etapas, chamadas aqui de "experimentos-algoritmos": a primeira se refere à origem do dado experimental (JASPAR - predominante ChIP-seq, ENCODE - ChIP-seq e PBM) e a segunda ao algoritmo de descoberta de motivos aplicado. Essas cores são replicadas mais abaixo novamente para indicar que o conjunto de SLFTs obtidos por cada um dos seis experimentos-algoritmos são utilizados como entrada tanto para as medidas de dependência como para treinamento e teste dos modelos.



Fonte: Guilherme Miura Lavezzo (2021).

3.1 Banco de dados JASPAR

O banco de dados JASPAR, em sua versão 2020, possui amostras de sítios exatos e não exatos para vários FTs, a depender do tipo de experimento que foi realizado. Em se tratando de dados *in vivo*, para muitos FTs há amostras contendo um número de sítios exatos suficiente para o treinamento e comparação dos modelos preditores. No entanto, para muitos outros FTs, a quantidade é insuficiente: menos de cem sequências por FT. Neste trabalho, foram utilizados dados de SLFTs para 289 FTs distintos, em que todos possuíam pelo menos cem SLFTs disponíveis, distribuídos para cinco organismos: *Arabidopsis thaliana*, *Mus Musculus*, *Homo Sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*. Para os dados do JASPAR, 273 FTs são de experimentos *ChIP-seq*, enquanto 16 são de *ChIP-chip*. Como a grande maioria é de *ChIP-seq*, o JASPAR foi referido como JASPAR ChIP-seq no capítulo de resultados (cap. 4). Além disso, 171 FTs são de *H. sapiens*, 58 FTs são de *Mus musculus*, 25 FTs são de *D. melanogaster*, 18 FTs são de *A. thaliana* e 17 FTs são de *C. elegans*.

Para os 289 dados coletados, algumas sequências são fornecidas como SLFTs exatos, enquanto outras são fornecidas como sequências longas de tamanho cem. Para esses últimos, bases de DNA representadas com letras maiúsculas eram os SLFTs exatos, enquanto letras minúsculas era o restante da sequência. Portanto, foi necessário apenas fazer o recorte das sequências nesse caso.

3.2 Banco de dados ENCODE

Devido à ausência de FTs para alguns organismos e baixo tamanho amostral de SLFTs nos dados do JASPAR, obteve-se também os sítios exatos de dados públicos experimentais *in vivo* — de ChIP-seq — com a finalidade de usá-los como amostra positiva. Dados de ChIP-seq têm sido o padrão ouro para estudar interações diretas entre FT e DNA, sendo produzido dados em larga escala obtidos por sequenciamento (*high throughput sequencing*).

Porém, diferente dos dados do JASPAR que fornecem as SLFTs de cada FT de interesse, isto é, os dados já processados, o ENCODE fornece o dado do experimento de ChIP-seq, com alguns pré-processamentos mas ainda não as sequências exatas. Portanto,

são necessárias etapas intermediárias de processamento destes dados experimentais com a finalidade de obter os SLFTs exatos.Para isso, foram utilizadas duas estratégias de descoberta de motivos: uma baseada em PWM (utilizando as ferramentas RSAT e STREME) e outra não baseada em PWM (utilizando a ferramenta InMoDe).

3.2.1 Processamento de dados ChIP-seq do ENCODE utilizando algoritmo de descoberta de motivos baseado em PWMs (RSAT-STREME)

Partindo de arquivos *narrowpeaks* uma série de etapas devem ser seguidas para atingir o objetivo de obtenção de SLFTs. Para tal, o protocolo do JASPAR 2020 (FORNES *et al.*, 2019) foi utilizado como uma base, sendo adaptado com adição de novas etapas e ferramentas, para a extração de SLFTs a partir de dados de ChIP-seq e utilizando um algoritmo de descoberta de motivos baseados em PWM. As etapas a seguir descrevem o procedimento:

1. obtenção dos dados experimentais de ChIP-seq em arquivos pré-processados do tipo BED *narrowpeak*. A figura 10 exemplifica um arquivo desse formato;
2. obtenção de todos os genomas mascarados[1] que foram originalmente utilizados pelo grupo ENCODE para mapear os dados de ChIP-seq. A saber : hg19, dm6 , mm10, ce11. Os genomas mascarados podem ser adquiridos no repositório da UCSC[2];
3. a partir das coordenadas dos picos de ChIP-seq, extração das sequências de DNA (em arquivos FASTA), usando a ferramenta BedTools (QUINLAN; HALL, 2010). Sequências essas de tamanho 101 e 501 bases e centralizadas em torno do *peak summit* (região central do pico). As sequências de tamanho 501 também são filtradas caso coincidam com sequências *blacklist* [3];
4. execução das ferramentas de descoberta de motivo RSAT e STREME, paralelamente, sobre as sequências de tamanho 101 ;

[1] Tradução para *hard masked genome*, que na prática significa substituir nucleotídeos por "N"onde houver artefatos de montagem e/ou repetições de bases no genoma (KONING *et al.*, 2011)

[2] A UCSC possui as anotações do genoma de diversos organismos. O endereço eletrônico é ¡https://genome.ucsc.edu/cgi-bin/hgGateway¿.

[3] Um estudo recente realizado pelo grupo do ENCODE (AMEMIYA; KUNDAJE; BOYLE, 2019) verificou que existem diversas regiões genômicas em dados *high-throughput NGS* que são potencialmente anomalias ou artefatos derivados da montagem e do mapeamento das anotações ao genoma. Regiões essas denominadas *blacklist* devem ser removidas em qualquer análise de ChIP-seq para evitar reportar falsos (ou distorcidos) resultados das análises posteriores. Tendo isso em mente, posterior à descoberta de motivos de qualquer dado *ChIP-seq*, regiões *blacklist* foram removidas de qualquer arquivo *narrowpeak* e FASTA derivado.

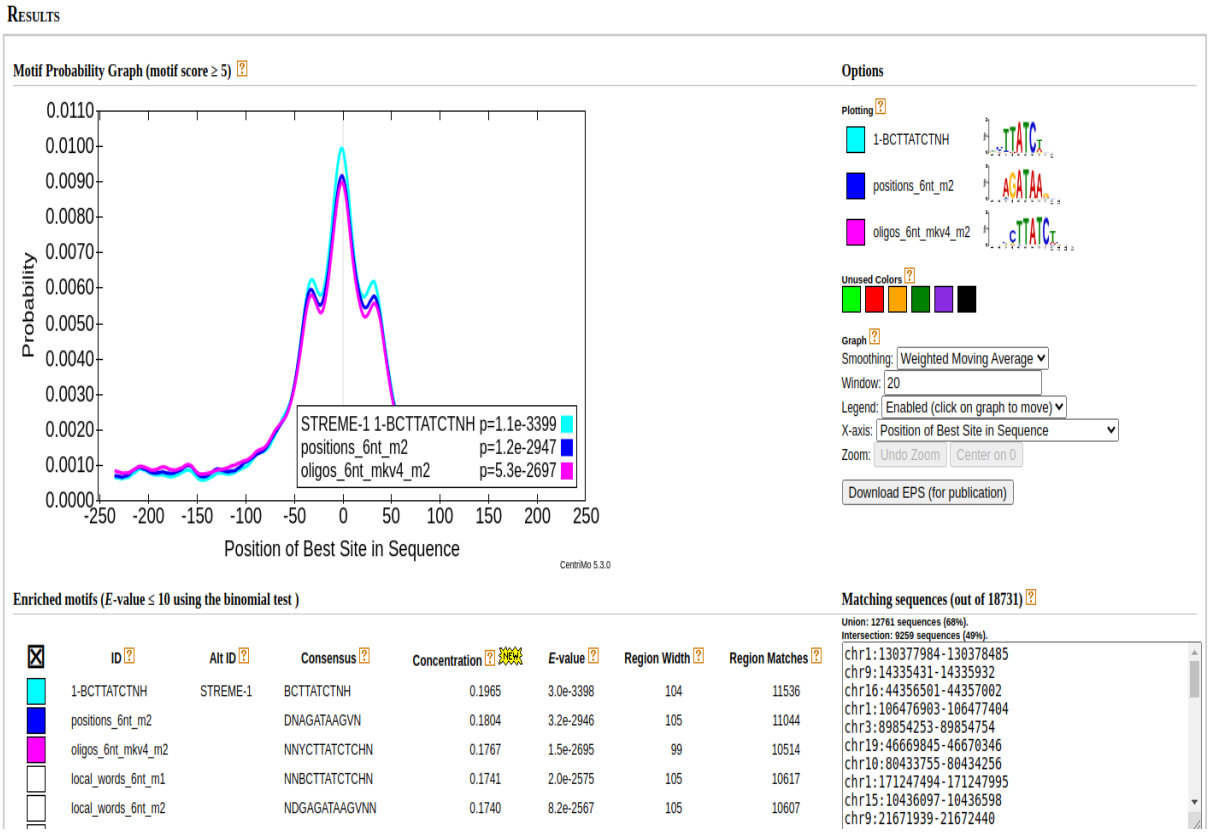
5. escolha do melhor motivo - dentre os diversos obtidos (e distintos) - para cada FT por meio da análise de enriquecimento de motivos (programa CENTRIMO (BAILEY; MACHANICK, 2012)). A figura 11 exemplifica a saída do programa CENTRIMO;
6. Obtenção de SLFTs de tamanho exatos, com uso da ferramenta FIMO (GRANT; BAILEY; NOBLE, 2011), usando como entrada o motivo obtido da etapa anterior e sequências de tamanho 501 bases. Os SLFTs encontrados, são filtrados por :
   - p-valor menor que 0.05 ;
   - o percentil dos 15% maiores escores *log-odd*;
7. todos os SLFTs que satisfizerem ambas as condições acima comporão a amostra de treinamento e teste dos modelos preditores.

Figura 10 – Arquivo formato BED *narrowpeak*. Cada linha representa um pico de ChIP-seq. A primeira coluna representa o nome do cromossomo (de acordo com o genoma utilizado), a segunda coluna delimita a posição de início do pico, a terceira coluna delimita o fim do pico, a penúltima coluna representa a medida de significância "q-valor"em $-log_{10}$, a última coluna delimita o *peaksummit*. Este último é a região mais significativa e central do pico, dado em coordenadas relativas ao início do pico. Exemplo : na primeira linha, o início desse pico é em 223459 e o *peaksummit* é de 58, logo a região central e mais significativa do pico, em coordenadas, é de 223459 + 58. As colunas não descritas não são essenciais para este trabalho, portanto não serão descritas para facilitar o entendimento.

chr2L	223459	223651	.	0	.	176.907510830703	-1	3.97822618167453	58
chr2L	414777	415207	.	0	.	137.538111120067	-1	3.97822618167453	215
chr2L	440901	441331	.	0	.	202.368627483907	-1	3.97822618167453	215
chr2L	442010	442163	.	0	.	164.982888456435	-1	3.97822618167453	92
chr2L	479940	480213	.	0	.	302.034604794466	-1	3.97822618167453	120
chr2L	480913	481275	.	0	.	761.689805483654	-1	3.97822618167453	165
chr2L	485743	485949	.	0	.	187.870136135276	-1	3.97822618167453	101
chr2L	510582	510701	.	0	.	256.992126984346	-1	3.97822618167453	112
chr2L	520710	520970	.	0	.	320.896881684551	-1	3.97822618167453	128

Fonte: Guilherme Miura Lavezzo (2021).

Figura 11 – Arquivo de saída do CENTRIMO para um FT. O gráfico mostra a relação média dos três motivos mais significativos obtidos pelas ferramentas STREME e RSAT com o enriquecimento de SLFTs em torno dos *peak summits* do arquivo FASTA centrado em 501 bases. O critério para significância de um motivo é dado pelo p-value de enriquecimento central (representado indiretamente pela medida *E-value* abaixo do gráfico). Observe que a intensidade do sinal é de fato maior para a porção central dada como coordenada relativa ao centro do pico. Note também que os três motivos exibidos no gráfico esquerdo são os mesmos, com algumas alterações na probabilidade das bases. Porém, motivos distintos também são encontrados, mas não enriquecidos centralmente.



ID	Alt ID	Consensus	Concentration	E-value	Region Width	Region Matches
1-BCTTATCTNH	STREME-1	BCTTATCTNH	0.1965	3.0e-3398	104	11536
positions_6nt_m2		DNAGATAAGVN	0.1804	3.2e-2946	105	11044
oligos_6nt_mkv4_m2		NNYCTTATCTCHN	0.1767	1.5e-2695	99	10514
local_words_6nt_m1		NNBCTTATCTCHN	0.1741	2.0e-2575	105	10617
local_words_6nt_m2		NDGAGATAAGVNN	0.1740	8.2e-2567	105	10607

Fonte: Guilherme Miura Lavezzo (2021).

3.2.2 Processamento de dados ChIP-seq utilizando algoritmo de descoberta de motivos não baseado em PWMs (InMoDe)

Os dados utilizados nesta etapa foram de 113 FTs do ENCODE, distribuídos para os mesmos quatro organismos utilizados anteriormente. Em relação à subseção anterior (3.2.1), há uma diferença nas últimas etapas desse algoritmo, simplificando o processo:

1. utilização dos genomas mascarados e dos 113 BED *narrowpeaks* do ENCODE selecionados[4];
2. conversão dos arquivos *narrowpeaks* em arquivos FASTA de tamanho 101 bases centralizados no *peak summit*, com o uso da ferramenta BedTools (QUINLAN; HALL, 2010), ;
3. obtenção de um único motivo [5] usando sequências de tamanaho 101 bases e a ferramenta de descoberta de motivos InMoDe (EGGELING, 2018), ;
4. diferente da ferramenta RSAT e STREME, o InMoDe já retorna todos os SLFTs que formaram a sequência motivo obtida. Portanto, todos esses SLFTs serviram como amostra positiva.

3.3 Dados de PBM públicos

Os dados de PBM foram coletados de (WEIRAUCH *et al.*, 2013). Neste artigo, foram testados diversos algoritmos de descoberta de motivos com o objetivo de treinar e testar modelos preditores do sinal de fluorescência emitido (vide detalhes da técnica PBM na seção 2.1.2). Para cada FT, havia duas réplicas do experimento. No artigo, os autores treinavam os modelos com uma réplica e tentavam predizer sobre a outra réplica. No caso deste trabalho, utiliza-se a informação de ambas as réplicas sem distinção. Como o propósito do artigo diverge do aqui proposto, adquiriu-se os dados brutos de PBM e estes foram processados para se adaptarem a aplicação de descoberta de motivos seguido da aquisição de SLFTs para treinamento dos modelos PWM e LAPFA. Em particular, dois algoritmos do artigo foram utilizados como base: team E e 8-mer align E.

3.3.1 Processamento dos dados de PBM a partir do algoritmo de descoberta de motivo baseado em PWM (RSAT-STREME)

Esse algoritmo, teve como base inicial o algoritmo team E (WEIRAUCH *et al.*, 2013), que foi adaptado para ser utilizado com as duas réplicas de cada FT fornecidas e

[4] A escolha pelos 113 FTs em vez de todos os 2047 utilizado na etapa do RSAT-STREME foi principalmente pelo custo mais elevado computacional da ferramenta InMoDe.

[5] Diferente das ferramentas RSAT e STREME, o InMoDe demanda muito mais tempo de processamento e necessita ajustar mais parâmetros relacionados à iteração do algoritmo. Por isso, a ferramenta retorna apenas um único motivo, exigindo que múltiplas iterações (dez, no caso) sejam realizadas para um resultado confiável.

sempre que mencionado neste trabalho serão utilizados os nomes RSAT/STREME em vez de team E adaptado, para facilitar o entendimento. Segue as etapas enumeradas:

1. coleta das 1000 sequências (de tamanho 35) cujo sinal de fluorescência seja o mais intenso, para cada réplica . Essas 2000 sequências serão utilizadas ao longo das próximos etapas;
2. aplicação das ferramentas RSAT e STREME simultaneamente, sobre essas sequências, ;
3. aplicação da ferramenta CENTRIMO e obtenção do motivo com melhor medida de enriquecimento (menor E-valor) ;
4. utilização da ferramenta FIMO e coleta de SLFTs, com o melhor motivo obtido;
5. uso como amostra positiva apenas o percentil dos 15% SLFTs cujo p-valor é menor que 0.05 e com maiores escores *log-odd*.

3.3.2 Processamento dos dados de PBM a partir do algoritmo de descoberta de motivo não baseado em PWM (InMoDe)

A seguir, está descrita a obtenção de SLFTs a partir de dados de PBM, fazendo uso adaptado do algoritmo *team E* (WEIRAUCH *et al.*, 2013), mas com substituição das ferramentas RSAT e STREME pela InMoDe.

1. coleta das 1000 sequências (de tamanho 35) cujo sinal de fluorescência seja o mais intenso, para cada réplica. Essas 2000 sequências serão utilizadas ao longo das próximos etapas;
2. aplicação da ferramenta InMoDe com 10 iterações, sobre essas sequências;
3. recuperação de todos os SLFTs que contruíram a sequência motivo (como já mencionado para dados de ChIP-seq aplicando o InMoDe), com o único motivo descoberto .

3.3.3 Processamento dos dados de PBM a partir do algoritmo 8-mer align E

O algoritmo 8-mer align E (WEIRAUCH *et al.*, 2013) foi motivado por este trabalho como uma maneira de extrair SLFTs sem a etapa intermediária de utilizar uma ferramenta bioinformática de descoberta de motivos. O princípio usufrui da ideia que todas as possíveis

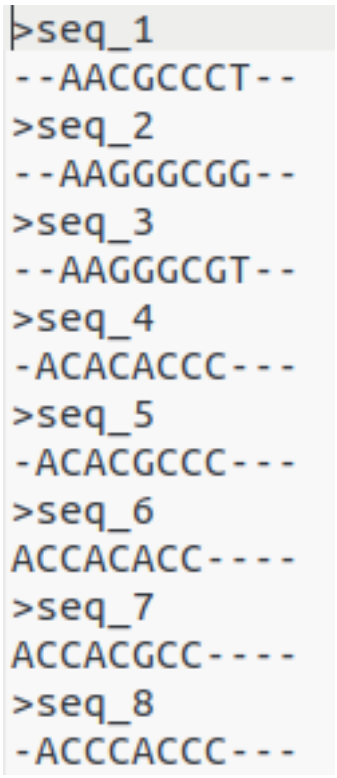
combinações de bases de tamanho 10 (10-mers[6]) estão presentes em um experimento de PBM. Assim, sequências de tamanho 8 (8-mer) também seriam cobertas e adicionalmente estariam em repetições, estas que são essenciais para o cálculo da medida *E-score*, de onde deriva o nome "E" do algoritmo. Os autores do artigo já forneciam todas as sequências 8-mers e a mediana do sinal de suas repetições, separadas pelas duas réplicas. Além da intensidade de sinal, cada sequência 8-mer possui um E-score[7], associado à relevância da interação entre o 8-mer e o FT. A seguir uma descrição do algoritmo o suficiente do que seja necessário para o entendimento do que foi aplicado neste trabalho:

1. obtenção da média de E-score entre as duas réplicas, para cada sequência 8-mer;
2. aquisição apenas das sequências 8-mers com E-score acima de 0.45;
3. alinhamento das sequências 8-mers, usando a ferramenta de alinhamento de sequências Clustal Omega (SIEVERS *et al.*, 2011), ;
4. com o alinhamento formado, existirão posições entre as sequências alinhadas que estarão vazias (figura 12). Se em uma determinada posição do alinhamento, metade das sequências não tiverem uma base preenchida, então essa posição é removida do alinhamento;
5. para as posições restante mas com alguns vazios (em menos da metade das sequências) nas 8-mers alinhadas, obtenção da frequência relativa de cada base naquela posição;
6. gerar E sequências repetidas para cada 8-mer, em que $E = \text{E-valor} \times 100$;
7. preenchimento do vazio do alinhamento (caso haja), com uma base aleatória da distribuição de frequências relativas daquela posição obtida na etapa 5. Desse modo, toda sequência alinhada terá todas as posições preenchidas com algum nucleotídeo.

O alinhamento das sequências 8-mers é necessário pois, não necessariamente a união de todos os 8-mers resultaria na sequência motivo do FT. Por exemplo, o FT poderia não ter uma sequência motivo de tamanho 8. Além disso, alguns FTs possuem poucas variações de sequências reconhecidas, e o 8-mer pode ser apenas a representação parcial desta interação conservada. O alinhamento é uma maneira de garantir que as sequências mais representativas possuem algum consenso entre si antes de comporem a sequência motivo[8].

[6] Termo derivado de *k-mer*, que quer dizer uma sequência de tamanho arbitrário k.
[7] Esse E-score mencionado é distinto do E-valor calculado pela ferramenta CENTRIMO.
[8] Para ferramentas de descoberta de motivos, o alinhamento é feito automaticamente.

Figura 12 – Exemplo de 8-mers alinhados com programa Clustal Omega no formato de arquivo FASTA. No exemplo, 8 sequências 8-mers foram alinhadas. O programa Clustal Omega tenta reposicionar cada 8-mer de maneira a melhor combinar as sequências dada suas composições de bases. Nota-se que existem posições que não possuem uma base correspondente em algumas sequências 8-mers alinhadas, representado por um "−" ("vazio"). Como apenas sequências SLFTs inteiras são desejadas neste trabalho, estes hífens são removidos caso estejam presentes em mais da metade dos 8-mers alinhados assim. Os restantes, são preenchidos com uma distribuição de bases relativa à frequência de bases de cada posição alinhada.



Fonte: Guilherme Miura Lavezzo (2020).

3.4 Medição de dependência entre posições de bases de um motivo e Conteúdo de Informação médio

O segundo objetivo específico consiste em propor uma ou mais estratégias para medir o nível de dependência entre posições de bases em uma dada amostra. Foram testados quatro critérios apresentados nesta subseção, que podem ser complementares.

Para isso, serão apresentadas algumas definições que servirão para todas as medidas aqui calculadas nesta seção:

No caso, considere que variável aleatória seja a frequência de bases de DNA para uma fixada posição da sequência motivo. Considere uma amostra positiva como sendo

um conjunto de n sequências de tamanho k. Cada nucleotídeo ocupado por uma das k posições tem um índice de 1 a k, distinto, que indica sua posição na sequência. Sejam então i e j posições fixadas.

Seja B uma variável aleatória que assume valores em $\{A, C, G, T\}$. Denotamos um valor assumido por B como b. Seja $P(b_i)$ a probabilidade da variável aleatória B assumir o valor b na posição fixada i.

A probabilidade $P(b_i)$ é estimada por Máxima Verossimilhança (TOMOVIC; OAKELEY, 2007) pela equação:

$$P(b_i) = \frac{N(b_i)}{n} \tag{15}$$

em que: $N(b_i)$ é a frequência absoluta de b na posição i, n é o número de sequências na amostra.

A probabilidade conjunta de duas bases ocorrerem em posições fixas i e j é definida como:

$$P(b_i, b_j) = \frac{N(b_i, b_j)}{n} \tag{16}$$

em que N é a frequência absoluta das duas bases b_i e b_j ocorrendo simultaneamente nas posições i e j, respectivamente.

A probabilidade conjunta para todas as combinações possíveis pode ser calculada a partir de uma tabela de contingência. A tabela de contingência mostra os valores (no caso, bases) que as variáveis aleatórias assumem e suas contagens obtidos empiricamente a partir da amostra considerada (conjunto de sequências). Nas linhas estão representados os possíveis valores da variável (B_i) e nas colunas os possíveis valores da outra (B_j), de tal forma que, em uma linha ou coluna fixada está a contagem total da ocorrência de um valor assumido pela variável aleatória fixada (ou as marginais). Portanto, em cada célula estão as contagens da intersecção (ou da contagem conjunta) dos valores assumidos por cada variável aleatória, e a soma de todas as células representa o tamanho da amostra considerada (n).

As medidas definidas nesta seção, utilizam uma tabela de contingência montada para o conjunto de SLFTs e da contagem de cada base em um dado par de posições desse conjunto de sequências. Por meio disso, é possível calcular todas as probabilidades empiricamente.

Notou-se que, para as medidas de dependência descritas a seguir, tanto a medida de Cramer V (subseção 3.4.1) quanto a medida de Theil U simétrico (subseção 3.4.2) possuem um denominador associado à probabilidade de ocorrência de uma base ($P(b_i)$), dada uma posição i. Caso essa probabilidade seja 0 para alguma base, as medidas podem se tornar indefinidas. Para corrigir tal questão, um pseudo-contador é adicionado para todas as contagens de nucleotídeos na tabela de contingência, no valor de $1/n$, com n já definido na equação 15.

3.4.1 Medida de associação Cramer V

A medida de associação de Cramer V, derivada da estatística de qui-quadrado (χ^2), indica a intensidade de associação entre duas variáveis aleatórias, variando entre zero e um. Zero significa nenhuma associação e um significa total correlação entre o par de variáveis aleatórias.

A estatística de qui-quadrado para cada par não ordenado de posições (i, j) é denotado por χ^2_{ij} e pode ser definida como :

$$\chi^2_{ij} = n \sum_{b_i, b_j} \frac{(P(b_i, b_j) - P(b_i)P(b_j))^2}{P(b_i)P(b_j)} \tag{17}$$

Conforme o n aumenta, maior o valor da estatística sem necessariamente estar atrelada alguma interpretação da intensidade de associação entre duas variáveis aleatórias (no caso entre duas posições da sequência motivo). A medida de Cramer V tenta corrigir esse efeito (KIM, 2017; SHARPE, 2015). Define-se a medida de Cramer V, para um par de posições (i, j) como:

$$V_{ij} = \sqrt{\frac{\chi^2_{ij}}{n \times min(c-1, l-1)}} \tag{18}$$

sendo χ^2_{ij} a estatística de qui-quadrado definida anteriormente, n o tamanho amostral, c o número de colunas da tabela de contingência e l o número de linhas da tabela de contingência. Observe que n é denominador do Cramer V, o que explica em partes a correção para o efeito de tamanho amostral.

Uma vez que para cada par não ordenado (i, j) do conjunto de SLFTs estão definidos uma medida de Cramer V, é desejável extrair uma estatística que resuma o conjunto de Cramer V calculados. Para tal, considerando todas as combinações (i, j) o máximo e a

média dos Cramer V calculados foram obtidos, chamados neste trabalho de Cramer V máximo e Cramer V médio, respectivamente.

3.4.2 Coeficiente de Incerteza simétrico

Outra medida de dependência a ser avaliada é a medida Theil U (ou Coeficiente de Incerteza simétrico), que pode ser visto como uma normalização da medida de Informação Mútua. O Coeficiente de Incerteza simétrico é calculado entre duas posições de uma sequência (assim como a medida anterior, vide equação 18). Antes de definir o Theil U, é necessário definir a medida de Informação Mútua.

Podemos definir a Informação Mútua $I(B_i, B_j)$ entre duas variáveis aleatórias B_i e B_j como:

$$I(B_i, B_j) = \sum_{a \in B_i} \sum_{b \in B_j} P(B_i = a, B_j = b) \log \left(\frac{P(B_i = a, B_j = b)}{P(B_i = a)P(B_j = b)} \right) \tag{19}$$

em que a e b são valores assumidos pela variável aleatória B nas posições i e j. $P(B_i = a, B_j = b)$ é a probabilidade conjunta, $P(B_i = a)$ e $P(B_j = b)$ são as probabilidades marginais.

A informação mútua $I(B_i, B_j)$ pode variar entre 0 e 2 bits de informação quando ambas as variáveis aleatórias representam as quatro bases com frequência maior que 0. Atinge o valor máximo (2 bits) quando as bases B_i e B_j estão perfeitamente correlacionadas. Quando as bases não possuem nenhuma relação de dependência, I assume valor zero.

Entretanto, não é incomum observar nas amostras que $P(B_i = a) = 0$ para algum $a \in \{A, C, T, G\}$ e para alguma posição i. Nesse caso, a medida máxima de $I(B_i, B_j)$ é menor que 2 bits. Novamente, como é desejável que as medidas de dependência entre posição de bases sejam comparáveis entre si, seria importante que todas as medidas tivessem a mesma interpretação (de associação) num mesmo intervalo de valores.

Para corrigir esse problema, foi proposto que em vez de usar a Informação Mútua, deveria se utilizar uma normalização da medida: o Coeficiente de Incerteza simétrico, ou Theil U simétrico (PRESS, 2007). A medida tem alcance sempre variando no intervalo de

0 a 1, sendo 0 indicando a não associação e 1 o valor máximo de associação. O Theil U é definido como:

$$U(B_i, B_j) = 2\left(\frac{I(B_i, B_j)}{H(B_i) + H(B_j)}\right) \tag{20}$$

sendo $I(B_i, B_j)$ a Informação Mútua definida anteriormente e, com $H(B_i)$, a entropia de B_i, definida como:

$$H(B_i) = -\sum_{B_i \in A,C,T,G} P(B_i) \log P(B_i) \tag{21}$$

Uma vez que para cada par não ordenado (i, j) do conjunto de SLFTs estão definidos uma medida de Theil U (como no Cramer V), é desejável extrair uma estatística que resuma o conjunto de Theil U calculados. Para tal, considerando todas as combinações (i, j) foram obtidos o máximo e a média dos Theil U calculados, chamados neste trabalho de Theil U máximo e Theil U médio, respectivamente.

3.4.3 Conteúdo de Informação médio

O conteúdo de informação é uma medida extraída diretamente dos pesos de uma PWM , PWM essa criada a partir do conjunto de sequências e considerando um determinado modelo nulo. Apesar de não ser uma medida de dependência, essa medida informa a preferência por bases em determinadas posições em relação ao modelo nulo utilizado. Em outras palavras é uma medida de qualidade da PWM (XIA, 2012).

Se o Conteúdo de Informação é utilizado para avaliar a qualidade de uma PWM, esta medida também pode ser informativa para avaliar se um modelo alternativo à PWM possa ser preferido para o conjunto de SLFTs considerados. Espera-se que modelos mais complexos e não baseados em PWM capturem a dependência entre posições de bases, e essa informação não é capturada pelo Conteúdo de Informação. Assim, caso haja dependência entre posições de bases, o Conteúdo de Informação pode diferir de quando uma PWM consegue modelar corretamente o conjunto de SLFTs, no caso de independência entre posições de bases.

Seja k o tamanho da sequência motivo e da PWM formada. Pode-se definir o conteúdo de informação (IC) como (HERTZ; STORMO, 1999):

$$IC = \sum_{b \in B} \sum_{i \in \{1,2,..,k\}} p(b,i) \times W_{b,i} \tag{22}$$

sendo $W_{b,i}$ definido na equação 6, o peso específico de uma base b na posição i de uma PWM e $p(b,i)$ definido na equação 7

O IC médio (ICm) pode ser definido como a média de IC sobre as posições da sequência motivo:

$$ICm = \frac{IC}{k} \tag{23}$$

3.5 Treinamento e estimação de desempenho modelos preditores baseados em PWM e GRE

O terceiro objetivo específico consiste em comparar os desempenhos de PWMs e GREs como modelos de predição de SLFTs. Enquanto esta seção descreve o treinamento e a estimação de desempenho de cada um desses modelos, a seção 3.6 descreve como seus desempenhos foram comparados à luz das medidas de dependência e outras características dos conjuntos de SLFT.

Para o treinamento dos modelos são necessários uma amostra positiva de sítios de ligação de um determinado FT e a definição de um modelo nulo. A obtenção das amostras positivas foi descrita nas seções 3.1, 3.2, 3.3. Para cada FT foi utilizado como modelo nulo a distribuição genômica da espécie à qual pertence o FT (vide seção 2.2.3). Já para a estimação de desempenho é necessária, além de uma amostra positiva de teste, uma amostra negativa cuja criação é descrita na seção 3.5.1. A estimação de desempenho realizada é baseada em vários passos de treinamento e teste utilizando a estratégia de validação cruzada 7x descrita genericamente na seção 2.2.2 e detalhada na seção 3.5.2.

3.5.1 Amostra negativa S^-

Para formar o conjunto de sequências não-SLFTs, chamada mais à frente de amostra negativa S^-, foi proposto gerar uma versão embaralhada de cada genoma utilizado. A

composição de bases de um genoma é heterogênea: existem regiões altamente repetitivas e com baixa variabilidade de nucleotídeos (KONING *et al.*, 2011). A heterogeneidade do genoma faz com que possam existir ambientes em que um FT esteja inserido que se distinguem de sequências puramente aleatórias. Dito isso, sequências não-SLFTs geradas de maneira puramente aleatória não refletem esses ambientes, isto é, este padrão de composição de bases, que circundam SLFTs. Para gerar a amostra negativa respeitando a composição de bases do genoma, optou-se por :

1. para cada espécie do FT testado: obter, do genoma dessa espécie, 15152 fragmentos de DNA de tamanho mil[9], escolhidos ao acaso;
2. cada fragmento é embaralhado de maneira pseudoaleatória (usando um $seed = 11$) para ser um experimento controlado e reproduzível;
3. Para cada conjunto de fragmentos de DNA, recortar sequências não-SLFTs de tamanhos equivalentes aos SLFTs da amostra positiva. O recorte é feito entre sequências disjuntas, sem sobreposição de bases entre cada recorte, até que de todo fragmento tenha sido extraído o maior número de recortes de não-SLFTs de tamanhos exatos;
4. se uma amostra positiva consiste de SLFTs de tamanho k, então a amostra negativa consiste das sequências obtidas na etapa anterior, de tamanho k;
5. se a amostra positiva contém x instâncias de SLFTs, então a amostra S^- usada na validação cruzada conterá $\min(500000, 100x)$ instâncias de sequências. Essa escolha delimita o número máximo de sequências em um conjunto amostral, evitando disparidades entre alocação de memória e outros recursos computacionais, principalmente na etapa de validação cruzada k-folds (subseção 3.5.2), em que múltiplas iterações são realizadas sobre esse mesmo conjunto.

Nota-se, por convenção em aprendizado de máquina, a representação de uma amostra negativa S^- que contrasta com a amostra positiva S^+. Contudo, a S^-, no contexto deste trabalho, se refere a sequências aleatórias, de acordo com alguma distribuição previamente escolhida. Tal distribuição define o modelo nulo, que corresponde à hipótese nula (vide seção 2.2.3).

[9] O número 15152 fragmentos de DNA de tamanho mil (cada) foi utilizado pois, observou-se que para SLFTs de tamanho 30, ao utilizar pelo menos esse número de fragmentos, seria possível obter pelo menos 500000 não-SLFTs para compor a amostra negativa. O problema de trabalhar com o genoma inteiro é o aumento drástico de memória e tempo de processamento computacional necessário. Limitando o número de sequências utilizadas, o problema pode ser circundado.

3.5.2 Validação Cruzada

A Validação Cruzada foi dividida em duas etapas: a primeira etapa consiste em calibrar os parâmetros do algoritmo LAPFA de treinamento das GREs e obter o melhor limiar de classificação para os modelos (PWM, GRE) com base na maximização da medida AP (*average precision*); a segunda etapa consiste em estimar o desempenho dos modelos calibrados usando os limiares da etapa anterior. Destaca-se que duas Validações Cruzadas foram executadas, porém, mantendo intacta as amostras positiva (S^+) e negativa (S^-) e as mesmas divisões dos folds. Destaca-se também que na primeira Validação Cruzada o fold de teste não é utilizado em nenhuma iteração, apenas na segunda Validação Cruzada.

A figura 13, exemplifica como ocorreu a iteração em cada um dos sete folds da validação cruzada na etapa de calibração:

Figura 13 – Validação Cruzada em sete folds para a etapa de calibração. A figura representa cada uma das sete iterações (linhas) e a respectiva divisão dos folds(colunas).

Teste	Calibração	treino	treino	treino	treino	treino
treino	Teste	Calibração	treino	treino	treino	treino
treino	treino	Teste	Calibração	treino	treino	treino
treino	treino	treino	Teste	Calibração	treino	treino
treino	treino	treino	treino	Teste	Calibração	treino
treino	treino	treino	treino	treino	Teste	Calibração
Calibração	treino	treino	treino	treino	treino	Teste

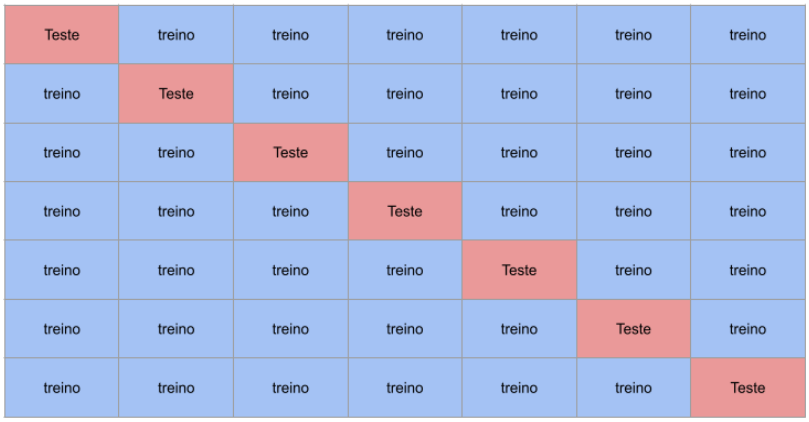
Fonte: Guilherme Miura Lavezzo (2021)

.

A figura 14, exemplifica a etapa de teste da calidação cruzada, note que a mesma divisão dos folds é utilizada:

A primeira Validação Cruzada (de calibração) é esquematizada pelo algoritmo 2. Por meio dessa execução é possível obter a melhor combinação de parâmetros dos modelos e o melhor limiar de classificação do mesmo. Para cada modelo, escolhe-se a combinação de parâmetros que possua o maior valor de AP, se comparado com as demais

Figura 14 – Validação Cruzada em sete folds para a etapa de teste. A figura representa cada uma das sete iterações (linhas) e a respectiva divisão dos folds(colunas).



Fonte: Guilherme Miura Lavezzo (2021)

.

combinações. Esse modelo será utilizado no algoritmo da segunda etapa de Validação Cruzada esquematizada no algoritmo 3.

Algoritmo 2 Pseudocódigo da etapa de calibração da Validação Cruzada *k-fold* e obtenção do limiar ótimo de classificação

procedure PRIMEIRA ETAPA DA VALIDAÇÃO CRUZADA(k, S^+, S^-)
    dividir S^+ em k conjuntos disjuntos ($fold^+$) de tamanhos iguais (ou o mais próximo disso)
    dividir S^- em k conjuntos disjuntos ($fold^-$) de tamanhos iguais (ou o mais próximo disso)
    para cada i, com $i = 1, 2, .., k$
        defina os *folds* de teste $fold^+i$ e $fold^-i$
        defina $fold^+(i+1)$ e $fold^-(i+1)$ comos os folds de calibração
        defina o fold de treinamento como $\bigcup_{j \in \{1,2..,k | j \notin \{i,i+1\}\}} fold^+j$
        para cada modelo (PWM, GRE-LAPFA):
            se o modelo for GRE-LAPFA, então **para cada** combinação c de parâmetros[10]:
                treinar os modelos com o $fold^+$ de treinamento
                aplicar os modelos treinados sobre os $fold^+(i+1)$ e $fold^-(i+1)$
                calcular a curva PR dos modelos e suas respectivas AP (*average precision*) com bases nos resultados da linha anterior
                encontrar o limiar de classificação ótimo l^* que maximize o f1-score
    calcular a AP média para cada modelo treinado e cada conjunto de parâmetros (sobre os valores calculados na linha 12) e, para cada modelo, escolher o conjunto de parâmetros que maximizou a AP média
    Definir l* como sendo: a média dos limiares ótimos l* de cada fold.

Fonte: Guilherme Miura Lavezzo, 2021.

Algoritmo 3 Pseudocódigo da etapa de avaliação da performance dos modelos usando Validação Cruzada *k-fold*

1: **procedure** SEGUNDA ETAPA DA VALIDAÇÃO CRUZADA(k, S^+, S^-)
2: Utilizando a mesma amostra S^+ e S^-, com a mesma divisão $fold^+$ e $fold^-$,
3: **para cada** i, com $i = 1, .., k$ da validação cruzada
4: defina $fold^+i$ e $fold^-i$ como fold de teste
5: defina o fold de treinamento como $\bigcup_{j \neq i} fold^+j$. (Ou seja, adicione o $fold^+i + 1$ de calibração ao $fold^+$ de treinamento)
6: **para cada** modelo com os respectivos parâmetros escolhidos na primeira etapa da Validação Cruzada (PWM, LAPFA):
7: treinar os modelos com o $fold^+$ de treinamento
8: testar os modelos treinados sobre os $fold^+i$ e $fold^-i$
9: calcular a curva PR dos modelos e suas respectivas AP (*average precision*) com bases nos resultados da linha anterior
10: com base no limiar de classificação ótimo l^* obtido na primeira etapa de Validação Cruzada, calcular a acurácia e outras medidas de desempenho
11: calcular a AP média para cada modelo treinado (sobre os valores calculados na linha 9) assim como os valores médios da acurácia e demais medidas de performance (sobre os valores calculados na linha 10)

Fonte: Guilherme Miura Lavezzo, 2021.

3.6 Comparação entre modelos e estudo da dependência entre posições de bases

A comparação de desempenho dos modelos PWM e GRE aprendida pelo LAPFA foi realizada utilizando a medida *Average Precision* (AP) média dos sete folds da validação cruzada (vide seções 2.2.1 e 3.5.2). Para cada conjunto de SLFT testado, notou-se que havia dois casos: ou os modelos tinham desempenhos semelhantes, ou o desempenho (AP) do modelo baseado em GRE era significativamente maior do que o do modelo baseado em PWM. Desse modo, foi utilizada a medida D de Cohen (seção 2.2.4) para diferenciar as duas categorias de comparação encontradas.

Considere $\overline{AP}_{PWM}$ e $\overline{AP}_{GRE}$, os AP médios do modelo PWM e GRE respectivamente, obtidos após a validação cruzada aplicada sobre uma mesma amostra de treinamento e teste. Então, o D de Cohen (mais detalhes na seção 2.2.4) entre modelos, é definido como:

$$D = \frac{\overline{AP}_{PWM} - \overline{AP}_{GRE}}{s} \tag{24}$$

sendo s, o desvio padrão entre os modelos, como já definido na equação 5.

Com $D >= 0.4$, observou-se que o LAPFA obtinha pelo menos 5% a mais de desempenho em relação à PWM, e com $D < 4$ justamente não havia imposição de nenhum modelo.

Com essas duas categorias distinguidas, $D >= 0.4$ ou $D < 0.4$, foi possível realizar um estudo com medidas de dependência e Conteúdo de Informação médio com o objetivo de conseguir obter regras para a preferência de um modelo ou outro. Essas medidas foram calculadas para todos os conjuntos de SLFTs dos seis grupos distintos, que são os grupos da combinação de base de dados e algoritmo de descoberta de motivos: JASPAR, ENCODE RSAT/STREME, ENCODE InMoDe, PBM RSAT/STREME, PBM InMoDe e PBM 8-mer align E. Uma vez feito isso, foi possível comparar as medidas obtidas por grupo, a fim de verificar se existem variabilidades para dados advindos de experimentos e/ou algoritmos de descoberta de motivos distintos. Foram analisadas cinco medidas: Cramer V médio, Cramer V máximo, Theil U (simétrico) médio, Theil U (simétrico) máximo e Conteúdo de Informação médio. As medidas de dependência (Cramer V e Theil U) são medidas definidas para pares de posição da sequência motivo. Por isso, média e máximo se referem às estatísticas obtidas depois de calculadas todas as possíveis combinações de pares numa sequência motivo (seções 3.4.1 e 3.4.2).

As comparações foram realizadas em três níveis: via *Violin Plot*, *effect size* da diferença entre medidas médias para cada par de grupos (experimento-algoritmo) e um teste de hipóteses de Wilcoxon. Essas últimas duas medidas são uma maneira mais assertiva de aferir a diferença de cada característica para cada experimento-algoritmo.

O *Violin Plot* é um tipo de gráfico que, assim como o *Box Plot*, mede a distribuição ordenada dos valores de algum dado. O traço interno mostra um *Box Plot* compacto. O círculo branco mostra a mediana. Os contornos que lembram um violino, mostram uma distribuição contínua aproximada de cada valor contido no dado. A largura mostra a quantidade de elementos que possuem a mesma magnitude, enquanto a altura mostra a dispersão dos dados.

Além disso, foram aplicadas técnicas de aprendizado computacional PCA (*Principal Component Analysis*) e Árvores de Decisão[11] na tentativa de criar uma regra que auxilie a escolha de, dado um conjunto de SLFT, qual modelo utilizar: PWM ou GRE aprendida pelo LAPFA. PCA é uma técnica que reduz a dimensionalidade dos dados com base em

[11] Tanto o PCA quanto a árvore de decisão foram implementadas em *Python*, com o pacote *Scikit-learn* (PEDREGOSA *et al.*, 2011).

uma transformação linear dos mesmos. Se ao reduzir a duas dimensões, por exemplo, é possível visualizar uma separação das classes, tem-se uma regra de separação. Já as árvores de decisão, após treinadas sobre os dados de entrada, encontra regras de separação entre as classes utilizando as características originais, favorecendo a interpretabilidade das regras.

Para a aplicação das duas técnicas, cada dado (instância do treinamento) foi um conjunto de sequências de SLFT de cada um dos seis grupos de origem/descoberta de motivo, conjunto esse representado por suas cinco características (quatro medidas de dependência e o IC médio) e sua classe ($D >= 0.4$ representando preferência por GRE-LAPFA, ou $D < 0.4$ representando indiferença de modelo). O objetivo foi: dado esses conjuntos de dados e o rótulo sobre cada um de pertencer a classe $D >= 0.4$ ou $D < 0.4$, criar uma regra de decisão que consiga optar automaticamente por uma classe ou outra, ou seja, um critério geral para se optar por GRE-LAPFA ou PWM. Nota-se que a opção por PWM é quando $D < 0.4$, pois neste caso os dois modelos possuem o mesmo desempenho, e por questões práticas e de custos computacionais, prefere-se um modelo mais simples ao complexo.

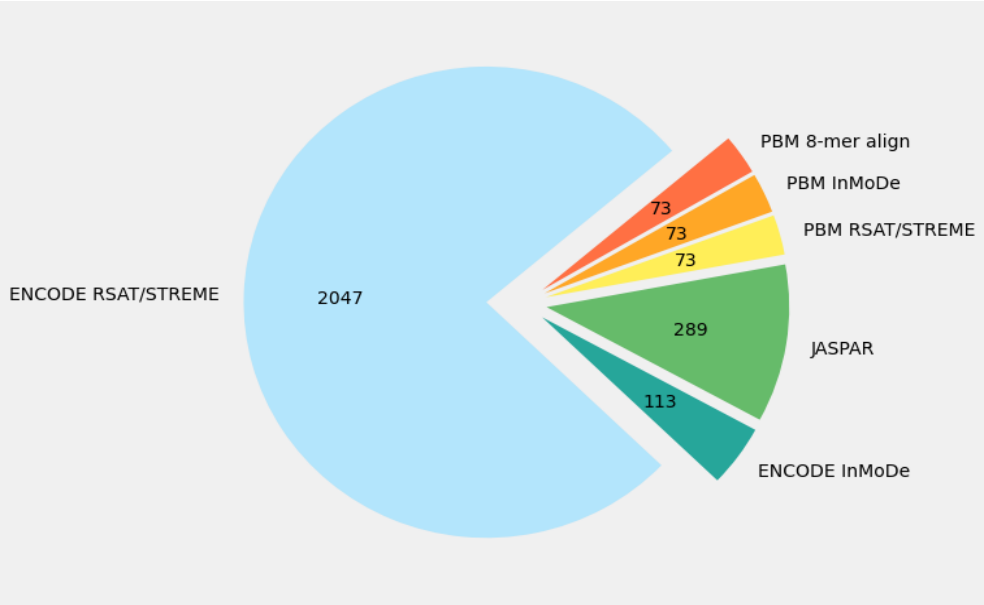
4 Resultados e discussão

Este capítulo está dividido em quatro seções. A seção 4.1 é uma apresentação de todos os dados obtidos por todas as combinações de experimentos e suas respectivas descobertas de motivos. A seção 4.2 contém a comparação de desempenho entre PWM e GRE-LAPFA, separada por experimento-algoritmo. A seção 4.3 contém comparações entre medidas de dependência e Conteúdo de Informação médio para cada experimento-algoritmo. A seção 4.4 mostra a Árvore de Decisão montada sobre todos os dados e a regra de decisão por um modelo ou outro, assim como a Análise de Componentes Principais.

4.1 Apresentação dos dados utilizados para a comparação entre PWM e GRE-LAPFA e análise de dependência entre posições de bases

Sobre a obtenção da amostra positiva nas seções 3.1, 3.2 e 3.3 foram considerados 2047 FTs de ChIP-seq do ENCODE, 289 FTs do JASPAR de ChIP-chip ou ChIP-seq e 73 FTs de PBM de (WEIRAUCH *et al.*, 2013). Sobre esses dados, distintos algoritmos de descoberta de motivos foram aplicados, como na figura 15.

Figura 15 – Combinação de dados e algoritmos de descoberta de motivos utilizados. O gráfico mostra três distintos bancos de dados obtidos: ENCODE, JASPAR, PBM. Para cada conjunto, um ou mais algoritmo de descoberta de motivos foi aplicado sobre o conjunto de dados. Os números mostram quantos FTs estão representados em cada grupo. Para dados do ENCODE, os FTs usados pelo InMoDe (descob. de motivo não baseada em PWM) são uma parcela dos FTs usados pelo RSAT/STREME (descob. de motivo baseada em PWM), assim como os três algoritmos aplicados sobre PBMs foram sobre o mesmo conjunto de 73 FTs de PBM. Os dados do JASPAR vieram apenas de ChIP-chip ou ChIP-seq, fazendo uso de descoberta de motivo baseado em PWMs, mas com curagem manual e verificação experimental da composição de bases do motivo.



Fonte: Guilherme Miura Lavezzo (2021)

.

Para cada FT em que foi aplicado a descoberta de motivo sobre algum dos dados públicos coletados, como na figura 15, foram treinados e testados as duas classes de modelos (PWM e GRE-LAPFA). Assim, foi possível obter um total de 2668 comparações entre os modelos, sob diversas condições.

Há um número maior de comparações entre PWMs e GRE-LAPFAs a partir de dados do ENCODE com uso das ferramentas RSAT e STREME, totalizando 2047 dados de FTs. A justificativa é que a ideia inicial do projeto era comparar o desempenho de PWM e GRE-LAPFA em dados *in vivo* de banco de dados públicos, principalmente de *ChIP-seq*. Que se saiba, não existe uma diferença clara na literatura entre PWMs como classificadores de SLFT e PWMs como o modelo utilizado com fins de descoberta de motivos. Somado a isso, existe pelo menos uma evidência na literatura de que obter SLFTs que foram classificados a partir de PWMs não limitaria a identificação de dependência entre bases de

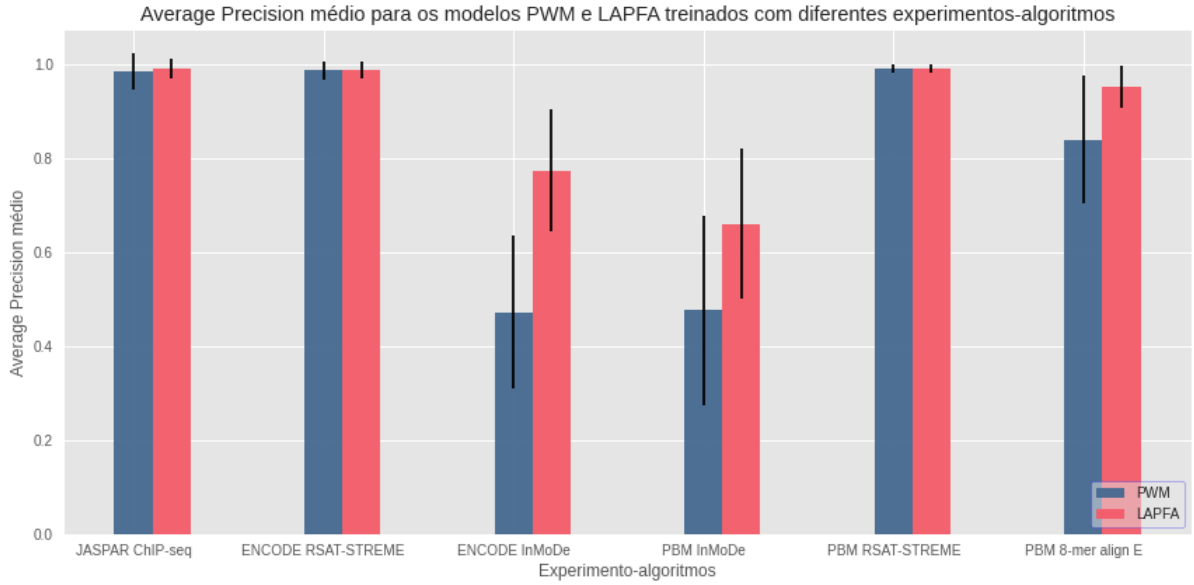
um motivo (YANG; CHANG, 2015). Conforme foram comparados PWM e GRE-LAPFA apenas nesses 2047 FTs, notou-se que poderia ser informativo testar diferentes algoritmos de descoberta de motivos para evitar qualquer conclusão precipitada sobre a análise final. Além disso, também poderia-se coletar algum experimento de alta resolução, como PBM. Em experimentos PBM é possível obter motivos e seus SLFTs derivados sem que haja uma etapa de aplicar uma ferramenta de descoberta de motivos, caso do algoritmo 8-mer align (vide subseção 3.3.3). Este enriquecimento de novos algoritmos poderia fornecer uma melhor generalização na comparação dos modelos, como será discutido e apresentado neste capítulo.

4.2 Comparação entre desempenhos dos modelos PWM e GRE-LAPFA

Para todas as 2668 amostras, a comparação de desempenho dos modelos foi feita usando a medida *Average Precision* (AP) média dos sete folds da validação cruzada (vide seções 2.2.1 e 3.5.2). Como observação, os modelos foram bastante consistentes no desempenho de cada fold, contendo um desvio padrão (da AP dos sete folds) na terceira ou quarta casa decimal.

A figura 16 apresenta o AP médio de cada PWM e GRE-LAPFA sendo comparado para cada um dos 2668 conjunto de SLFTs, estes divididos na forma combinada "experimento e algoritmo de descoberta de motivo respectivo". Nota-se que o desempenho da PWM e do GRE-LAPFA é dependente do experimento e metodologia utilizada. Essencialmente, o que mais afeta o desempenho dos modelos foi o algoritmo de descoberta de motivo e não tanto o tipo de experimento. Isso é evidenciado ao se comparar que há uma perda muito significativa de desempenho de PWMs quando aplicado o InMoDe como algoritmo. Por outro lado, PWMs obtidas pelos algoritmos RSAT/STREME tiveram excelente performance seja em dados de ChIP-seq do ENCODE ou em dados de PBM.

Figura 16 – Comparação de desempenho entre PWM e GRE-LAPFA para todo o conjunto de dados considerado. O gráfico mostra o AP médio de sete folds da validação cruzada para PWM e GRE-LAPFA. O valor do AP, representado pela altura da barra varia de 0 a 1. Barras laranjas representam o modelo GRE-LAPFA e barras cinzas representam PWMs. Os traços contidos em cada barra representam o desvio padrão do modelo sob aquela categoria. Cada categoria no eixo horizontal representa a maneira como foi obtida o conjunto de SLFTs, sendo uma combinação entre um tipo de experimento e um algoritmo de descoberta de motivos em que foi aplicado o primeiro. Note que cada categoria indica um experimento-algoritmo utilizado. Assim, barras de categorias distintas não devem ser comparadas de maneira direta, pois se referem a sítios de diferentes FTs. Assim, um total de 2668 conjuntos de SLFTs estão sendo representados com o mesmo número em comparações dos modelos classificadores. RSAT/STREME são ferramentas cuja descoberta de motivos é baseada em PWMs. InMoDe é uma ferramenta não baseada em PWMs. 8-mer align é um algoritmo não automatizado que permite acessar a descoberta de motivos não baseada em algum modelo prévio. Dados do JASPAR são exclusivamente de ChIP-chip ou ChIP-seq.



Fonte: Guilherme Miura Lavezzo (2021)

.

Para o GRE-LAPFA, nota-se que seu desempenho é semelhante ou superior às PWMs. A princípio, para algoritmos como do RSAT/STREME e de amostras do JASPAR, o alto desempenho de PWMs e GRE-LAPFA é bastante consistente. Por outro lado, para algoritmos como InMoDe e 8-mer align, nota-se uma diferença significativa entre os modelos PWM e GRE-LAPFA. De preferência, espera-se que caso seja optado fazer uso de um desses dois algoritmos, é recomendado o uso de modelos classificadores mais complexos que PWMs, para favorecer o desempenho e evitar erros de classificação. Para algoritmos como

RSAT/STREME ou dados coletados do JASPAR de ChIP-chip ou ChIP-seq, a escolha por um modelo classificador pode ser arbitrária, pois ambos PWM e GRE-LAPFA garantem o excelente desempenho. Por praticidade, PWMs são recomendadas por serem mais fáceis de implementar, amplamente difundidas em diversas ferramentas de bioinformática e menos exigentes com desempenho computacional e tempo de processamento.

Apesar de mostrados o desempenho dos modelos para cada possível combinação experimento-algoritmo, a comparação entre modelos para categorias distintas, por exemplo entre PWMs do ENCODE InMoDe com PWMs de PBM 8-mer align, é uma comparação indireta. Cada experimento-algoritmo representa uma maneira distinta de obter conjuntos de SLFTs. Mesmo para um FT em comum, o motivo obtido é distinto, portanto, a amostra positiva também é.Por outro lado, a comparação entre PWM e GRE-LAPFA do mesmo experimento-algoritmo é uma comparação direta.

A maior diferença entre desempenhos de modelos está principalmente no uso do algoritmo InMoDe como descoberta de motivos. Neste caso, o GRE-LAPFA possui desempenho consideravelmente superior, enquanto PWM possui um desempenho muito baixo. Isso provavelmente se deve ao fato do InMoDe ser um algoritmo baseado em modelos de Markov flexíveis. Como GREs são modelos similares a cadeias de Markov, a relação de dependência entre posições de bases é capturada por ambos. Por outro lado, PWMs treinadas com o uso prévio do InMoDe não conseguem associar a relação entre posições de bases estabelecida.

4.3 Análise de medidas de dependência entre posição de bases

4.3.1 Comparação de medidas de dependência entre experimentos-algoritmos

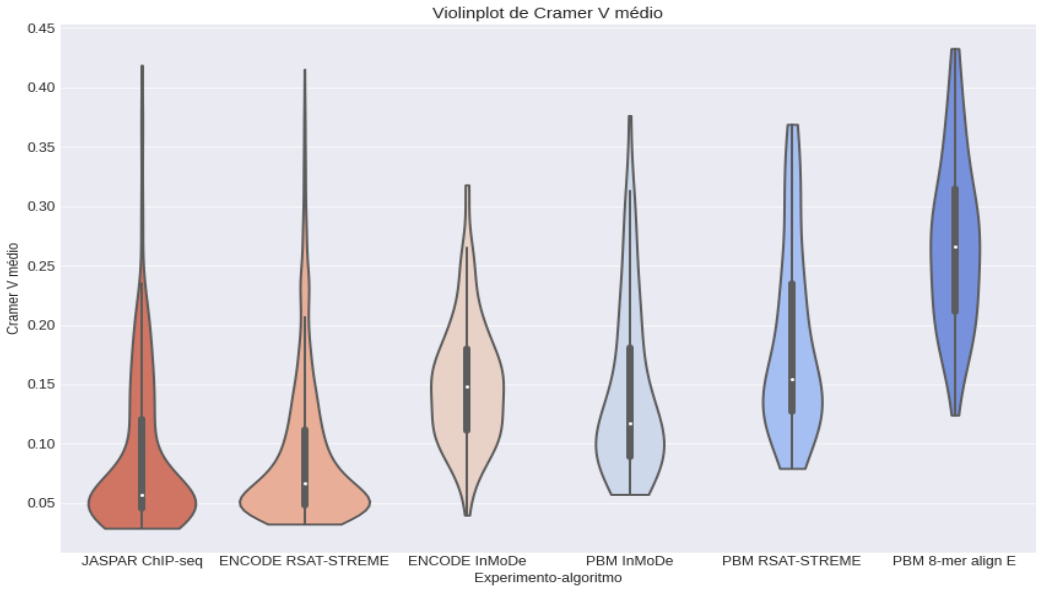
Esta subseção utiliza os mesmos 2668 conjuntos de SLFTs. O objetivo é verificar se as medidas de dependência definidas na seção 3.4 possuem relação com o desempenho dos modelos nos vários conjuntos experimento-algoritmo.

Como já mencionadas, as medidas de dependência são definidas entre duas posições da sequência motivo. Por isso, para um motivo existirão diversos valores medidos, um para cada combinação de posições. Foram testadas as medidas de Cramer V média, Cramer V máxima, Theil U médio (ou coeficiente de incerteza simétrico), Theil U máximo e

Conteúdo de Informação médio. Esse conjunto de medidas serão denominadas medidas características (*features*).

Nas figuras 17, 18, 19, 20, 21 , *Violin Plots* são mostrados comparando as características entre experimentos-algoritmos.

Figura 17 – *Violin Plot* da medida de Cramer V médio para cada grupo de experimento-algoritmo combinado.



Fonte: Guilherme Miura Lavezzo (2021)

.

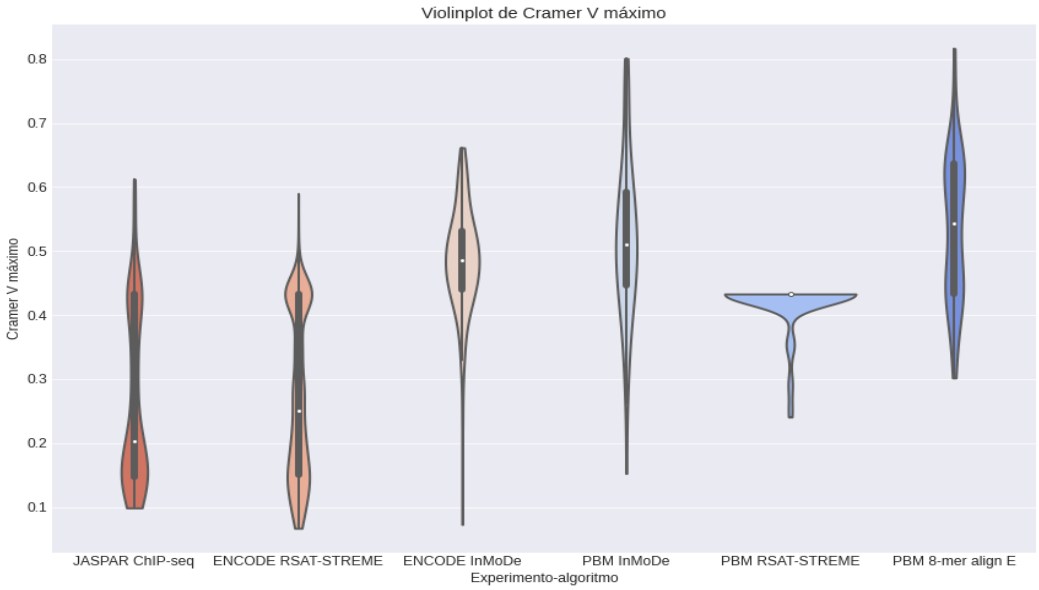
Pela figura 17, observa-se que há uma semelhança entre a medida de Cramer V médio dos dados do JASPAR com os dados do ENCODE RSAT/STREME. Isso era esperado, pois o protocolo de descoberta de motivo e extração de SLFTs do ENCODE RSAT/STREME foi uma adaptação do protocolo do JASPAR, publicado na seção de métodos suplementares (FORNES *et al.*, 2019). A diferença maior, é que o JASPAR possui uma maneira de curagem manual dos SLFTs, seguido da validação experimental em evidências na literatura para validar o motivo. Isto não significa que o motivo do JASPAR seja o correto, mas possui ao menos alguma evidência para tal. Além disso, mesmo que a composição de um motivo seja sabida, a probabilidade real de ocorrência das bases em cada posição não o é necessariamente. Também não há muitas validações experimentais que confirmem a dependência entre bases para sequências reconhecidas por um FT. Algumas poucas validações experimentais na literatura são experimentos *in vitro* (MAN, 2001; UDALOVA *et al.*, 2002; BULYK *et al.*, 2001; WOLFE *et al.*, 1999). Como sabido, experimentos *in vitro* possuem a desvantagem de não informar a condição

in natura da interação FT-DNA, portanto, algumas possíveis relações de dependência de um conjunto de SLFTs que foram verificadas, podem ser afuncionais (ou inexistentes) *in vivo*. Além disso, a confirmação da dependência entre posições de bases depende de outros experimentos que corroborem com o mesmo achado, e de dados sobre a composição e estrutura tridimensional da proteína, por exemplo, cristalografia. Ou seja, a verificação de dependência entre posições de bases *in vivo* ainda é um assunto a ser investigado, conforme novas técnicas com maior resolução possam ser introduzidas e por isso, esse assunto permanece em aberto.

PBM 8-mer foi o conjunto de dados com maior Cramer V médio, evidenciado pela mediana de aproximadamente 0.26. No geral, a média de Cramer V foi baixa, considerando que a medida varie de 0 a 1. No entanto a medida Cramer V não varia com escala linear, ou seja, de que 0.5 signifique metade da magnitude de associação. Isso torna difícil interpretar os valores obtidos, sem uma referência.

O Cramer V médio obtido entre grupos ora foi intuitivo de ser interpretado ora não. Estimava-se que a descoberta de motivos baseados em PWM pudesse, de maneira generalizada, gerar motivos e consequentemente conjunto de SLFTs com menor dependência entre posições de bases. Isso foi evidenciado pelos valores de Cramer V médio do JASPAR e do ENCODE RSAT/STREME, mas curiosamente não para PBM RSAT/STREME. Para este último, mesmo utilizando descoberta de motivos baseado em PWMs, em mediana, o Cramer V médio foi o segundo maior das comparações, com valor de 0.15 aproximadamente. A justificativa pode estar no fato que foram utilizados apenas as 2000 sequências 35-mers de maior sinal de fluorescência (vide seção 3.3.1). Além disso, em sequências ChIP-seq são esperadas que ocorram muitas redundâncias de sequências que compõe a sequência motivo, uma vez que é esperado que há ocorrências múltiplas de SLFTs de mesma composição, mas em diferentes coordenadas genômicas.

Figura 18 – *Violin Plot* da medida de Cramer V máximo para cada grupo de experimento-algoritmo combinado.



Fonte: Guilherme Miura Lavezzo (2021)

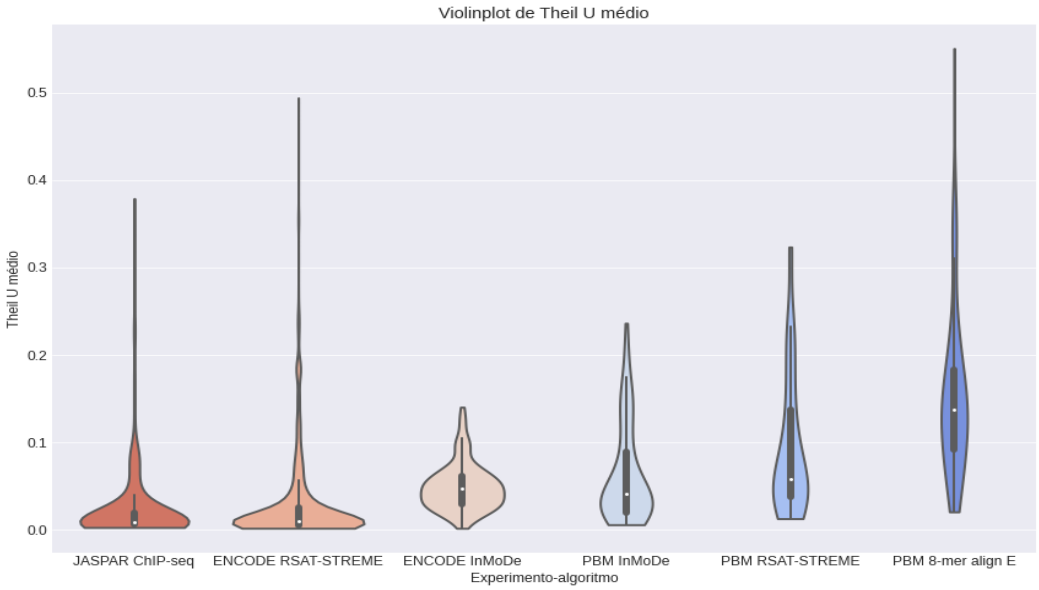
.

Nota-se no geral, uma distribuição muito mais difusa que em relação ao Cramer V médio, com valores assumindo distribuições quase uniformes para uma janela de até 0.5 de diferença entre o máximo e o mínimo valor. A excessão é o Cramer V máximo do conjunto PBM RSAT/STREME, que obteve uma distribuição densa em torno da mediana de 0.43 aproximadamente. No geral, as medidas de Cramer V máximo maiores foram, em mediana, de grupos que fizeram uso do algoritmo InMoDe ou PBM 8-mer.

Além do Cramer V médio ter sido mais baixo para JASPAR e ENCODE RSAT/STREME, o mesmo ocorreu para a medida de Cramer V máximo (figura 18). Ou seja, poderia ser inferido que obter conjuntos de SLFTs com ferramentas de descoberta de motivos baseadas em PWMs possa dificultar análises sobre associação entre bases, o que poderia ir contra o estudo de (YANG; CHANG, 2015) em dados *in vivo*.

Por outro lado, o uso do algoritmo InMoDe, algoritmo não baseado em PWMs, favoreceu valores mais altos de Cramer V máximo, assim como o algoritmo 8-mer align. Novamente, PBM RSAT/STREME possui um valor em mediana maior do que o esperado, sendo mais semelhante às técnicas não baseadas em PWM do que o contrário.

Figura 19 – *Violin Plot* da medida de Theil U simétrico médio para cada grupo de experimento-algoritmo combinado.



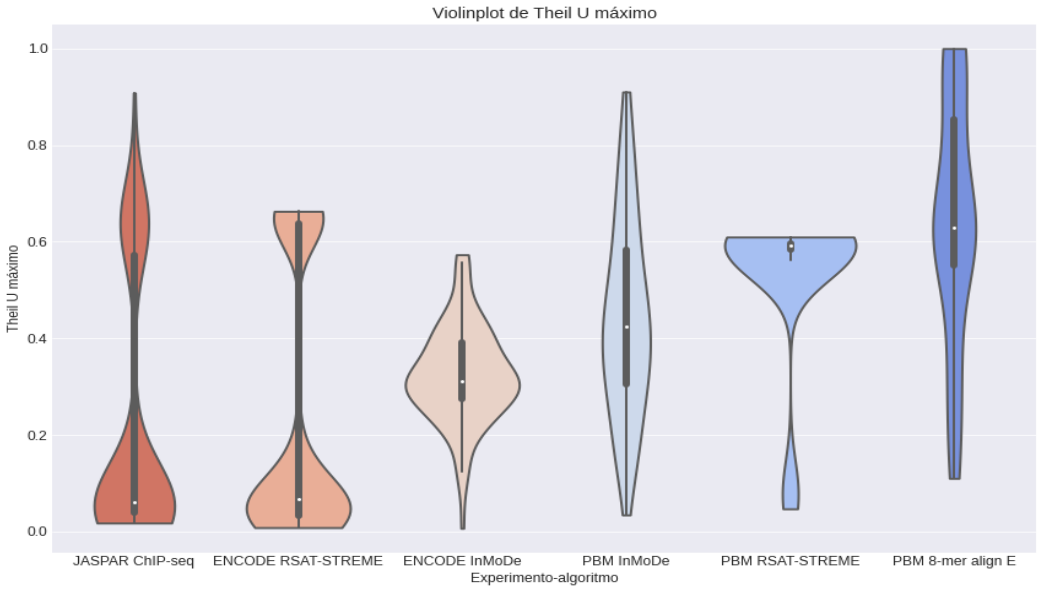
Fonte: Guilherme Miura Lavezzo (2021)

.

A medida de Theil U médio (figura 19) segue a mesma ideia da medida do Cramer V médio, em termos da diferença relativa entre grupos comparados, porém numa escala ainda menor. Observa-se, de maneira geral, que a distribuição de Theil U médio tendeu a valores muito próximos de 0 e 0.1. Como o Cramer V, o Theil U simétrico não escala linearmente, dificultando interpretações do quanto a magnitude de 0.2 possa significar em termos de "força" de associação entre duas posições de bases.

Em termos relativos, o ENCODE InMoDe e o PBM InMoDe possuem maiores valores, em mediana, de Theil U médio do que JASPAR ChIP-seq e ENCODE RSAT/STREME, como esperado. Esses dois últimos, apesar de serem valores muito próximos de 0, em mediana, possuem uma cauda longa em 0.4 e 0.5, evidenciando que para alguns poucos FTs, houve uma dependência de bases muito maior. Entretanto, são raros os casos, pois o achatamento da distribuição próximo de 0.02 evidencia em grande maioria os dados do JASPAR e ENCODE RSAT/STREME possuam baixíssimos valores de Theil U médio. Novamente, PBM 8-mer segue sendo a maior mediana da medida de dependência em questão, contando também uma cauda longa vertical.

Figura 20 – *Violin Plot* da medida de Theil U simétrico máximo para cada grupo de experimento-algoritmo combinado.

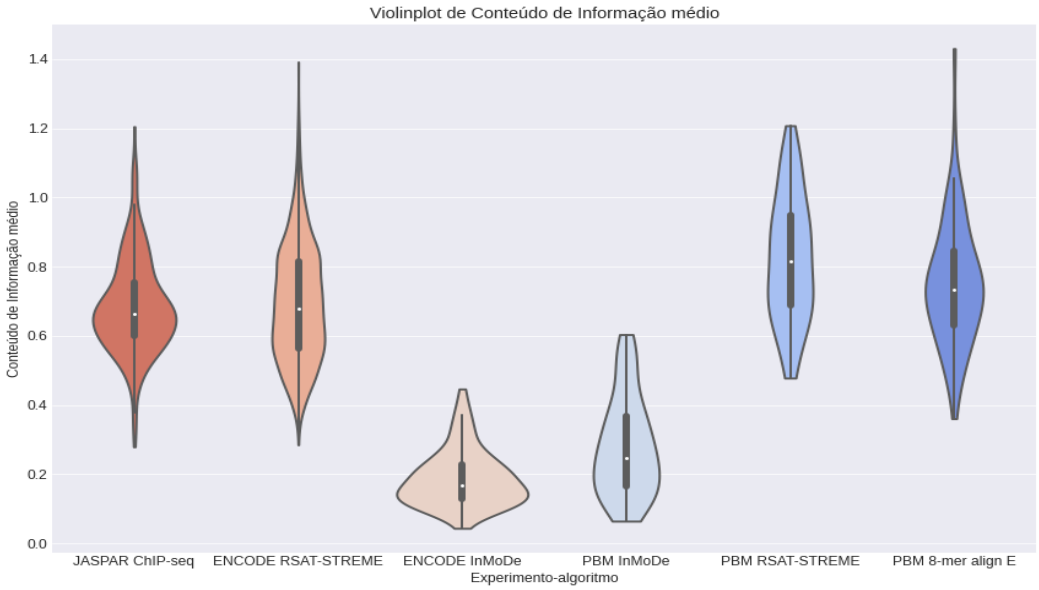


Fonte: Guilherme Miura Lavezzo (2021)

.

Considerando a medida de Theil U máximo (figura 20), em primeira instância, nota-se similaridade dos formatos das distribuições com os aferidos com o Cramer V máximo (figura 18). Todavia, no Theil U máximo, o grupo PBM RSAT/STREME possui mediana visivelmente maior que em experimento com InMoDe. Observa-se também que alguns motivos obtiveram valores muito próximo ao valor 1.0, limite superior de Theil U. Portanto, sendo uma medida com maior cobertura de valores calculados.

Figura 21 – *Violin Plot* da medida de Conteúdo de Informação médio para cada grupo de experimento-algoritmo combinado.



Fonte: Guilherme Miura Lavezzo (2021)

.

Na figura 21 observou-se que a aplicação do algoritmo InMoDe resultou em motivos com baixo Conteúdo de Informação médio, enquanto os demais algoritmos se mantiveram em distribuição semelhantes. Um baixo conteúdo de Informação médio implicaria em um modelo mais generalizado e portanto com maior tendência a falsos positivos. Isso explicaria o menor desempenho de PWMs, modelos que são conversões matemáticas direta das sequências motivos. O Conteúdo de Informação médio alto para JASPAR e experimentos com RSAT/STREME justifica também que PWM e GRE-LAPFAs estejam com desempenho semelhante. Todavia, em PBM 8-mer, houve diferença entre o desempenho de PWMs e LAPFAs, em média, mas o Conteúdo de Informação médio é alto. Portanto somente essa medida não é suficiente para decidir por um modelo ou outro.

4.3.2 Relação entre as medidas de dependência e de informação dos conjuntos de sítios e a diferença de desempenho entre PWM e GRE-LAPFA

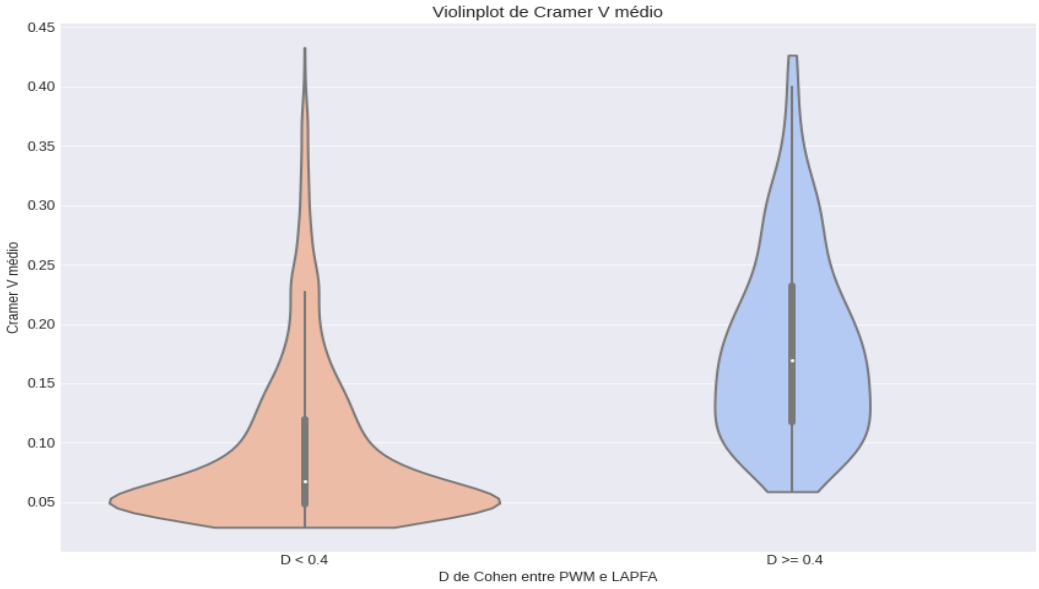
Como já adiantado na seção 3.6, o GRE-LAPFA obteve um resultado ou muito próximo ao desempenho de uma PWM ou melhor para os conjuntos de sítios testados. Para medir essa diferença entre desempenhos, fez-se uso da medida D de Cohen (*effect size*), que quantifica a diferença entre médias padronizadas de dois grupos (no caso, modelos) (vide seção 2.2.4). Observou-se que quando D é menor que 0.4 ($D < 0.4$), os modelos

possuem desempenho semelhante, enquanto para valores D iguais ou acima de 0.4 o desempenho do GRE-LAPFA havia sido pelo menos 5% maior que o desempenho da PWM. Não foi encontrado nenhum caso em que com $D >= 0.4$ o modelo PWM tenha sido o modelo melhor para um FT se comparado ao desempenho do GRE-LAPFA sob as mesmas condições. Portanto, quando referido ao valor de D, será usado a expressão $D < 0.4$ para mencionar que não houve diferença relevante no desempenho de PWM e GRE-LAPFA, e $D >= 0.4$ para mencionar que o GRE-LAPFA foi melhor que a PWM em magnitudes relevantes.

Na totalidade dos 2668 motivos obtidos, 224 pertencem ao grupo $D >= 0.4$ e 2444 pertencem ao grupo $D < 0.4$.

A figura 22 apresenta a comparação entre Cramer V médio para as classes $D < 0.4$ e $D >= 0.4$ desses 2668 dados.

Figura 22 – *Violin Plot* da medida de Cramer V médio para as classes $D >= 0.4$ e $D < 0.4$ de todos os 2668 dados.



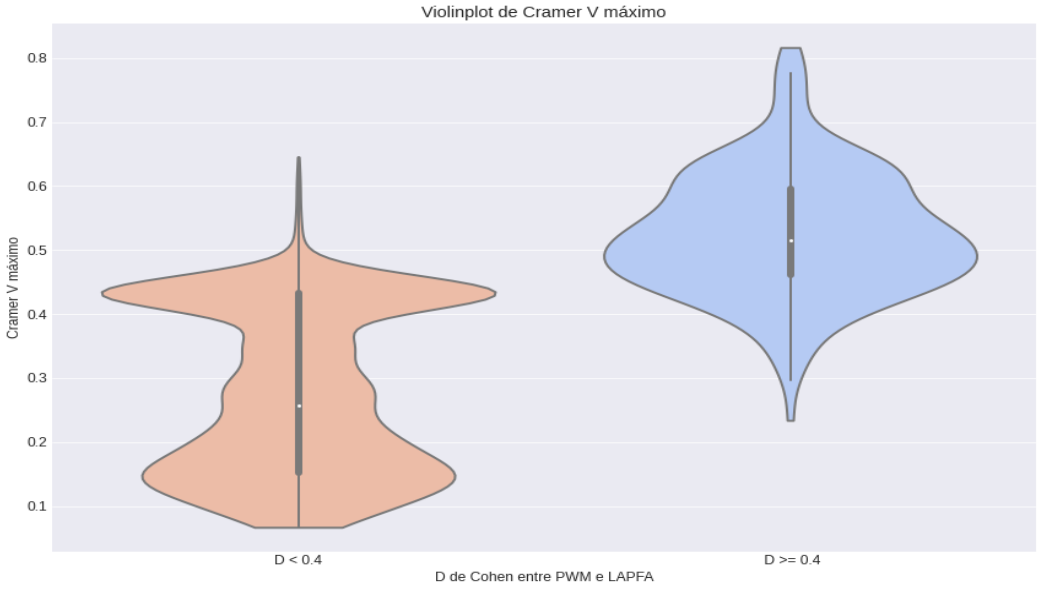
Fonte: Guilherme Miura Lavezzo (2021)

.

Observa-se, como desejado, que a classe $D >= 0.4$, ou seja, quando o GRE-LAPFA é um modelo significativamente melhor que PWM, que a média de Cramer V também é, em mediana. Isso implica que, quando o conjunto de SLFTs possui alguma magnitude de dependência entre bases, modelos que conseguem capturar essa dependência podem ter desempenho melhor. Apesar do constatado, as distribuições não estão totalmente

separadas, existem valores em $D < 0.4$ que possuem valores mais altos de Cramer V médio assim como existem valores mais baixos da medida na classe $D >= 0.4$.

Figura 23 – *Violin Plot* da medida de Cramer V máximo para as classes $D >= 0.4$ e $D < 0.4$ de todos os 2668 dados.



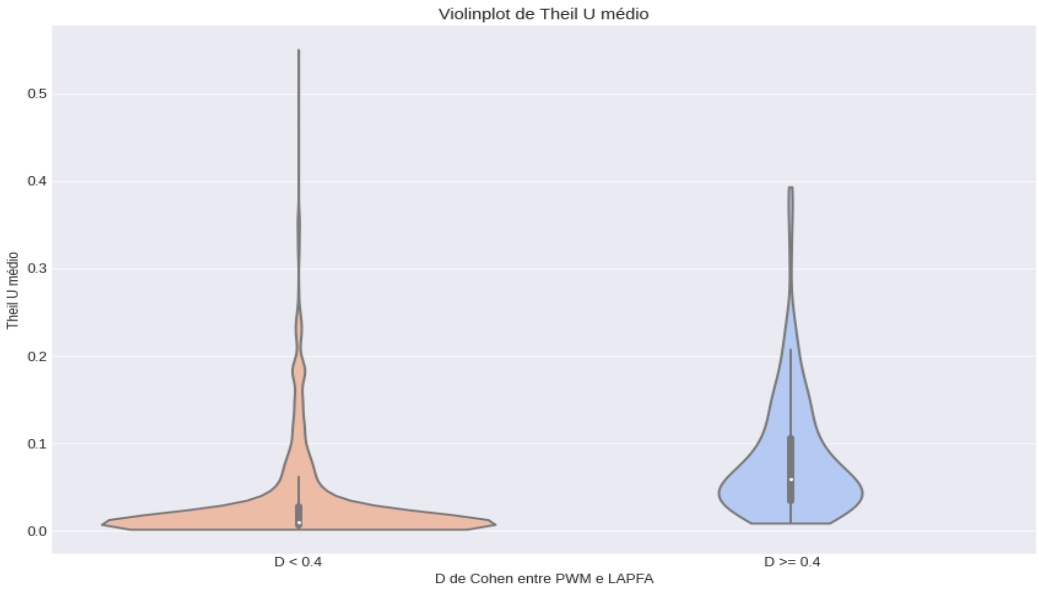
Fonte: Guilherme Miura Lavezzo (2021)

.

Para Cramer V máximo (figura 23), observa-se novamente que a medida máxima de Cramer V tende a ser mais alta para os motivos para os quais GRE-LAPFA é o melhor modelo. Observa-se que, quando PWM e GRE-LAPFA são semelhantes em desempenho, a medida de Cramer V máximo tende a um valor próximo de 0.43 ou menos; com algumas poucas exceções na distribuição (em vermelho). Quando o GRE-LAPFA é melhor que uma PWM, observa-se que a distribuição do Cramer V máximo está acima de 0.43 em maioria e em mediana, com algumas poucas exceções que atingem valores próximos a 0.3. Essa disparidade nas medianas dos grupos e na maneira como as distribuições estão verticalmente separadas de maneira razoável pode ser útil como um critério para decisão da Árvore de Decisão, como será abordado na seção seguinte.

A medida de Cramer V máximo, diferente do Cramer V médio (figura 17), obteve maior separação entre $D < 0.4$ e $D \geq 0.4$. Isso implica que, para se beneficiar do GRE-LAPFA, bastaria que exista maior relação de dependência entre posições de bases, desde que seja significativa.

Figura 24 – *Violin Plot* da medida de Theil U médio para as classes $D >= 0.4$ e $D < 0.4$ de todos os 2668 dados.



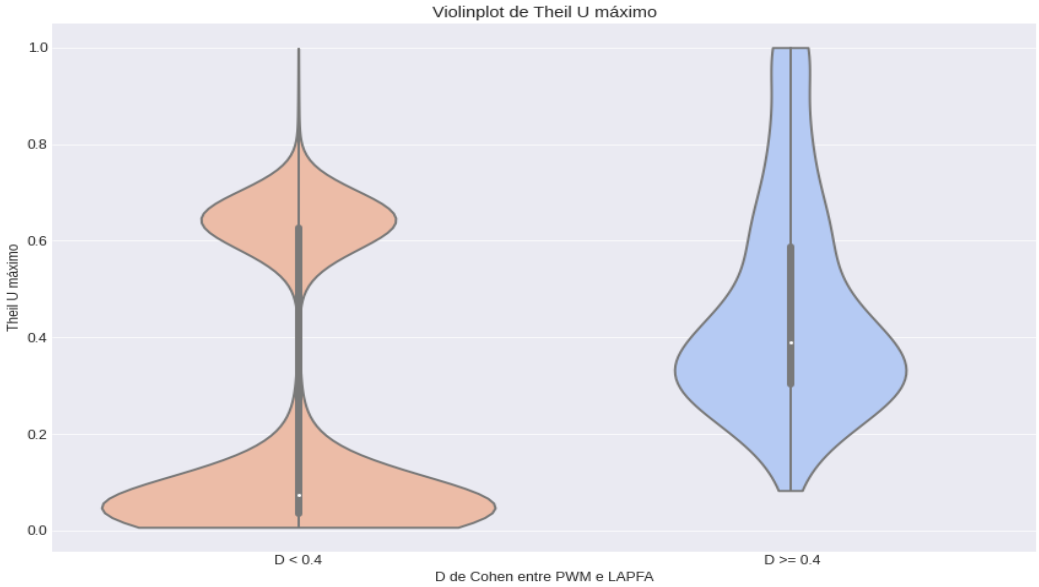
Fonte: Guilherme Miura Lavezzo (2021)

.

Os formatos das distribuições do Theil U médio (figura 24) é semelhante ao do Cramer V médio, mas com valores muito menores, próximos a 0 e 0.1. Observa-se que a medida Theil U é mais estringente na maneira de definir dependência entre posições de bases, atribuindo que, no geral, praticamente não há muita dependência entre bases a ser considerada. Contudo, existe uma diferença em mediana e no arranjo da distribuição para as duas classes comparadas. Novamente, quando o LAPFA é melhor que PWMs, o Theil U médio também é maior, mas com uma magnitude mais discreta.

Para a medida Theil U máximo (figura 25), curiosamente, a classe $D < 0.4$ apresenta uma distribuição dividida dos dados: existe uma parte que possui Theil U máximo em 0.5 ou mais, enquanto uma maioria está entre 0 e 0.2; com mediana inferior a 0.1. Uma vez que a classe $D >= 0.4$ possui valores maiores no geral, o que era esperado, parte dos valores de Theil U máximo de $D < 0.4$ se sobrepõem aos valores de $D >= 0.4$, o que não seria desejável como um indicador de dependência entre bases.

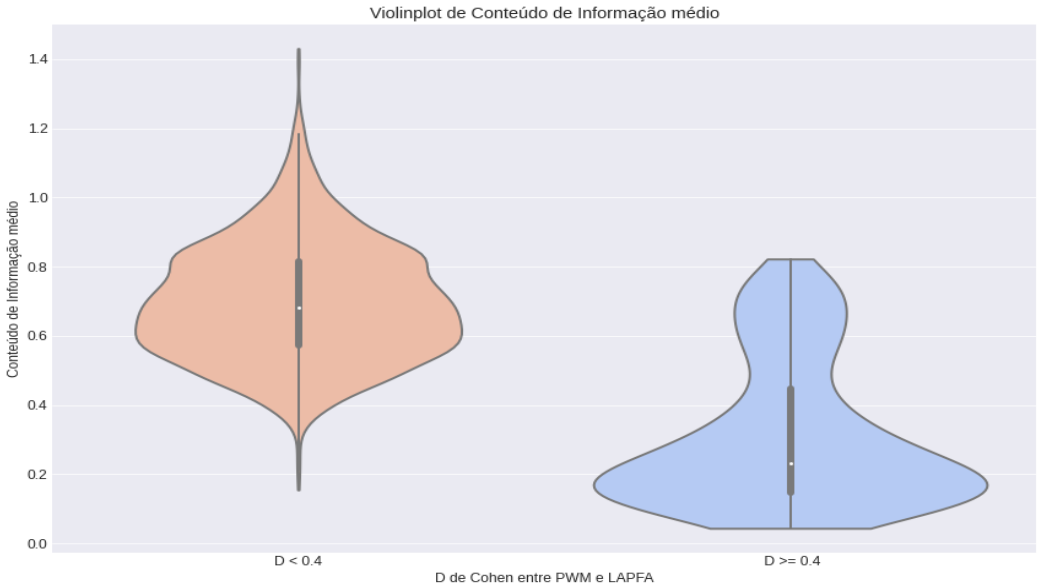
Figura 25 – *Violin Plot* da medida de Theil U máximo para as classes $D >= 0.4$ e $D < 0.4$ de todos os 2668 dados.



Fonte: Guilherme Miura Lavezzo (2021)

.

Figura 26 – *Violin Plot* da medida de Conteúdo de Informação médio para as classes $D >= 0.4$ e $D < 0.4$ de todos os 2668 dados.



Fonte: Guilherme Miura Lavezzo (2021)

.

Os *Violin Plots* da medida de Conteúdo de Informação são apresentados na figura 26. Observa-se que tanto pela mediana quanto pela distribuição das classes, há uma diferença relevante. A classe $D < 0.4$ possui Conteúdo de Informação médio maior que a classe $D \geq 0.4$. De fato, como o Conteúdo de Informação médio mede a qualidade de

uma PWM, é esperado que essa medida esteja maior no caso em que os modelos PWM e GRE-LAPFA possuam desempenhos equiparáveis. No caso em que essa medida é baixa, é provável que os algoritmos de descoberta de motivo, principalmente InMoDe e 8-mer align E, tenham capturado relações de dependência entre as posições de bases, as quais GRE-LAPFA se beneficiaria enquanto PWM não, pela definição do modelo.

No apêndice C, estão mostradas todas as comparações via *effect size* e testes de Wilcoxon, realizadas aqui como *Violin Plots*. Observou-se que todas as medidas possuíram um D de Cohen altíssimo (próximo de 1 ou maior), com p-valores do teste de medianas muito abaixo de 0.05. As medidas Cramer V máximo e Conteúdo de Informação médio obtiveram os maiores valores de D (1.96 e 2.21 respectivamente), e também os menores p-valores (9.99e-63 e 1.32e-85) de todas as comparações, reforçando a ideia da importância dessas medidas, como será visto em 4.4.

4.4 Obtenção de uma regra de decisão para escolher um modelo classificador

Na tentativa de obter uma regra de decisão para, dado um conjunto de sítios, selecionar PWM ou GRE-LAPFA como modelo preditor, foram testadas duas estratégias: análise de componentes principais e árvores de decisão.

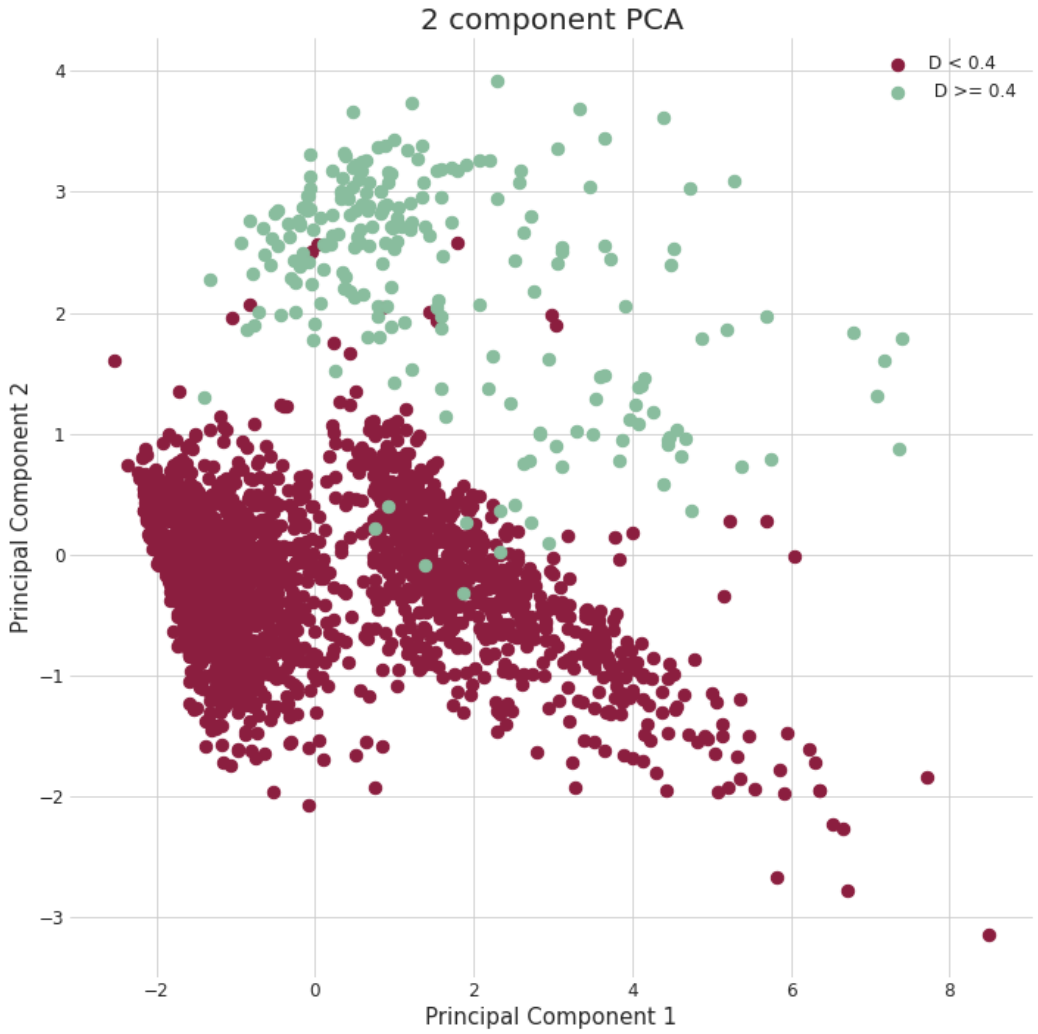
4.4.1 Análise de Componentes Principais

Foi realizada uma Análise de Componentes Principais (PCA) para verificar se seria possível visualizar agrupamentos distintos, em 2 dimensões, das classes. Caso houvesse a separação, seria possível criar uma regra de decisão para os modelos. Neste PCA todas as cinco medidas aqui discutidas foram utilizadas.

A figura 27 apresenta a visualização dos dados em duas dimensões (os dois primeiros componentes principais) após a aplicação do PCA. Observa-se uma clara divisão entre as duas classes, notada pela formação de *clusters* em cores distintas. Existem alguns pontos roxos que aglomeraram com os pontos verdes e vice-versa, mas isso é aceitável uma vez que não obtivemos uma medida perfeita que consiga separar as duas classes. Além disso, como o critério de formação das classes foi de D maior ou menor que 0.4, também é possível

que esses pontos distantes dos respectivos *clusters* possam ser valores que obtiveram D próximo a 0.4.

Figura 27 – Análise de Componentes Principais bidimensional utilizando as 5 medidas características definidas. Cada ponto em roxo representa um elemento do conjunto de sequências motivos cujo desempenho entre PWM e GRE-LAPFA são semelhantes. Em verde, o caso em que o modelo GRE-LAPFA é melhor que PWM em uma diferença relevante. Existem 224 pontos verdes e 2444 pontos roxos.



Fonte: Guilherme Miura Lavezzo (2021)

.

A matriz de componentes principais (25) que foi calculada para a figura 27, mostra os dois primeiros autovetores da matriz de transformação do PCA, ou seja, os coeficientes da transformação linear das cinco características para os dois componentes principais. As cinco características originais são, nesta ordem, Cramer V médio, Cramer V máximo, Theil U médio, Theil U máximo, Conteúdo de Informação médio. Logo, cada posição desses autovetores corresponde ao peso da respectiva característica na transformação linear.

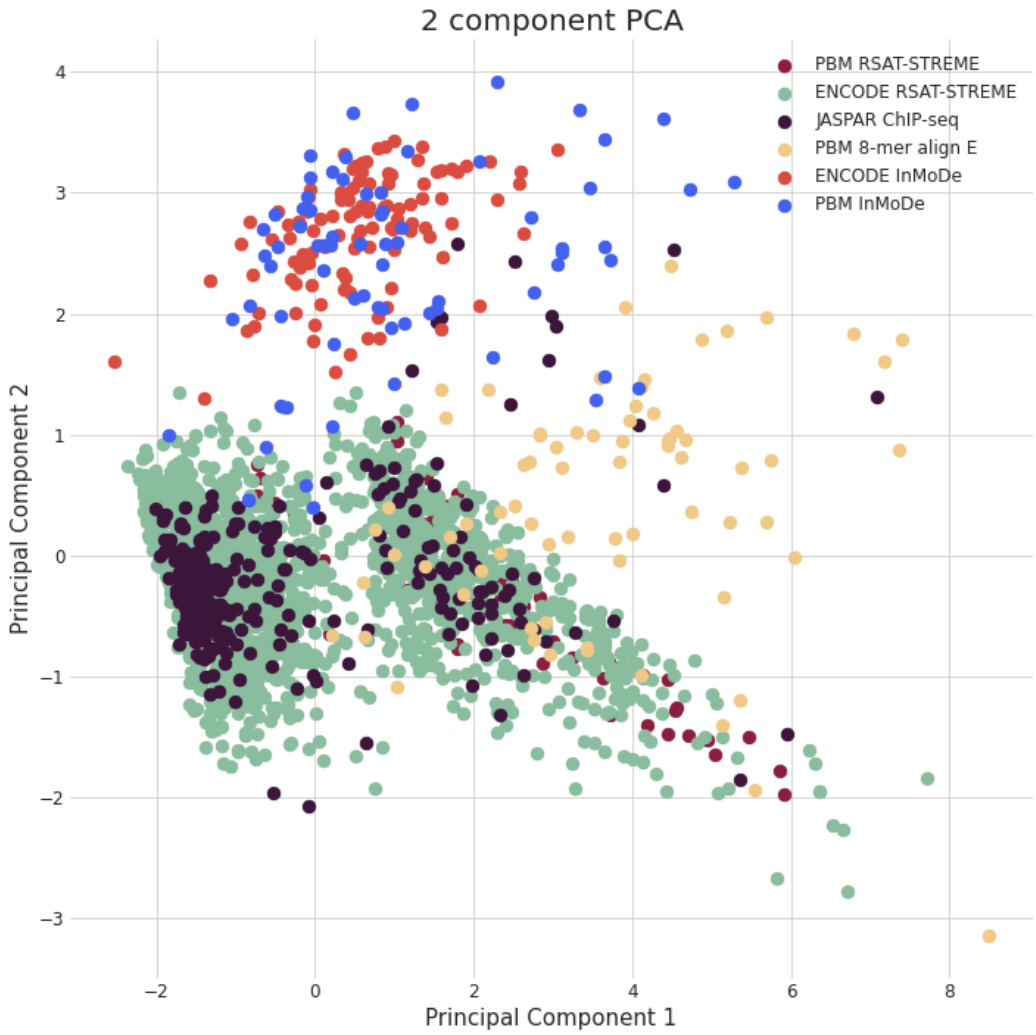
$$\begin{pmatrix} 0.5 & 0.468 & 0.471 & 0.49 & 0.261 \\ 0.058 & 0.36 & -0.047 & 0.133 & -0.921 \end{pmatrix} \tag{25}$$

A primeira linha (de cima para baixo) mostra o autovetor do primeiro componente principal, com seus coeficientes para cada medida característica. Quanto maior o coeficiente, em módulo, mais importante a medida característica para explicar a variabilidade dos dados e consequentemente para a distinção dos dados em *clusters*. Nota-se que no primeiro componente principal, as medidas mostram intensidades mais ou menos semelhantes. Apenas no segundo componente principal que é revelada a contribuição diferencial dessas medidas características: para o Cramer V médio, Theil U médio e Theil U máximo, seus valores estão muito baixos, próximos a zero, enquanto o Cramer V máximo e o Conteúdo de Informação médio mostram valores maiores, contribuindo mais para explicar os dados.

A matriz de componentes principais possui uma interpretabilidade dificultada e, por isso, utilizou-se uma técnica mais explicável: Árvore de Decisão (seção 4.4.2), para se conseguir extrair não somente as medidas mais importantes para decidir pelas classes $D < 0.4$ e $D \geq 0.4$, como também criar uma regra de decisão mais interpretável e de uso prático.

Para esclarecer a origem de cada ponto no PCA acima, o mesmo PCA pode ser visualizado com as cores representando o experimento-algoritmo ao invés do D de Cohen (figura 28). Nota-se que os *clusters* são formados basicamente de dados aplicando InMoDe e dados com aplicação do RSAT/STREME ou do JASPAR. O algoritmo 8-mer align E apresentou uma maior dispersão entre os *clusters*, o que era esperado de um algoritmo que não possui nem uma descoberta de motivos baseadas em PWM (RSAT/STREME) nem em modelos mais complexos que PWM (InMoDe).

Figura 28 – Análise de Componentes Principais bidimensional utilizando as 5 medidas características definidas para os 6 experimentos-algoritmos. O mesmo PCA da figura 27 é mostrado, mas com legendas indicando a qual experimento-algoritmo cada ponto pertence. Note que existem dois *clusters* principais. O primeiro constituído de dados do ENCODE InMoDe e de PBM InMoDe. O segundo constituído de dados do JASPAR, ENCODE RSAT/STREME e PBM RSAT/STREME. O PBM 8-mer align E ficou entre os dois *clusters*.



Fonte: Guilherme Miura Lavezzo (2021)

.

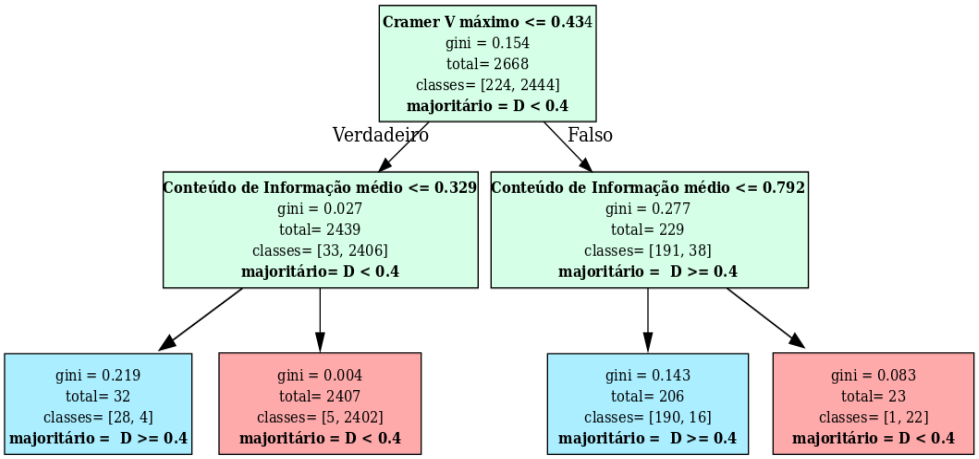
4.4.2 Árvore de Decisão

Uma vez calculadas e analisadas todas as cinco medidas características, é desejado obter uma regra que decida por um modelo ou outro, dado a sequência motivo e o conjunto de SLFTs. Como já mencionado anteriormente, ao considerar a classe $D < 0.4$, entende-se que PWM e LAPFA possuem desempenhos semelhantes. Logo, é preferível escolher PWMs

por efeito de praticidade. Além da técnica de PCA, foi também investigado se uma árvore de decisão poderia ser utilizada para realizar tal escolha, por possuir a vantagem de fornecer uma regra mais interpretável que o PCA.

A Árvore de Decisão inferida sobre o conjunto de 2668 conjunto de SLFTs é representada na figura 29. O objetivo da árvore era classificar as duas classes de maneira mais distinguível possível, com o menor número possível de etapas. A figura mostra, que, com apenas 2 níveis da árvore já é possível obter uma separação entre as duas classes.

Figura 29 – Árvore de Decisão criada sobre os 2668 conjunto de SLFTs visando distinguir as classes em que há semelhança ou não entre desempenho de modelos. Para a construção das árvores, apenas as cinco medidas características foram utilizadas (IC médio e as quatro de dependência – Theil U e Cramer V, ambos médio e máximo). Cada retângulo mostra uma regra de decisão que foi obtida pela árvore, com um critério considerado verdadeiro ou falso. Se o critério for verdadeiro, é seguido o caminho da esquerda (de cima para baixo), enquanto se falso o caminho da direita. O campo "classe" mostra a quantidade de elementos dos grupos $D >= 0.4$ e $D < 0.4$, respectivamente, antes da aplicação da regra no mesmo retângulo. Esses valores possuem a soma que resulta no indicado no campo "total". O campo "majoritário" representa a classe majoritária presente no retângulo. Note que inicialmente está indicada por $D < 0.4$ já que neste trabalho, esta classe é majoritária. O índice gini representa o grau de pureza da amostra considerada no retângulo. Quando o índice gini é igual a 0, há apenas a representação de uma classe, e quando é igual a 1, há o mesmo número de elementos de cada classe. Observe que com apenas dois níveis de decisões já foi possível, de maneira generalizada, separar as classes $D >= 0.4$ e $D < 0.4$.



Fonte: Guilherme Miura Lavezzo (2021)

.

Cada retângulo da árvore é a criação de uma decisão, com exceção dos retângulos "folhas", que são aqueles dos quais não partem novos retângulos para o nível inferior. A árvore inicia de cima para baixo, e deve ser seguida nesta ordem. Assim, a primeira decisão a ser tomada é se o Cramer V máximo, para o motivo em questão, é acima ou abaixo/igual a 0.434. Cada retângulo mostra informações adicionais, como o índice gini, que é um índice que mostra o grau de pureza do grupo representado pelo retângulo. O grau de pureza, informalmente, é o quanto há de mistura restante entre as duas classes. Quanto maior a mistura entre classes, maior o índice gini, variando de 0 a 1. Abaixo do índice gini, há também "total", o número de amostras consideradas antes da aplicação da decisão. "Classes" se refere ao número de elementos presentes em cada grupo, antes da aplicação da decisão e tal que a soma dos valores em "classes" resulta em "total". Como a árvore foi podada[1] após o segundo nível de decisão, os retângulos da base, destacados com as cores vermelho e azul não mais serão utilizados como decisão, mas para rotular os dados que ali incidirem. Isto é, se por exemplo, após a aplicação das duas primeiras regras, um dado incidir no primeiro retângulo (à esquerda) de cor azul, ele será classificado como sendo da classe $D \geq 0.4$, indicando que um modelo GRE-LAPFA deve ser utilizado.

Segundo a figura 29, a árvore de decisão pode ser utilizada da seguinte maneira para decidir pela escolha de PWM ou LAPFA. Para um dado conjunto de sítios de ligação de um fator de transcrição, calcula-se o conteúdo de informação médio e a medida de dependência Cramer V máximo do mesmo. A seguir, aplica-se a seguinte regra:

- se Cramer V máximo > 0.434 e o Conteúdo de Informação médio > 0.329, ou se Cramer V máximo ≤ 0.434 e o Conteúdo de Informação médio > 0.792, escolha PWM como modelo preditor;
- caso contrário, escolha o modelo GRE-LAPFA.

O fato da árvore de decisão basear-se nas características de Cramer V máximo e Conteúdo de Informação médio, sendo Cramer V máximo a primeira decisão a ser tomada, já informa sobre o valor dessas características em relação às demais. Mas é também possível calcular quantitativamente a importância delas por meio de uma medida de *feature importance*.

[1] Quando uma árvore de decisão é induzida até que os nós folhas possuam elementos apenas de uma classe, ela tende a perder poder de generalização, o que é conhecido como sobreajuste ou *overfitting*. Logo, faz parte do procedimento de indução de árvores de decisão realizar uma poda na árvore, que consiste em interromper a divisão de um nó quando, por exemplo, a mistura de classes contida nele fica abaixo de um limiar.

Feature importance é uma técnica que atribui um escore para cada medida característica (*feature*) utilizada, de modo que a melhor *feature* contribui mais para a capacidade preditora do modelo considerado. No caso de árvores de decisão, a medida de *feature importance* utilizada foi a *Gini importance*. A *Gini importance* é derivada da medida *Gini impurity* (G), que é definida como a probabilidade de uma classe i ocorrer no subconjunto em que foi aplicada a regra de decisão, multiplicado pela probabilidade de não ser essa classe. No caso, seja C o conjunto de classes e $p(i)$ a probabilidade empírica da classe i no subconjunto considerado, então define-se G como (JAMES; HASTIE; TIBSHIRANI, 2013):

$$G = \sum_{i \in C} p(i) \times (1 - p(i)) \tag{26}$$

A medida *Gini impurity* pode ser interpretada como a probabilidade de se escolher uma classe ao acaso, do subconjunto em que foi aplicado a regra de decisão, e essa classe ser incorretamente classificada. Valores baixos, próximos a zero, de G indicam predominância de uma classe naquele subconjunto.

O *Gini importance* de uma *feature* m é a soma total da redução de G (normalizada pela fração do tamanho do subconjunto[2] pelo tamanho total do conjunto) e dos seus nós filhos[3]. Feito isso para cada *feature*, o *Gini importance* é normalizado pelo número de *features* para ser exibido como uma porcentagem.

Seja τ o nó da árvore em questão, com τ_{dir} e τ_{esq} os nós filhos de τ. Seja q_{esq} e q_{dir} como a fração amostral de τ_{esq} e τ_{dir} resultado da divisão após a aplicação do critério de decisão no nó τ. Define-se $q_{filho} = \frac{n_{filho}}{n}$, com n o tamanho da amostra em τ e com $filho = \{esq, dir\}$, tal que $n_{esq} + n_{dir} = n$. Seja, $G^{\tau}(m)$ o *Gini impurity* localizado no nó τ para a *feature* m e, $G^{\tau_{esq}}$ e $G^{\tau_{dir}}$ como G dos nós filhos de $G^{\tau}(m)$. Define-se $\Delta^{\tau}(m)$ como sendo a redução da impureza, na *feature* m no nó τ da árvore, como (JIANG *et al.*, 2009):)

$$\Delta^{\tau}(m) = G^{\tau}(m) - q_{esq} G^{\tau_{esq}} - q_{dir} G^{\tau_{dir}} \tag{27}$$

[2] Cada aplicação de uma regra de decisão divide a amostra em dois subconjuntos, o primeiro em ser verdadeiro a regra, o segundo em ser falso.

[3] Essa medida foi calculada e implementada em Python, pelo pacote Scikit-learn (PEDREGOSA *et al.*, 2011)

, o *Gini importance* da *feature* m, IG(m), é definido como (JIANG *et al.*, 2009; MENZE *et al.*, 2009):

$$IG(m) = \sum_{\tau} \Delta^{\tau}(m) \tag{28}$$

Segundo a árvore de decisão construída, notou-se que as medidas mais importantes a se considerar para escolher por PWM ou LAPFA são Cramer V máximo e Conteúdo de Informação médio, como esperado segundo as regras de decisão criadas pela árvore na figura 29. Seus *Gini importances* foram respectivamente 77% e 23%, totalizando a importância das *features*. As medidas Cramer V médio, Theil U médio e máximo aparecem com nenhuma influência na tomada de decisões. Isso não significa na prática que essas últimas medidas mencionadas não sejam importantes. Na verdade, a importância do Cramer V máximo e do Conteúdo de Informação médio foi tamanha, que omitiu a necessidade de mais medidas para uma decisão que torne distinguível as classes consideradas. Além disso, a árvore de decisão toma decisões em ordem, isto é, uma vez optado pela decisão do Cramer V máximo, a separação dos dados resultante deve optar por uma nova decisão condicionada. Seria possível construir outras regras de decisão em ordens distintas, mas para os 2668 dados considerados, essas regras foram as que mais se ajustaram.

Os apêndices contêm as árvores de decisões induzidas a partir de dados exclusivamente de ChIP-seq (apêndice D) ou PBM (apêndice E). Nota-se que as regras são muito semelhantes e as duas medidas mais importantes são Cramer V máximo e Conteúdo de Informação médio.

Como também foi visto nos *Violin Plots* das figuras (23 e 26), as medidas Cramer V máximo e Conteúdo de Informação médio são medidas cujas distribuições e medianas conseguem separar bem as classes $D < 0.4$ e $D \geq 0.4$. Por isso, as medidas mais importantes obtidas pelo cálculo de *feature importance*, matriz de componentes principais, *effect size* entre classes e teste de Wilcoxon entre classes (apêndice C) também corroboram os achados.

5 Conclusão

Neste trabalho, o objetivo principal foi comparar modelos preditores de SLFTs com base na informação de dependência entre bases (GRE-LAPFA) ou não (PWM). Para tal, seis algoritmos de descoberta de motivos foram aplicados em três conjuntos de dados experimentais públicos, com o objetivo de adquirir conjuntos de SLFTs exatos para treinar e testar os modelos.

Observou-se que o desempenho de GREs-LAPFA é equivalente ou superior ao de PWMs, a depender da combinação experimento-algoritmo utilizada, mas principalmente se o algoritmo de descoberta de motivos não era baseado em PWMs. A diferença entre modelos era mais evidente calculando o D de Cohen, uma medida de *effect size*. Observou-se também que com o cálculo de medidas como Cramer V máximo e Conteúdo de Informação médio, era possível criar uma regra de decisão que faça a distinção entre optar por PWMs (classe $D < 0.4$) ou optar por GREs (classe $D \geq 0.4$). A matriz e o gráfico de duas dimensões do PCA também corroboram os achados, assim como é possível distinguir essas duas classes visualmente por *Violin Plots.*

Além disso, outro modelo GRE, o AMNESIA (RON; SINGER; TISHBY, 1997), inicialmente foi testado para uma parte do conjunto de dados e não se equiparou nem ao desempenho de PWMs nem de GRE-LAPFAS. Mais informações no apêndice A.

5.1 Principais contribuições

Este trabalho apresenta três principais contribuições que estão diretamente relacionadas com os três objetivos específicos propostos:

1. a proposta de formas de obtenção das sequências exatas de SLFTs para dados de ChIP-seq e PBM utilizando diferentes ferramentas de descoberta de motivos;
2. a proposta e avaliação de medidas do nível de dependência entre posições de bases em uma dada amostra;
3. a criação de uma regra de decisão que, dado um conjunto de SLFTs, escolhe qual modelo preditor, PWM ou GRE-LAPFA, deve ser mais adequado.

Com relação à primeira principal contribuição, este trabalho contribuiu para ajudar a comunidade científica a processar dados do tipo *high throughput* com intuito de obter

sequências motivos e depois modelos preditores de SLFTs. Como pôde ser visto, existem inúmeras maneiras de se coletar dados públicos, combinado com diversas maneiras de se obter sequências motivos. A escolha final é do pesquisador interessado, mas a maneira como ele(a) escolhe pode afetar a aplicação em predizer SLFTs acuradamente.

Foram exemplificadas algumas maneiras de se obter o conjunto de SLFTs exatos (seções 3.1, 3.2 e 3.3). Existem diversas outras maneiras para tal, mas aqui, as ideias foram adaptadas de métodos já implementados (FORNES *et al.*, 2019; WEIRAUCH *et al.*, 2013). Essas adaptações e o processo inteiro de obter SLFTs estão protocoladas por completo, tornando o processo facilmente reproduzível.

Existem diversos algoritmos de descoberta de motivos, alguns consideram dependência entre bases (InMoDe), outros não (RSAT e STREME). Por isso, este trabalho ajudou a investigar por motivos com diferentes características (com diferentes níveis de dependência), dependentes apenas do conjunto iniciais de dados experimentais.

Com relação à contribuição da proposta e avaliação de medidas do nível de dependência entre posições de bases em uma dada amostra, destaca-se que este trabalho propôs a utilização de quatro novas medidas de dependência que não tinham sido usadas antes pela literatura, que se saiba: Cramer V médio, Cramer V máximo, Theil U médio, Theil U máximo. Os autores Tomovic e Oakeley (2007) definiram medidas de dependência via teste de hipóteses χ^2, sob teste randomizado com Monte Carlo e sob a medida de Informação Mútua. Para os testes de hipóteses, ambas as formas possuem o problema principal de reportarem p-valores. Conforme múltiplos testes são realizados, a chance de uma hipótese nula ser rejeitada ao acaso aumenta consideravelmente (STOREY; TIBSHIRANI, 2003). O problema de teste múltiplos é conhecido na literatura (JAFARI; ANSARI-POUR, 2018; STOREY; TIBSHIRANI, 2003), e a correção dos p-valores acaba sendo uma medida paliativa para contornar o problema. Além disso, em particular o teste de χ^2, conforme o tamanho amostral aumenta, reduz a magnitude do p-valor. Isso fica evidente redefinindo a estatística de χ^2 para probabilidades empíricas em vez da frequência observada (ver equação 17). Com uma estatística aumentada, consequentemente o p-valor é menor, sem necessariamente implicar em relevância. Esse efeito acaba prejudicando a interpretação do teste de hipóteses. Para a Informação Mútua, o problema está no fato que o limite superior dessa medida está no número possível de valores que as duas variáveis podem assumir. No caso particular, as variáveis são posições da sequência motivo. Assim, existem no máximo quatro valores assumidos, isto é, dos quatro possíveis nucleotídeos. Porém,

algumas posições específicas da sequência motivo são muito conservadas. Logo, não são assumidas todas as possíveis bases em uma posição e, nesse caso particular, a informação mútua já não assume mais valores no mesmo intervalo e não são mais comparáveis entre diferentes pares de posições.

Esse trabalho também ajudou a esclarecer que o nível de dependência entre posições de bases de um conjunto de SLTFs diverge dependendo do algoritmo de descoberta de motivos utilizado. Com isso, a escolha do algoritmo de descoberta de motivos a partir de um experimento deve ser analisada previamente. Uma vez aplicado, o desempenho do modelo preditor de SLFTs pode ser afetado drasticamente, a depender do algoritmo considerado.

A terceira grande contribuição foi que, utilizando algumas técnicas estatísticas e de aprendizado de máquina, foi possível propor uma regra de decisão para escolher por um modelo preditor de SLFTs, com base nas características da amostra de SLTFs a ser utilizada como treinamento. Para os 2668 conjuntos de SLFTs aqui considerados, sob experimentos de ChIP-seq, PBM e alguns pouquíssimos de ChIP-chip, foi possível observar que com apenas duas medidas, Cramer V máximo e Conteúdo de Informação médio, já é possível distinguir a preferência por um modelo sobre outro.

Na prática, com as regras de decisão obtidas, o pesquisador não necessitaria testar e treinar exaustivamente diversos modelos preditores de SLFTs. Com o cálculo de algumas medidas como Cramer V máximo e Conteúdo de Informação média, já é possível ser orientado para optar ou não por PWMs. Dado um FT, existem diversas fontes de variabilidade para se obter um conjunto de SLFTs: a depender do tipo de experimento, da própria variabilidade experimental, o processamento dos dados e aplicação de um algoritmo de descoberta de motivos a critério do pesquisador. Que se saiba, e justamente pelas variabilidades mencionadas, não existem motivos bem definidos para um dado FT. Assim, a vantagem de se optar por um modelo preditor ou outro é independente do experimento e do algoritmo de descoberta de motivos escolhido, ou seja, sendo procedimentos flexíveis em seu uso.

De maneira geral, o público alvo para o qual este trabalho possa ter contribuído são o de pesquisadores, principalmente biólogos e bioinformatas que desejam predizer SLFTs ao longo de sequências de DNA (potencialmente do genoma todo). Ao se deparar na literatura com inúmeros algoritmos de descoberta de motivos e diferentes modelos classificadores, o pesquisador se vê na situação difícil de entender os conceitos, palavras

chaves e qual o melhor modelo a ser considerado. Existem diversos artigos na literatura que comparam o desempenho de PWMs, mas que se saiba, nenhuma conclusão foi obtida, pois a literatura ainda diverge muito no uso ou não de PWMs. Mais do que isso, artigos que ofertam modelos alternativos às PWMs costumam criar suas próprias metodologias para comparar o desempenho dos modelos.

5.2 Projetos Futuros

Idealmente, gostaria-se de que mais dados experimentais pudessem ser coletados e mais algoritmos de descoberta de motivos pudessem ter sido aplicados. Com um ainda maior número amostral, talvez possa ser criada uma regra mais universal que valha para experimentos diferentes, não apenas de ChIP-seq e PBM, mas também de HT-SELEX (*high throughput systematic evolution of ligands by exponential enrichment*) e outros. O HT-SELEX é uma técnica *in vitro* que testa a afinidade do FT com moléculas sintéticas de DNA de tamanho definido. A técnica faz uso de PCR para amplificar as sequências que obtiveram maior afinidade ao FT e, a cada ciclo refinar o processo. Infelizmente, dado o tempo de processamento e o número de etapas envolvidas, não foi possível ainda coletar mais dados. Com um maior volume de dados, seria possível aplicar o algoritmo aqui definido para tentar criar uma regra de decisão que possa universalmente optar pelo uso ou não de PWMs. O algoritmo definido neste trabalho pode ser facilmente estendido conforme mais dados são coletados.

Uma vez que possa ser medido o desempenho dos modelos, e também escolhido o melhor modelo preditor *in silico*, espera-se contribuir para uma aplicação em predizer Módulos cis-Reguladores (MCRs), principalmente do organismo *D. melanogaster* embrião blastoderma foco de nosso grupo de pesquisa. Apesar de existirem MCRs validados experimentalmente, a predição *in silico* de SLFTs ainda é essencial para obter com alta resolução a localização, a composição e quantidade dos SLFTs de cada FT de interesse, o que ainda não se sabe completamente. Realizando uma simples busca por MCRs em um banco de dados como o *REDfly* (RIVERA *et al.*, 2018), observa-se que majoritariamente MCRs validados não confirmam a presença de todos os FTs envolvidos e nem de todos os SLFTs funcionais. Essas etapas poderão futuramente contribuir com estudos sobre regulação transcricional de genes com maior precisão, uma vez que estudos experimentais

ainda não conseguem capturar com resolução de nucleotídeos todos os SLFTs envolvidos funcionalmente. A aplicação de estudos para predizer SLFTs e, consequentemente, inferir MCRs em *D. melanogaster* está em andamento pelo autor deste trabalho.

Além disso, é desejado também realizar um estudo de espaçamento entre SLFTs de diferentes (ou mesmos) FTs. Isso só é possível uma vez que tenham sido obtidos modelos mais acurados possíveis para obter SLFTs ao longo de sequências de DNA *in vivo*. Já existe uma ferramenta chamada SPAMO, cuja ideia é obter todos os SLFTs entre dois motivos e suas respetivas contagens para cada distância observada. Por hipótese, se dois FTs (distintos ou não) estão interagindo mediante um complexo proteico, a distância entre os dois SLFTs terá maior número de contagens do que ao acaso. O programa SPAMO testa isso com um teste hipóteses para cada distância. Apesar da ideia ser interessante, alguns ajustes, principalmente com relação às estatísticas empregadas, poderiam ser implementados. O fato do SPAMO testar cada distância faz com que o ajuste final de p-valor seja muito estringente. Além disso, fica difícil controlar p-valores reportados ao acaso conforme mais distâncias são testadas. Já que o p-valor é uma probabilidade, está passível de erro também e quanto mais testes são aplicados, maior a chance de se reportar um evento como sendo significativo quando na verdade não era. Alternativas à estratégia de contagens e ao teste de hipóteses podem melhorar e fornecer um programa superior com o mesmo propósito.

Este trabalho possui uma limitação que está principalmente no fato que GREs e o seu algoritmo de aprendizado LAPFA não são uma ferramenta de descoberta de motivos. Por isso, o modelo precisa estar acoplado a algoritmos que se propõem a essa etapa. Seria desejável que fosse implementada uma ferramenta de descoberta de motivo baseada no LAPFA, podendo assim, descobrir motivos enquanto treina um modelo preditor baseado em GRE. Isso seria o mais eficiente a ser feito e economizaria etapas de processamento computacionais que são só necessárias pelo fato do LAPFA precisar ser treinado com SLFTs de tamanhos exatos. Para tal, foi pensado em utilizar duas informações a *priori*. A primeira, é de que já existem algoritmos de descoberta de motivos implementados em PWMs, por exemplo uma técnica muito conhecida e utilzada para tal é a *Expectation Maximization*(EM) (BAILEY, 2002). Esse algoritmo tenta descobrir tanto a localização do motivo quanto sua composição, de maneira iterativa usando uma PWM. A segunda, é que uma vez que já existem tanto algoritmos implementados e PWMs treinadas em bancos de dados públicos como JASPAR, a tarefa de implementar uma descoberta de motivo

usando o LAPFA pode ser iniciada a partir do conhecimento a *priori* sobre o motivo. Isso facilitaria, por exemplo implementar algoritmos como EM.

Apesar de uma validação cruzada ter sido utilizada para testar o desempenho de dois modelos, existem outras alternativas não exploradas ainda. Alguns estudos sugerem que o teste de desempenho seja feito diretamente com dados *in vivo* ou produzidos sinteticamente (TOMPA *et al.*, 2005). Assumindo que sejam conhecidos os SLFTs e a localização de cada um, os modelos se encarregariam de tentar encontrá-los em sequências de DNA maiores, como de picos de ChIP-seq, enquanto evitam reportar falsos positivos.

Atualmente, com um maior número de técnicas experimentais, também é desejável acoplar experimentos que possam complementar os estudos. Apesar de não ter sido utilizado neste trabalho, experimentos que possam informar ocupação de histonas no DNA, como ATAC-seq e DNAse-seq, possa ser altamente valioso para inferir regiões de cromatina aberta ou fechada e que portanto possam inferir o estado transcricional de um gene próxima àquela coordenada genômica. Não só isso, mas dados de RNA-seq acoplados ao ChIP-seq podem, desde que sob as mesmas condições experimentais, auxiliar com a estratégia de inferir MCRs indicando se um gene está sendo transcrito ou não. Comparar duas ou mais condições experimentais é ainda mais assertivo. Por exemplo, com o silenciamento (*knockdown*) de um gene alvo ou FT alvo, pode-se sugerir o papel funcional deste alvo silenciado sobre possíveis genes regulados pelo alvo.

Apesar do trabalho ter dado foco a GREs aprendidas pelo algoritmo LAPFA, também podem ser testados outros modelos preditores sob as mesmas condições. Além disso, medidas de dependência poderiam ser estendidas para funcionarem com mais de duas posições aferidas, sugestão essa dada pelos autores Tomovic e Oakeley (2007) com a medida de Informação Mútua.

Referências[1]

ALBERTS, B.; JOHNSON, A.; LEWIS, J. Molecular biology of the cell. New York, NY: Garland Science, Taylor and Francis Group, 2015. ISBN 978-0815345244. Citado 2 vezes nas páginas 19 e 20.

AMEMIYA, H. M.; KUNDAJE, A.; BOYLE, A. P. The ENCODE blacklist: Identification of problematic regions of the genome. Scientific Reports, Springer Science and Business Media LLC, v. 9, n. 1, jun. 2019. Disponível em: ⟨https://doi.org/10.1038/s41598-019-45839-z⟩. Citado na página 57.

ANDERSSON, R.; SANDELIN, A. Determinants of enhancer and promoter activities of regulatory elements. Nature Reviews Genetics, Springer Science and Business Media LLC, v. 21, n. 2, p. 71–87, out. 2019. Disponível em: ⟨https://doi.org/10.1038/s41576-019-0173-8⟩. Citado 2 vezes nas páginas 20 e 22.

ANDRIOLI, L. P. M. *et al.* Anterior repression of a drosophila stripe enhancer requires three position-specific mechanisms. The Company of Biologists, v. 129, n. 21, p. 4931–4940, nov. 2002. Disponível em: ⟨https://doi.org/10.1242/dev.129.21.4931⟩. Citado na página 22.

BAILEY, T. L. Discovering novel sequence motifs with MEME. Wiley, v. 00, n. 1, nov. 2002. Disponível em: ⟨https://doi.org/10.1002/0471250953.bi0204s00⟩. Citado na página 102.

BAILEY, T. L. Streme: Accurate and versatile sequence motif discovery. 2020. Citado 2 vezes nas páginas 28 e 31.

BAILEY, T. L.; GRANT, C. E. SEA: Simple enrichment analysis of motifs. Cold Spring Harbor Laboratory, ago. 2021. Disponível em: ⟨https://doi.org/10.1101/2021.08.23.457422⟩. Citado na página 28.

BAILEY, T. L.; MACHANICK, P. Inferring direct DNA binding from ChIP-seq. Nucleic Acids Research, Oxford University Press (OUP), v. 40, n. 17, p. e128–e128, maio 2012. Disponível em: ⟨https://doi.org/10.1093/nar/gks433⟩. Citado 2 vezes nas páginas 28 e 58.

BULYK, M. L. *et al.* Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proceedings of the National Academy of Sciences, v. 98, n. 13, p. 7158–7163, jun. 2001. Disponível em: ⟨https://doi.org/10.1073/pnas.111163698⟩. Citado na página 80.

CASAMASSIMI, A.; CICCODICOLA, A. Transcriptional regulation: Molecules, involved mechanisms, and misregulation. International Journal of Molecular Sciences, MDPI AG, v. 20, n. 6, p. 1281, mar. 2019. Disponível em: ⟨https://doi.org/10.3390/ijms20061281⟩. Citado na página 19.

COHEN, J. Statistical Power Analysis for the behavioral sciences. [S.l.]: Academic Press, 1977. Citado na página 41.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

DABROWSKI, M. *et al.* Optimally choosing PWM motif databases and sequence scanning approaches based on ChIP-seq data. BMC Bioinformatics, Springer Science and Business Media LLC, v. 16, n. 1, maio 2015. Disponível em: ⟨https://doi.org/10.1186/s12859-015-0573-5⟩. Citado 3 vezes nas páginas 21, 43 e 44.

DEFRANCE, M. Local-word-analysis. 2011. Disponível em: ⟨http://rsat.sb-roscoff.fr/local-word-analysis_form.cgi⟩. Citado na página 30.

DONIGER, S. W. Identification of functional transcription factor binding sites using closely related saccharomyces species. Cold Spring Harbor Laboratory, v. 15, n. 5, p. 701–709, abr. 2005. Disponível em: ⟨https://doi.org/10.1101/gr.3578205⟩. Citado na página 22.

DUDA, R.; HART, P.; STORK, D. Pattern classification. New York: Wiley, 2001. ISBN 978-0471056690. Citado na página 33.

DUNKLER, D. *et al.* To test or to estimate? p-values versus effect sizes. Wiley, v. 33, n. 1, p. 50–55, out. 2019. Disponível em: ⟨https://doi.org/10.1111/tri.13535⟩. Citado na página 41.

EGGELING, R. Disentangling transcription factor binding site complexity. Nucleic Acids Research, Oxford University Press (OUP), ago. 2018. Disponível em: ⟨https://doi.org/10.1093/nar/gky683⟩. Citado 9 vezes nas páginas 16, 22, 26, 28, 32, 43, 44, 45 e 60.

FISHER, W. W. *et al.* DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in drosophila. Proceedings of the National Academy of Sciences, v. 109, n. 52, p. 21330–21335, dez. 2012. Disponível em: ⟨https://doi.org/10.1073/pnas.1209589110⟩. Citado na página 22.

FORNES, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research, Oxford University Press (OUP), nov. 2019. Disponível em: ⟨https://doi.org/10.1093/nar/gkz1001⟩. Citado 5 vezes nas páginas 16, 27, 57, 80 e 99.

FRIETZE, S.; FARNHAM, P. J. Transcription factor effector domains. In: Subcellular Biochemistry. Springer Netherlands, 2011. p. 261–277. Disponível em: ⟨https://doi.org/10.1007/978-90-481-9069-0_12⟩. Citado 2 vezes nas páginas 16 e 19.

FUREY, T. S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nature Reviews Genetics, Springer Science and Business Media LLC, v. 13, n. 12, p. 840–852, out. 2012. Disponível em: ⟨https://doi.org/10.1038/nrg3306⟩. Citado na página 26.

FURLONG, E. E. M.; LEVINE, M. Developmental enhancers and chromosome topology. Science, American Association for the Advancement of Science (AAAS), v. 361, n. 6409, p. 1341–1345, set. 2018. Disponível em: ⟨https://doi.org/10.1126/science.aau0320⟩. Citado na página 22.

GASPAR, J. M. Improved peak-calling with MACS2. Cold Spring Harbor Laboratory, dez. 2018. Disponível em: ⟨https://doi.org/10.1101/496521⟩. Citado na página 27.

GERTZ, J. *et al.* Genistein and bisphenol a exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. Genome Research, Cold Spring Harbor Laboratory, v. 22, n. 11, p. 2153–2162, set. 2012. Disponível em: ⟨https://doi.org/10.1101/gr.135681.111⟩. Citado na página 16.

GILBERT, S. Developmental biology. Oxford: Oxford University Press, 2018. ISBN 978-1605357386. Citado na página 15.

GRANT, C. E.; BAILEY, T. L.; NOBLE, W. S. FIMO: scanning for occurrences of a given motif. Bioinformatics, Oxford University Press (OUP), v. 27, n. 7, p. 1017–1018, fev. 2011. Disponível em: ⟨https://doi.org/10.1093/bioinformatics/btr064⟩. Citado na página 58.

GRAU, J.; NETTLING, M.; KEILWAGEN, J. DepLogo: visualizing sequence dependencies in r. Bioinformatics, Oxford University Press (OUP), v. 35, n. 22, p. 4812–4814, jun. 2019. Disponível em: ⟨https://doi.org/10.1093/bioinformatics/btz507⟩. Citado na página 44.

GRAY, S.; SZYMANSKI, P.; LEVINE, M. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes & Development, Cold Spring Harbor Laboratory, v. 8, n. 15, p. 1829–1838, ago. 1994. Disponível em: ⟨https://doi.org/10.1101/gad.8.15.1829⟩. Citado na página 22.

HALFON, M. S. Silencers, enhancers, and the multifunctional regulatory genome. Trends in Genetics, Elsevier BV, v. 36, n. 3, p. 149–151, mar. 2020. Disponível em: ⟨https://doi.org/10.1016/j.tig.2019.12.005⟩. Citado na página 22.

HELDEN, J. v. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Research, Oxford University Press (OUP), v. 28, n. 8, p. 1808–1818, abr. 2000. Disponível em: ⟨https://doi.org/10.1093/nar/28.8.1808⟩. Citado na página 30.

HELDEN, J. van; ANDRÉ, B.; COLLADO-VIDES, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies 1 1edited by g. von heijne. Journal of Molecular Biology, Elsevier BV, v. 281, n. 5, p. 827–842, set. 1998. Disponível em: ⟨https://doi.org/10.1006/jmbi.1998.1947⟩. Citado na página 30.

HERTZ, G. Z.; STORMO, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Oxford University Press (OUP), v. 15, n. 7, p. 563–577, jul. 1999. Disponível em: ⟨https://doi.org/10.1093/bioinformatics/15.7.563⟩. Citado na página 68.

HIGUERA, C. Grammatical inference : learning automata and grammars. Cambridge New York: Cambridge University Press, 2010. ISBN 978-0521763165. Citado 2 vezes nas páginas 45 e 47.

INUKAI, S.; KOCK, K. H.; BULYK, M. L. Transcription factor–DNA binding: beyond binding site motifs. Current Opinion in Genetics & Development, Elsevier BV, v. 43, p. 110–119, abr. 2017. Disponível em: ⟨https://doi.org/10.1016/j.gde.2017.02.007⟩. Citado 2 vezes nas páginas 16 e 27.

JAFARI, M.; ANSARI-POUR, N. Why, when and how to adjust your p values? Cell J (Yakhteh), Royan Institute, Iranian Academic Center for Education Culture and Research (ACECR), v. 20, n. 04, ago. 2018. Disponível em: ⟨https://doi.org/10.22074/cellj.2019.5992⟩. Citado na página 99.

JAMES, G.; HASTIE, T.; TIBSHIRANI, R. An introduction to statistical learning : with applications in R. New York, NY: Springer, 2013. ISBN 978-1461471370. Citado 2 vezes nas páginas 33 e 96.

JAYARAM, N.; USVYAT, D.; MARTIN, A. C. R. Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics, Springer Science and Business Media LLC, v. 17, n. 1, nov. 2016. Disponível em: ⟨https://doi.org/10.1186/s12859-016-1298-9⟩. Citado 2 vezes nas páginas 26 e 27.

JIANG, R. *et al.* A random forest approach to the detection of epistatic interactions in case-control studies. Springer Science and Business Media LLC, v. 10, n. S1, jan. 2009. Disponível em: ⟨https://doi.org/10.1186/1471-2105-10-s1-s65⟩. Citado 2 vezes nas páginas 96 e 97.

KATERENCHUK, D.; ROSENBERG, A. Rankdcg: Rank-ordering evaluation measure. CoRR, abs/1803.00719, 2018. Disponível em: ⟨http://arxiv.org/abs/1803.00719⟩. Citado na página 37.

KIM, H.-Y. Statistical notes for clinical researchers: Chi-squared test and fishers exact test. Restorative Dentistry & Endodontics, The Korean Academy of Conservative Dentistry, v. 42, n. 2, p. 152, 2017. Disponível em: ⟨https://doi.org/10.5395/rde.2017.42.2.152⟩. Citado na página 65.

KIM, S. *et al.* Probing allostery through DNA. Science, American Association for the Advancement of Science (AAAS), v. 339, n. 6121, p. 816–819, fev. 2013. Disponível em: ⟨https://doi.org/10.1126/science.1229223⟩. Citado na página 23.

KONING, A. P. J. de *et al.* Repetitive elements may comprise over two-thirds of the human genome. Public Library of Science (PLoS), v. 7, n. 12, p. e1002384, dez. 2011. Disponível em: ⟨https://doi.org/10.1371/journal.pgen.1002384⟩. Citado 2 vezes nas páginas 57 e 69.

KULAKOVSKIY, I. V.; MAKEEV, V. J. DNA sequence motif. Elsevier, p. 135–171, 2013. Disponível em: ⟨https://doi.org/10.1016/b978-0-12-411637-5.00005-6⟩. Citado na página 28.

LAMBERT, S. A. *et al.* The human transcription factors. Cell, Elsevier BV, v. 172, n. 4, p. 650–665, fev. 2018. Disponível em: ⟨https://doi.org/10.1016/j.cell.2018.01.029⟩. Citado 5 vezes nas páginas 15, 16, 19, 21 e 27.

LANDT, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research, Cold Spring Harbor Laboratory, v. 22, n. 9, p. 1813–1831, set. 2012. Disponível em: ⟨https://doi.org/10.1101/gr.136184.111⟩. Citado na página 27.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with bowtie 2. Nature Methods, Springer Science and Business Media LLC, v. 9, n. 4, p. 357–359, mar. 2012. Disponível em: ⟨https://doi.org/10.1038/nmeth.1923⟩. Citado na página 26.

LEVINE, M.; MANLEY, J. L. Transcriptional repression of eukaryotic promoters. Cell, Elsevier BV, v. 59, n. 3, p. 405–408, nov. 1989. Disponível em: ⟨https://doi.org/10.1016/0092-8674(89)90024-x⟩. Citado na página 23.

LIU, E. T.; POTT, S.; HUSS, M. Q&a: ChIP-seq technologies and the study of gene regulation. BMC Biology, Springer Science and Business Media LLC, v. 8, n. 1, maio 2010. Disponível em: ⟨https://doi.org/10.1186/1741-7007-8-56⟩. Citado na página 26.

MACARTHUR, S. *et al.* Developmental roles of 21 drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biology, Springer Science and Business Media LLC, v. 10, n. 7, p. R80, 2009. Disponível em: ⟨https://doi.org/10.1186/gb-2009-10-7-r80⟩. Citado na página 22.

MACHADO-LIMA, A.; KASHIWABARA, A.; DURHAM, A. Decreasing the number of false positives in sequence classification. BMC Genomics, Springer Science and Business Media LLC, v. 11, n. Suppl 5, p. S10, 2010. Disponível em: ⟨https://doi.org/10.1186/1471-2164-11-s5-s10⟩. Citado na página 40.

MAN, T.-K. Non-independence of mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Oxford University Press (OUP), v. 29, n. 12, p. 2471–2478, jun. 2001. Disponível em: ⟨https://doi.org/10.1093/nar/29.12.2471⟩. Citado na página 80.

MATHELIER, A.; WASSERMAN, W. W. The next generation of transcription factor binding site prediction. PLoS Computational Biology, Public Library of Science (PLoS), v. 9, n. 9, p. e1003214, set. 2013. Disponível em: ⟨https://doi.org/10.1371/journal.pcbi.1003214⟩. Citado na página 44.

MENZE, B. H. *et al.* A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. Springer Science and Business Media LLC, v. 10, n. 1, jul. 2009. Disponível em: ⟨https://doi.org/10.1186/1471-2105-10-213⟩. Citado na página 97.

MIRNY, L. A. Nucleosome-mediated cooperativity between transcription factors. Proceedings of the National Academy of Sciences, Proceedings of the National Academy of Sciences, v. 107, n. 52, p. 22534–22539, dez. 2010. Disponível em: ⟨https://doi.org/10.1073/pnas.0913805107⟩. Citado na página 23.

MOORMAN, C. *et al.* Hotspots of transcription factor colocalization in the genome of drosophila melanogaster. Proceedings of the National Academy of Sciences, Proceedings of the National Academy of Sciences, v. 103, n. 32, p. 12027–12032, jul. 2006. Disponível em: ⟨https://doi.org/10.1073/pnas.0605003103⟩. Citado na página 23.

MORGUNOVA, E.; TAIPALE, J. Structural perspective of cooperative transcription factor binding. Current Opinion in Structural Biology, Elsevier BV, v. 47, p. 1–8, dez. 2017. Disponível em: ⟨https://doi.org/10.1016/j.sbi.2017.03.006⟩. Citado na página 23.

NAKATO, R.; SHIRAHIGE, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Briefings in Bioinformatics, Oxford University Press (OUP), p. bbw023, mar. 2016. Disponível em: ⟨https://doi.org/10.1093/bib/bbw023⟩. Citado na página 26.

NETO, A. F. Predição computacional de sítios de ligação de fatores de transcrição baseada em gramáticas regulares estocásticas. Tese (Mestrado em Bioinformática) — Universidade de São Paulo, São Paulo, jan. 2018. Disponível em: ⟨http://www.teses.usp.br/teses/disponiveis/95/95131/tde-02012018-144349/⟩. Citado na página 17.

NGUYEN, N. T. T. *et al.* RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Research, Oxford University Press (OUP), v. 46, n. W1, p. W209–W214, maio 2018. Disponível em: ⟨https://doi.org/10.1093/nar/gky317⟩. Citado na página 28.

NüSSLEIN-VOLHARD, C.; WIESCHAUS, E. Mutations affecting segment number and polarity in drosophila. Nature, Springer Science and Business Media LLC, v. 287, n. 5785, p. 795–801, out. 1980. Disponível em: ⟨https://doi.org/10.1038/287795a0⟩. Citado na página 15.

PAPATSENKO, D.; GOLTSEV, Y.; LEVINE, M. Organization of developmental enhancers in the drosophila embryo. Nucleic Acids Research, Oxford University Press (OUP), v. 37, n. 17, p. 5665–5677, ago. 2009. Disponível em: ⟨https://doi.org/10.1093/nar/gkp619⟩. Citado na página 23.

PARK, P. J. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics, Springer Science and Business Media LLC, v. 10, n. 10, p. 669–680, set. 2009. Disponível em: ⟨https://doi.org/10.1038/nrg2641⟩. Citado na página 26.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Citado 3 vezes nas páginas 37, 73 e 96.

PRESS, W. Numerical recipes : the art of scientific computing. Cambridge, UK New York: Cambridge University Press, 2007. ISBN 978-0-521-88068-8. Citado na página 66.

QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, Oxford University Press (OUP), v. 26, n. 6, p. 841–842, jan. 2010. Disponível em: ⟨https://doi.org/10.1093/bioinformatics/btq033⟩. Citado 2 vezes nas páginas 57 e 60.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: _____. Encyclopedia of Database Systems. Boston, MA: Springer US, 2009. p. 532–538. ISBN 978-0-387-39940-9. Disponível em: ⟨https://doi.org/10.1007/978-0-387-39940-9_565⟩. Citado na página 38.

RIVERA, J. *et al.* REDfly: the transcriptional regulatory element database for drosophila. Oxford University Press (OUP), v. 47, n. D1, p. D828–D834, out. 2018. Disponível em: ⟨https://doi.org/10.1093/nar/gky957⟩. Citado na página 101.

ROGERS, J. M.; BULYK, M. L. Diversification of transcription factor-DNA interactions and the evolution of gene regulatory networks. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, Wiley, v. 10, n. 5, p. e1423, abr. 2018. Disponível em: ⟨https://doi.org/10.1002/wsbm.1423⟩. Citado na página 22.

RON, D.; SINGER, Y.; TISHBY, N. The power of amnesia: Learning probabilistic automata with variable memory length. Machine Learning, Springer Science and Business Media LLC, v. 25, n. 2-3, p. 117–149, 1997. Disponível em: ⟨https://doi.org/10.1007/bf00114008⟩. Citado 5 vezes nas páginas 45, 46, 47, 98 e 114.

RON, D.; SINGER, Y.; TISHBY, N. On the learnability and usage of acyclic probabilistic finite automata. Journal of Computer and System Sciences, Elsevier BV, v. 56, n. 2, p. 133–152, abr. 1998. Disponível em: ⟨https://doi.org/10.1006/jcss.1997.1555⟩. Citado 5 vezes nas páginas 45, 46, 47, 53 e 114.

ROSENFELD, M. G. Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. Genes & Development, Cold Spring Harbor Laboratory, v. 20, n. 11, p. 1405–1428, jun. 2006. Disponível em: ⟨https://doi.org/10.1101/gad.1424806⟩. Citado na página 16.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, Public Library of Science (PLoS), v. 10, n. 3, p. e0118432, mar. 2015. Disponível em: ⟨https://doi.org/10.1371/journal.pone.0118432⟩. Citado na página 37.

SAKAKIBARA, Y. Grammatical inference: An old and new paradigm. In: Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1995. p. 1–24. Disponível em: ⟨https://doi.org/10.1007/3-540-60454-5_25⟩. Citado na página 46.

SAWILOWSKY, S. S. New effect size rules of thumb. Wayne State University Library System, v. 8, n. 2, p. 597–599, nov. 2009. Disponível em: ⟨https://doi.org/10.22237/jmasm/1257035100⟩. Citado na página 41.

SCHMITGES, F. W. *et al.* Multiparameter functional diversity of human c2h2 zinc finger proteins. Genome Research, Cold Spring Harbor Laboratory, v. 26, n. 12, p. 1742–1752, nov. 2016. Disponível em: ⟨https://doi.org/10.1101/gr.209643.116⟩. Citado na página 16.

SCHOENFELDER, S.; FRASER, P. Long-range enhancer–promoter contacts in gene expression control. Nature Reviews Genetics, Springer Science and Business Media LLC, v. 20, n. 8, p. 437–455, maio 2019. Disponível em: ⟨https://doi.org/10.1038/s41576-019-0128-0⟩. Citado na página 20.

SHARPE, D. Chi-square test is statistically significant: Now what? University of Massachusetts Amherst, 2015. Disponível em: ⟨https://scholarworks.umass.edu/pare/vol20/iss1/8/⟩. Citado na página 65.

SIEVERS, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. EMBO, v. 7, n. 1, p. 539, jan. 2011. Disponível em: ⟨https://doi.org/10.1038/msb.2011.75⟩. Citado na página 62.

SIPSER, M. Introduction to the theory of computation. Boston, MA: Cengage Learning, 2013. ISBN 978-1133187790. Citado 2 vezes nas páginas 45 e 47.

SMALL, S. *et al.* Transcriptional regulation of a pair-rule stripe in drosophila. Cold Spring Harbor Laboratory, v. 5, n. 5, p. 827–839, maio 1991. Disponível em: ⟨https://doi.org/10.1101/gad.5.5.827⟩. Citado na página 22.

SPITZ, F.; FURLONG, E. E. M. Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics, Springer Science and Business Media LLC, v. 13, n. 9, p. 613–626, ago. 2012. Disponível em: ⟨https://doi.org/10.1038/nrg3207⟩. Citado na página 20.

STADEN, R. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Research, Oxford University Press (OUP), v. 12, n. 1Part2, p. 505–519, 1984. Disponível em: ⟨https://doi.org/10.1093/nar/12.1part2.505⟩. Citado na página 41.

STOREY, J. D.; TIBSHIRANI, R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, v. 100, n. 16, p. 9440–9445, jul. 2003. Disponível em: ⟨https://doi.org/10.1073/pnas.1530509100⟩. Citado na página 99.

SU, W.; YUAN, Y.; ZHU, M. A relationship between the average precision and the area under the ROC curve. ACM, set. 2015. Disponível em: ⟨https://doi.org/10.1145/2808194.2809481⟩. Citado na página 37.

SULLIVAN, G. M.; FEINN, R. Using effect size—or why the p value is not enough. Journal of Graduate Medical Education, v. 4, n. 3, p. 279–282, set. 2012. Disponível em: ⟨https://doi.org/10.4300/jgme-d-12-00156.1⟩. Citado na página 41.

SURYAMOHAN, K.; HALFON, M. S. Identifying transcriptional cis -regulatory modules in animal genomes. Wiley, v. 4, n. 2, p. 59–84, dez. 2014. Disponível em: ⟨https://doi.org/10.1002/wdev.168⟩. Citado na página 22.

THOMAS-CHOLLIER, M. *et al.* RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Research, Oxford University Press (OUP), v. 40, n. 4, p. e31–e31, dez. 2011. Disponível em: ⟨https://doi.org/10.1093/nar/gkr1104⟩. Citado na página 30.

TOMOVIC, A.; OAKELEY, E. J. Position dependencies in transcription factor binding sites. Bioinformatics, Oxford University Press (OUP), v. 23, n. 8, p. 933–941, fev. 2007. Disponível em: ⟨https://doi.org/10.1093/bioinformatics/btm055⟩. Citado 6 vezes nas páginas 16, 39, 45, 64, 99 e 103.

TOMPA, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology, Springer Science and Business Media LLC, v. 23, n. 1, p. 137–144, jan. 2005. Disponível em: ⟨https://doi.org/10.1038/nbt1053⟩. Citado na página 103.

UDALOVA, I. A. *et al.* Quantitative prediction of NF- b DNA- protein interactions. Proceedings of the National Academy of Sciences, v. 99, n. 12, p. 8167–8172, jun. 2002. Disponível em: ⟨https://doi.org/10.1073/pnas.102674699⟩. Citado na página 80.

VIEIRA, D.; DURHAM, A. Geração de classificadores de sequências genéticas utilizando inferência de linguagens regulares. 01 2005. Citado 2 vezes nas páginas 47 e 53.

VOSS, T. C.; HAGER, G. L. Dynamic regulation of transcriptional states by chromatin and transcription factors. Nature Reviews Genetics, Springer Science and Business Media LLC, v. 15, n. 2, p. 69–81, dez. 2013. Disponível em: ⟨https://doi.org/10.1038/nrg3623⟩. Citado na página 16.

WASSERMAN, W. W.; SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics, Springer Science and Business Media LLC, v. 5, n. 4, p. 276–287, abr. 2004. Disponível em: ⟨https://doi.org/10.1038/nrg1315⟩. Citado 4 vezes nas páginas 16, 42, 43 e 44.

WEIRAUCH, M. T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. Nature Biotechnology, Springer Science and Business Media LLC, v. 31, n. 2, p. 126–134, jan. 2013. Disponível em: ⟨https://doi.org/10.1038/nbt.2486⟩. Citado 8 vezes nas páginas 16, 43, 44, 54, 60, 61, 75 e 99.

WOLFE, S. A. *et al.* Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code11edited by p. e. wright. Journal of Molecular Biology, v. 285, n. 5, p. 1917–1934, 1999. ISSN 0022-2836. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0022283698924214⟩. Citado na página 80.

XIA, X. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica, Hindawi Limited, v. 2012, p. 1–15, 2012. Disponível em: ⟨https://doi.org/10.6064/2012/917540⟩. Citado 2 vezes nas páginas 42 e 67.

YANG, C.; CHANG, C.-H. Exploring comprehensive within-motif dependence of transcription factor binding in escherichia coli. Springer Science and Business Media LLC, v. 5, n. 1, nov. 2015. Disponível em: ⟨https://doi.org/10.1038/srep17021⟩. Citado 2 vezes nas páginas 77 e 82.

ZHAO, Y. *et al.* Improved models for transcription factor binding site identification using nonindependent interactions. Genetics, Genetics Society of America, v. 191, n. 3, p. 781–790, abr. 2012. Disponível em: ⟨https://doi.org/10.1534/genetics.112.138685⟩. Citado na página 44.

ZHAO, Y.; STORMO, G. D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nature Biotechnology, Springer Science and Business Media LLC, v. 29, n. 6, p. 480–483, jun. 2011. Disponível em: ⟨https://doi.org/10.1038/nbt.1893⟩. Citado na página 44.

Apêndices

Apêndice A – AMNESIA como modelo preditor de SLFTs

AMNESIA é um outro algoritmo para aprendizado de gramáticas regulares estocásticas (RON; SINGER; TISHBY, 1997; RON; SINGER; TISHBY, 1998). Brevemente, o AMNESIA é um modelo baseado em Cadeias de Markov de alcance variável estacionárias. Em outras palavras, é um modelo que não consegue capturar informações da posição da sequência motivo, e por isso, não obteve vantagem nenhuma em relação à PWM e ao LAPFA, ambos modelos que possuem essa característica. Por esse motivo, seus resultados foram excluídos das análises posteriores. Para efeito de completude, tais resultados são sumarizados nesta seção.

Na comparação entre os três modelos, foram considerados 2039 FTs provenientes do ENCODE e 288 FTs do JASPAR, totalizando 2327 FTs em que todos os modelos foram treinados e testados (vide algoritmos 2 e 3). Para vias de comparação, segue uma tabela contendo os modelos com maior AP médio (entre 7 folds) para cada FT, isto é, os modelos com melhor performance.

Tabela 1 – Contagem de melhores modelos selecionados via *Average Precision* médio para 2327 fatores de transcrição

Modelo(s) com maior(es) AP média por fator de transcrição	**Contagem**
LAPFA	1471
PWM	832
LAPFA e PWM	22
LAPFA, PWM e AMNESIA	2
TOTAL	**2327**

Inicialmente, um número menor de dados experimentais tinham sido obtidos, todos referentes a dados *in vivo* do ENCODE e do JASPAR. Na maioria dos casos (1471 FTs), o LAPFA foi considerado o melhor modelo, a PWM foi considerada o melhor modelo em 832 dados de FTs. Houve 22 empates entre LAPFA e PWM e 2 empates entre LAPFA, PWM e AMNESIA.

Uma vez que o AMNESIA obteve o pior desempenho para praticamente todos os 2327 dados de FTs considerados, as análises posteriores que seguem focam apenas na comparação direta entre PWM e LAPFA.

Apêndice B – Testes de Hipóteses e *effect size* de medidas características comparadas

Nesta seção, estão todas as medidas comparativas referentes a cada figura *Violin Plot* mencionada no capítulo de resultados. Cada teste de hipótese e D de Cohen são a comparação entre todos os pares possíveis de experimento-algoritmos, ou seja, 15. Isto é, uma combinação de pares de todos os 6 experimentos-algoritmos($C(6,2) = 15$). A medida de D de Cohen é calculada a partir das diferenças das médias de cada grupo do tipo experimento-algoritmo. Já, o teste de Wilcoxon verifica se a mediana de cada par testado é igual ou diferente entre si. O p-valor do teste de Wilcoxon está em notação científica por ser um valor muito abaixo de 0 na maioria das vezes.

Tabela 2 – Comparação via teste de Wilcoxon e D de Cohen entre ENCODE InMoDe e PBM InMoDe

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	0.118	3.486e-2
Cramer V máximo	0.292	9.561e-2
Theil U médio	0.346	7.707e-1
Theil U máximo	0.773	3.144e-5
Conteúdo de Informação médio	0.838	7.438e-6

Tabela 3 – Comparação via teste de Wilcoxon e D de Cohen entre ENCODE InMoDe e ENCODE RSAT/STREME

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	0.957	5.00e-30
Cramer V máximo	1.705	8.02e-52
Theil U médio	0.412	7.01e-29
Theil U máximo	0.358	1.83e-12
Conteúdo de Informação médio	3.058	1.69e-71

Tabela 4 – Comparação via teste de Wilcoxon e D de Cohen entre JASPAR e ENCODE RSAT/STREME

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	0.038	1.72e-2
Cramer V máximo	0.038	1.66e-1
Theil U médio	0.077	6.04e-1
Theil U máximo	0.039	6.89e-1
Conteúdo de Informação médio	0.060	5.91e-1

Tabela 5 – Comparação via teste de Wilcoxon e D de Cohen entre PBM 8-mer align e ENCODE RSAT/STREME

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	2.786	2.14e-40
Cramer V máximo	2.071	1.28e-32
Theil U médio	2.381	4.66e-36
Theil U máximo	1.521	2.97e-20
Conteúdo de Informação médio	0.330	8.87e-3

Tabela 6 – Comparação via teste de Wilcoxon e D de Cohen entre PBM InMoDe e ENCODE RSAT/STREME

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	0.832	2.84e-15
Cramer V máximo	1.917	2.48e-33
Theil U médio	0.686	1.54e-18
Theil U máximo	0.782	5.45e-13
Conteúdo de Informação médio	2.465	3.31e-42

Tabela 7 – Comparação via teste de Wilcoxon e D de Cohen entre ENCODE InMoDe e JASPAR

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	1.000	6.11e-22
Cramer V máximo	1.787	1.22e-36
Theil U médio	0.568	2.67e-26
Theil U máximo	0.453	1.74e-11
Conteúdo de Informação médio	4.004	1.46e-54

Tabela 8 – Comparação via teste de Wilcoxon e D de Cohen entre JASPAR e PBM 8-mer align

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	2.675	6.86e-34
Cramer V máximo	2.058	7.96e-28
Theil U médio	2.117	3.42e-32
Theil U máximo	1.597	5.08e-20
Conteúdo de Informação médio	0.451	1.47e-3

Tabela 9 – Comparação via teste de Wilcoxon e D de Cohen entre JASPAR e PBM InMoDe

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	0.821	1.84e-13
Cramer V máximo	1.879	1.57e-26
Theil U médio	0.781	1.35e-18
Theil U máximo	0.866	2.81e-12
Conteúdo de Informação médio	2.929	1.02e-36

Tabela 10 – Comparação via teste de Wilcoxon e D de Cohen entre PBM 8-mer align e ENCODE InMoDe

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	1.952	1.50e-21
Cramer V máximo	0.505	6.97e-3
Theil U médio	1.614	9.80e-20
Theil U máximo	1.776	2.61e-16
Conteúdo de Informação médio	4.385	1.53e-30

Tabela 11 – Comparação via teste de Wilcoxon e D de Cohen entre PBM 8-mer align e PBM InMoDe

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	1.734	1.22e-15
Cramer V máximo	0.157	3.14e-1
Theil U médio	1.128	2.40e-11
Theil U máximo	0.871	1.46e-6
Conteúdo de Informação médio	2.917	7.03e-24

Tabela 12 – Comparação via teste de Wilcoxon e D de Cohen entre PBM RSAT/STREME e ENCODE InMoDe

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	0.550	1.09e-2
Cramer V máximo	0.913	1.56e-12
Theil U médio	0.829	6.05e-4
Theil U máximo	1.642	2.93e-19
Conteúdo de Informação médio	4.885	1.26e-30

Tabela 13 – Comparação via teste de Wilcoxon e D de Cohen entre PBM RSAT/STREME e ENCODE RSAT/STREME

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	1.489	1.87e-26
Cramer V máximo	1.188	5.89e-9
Theil U médio	1.278	1.27e-25
Theil U máximo	1.141	1.85e-8
Conteúdo de Informação médio	0.759	1.22e-8

Tabela 14 – Comparação via teste de Wilcoxon e D de Cohen entre PBM RSAT/STREME e JASPAR

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	1.417	9.40e-21
Cramer V máximo	1.268	4.68e-10
Theil U médio	1.253	3.86e-24
Theil U máximo	1.270	5.23e-10
Conteúdo de Informação médio	0.943	1.85e-9

Tabela 15 – Comparação via teste de Wilcoxon e D de Cohen entre PBM RSAT/STREME e PBM 8-mer align

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	1.096	2.05e-9
Cramer V máximo	1.348	3.92e-14
Theil U médio	0.657	7.21e-6
Theil U máximo	0.493	5.47e-6
Conteúdo de Informação médio	0.405	1.69e-2

Tabela 16 – Comparação via teste de Wilcoxon e D de Cohen entre PBM RSAT/STREME e PBM InMoDe

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	0.558	6.28e-5
Cramer V máximo	1.019	2.48e-9
Theil U médio	0.446	4.86e-3
Theil U máximo	0.519	2.79e-4
Conteúdo de Informação médio	3.331	9.42e-25

Como observação geral, note que o Cohen D e o p-valor do teste de Wilcoxon tendem a ser inversamente proporcionais: quanto maior a diferença entre as médias dos grupos avaliados, menor tende a ser o p-valor.

Além disso, as maiores medidas D e menores p-valores estão concentrados quando a comparação está entre os algoritmos InMoDe e RSAT/STREME ou InMoDe e JASPAR. Essa diferença pode ser visualmente conferida nos *Violin Plots* pela posição das distribuições e as medianas de cada grupo.

Observa-se também, como determinado pela Árvore de Decisão (seção 4.4.2), que em geral o maior Cohen D e os menores p-valores são das medidas de Cramer V máximo e de Conteúdo de Informação médio.Os valores de Cohen D para essas medidas costumam se apresentar acima de 1, o que significa um altíssimo grau de diferença entre as médias dos dois grupos, assim como o p-valor se mostra como drasticamente menor que 0.05. Nenhum p-valor foi corrigido por algum critério de correção.

Apêndice C – Testes de Hipóteses e *effect size* de medidas características para as classes $D < 0.4$ e $D \geq 0.4$

Neste apêndice, foram realizados testes de hipóteses e medidas de D de Cohen para cada grupo plotado no Violin Plot da seção 4.3.2, referente a comparação entre as classes $D < 04$ e $D \geq 0.4$. A medida de D de Cohen é calculada a partir das diferenças das médias, de cada *feature*, entre as classes $D < 0.4$ e $D \geq 0.4$. Já, o teste de Wilcoxon verifica se a mediana das classes, para cada *feature*, é igual ou diferente entre si.

Tabela 17 – Comparação via teste de Wilcoxon e D de Cohen entre $D < 0.4$ e $D \geq 0.4$

Medidas características	**Cohen D**	**Wilcoxon rank test p-valor**
Cramer V médio	1.27	9.993e-63
Cramer V máximo	1.96	3.102e-108
Theil U médio	0.93	1.150e-64
Theil U máximo	0.82	9.549e-35
Conteúdo de Informação médio	2.21	1.322e-85

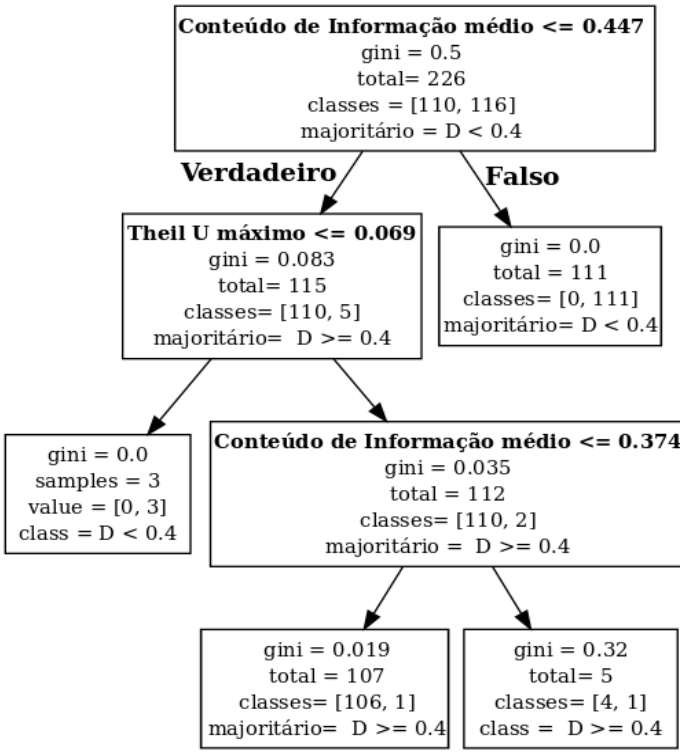
Observou-se que, todas as medidas possuem um alto valor de *effect size*, assim como todos os testes de hipóteses confirmam p-valores muito abaixos de 0.05, corroborando com o evidenciado pelos Violin Plots.

Apêndice D – Árvore de Decisão para dados do ENCODE

Assim como na seção 4.4.2 em que desejava-se criar uma regra de decisão para separar os modelos GRE-LAPFA e PWM segundo as medidas características, esta seção tem o mesmo objetivo, mas restringindo-se aos dados obtidos com um mesmo experimento (ENCODE ou PWM) para verificar se há uma diferença das regras induzidas. A justificativa aqui é ser mais específico para o tipo de dado. Na seção 4.4.2, foi visto que mesmo a união de todos os 2668 motivos permitiu criar uma regra de decisão unificada. Mas também é possível replicar um algoritmo de árvore de decisão que consiga separar os modelos em dados específicos sem generalizar pela diversidade experimental.

Na figura 34, foram considerados 113 FTs do ENCODE, em que foram aplicados os algoritmos RSAT/STREME e o InMoDe, portanto 226 motivos obtidos. Dessa quantia, 110 estavam na classe $D >= 0.4$ e 116 em $D < 0.4$.

Figura 30 – Árvore de Decisão utilizando 113 FTs do ENCODE para os algoritmos RSAT/STREME e InMoDe. Estão mostrados somente os três níveis da árvore. Note que inicialmente, 110 motivos foram considerados com $D >= 0.4$ e 116 com $D < 0.4$.



Fonte: Guilherme Miura Lavezzo (2021)

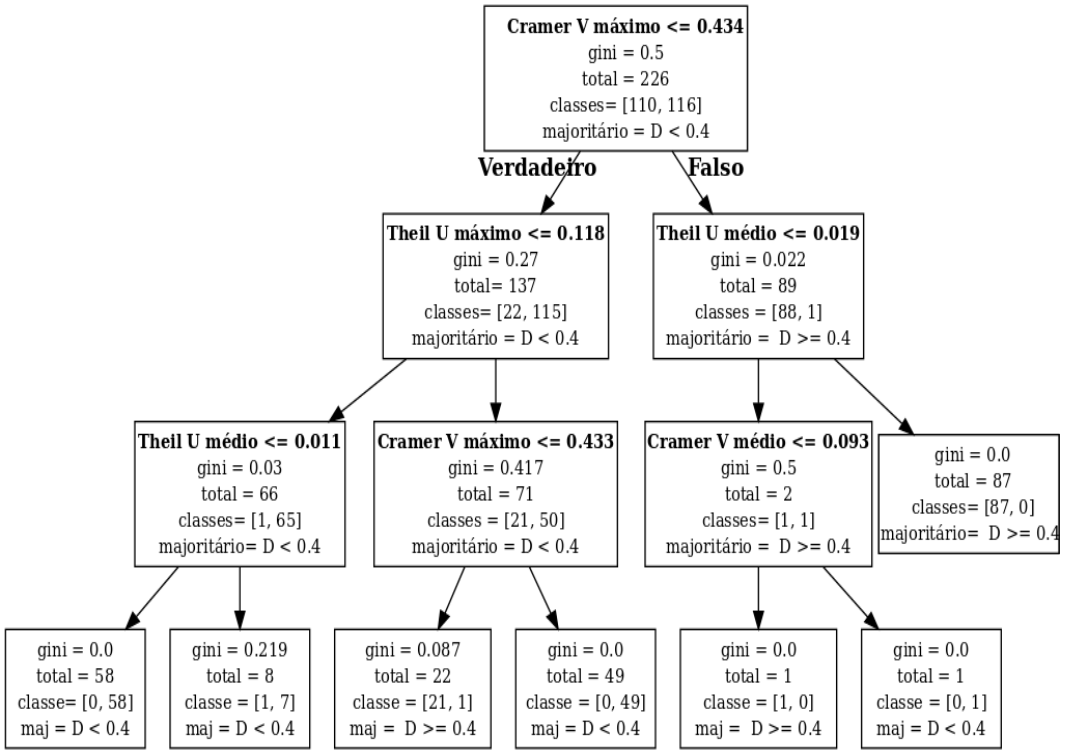
.

As medidas mais importantes (*feature importance*) em porcentagem são Conteúdo de Informação médio (95%)e Cramer V máximo (5%).

Foi observado que para dados do ENCODE, a característica mais importante que ajuda a decidir entre o modelo PWM e GRE-LAPFA está principalmente no Conteúdo de Informação médio. Isso, pois para o algoritmo InMoDe, a representação de motivo precisou ser estendida, uma vez que o modelo consegue capturar dependência entre posições de bases que não são visualizadas na sequência motivo original. Modelos que conseguem capturar a dependência apresentam uma clara vantagem sobre PWMs ao se utiliar o InMoDe. Isso pode ser observado pelo *effect size* de $D > 0.4$.

Para mostrar que o Conteúdo de Informação médio está mascarando a importância da medida Cramer V máximo, uma segunda árvore de decisão foi construída a partir das medidas características utilizadas neste trabalho exceto pelo Conteúdo de Informação médio (figura 31).

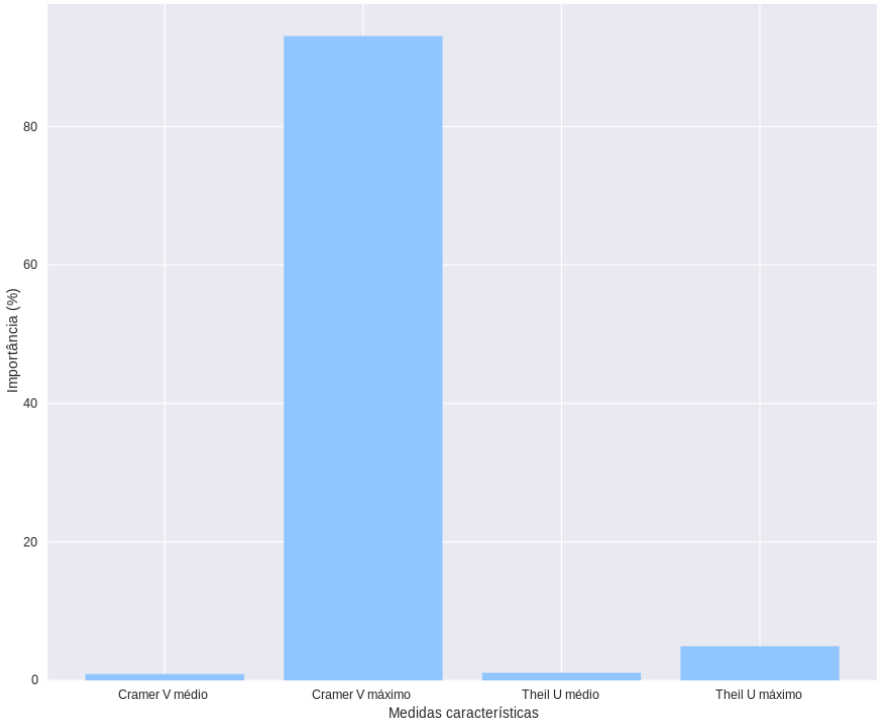
Figura 31 – Árvore de Decisão utilizando 113 FTs do ENCODE para os algoritmos RSAT/STREME e InMoDe, com remoção da medida Conteúdo de Informação médio. Estão mostrados somente os três níveis da árvore. Note que assim que a medida Conteúdo de Informação médio é removido, o Cramer V máximo aparece como a medida mais importante para a regra de decisão.



Fonte: Guilherme Miura Lavezzo (2021)

.

Figura 32 – Medidas características mais importantes para 113 FTs do ENCODE, com exclusão da medida Conteúdo de Informação médio. Note como a exclusão do Conteúdo de Informação médio faz com que a medida mais importante (em torno de 95%) seja o Cramer V máximo.



Fonte: Guilherme Miura Lavezzo (2021)

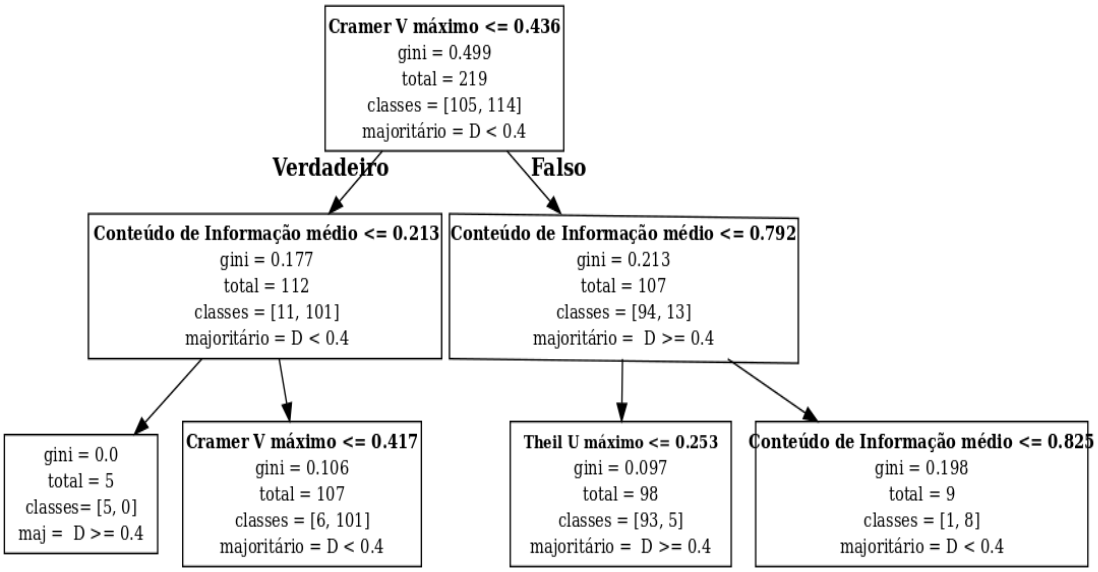
.

Foi observado que quando se é excluído de análise a medida Conteúdo de Informação médio, a medida Cramer V máximo se sobressai, com mais de 95% de importância. Como mostrado na seção 4.4.2, esses resultados são condizentes com os achados na união de todos os 2668 motivos obtidos de diversos experimentos-algoritmos.

Apêndice E – Árvore de Decisão para dados de PBM

Da mesma maneira que foi implementada uma árvore de decisão para apenas os dados do ENCODE, também o foi para dados de PBM. Nos dados de PBM, três experimentos-algoritmos foram utilizados, com 105 pertencentes à classe $D >= 0.4$ e 114 pertencentes à $D < 0.4$.

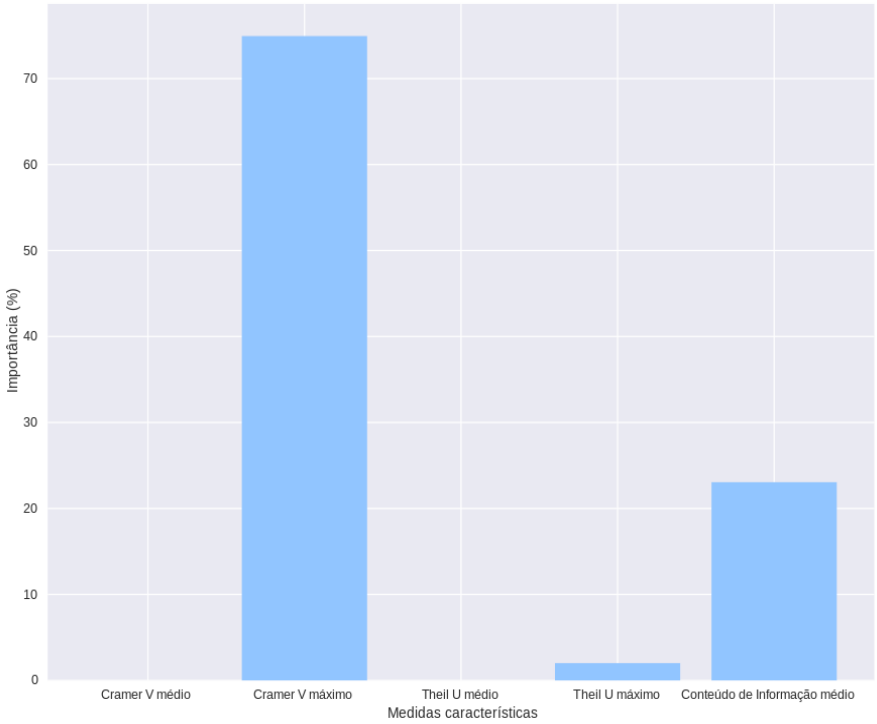
Figura 33 – Árvore de Decisão utilizando 73 FTs de PBM para os algoritmos RSAT/STREME, InMoDe e 8-mer align E. Estão mostrados somente os três níveis da árvore. Note que as medidas Conteúdo de Informação médio e Cramer V máximo foram suficientes para separar as duas classes.



Fonte: Guilherme Miura Lavezzo (2021)

.

Figura 34 – Medidas características mais importantes para 73 FTs de PBM. 219 motivos foram descobertos e utilizados nesta análise, para os algoritmos 8-mer align E, RSAT/STREME e InMoDe.



Fonte: Guilherme Miura Lavezzo (2021)

.

Corroborando com os achados no apêndice D e na seção 4.4.2, as medidas com maior importância são Cramer V máximo e Conteúdo de Informação médio.