UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Giovanni Marques de Castro

PREDIÇÃO E VALIDAÇÃO DE GENES ESSENCIAIS EM PROCARIOTOS E EUCARIOTOS UTILIZANDO APRENDIZADO DE MÁQUINA E ATRIBUTOS INTRÍNSECOS À SEQUÊNCIA

Belo Horizonte
Junho 2021

Giovanni Marques de Castro

PREDIÇÃO E VALIDAÇÃO DE GENES ESSENCIAIS EM PROCARIOTOS E EUCARIOTOS UTILIZANDO APRENDIZADO DE MÁQUINA E ATRIBUTOS INTRÍNSECOS À SEQUÊNCIA

Tese apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito para a obtenção do título de Doutor em Bioinformática.

Orientador: Francisco Pereira Lobo

Belo Horizonte
Junho 2021

043 Castro, Giovanni Marques de.
Predição e validação de genes essenciais em procariotos e eucariotos utilizando aprendizado de máquina e atributos intrínsecos à sequência [manuscrito] / Giovanni Marques de Castro. – 2021.
112 f. : il. ; 29,5 cm.

Orientador: Francisco Pereira Lobo.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Genes Essenciais. 3. Aprendizado de Máquina I. Lobo, Francisco Pereira. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

Bi
Giova

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

PARECER Nº **19/2021**
PROCESSO Nº 23072.238444/2021-19

FOLHA DE APROVAÇÃO

"Predição e validação de genes essenciais em procariotos e eucariotos utilizando aprendizado de máquina e atributos intrínsecos à sequência"

Giovanni Marques de Castro

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Francisco Pereira Lobo - Orientador
Universidade Federal de Minas Gerais

Profa Glória Regina Franco
Universidade Federal de Minas Gerais

Profa Mariana Torquato Quezado de Magalhaes
Universidade Federal de Minas Gerais

Prof. Eric Roberto Guimarães Rocha Aguiar
Universidade Estadual de Santa Cruz

Prof. Fabiano Sviatopolk-Mirsky Pais
Fiocruz/Minas

Belo Horizonte, 30 de julho de 2021.

Documento assinado eletronicamente por **Mariana Torquato Quezado de Magalhaes**, **Professora do Magistério Superior**, em 30/07/2021, às 19:52, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eric Roberto Guimaraes Rocha Aguiar**, **Usuário Externo**, em 30/07/2021, às 19:53, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabiano Sviatopolk Mirsky Pais**, **Usuário Externo**, em 31/07/2021, às 20:03, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gloria Regina Franco**, **Professora do Magistério Superior**, em 05/08/2021, às 09:04, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 20/08/2021, às 17:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0862004** e o código CRC **2FF130C7**.

Referência: Processo nº 23072.238444/2021-19 SEI nº 0862004

Agradecimentos

Agradeço imensamente a minha família, a minha mãe e a minha irmã, sem a ajuda delas eu jamais teria começado este caminho. Também agradeço meu amor, a minha esposa Aline Alves Lopes, que além do suporte emocional também elevou o meu conhecimento do mundo.

Aos integrantes que são ou já passaram pelo LAB, Agnelo, Aline, Thieres, Maycon, Amanda, Anderson (Raul), Zandora, Thiago, Dani, Dalbert, Marcos, Tarcisio, Leonardo e Igor. Que me ensinaram muito durante o tempo que passamos juntos, com discussões proveitosas e inúmeras vezes hilárias. E em especial o meu orientador Francisco Pereira Lobo, que sempre me apoiou, mesmo nas iniciativas fora da academia.

Também deixo meus agradecimentos a professora Glória Regina Franco, que nos acolheu quando ainda não tínhamos um espaço e nos ensinou muito. Ao Heron, Stellamaris, Gabriel, Rondon, Nayara e Paula, os quais convivi por um tempo um pouco menor, mas foram tão importantes para mim. E ao Thiago Mafra e Thomaz, que embora não tenhamos seguido em frente com a *startup*, foi uma grande experiencia que tivemos.

Agradeço muito ao Professor Aristóteles Góes Neto, sua orientanda Carolina e seu grupo, assim como a Professora Andréia Maria Amaral Nascimento e suas orientandas, a Marcela e a Maria Luiza, pelas oportunidades e colaborações que tivemos durante esse tempo.

Por fim, agradeço ao Prof. Miguel por manter o servidor Sagarana, ao Programa Interunidades de Pós-Graduação em Bioinformática da UFMG, a Universidade Federal de Minas Gerais e a agência de fomento Capes, por providenciarem e manterem os recursos utilizados.

RESUMO

Genes essenciais são aqueles cuja ausência do produto funcional é incompatível com a viabilidade do organismo. Genes não-essenciais, ao contrário, são aqueles cuja ausência ainda produz indivíduos fenotipicamente viáveis. A caracterização em larga escala destes genes provê a descrição de genomas mínimos compatíveis com a vida celular, bem como sugere alvos moleculares interessantes para o desenvolvimento de bioinseticidas mais específicos e com menor impacto ambiental. Entretanto, sua caracterização experimental é custosa e demorada, diversas estratégias computacionais têm sido utilizadas para a predição de genes essenciais. Dentre estas, destaca-se o aprendizado de máquina. Os algoritmos de aprendizado de máquina utilizados para a predição de genes essenciais utilizam dois tipos de atributos gene-cêntricos: 1) extrínsecos à sequência, definidos como aqueles que utilizam informação que não está contida na própria sequência gênica (e.g. perfil de expressão gênica, anotação); 2) intrínsecos à sequência, definidos como aqueles que são computados a partir da sequência gênica, somente (e.g. frequência de k-mers, entropia). Embora os preditores de genes essenciais que usam atributos extrínsecos sejam superiores aos que utilizam somente atributos intrínsecos, estes carecem de generalização, uma vez que não podem ser utilizados em organismos não-modelo que não possuam informações extrínsecas. Nesse trabalho, desenvolvemos e validamos uma rotina computacional completa para a predição de genes essenciais em procariotos e eucariotos. Especificamente, um pacote R que integra e calcula 5093 atributos nucleotídicos e 9815 protéicos, totalizando 14908 atributos intrínsecos. Estes atributos, em conjunto com os rótulos foram utilizados para treinar modelos de florestas aleatórias e *gradient boosting*, levando-se em consideração o estado-da-arte para a avaliação de desempenho dos modelos produzidos. Validamos nossa metodologia inicialmente construindo bancos de dados de alta qualidade de genes essenciais e não-essenciais para duas espécies de procariotos filogeneticamente distantes (*Acinetobacter baylyi*, Proteobacteria; *Staphylococcus aureus*, Firmicutes) e para duas espécies de inseto (*Drosophila melanogaster*, Diptera; *Tribolium castaneum*, Coleoptera). Posteriormente, utilizamos estes bancos para treinar classificadores para cada uma das quatro espécies. Como validação, demonstramos que classificadores treinados com dados de uma espécie de procarioto/inseto são capazes de predizer genes essenciais em outra espécie de procarioto/inseto, o que emula o uso cotidiano da ferramenta. O código-fonte do projeto, e os bancos de dados de genes essenciais e não-essenciais desenvolvidos nesse estudo encontram-se disponíveis em https://github.com/g1o/GeneEssentiality.

Palavras-chaves: *Drosophila melanogaster*, Aprendizado de máquina, *Tribolium castaneum*, Genes essenciais, Procariotos, bases de dados.

ABSTRACT

Essential genes are defined as those whose absence of the functional product is incompatible with the organism's viability. Non-essential genes, in contrast, are those where this absence still generates phenotypically viable individuals. The large-scale characterization of these genes provides the description of minimal genomes compatible with cellular life, as well as suggests potential molecular targets for the development of specific biopesticides with a smaller ecological footprint. However, since the experimental characterization of these genes is a costly and time-consuming process, several computational strategies have been used for the prediction of essential genes. A common approach in this direction is the usage of machine learning algorithms, since these programs are expected to learn from experience and the use of data. Machine learning algorithms developed to predict essential genes can use two types of gene-centric attributes: 1) extrinsic, defined as those that use information not contained in the gene sequence itself (e.g. gene expression profile, annotation); 2) intrinsic, defined as those computed from the gene sequence only (e.g. frequency of k-mers, entropy). Even though essential gene predictors that use extrinsic attributes have superior performance compared to those that use only intrinsic attributes, they former lack generalization, since they cannot be used in non-model organisms that do not have extrinsic information. In this work, we developed and validated a complete computational routine for the prediction of essential genes in prokaryotes and eukaryotes. Specifically, we developed an R package that integrates and calculates 5093 nucleotide and 9815 protein attributes, totaling 14908 intrinsic attributes. These attributes, together with the labels (essential genes versus non-essential genes), are then used to train random forest models and gradient boosting models while taking into account the state-of-the-art for model performance evaluation. We validated our methodology by gathering high-quality sequence information data of essential and non-essential genes for two phylogenetically distant prokaryote species (*Acinetobacter baylyi*, Proteobacteria; *Staphylococcus aureus*, Firmicutes) and for two insect species (*Drosophila melanogaster*, Diptera; *Tribolium castaneum*, Coleoptera). We used these data to train individual classifiers for each species. As validation, we demonstrate that classifiers trained with data from one species of prokaryote/insect are able to predict essential genes in another species of prokaryote/insect, which emulates the daily use of the tool in new organisms. The source code for the calculation of attributes and models training, as well as the databases of essential and non-essential genes used in this study are available at https://github.com/g1o/GeneEssentiality.

Keywords: *Drosophila melanogaster*, machine learning, *Tribolium castaneum*, Essential genes, prokaryotes, databases.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GE: Gene essencial

GNE: Gene não essencial

DEG: *Database of Essential Genes*

OGEE: *Online Gene Essentiality*

CV: Validação Cruzada (CV, do inglês, *Cross Validation*)

DNA: Ácido desoxirribonucleico

RNA: Ácido ribonucleico

RNAm: RNA mensageiro

RNAi: RNA interferente

dsRNA: RNA dupla-fita (dsRNA, do inglês, *double stranded* RNA)

siRNA: Pequenos RNAs interferentes (siRNA, do inglês, *small interfering* RNA)

CDS: Sequências codificadoras (CDS, do inglês, *CoDing Sequences*)

ROC: Característica de Operação do Receptor (ROC, do inglês, *Receiver Operating Characteristic*)

AUC: Área sob a curva (do inglês, *Area Under the Curve*)

ZR: Classificador de regra zero (ZR, do inglês, *Zero Rule*)

LOF: Perda de função (do inglês, *Loss Of Function*)

EGS: busca de genes essenciais (do inglês, *Essential Genes Search*)

Sumário

ORGANIZAÇÃO DO DOCUMENTO

O presente documento compreende 1) uma introdução geral sobre a predição de genes essenciais via aprendizado de máquina; 2) dois capítulos a serem publicados individualmente na forma de artigos científicos, com o segundo já publicado e 3) uma conclusão geral. As referências da introdução e capítulo 1 são compartilhadas, enquanto as referências do capítulo 2 encontram-se ao final dele. O capítulo 1 descreve o desenvolvimento de um protocolo geral de aprendizado de máquina para a predição de genes essenciais a partir de dados de sequência, somente, desde o cálculo de atributos até o treinamento dos modelos, bem como a sua validação em um conjunto de dados compreendendo organismos procarióticos. Adicionalmente, reportarmos um erro sistemático encontrado no banco de dados DEG, e demonstramos como esse erro possivelmente inflou artificialmente o desempenho de algoritmos para a predição de genes essenciais já publicados. O capítulo 2 descreve a obtenção de conjuntos de genes essenciais e não-essenciais para dois insetos: a mosca-da-fruta *Drosophila melanogaster* (Diptera) e o besouro *Tribolium castaneum* (Coleoptera), bem como o desenvolvimento e a validação de algoritmos visando a predição de genes essenciais nestes insetos. Posteriormente, demonstramos que algoritmos treinados em uma espécie de inseto são capazes de predizer genes essenciais na outra espécie. Adicionalmente, demonstramos que o algoritmo treinado em *T. castaneum* é capaz de predizer genes essenciais linhagem-específicos na mosca, demonstrando que tal abordagem permite a identificação de candidatos para a produção de bioinseticidas linhagem-específicos potencialmente com maior especificidade e menor impacto ambiental em função de ações *off-target*.

1. Introdução Geral

1.1. Genes Essenciais

Alguns genes possuem funções que são indispensáveis para a viabilidade de um ser vivo. Logo, a interrupção ou inativação de alguma das funções indispensáveis codificadas por esses genes leva a um fenótipo letal. Os genes que codificam essas funções vitais são chamados de genes essenciais (GEs), também sendo referenciados na literatura científica como genes vitais ou genes letais (LLOYD et al., 2015). Os genes que não codificam funções essenciais para um organismo são chamados de genes não essenciais (GNEs).

No entanto, conforme exposto anteriormente, a expressão de um fenótipo pode não depender somente do seu componente genético, podendo haver fatores ambientais, bem como fatores que derivam da interação entre genótipo e ambiente, que também influenciam no fenótipo de um organismo. No caso de GEs, alguns genes são essenciais somente em condições ambientais específicas: genes que estão envolvidos no escape do sistema imune, por exemplo, são vitais para a persistência da bactéria *Acinetobacter baumannii* quando infectando um camundongo, mas não o são em um meio de cultura (WANG et al., 2014).

O mesmo conceito pode ser utilizado para genes fazem parte da via de biossíntese de aminoácidos. Se os aminoácidos estão disponíveis no ambiente para o organismo, e ele consegue interiorizá-los em suas células para utilizá-los em seu metabolismo, consequentemente não é necessário sintetizá-los. Logo, genes que fazem parte da via de síntese desse aminoácido não são essenciais, desde que o aminoácido seja um nutriente disponível no ambiente para o organismo. Tais genes, que são ou não essenciais dependendo do ambiente, são chamados de genes condicionalmente essenciais (MOLINA-HENARES et al., 2010)

Um outro tipo de classificação dos genes em função de sua essencialidade compreende os genes que não são essenciais individualmente, mas reduzem em diversos níveis o crescimento celular. Estes genes impactam no crescimento robusto e, ao limitar a taxa de replicação celular, podem inviabilizar o crescimento de mutantes que os contenham. Tais genes são chamados de quasi-essenciais (HUTCHISON et al., 2016).

Algumas funções essenciais podem ser fornecidas por mais de um gene, que podem ou não ser parálogos (genes que passaram por um evento de duplicação e estão mantidos no genoma (LI, 2003)). Supondo que há dois genes que apresentem redundância funcional, a interrupção da função de somente um destes pode não causar um fenótipo letal, uma vez que há outro gene que é funcionalmente redundante e, consequentemente, pode complementar a

ausência de função do gene interrompido/removido. Portanto, é necessário que as funções dos dois genes sejam interrompidas/removidas simultaneamente para se obter um fenótipo letal. Essa combinação letal de genes é chamada de par "sintético letal" (DOBZHANSKY, 1946). A existência dessa redundância para funções essenciais é uma das limitações para a descoberta de GEs em estudos de larga escala que visam a deleção de genes individuais. Assim, verifica-se que o universo de GEs de um organismo depende de diversos fatores ambientais e genômicos, e a sua detecção é uma tarefa não-trivial, envolvendo o uso de diversas ferramentas genéticas e genômicas.

Após a obtenção dos primeiros genomas completos de organismos celulares, diversos estudos pós-genômicos tiveram como objetivo determinar quais seriam os GEs mínimos necessários para a manutenção da vida, permitindo assim o conhecimento em nível molecular dos processos essenciais para o funcionamento celular e a determinação de "genomas mínimos". Os estudos iniciais da caracterização nesse sentido foram conduzidos com o microrganismo *Mycoplasma genitalium*, uma bactéria parasita intracelular obrigatória, visando determinar quais dos genes desse organismo seriam essenciais via mutagênese aleatória mediada por transposons (HUTCHISON et al., 1999). Cabe ressaltar que tais genomas mínimos são ambiente-dependentes, uma vez que um gene responsável pela síntese de uma molécula essencial para a viabilidade celular pode não ser essencial caso esta molécula esteja disponível, por exemplo, via suplementação em meio de cultura.

Após esse estudo pioneiro, os GEs de diversos microrganismos começaram a ser descritos em seus respectivos meios de cultivo controlados (GERDES et al., 2003; LIBERATI et al., 2006; RUBIN et al., 2015; SASSETTI CHRISTOPHER M.; BOYD DANA H.; RUBIN ERIC J., 2003). Isto culminou com o desenvolvimento de uma bactéria, contendo um genoma artificial, baseado no genoma da espécie *Mycoplasma mycoides*, totalmente sintetizado *ex vivo* e que contém o mínimo de genes sabidamente necessários para o crescimento no meio de cultura estabelecido (HUTCHISON et al., 2016). Este microrganismo, nomeado JCVI-syn3.0 e construído explicitamente para sobreviver com o mínimo possível de genes, permitiu a avaliação objetiva das vantagens e desvantagens de um genoma mínimo construído racionalmente. Interessantemente, JCVI-syn3.0 tem o seu tempo de replicação acelerado em 5 vezes quando comparado ao organismo selvagem, mas sua replicação é limitada a um meio de cultura específico. Surpreendentemente, dentre os 473 genes encontrados nesse genoma mínimo e que são experimentalmente comprovados como essenciais para a sobrevivência celular, 84 genes foram anotados com funções genéricas (eg. quinase) e 65 genes não possuem função biológica conhecida, totalizando 149 genes (31,5%) sem função biológica descrita em

detalhes quando anotados por homologia/similaridade com genes de função biológica conhecida. Portanto, mesmo para o menor genoma já produzido artificialmente e extensivamente estudado, ainda há uma fração considerável de GEs com funções biológicas desconhecidas.

Diversas aplicações biotecnológicas e científicas são possíveis com o conhecimento de GEs. Uma delas é a de se ter um organismo celular com apenas genes necessários para crescer em um meio sem estresses e que forneça todos nutrientes necessários (GLASS et al., 2006). Tal organismo com um genoma mínimo seria tão simples que eventualmente poderia ser possível saber a função de cada gene e as interações de todos os produtos gênicos, facilitando o desenvolvimento de um modelo computacional do funcionamento celular (KARR et al., 2012). De posse de tal modelo seria possível, por exemplo, predizer computacionalmente e avaliar experimentalmente as consequências de funções biológicas e vias metabólicas adicionadas gradativamente (HUTCHISON et al., 2016).

Uma vez que a interrupção de uma função essencial causa a morte celular, tais genes também podem servir para priorização de alvos para drogas em bactérias e fungos patogênicos para animais e plantas (HU et al., 2007; LU et al., 2014) e para o desenvolvimento de fármacos para câncer (CHEN et al., 2017). Seguindo a mesma lógica de letalidade, diversos inseticidas para insetos transmissores de doenças e pragas de plantações já foram desenvolvidos baseados em GEs conhecidos (AIRS; BARTHOLOMAY, 2017; BAUM et al., 2007). Além dessas características, os GEs também podem estar envolvidos no processo de especiação, interrompendo a mitose em machos híbridos e causando isolamento reprodutivo (PHADNIS et al., 2015).

Uma das grandes dificuldades em se prever GEs, especialmente em organismos multicelulares, é o fato que diversos genes são essenciais somente durante uma etapa na vida do organismo, incluindo o seu desenvolvimento embrionário, ou sob alguma condição específica. Em *D. melanogaster*, vários genes são essenciais durante sua embriogênese, mas podem não possuir consequências fenotípicas importantes durante sua vida adulta. Adicionalmente, tais genes podem não ser importantes para o funcionamento de células individuais, o que impossibilita a sua detecção utilizando ensaios de varredura em larga escala em culturas de células. Já em microrganismos, uma outra complicação é transferência horizontal de genes que realizam uma função essencial para a sobrevivência do organismo em um ambiente específico (genes nicho-específicos), mas não tendo relação evolutiva direta com os GEs que já estavam presentes anteriormente no organismo, aumentando assim a variabilidade no conjunto de GEs (MARTÍNEZ-CARRANZA et al., 2018).

1.2. Genes essenciais e controle de pragas

A relativa ausência de estudos em larga escala para a busca por GEs em insetos nos bancos de dados DEG e OGEE contrasta com os diversos estudos sobre o silenciamento da expressão de genes potencialmente essenciais para o controle de insetos praga usando da tecnologia de RNAi (BAUM et al., 2007; KNORR et al., 2018; ULRICH et al., 2015). Em situação fisiológica, esse mecanismo promove a resistência à parasitas genéticos endógenos, como transposons, bem como à RNAs patogênicos exógenos, como os vírus de RNA. A maquinaria do RNAi foi inicialmente descrita em *C. elegans* (FIRE et al., 1998; GRISHOK, 2005), sendo posteriormente observada em diversas outras espécies e conservado nas principais linhagens de eucariotos (SHABALINA; KOONIN, 2008).

A extrema especificidade da resposta mediada por RNAi, causada pelo fato de que a degradação do RNA-alvo depende do dsRNA carregado pela maquinaria proteica responsável pelo silenciamento de RNA, levou à sugestão de que essa metodologia possa ser utilizada no lugar de pesticidas convencionais para o controle de insetos vetores e pragas agrícolas (AIRS; BARTHOLOMAY, 2017; KANAKALA; GHANIM, 2016).

Como exemplo de tal aplicação, o banco de dados iBeetle contém informações sobre a essencialidade gênica no inseto *Tribolium castaneum* (Coleoptera) (ULRICH et al., 2015). Usando os dados disponíveis no iBeetle, os ortólogos de 50 GEs em *T. castaneum* foram silenciados via RNAi em *Diabrotica virgifera virgifera* (Coleoptera) (KNORR et al., 2018), um inseto-praga também conhecido como lagarta-da-raiz do milho, que causa perdas estimadas de mais de 500 milhões de euros por ano para a União Europeia (WESSELER; FALL, 2010). Todos os 50 GEs silenciados em *D. virgifera virgifera* foram verificados como essenciais, com diferentes eficiências de letalidade. Interessantemente, 4 genes que tiveram alta letalidade em *D. virfifera virgifera* também foram silenciados e verificados como essenciais em *Meligethes aeneus* (Coleoptera) (KNORR et al., 2018), um outro inseto-praga que pode reduzir a produção de plantações de colza (*Brassica napus*) em até 80% (HANSEN, 2004). Assim, demonstra-se que a transferência de informações de essencialidade via homologia e seu uso via RNAi podem ser abordagens interessantes para o controle de insetos.

Como vantagens do uso de metodologias de RNAi para o controle de pragas de insetos, menciona-se que as mesmas podem ser feitas com poucos ou eventualmente nenhum alvo diferente do previsto dentro do organismo-alvo (*off-targets*) (HORN; SANDMANN; BOUTROS, 2010). Também foi demonstrado que, mesmo considerando-se a existência de espécies filogeneticamente próximas às espécies-alvo, é possível ter alta especificidade, o que

evita a letalidade em espécies não-alvo, embora haja eventuais casos de letalidade parcial (BACHMAN et al., 2013; BAUM et al., 2007; WHYARD; SINGH; WONG, 2009).

Uma vez que a utilização da tecnologia de RNAi precisa de um dsRNA, o mesmo pode ser inserido como um transgene em plantas para induzir a proteção das mesmas contra pragas. Outra possibilidade de utilização do dsRNA é a sua aplicação em formulações para serem aplicadas em plantas, sem envolver a produção de organismos transgênicos. Entretanto, embora estas formulações sejam facilmente degradadas no ambiente e sejam feitas com custos cada vez menores, este pode ser um procedimento relativamente caro para aplicações comerciais (aproximadamente U$4000,00 o quilograma em setembro de 2017) (ZOTTI et al., 2018).

Apesar das vantagens citadas acima, as estratégias de RNAi possuem algumas desvantagens em potencial, bem como questões adicionais a serem consideradas. Como exemplos, podemos citar: 1) há relativamente poucos estudos de campo para avaliar as consequências ecológicas/ambientais a médio e longo prazo da adoção dessa tecnologia; 2) uma única variação de ponto em um alelo encontrado em alguma população pode promover a resistência em indivíduos e, consequentemente, diminuir os efeitos desejados do RNAi; 3) possíveis efeitos ecológicos em espécies não-alvo não foram adequadamente avaliados; 4) é necessária a exposição prolongada para se obter o efeito desejado (ZOTTI et al., 2018).

Uma consulta realizada em junho de 2018 revelou 297 genomas de insetos (busca por *Insecta [Organism]*) no NCBI, representando 296 espécies. Já para os dados no SRA de sequenciamento de RNA, há dados de sequência disponíveis para 1511 espécies, das quais 1254 são representadas por até 2 entradas no SRA. Em maio de 2021 foi feita uma atualização para consulta dos genomas, e obteve-se 750 genomas nucleares de insetos. A falta de dados para espécies pouco estudadas é evidente, uma vez que é estimado que existam mais de 1 milhão de espécies de insetos (CHAPMAN, 2009). A enorme diversidade de espécies de insetos mostra que não seria trivial desenvolver um sRNA para inativar um eventual GE em um inseto-praga não afetasse outra espécie ainda desconhecida de maneira inespecífica (off-target).

Uma possibilidade de diminuir a chance da ação inespecífica do RNAi como bioinseticida seria o desenvolvimento de alvos táxon-específicos, preferencialmente genes ou produtos gênicos essenciais existentes somente para aquele táxon (GEs linhagem-específicos). Entretanto, qualquer estratégia dependente de homologia para a predição de GEs teria grandes limitações para caracterizar GEs que são táxon-específicos, uma vez que há a necessidade de que algum gene do táxon em questão já tenha sido previamente caracterizado experimentalmente como essencial para transferir a anotação com alguma confiança, o que

geralmente acontece somente para organismos-modelo. Uma possível solução para essa questão seria utilizar atributos gênicos que sejam independentes de homologia (eg. tamanho do gene, composição nucleotídica etc.) de modo a treinar algum software para distinguir potenciais GEs linhagem-específicos.

1.3. Aprendizado de Máquina

O aprendizado de máquina (AM) é um campo do conhecimento que visa desenvolver e aplicar algoritmos computacionais que melhoram seu desempenho com a experiência. Os algoritmos que fazem uso de AM possuem grande aplicação nos mais diversos campos científicos, que vão do processamento de imagens (KUBAT; HOLTE; MATWIN, 1998; VIOLA; JONES, 2001) e tradução de automática de textos (HINTON et al., 2012; MIKOLOV et al., 2011) à física de partículas (BALDI; SADOWSKI; WHITESON, 2014).

Dentro do domínio da biologia computacional/bioinformática, o AM vem sendo utilizado com sucesso em tarefas tão heterogêneas como predição de função gênica (BARUTCUOGLU; SCHAPIRE; TROYANSKAYA, 2006), predição de estrutura secundária da proteína (WANG et al., 2016), predição de estrutura terciária da proteína (JO et al., 2015), alvos de miRNA (KIM et al., 2006), localização subcelular (KAUNDAL; SAINI; ZHAO, 2010), predição de função enzimática (LI et al., 2018), caracterização da estrutura terciaria de RNA (YANG et al., 2017), classificação de sequências de metagenomas (VERVIER et al., 2016; XING; LIU; ZHONG, 2017) e estudos do comportamento de animais (VALLETTA et al., 2017).

Algoritmos de AM podem ser classificados de diversas maneiras em função de seu funcionamento. Uma classificação usual consiste em dividi-los em algoritmos supervisionados (que dependem do uso de dados rotulados previamente para a obtenção dos classificadores) e não-supervisionados (que derivam os possíveis rótulos e padrões de classificação em função dos dados de entrada)(JAMES et al., 2013). A Figura 1 contém um exemplo da construção de um classificador de GEs/GNEs via aprendizado supervisionado, e será utilizada para descrever o funcionamento geral de tais algoritmos.

Cada dado (registro) usado em um modelo de AM supervisionado compreendem um exemplo positivo (GE) ou negativo (GNE), e é descrito por diversos valores que são chamados de atributos (Figura 1, colunas das matrizes representam atributos, sendo a intensidade e cor para cada célula a representação visual de um valor). Atributos podem ser numéricos (discretos ou contínuos) ou qualitativos. Um dos atributos é considerado o atributo de saída, também

chamado de atributo-resposta ou variável dependente, e compreende os rótulos dos genes nas duas classes de interesse (Figura 1, essencial (E) ou não essencial (NE)). Os atributos de saída podem ser estimados utilizando os valores dos demais atributos, denominados atributos de entrada. Em problemas de classificação, o atributo de saída é considerado como o rótulo ou etiqueta do objeto a ser classificado, podendo ter diversas classes possíveis (JAMES et al., 2013).

O processo do aprendizado pode ser conceitualmente dividido em 3 etapas. Primeiro, os dados devem ser obtidos e conhecidos, assim, um algoritmo de AM pode ser escolhido ou desenvolvido em função da tarefa que se deseja realizar. Diversos pacotes em diversas linguagens de programação já fornecem implementações eficientes das principais classes de algoritmos de AM, tais como *random forests*, *gradiente boosting* e *support vector machines*, dentre outros (HALL et al., 2009; KUHN, 2018; MARTÍN ABADI et al., 2015). Após a escolha/desenvolvimento do algoritmo, a segunda etapa consiste em utilizar os dados como entrada para o treinamento do modelo. Na Figura 1, esses dados compreendem rótulos descrevendo cada gene como GE ou GNE. O algoritmo então processa esses dados, treinando um modelo com base nos diversos atributos fornecidos, selecionando possíveis combinações de atributos e valores que permitam distinguir satisfatoriamente entre as duas classes de genes (GEs versus GNEs).

A terceira etapa consiste na avaliação objetiva do desempenho do modelo (etapa de avaliação). Para tal, o padrão-ouro consiste na utilização de dados novos, com rótulos conhecidos, e que não foram utilizados na etapa de treinamento do modelo. Estes dados terão seus rótulos removidos e são fornecidos para o algoritmo realizar a predição dos rótulos (OHLER et al., 2002). Usualmente, tais dados novos compreendem uma fração dos dados do mesmo organismo utilizado no treinamento, fração esta, reservada antes do treinamento do modelo e utilizada somente para a validação final do mesmo.

Opcionalmente, uma estratégia de validação biologicamente interessante de AM consiste na utilização de uma espécie que possua dados de GEs/GNEs para treinar um modelo e na utilização de outra espécie que também possua estes dados, mas que não tenha sido utilizada para treinar o modelo (*leave-one-organism-out*) (NIGATU et al., 2017). Esse tipo de abordagem permite avaliar o desempenho do classificador em uma situação que emula o seu uso para predizer GEs em organismos novos.

Como as classes do conjunto de teste são conhecidas, pode-se avaliar o algoritmo objetivamente através de métricas como sensibilidade, especificidade e *F-measure*, com a sua taxa de sucesso dependendo de quantos dados do conjunto de teste tiveram suas classes preditas

de forma correta (LIBBRECHT; NOBLE, 2015). Em uma dada tarefa de classificação de um conjunto ouro com classes conhecidas, quatro valores podem ser computados: verdadeiro-positivos (VP), verdadeiro-negativos (VN), falso-positivos (FP) e falso-negativos (FN). A sensibilidade é definida como a fração de verdadeiros positivos recuperada pelo classificador (VP / (VP + FN)), enquanto a especificidade é definida como a fração dos resultados preditos como positivos que são realmente relevantes (VP / (VP + FP)).

A sensibilidade é a taxa de verdadeiros positivos (teste), e representa o número dos que foram corretamente identificados como sendo positivos em relação a todos os verdadeiros positivos, enquanto a especificidade é a taxa de verdadeiros negativos, e representa o número dos que foram corretamente identificados como negativos em relação a todos os verdadeiros negativos. Uma maneira comumente utilizada para se visualizar a evolução da sensibilidade com relação a especificidade é utilizando a curva ROC (Característica de Operação do Receptor – ROC, do inglês, *Receiver operating characteristic*). A terceira etapa na Figura 1 mostra um exemplo simples de uma curva ROC. A ROC mostra os valores para cada ponto de corte, neste exemplo, a curva é feita de 6 pontos, a qual consiste na visualização dos valores da taxa de verdadeiros positivos corretamente classificados (sensibilidade) em função da taxa de falso positivos (1 – especificidade). Diversos algoritmos de AM visam maximizar a área sob a curva ROC (*area under the curve*, AUC), uma vez que valores maiores desse parâmetro indicam modelos com elevada sensibilidade e especificidade.

Há dois tipos de erros usualmente verificados no momento da avaliação do modelo: o erro do teste e o erro do treino. O erro do teste é a média do erro para predizer uma observação nova, ou seja, que não foi usada para o treino. Já o erro do treino é calculado usando as observações usadas para treinar o modelo, geralmente tendo um resultado que subestima o erro do teste, pois o conjunto usado já é conhecido pelo modelo.

A validação cruzada (CV - *cross validation*) é um dos métodos de reamostragem mais comumente utilizado nos treinamentos de modelos de AM. Ela é utilizada para se estimar o erro associado a um método e, consequentemente, para avaliar sua performance, bem como o nível de flexibilidade/robustez apresentado pela solução. A validação cruzada é feita com a separação dos dados de treinamento em duas partes, uma que será usada para o treino e a outra fica reservada para o teste. Uma vez que se sabe de antemão os rótulos das observações reservadas, os erros e acertos do modelo podem ser utilizados para calcular sua taxa de erro (JAMES et al., 2013).

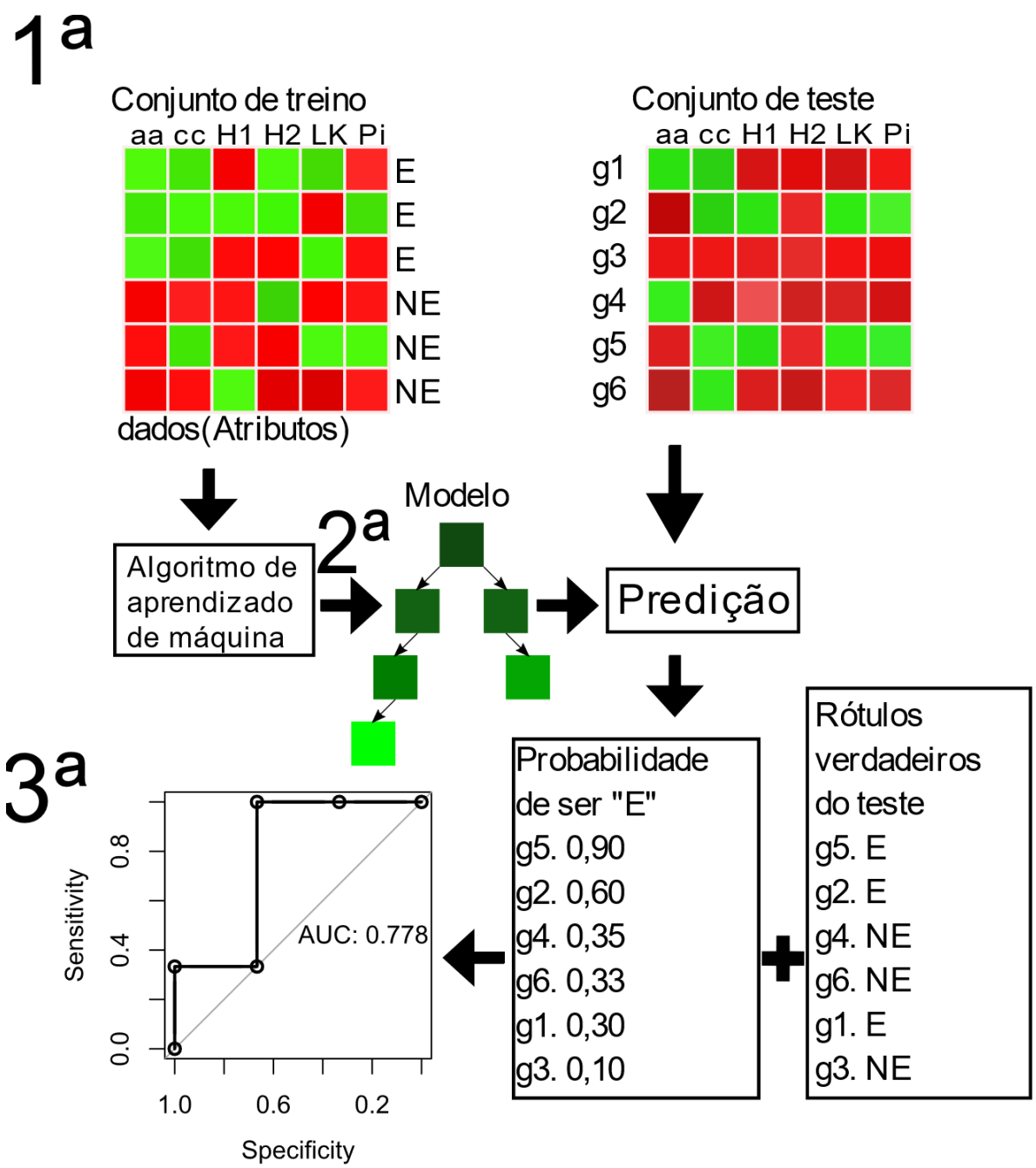


Figura 1:Esquema de uma aplicação de AM para predizer se uma sequência é essencial (E) ou não essencial (NE). *1ª) Dados a serem utilizados na etapa de treinamento. A matriz contendo células vermelhas e verdes representa os dados utilizados para treinar um classificador (colunas representam atributos, linhas representam amostras). 2ª) Etapa de treinamento. Nesse momento, os dados representados em "1" são utilizados para treinar modelos capazes de selecionar os melhores conjuntos de valores de atributos capazes de separar as duas classes. 3ª) Etapa de validação/predição: nessa etapa, um conjunto de dados não utilizado para treinar o modelo (e.g. uma espécie-teste), e para o qual se sabe os rótulos, é utilizada para a avaliação do desempenho do mesmo. Nessa etapa, usualmente escolhe-se um ponto de corte dos valores de predição de maneira a maximizar a sua capacidade de discriminação.*

Dos diversos métodos de CV existentes, o de *k*-iterações (*k-fold Cross Validation)* é amplamente utilizado para se estimar o erro do teste. No método de validação cruzada de *k*-iterações, o conjunto de treino é dividido em *k* partes (e.g. 10 partes) aleatórias, de tamanho semelhante, e sem sobreposições. Em seguida, uma parte *k* (e.g. 1ª) é reservada para se estimar o erro do teste, enquanto as outras partes (e.g. 2ª até 10ª) são utilizadas para treinar o modelo. Em uma segunda rodada de treinamento, a próxima parte (e.g. 2ª) é reservada para o teste, enquanto as outras (e.g. 1ª e 3ª até a 10ª) treinam outro modelo. Esse processo é feito sucessivamente para todas as partes, sendo que o modelo final, treinado com todos os dados, utilizará os parâmetros com os menores erros verificados durante a etapa de teste.

Um outro método de validação cruzada, chamado de validação cruzada repetida (*repeatedcv*), pode ser usado para se reduzir um possível viés causado pela eventual dependência entre as *k* partes, o que tornaria as estimativas enviesadas (BENGIO; GRANDVALET, 2004). A *repeatedcv* funciona de forma semelhante ao CV de *k*-iterações, mas com o processo repetido n vezes, gerando conjuntos de diferentes para cada repetição, reduzindo assim a dependência com os conjuntos de repetições diferentes.

1.4. Algoritmos de aprendizado de máquina utilizados

1.4.1. Floresta Aleatória (*Random Forest*)

Uma árvore de decisão é uma estrutura de grafo na qual, em cada nó, se tem o teste de algum atributo, com os resultados possíveis desse teste levando a outro nó ou à decisão final. O algoritmo de Florestas Aleatórias (BREIMAN, 2001) é uma combinação de diversas árvores de decisão. Para se gerar as diversas árvores da floresta, realiza-se um procedimento de *bootstrap*, onde cada árvore da floresta é uma reamostragem com reposição do conjunto original. Em cada nó de uma árvore da Floresta Aleatória um número, *mtry*, dos atributos gerais são selecionados aleatoriamente, e é este conjunto de atributos do nó que será usado para decidir o resultado. O pacote ranger, escrito na linguagem R, contém uma implementação eficiente do algoritmo de florestas aleatórias (WRIGHT; ZIEGLER, 2017).

1.4.2. *Extreme Gradient Boosting* (XGBoost)

O algoritmo de *gradient boosting* é utilizado tanto para regressão como para classificação (FRIEDMAN, 2001; JEROME FRIEDMAN; TREVOR HASTIE; ROBERT TIBSHIRANI, 2000). O modo como o algoritmo funciona visa gerar um modelo preditivo mais eficiente por meio da combinação de vários modelos individualmente menos eficientes. Se os

modelos menos eficientes forem árvores de decisão, então o modelo é chamado de *gradient boosting tree*. De forma resumida, a cada iteração o modelo recém-gerado é somado ao novo modelo anterior seguindo uma função para minimizar a função de perda (função de custo)[1]. No algoritmo de *boosting*, cada árvore cresce usando a informação da árvore anterior, o que compreende uma abordagem diferente da Floresta Aleatória, que não usa da amostragem de bootstrap (JAMES et al., 2013). O pacote XGBoost é uma implementação eficiente e escalável deste algoritmo na linguagem R (CHEN; GUESTRIN, 2016).

1.5. Aprendizado de máquina na predição de genes essenciais

Devido aos elevados custos envolvidos nas abordagens experimentais para detectar GEs em escala genômica, as quais usualmente envolvem varredura em larga escala por fenótipos letais, diversos métodos computacionais têm sido propostos como para predição de GEs, com diversos atributos associados à essencialidade. Tradicionalmente, tais atributos podem ser categorizados como intrínsecos às sequências (e.g. conteúdo GC, entropia, composição de aminoácidos, propriedades físico-químicas), os quais não dependem de informações relativas à função do produto gênico (anotação) para serem computados e, consequentemente, não dependem da existência de homólogos anotados em outras espécies (LIU et al., 2017; NING et al., 2014; SONG; TONG; WU, 2014; YU et al., 2017).

Além dos atributos intrínsecos às sequências, diversos outros atributos só podem ser computados para os genes cuja função biológica seja parcial ou totalmente conhecida, ou para aqueles que tenham sido caracterizados experimentalmente de outras maneiras, como, por exemplo, em ensaios de expressão gênica. Como exemplos desses tipos de atributos, pode-se citar: a topologia de redes de interação proteína-proteína (e.g. grau de centralidade e coeficiente de agrupamento) (ACENCIO; LEMKE, 2009; CHENG et al., 2014; LU et al., 2014; PLAIMAS; EILS; KÖNIG, 2010), predição de homologia com genes de função conhecida (CHENG et al., 2013; SONG; TONG; WU, 2014), perfis de expressão gênica (e.g. nível de expressão dos RNAms e flutuações da expressão gênica em diferentes situações experimentais) (CHENG et al., 2014; DENG et al., 2011), localização celular (e.g. pontuação para posição no citoplasma e na membrana externa) (CHENG et al., 2014; LIU et al., 2017), domínio funcional (e.g. enriquecimento de domínios) (DENG et al., 2011).

Alguns dos atributos já citados (e.g. nível de expressão, topologia de redes) são

[1] Uma função de perda é usada para se obter um valor que seja relacionado ao um custo, como a diferença entre os valores reais e estimados ou penalidades para classificações incorretas. Assim, é possível se ter esse valor como uma referência a ser minimizado.

extraídos de dados experimentais, os quais muitas vezes são inexistentes para organismos não-modelo. Adicionalmente, sua obtenção nem sempre é viável, devido ao elevado custo envolvido. Os atributos dependentes de homologia, embora sejam extensivamente utilizados para transferência de anotação e inferência de função gênica, uma vez que requerem a existência de um gene homólogo anotado no banco de dados, terminam por inviabilizar a análise de genes essenciais táxon-específicos, para os quais possivelmente não existem homólogos já caracterizados experimentalmente em organismos-modelo (HUERTA-CEPAS; DOPAZO; GABALDÓN, 2010; KNORR et al., 2018; NAGY et al., 2020).

Um ponto fundamental para o desenvolvimento de preditores de genes essenciais via AM supervisionado é a existência de um banco de dados contendo um número suficientemente grande de genes classificados como essenciais ou não-essenciais. Eucariotos multicelulares possuem GEs que podem exercer sua função em nível celular, tais como componentes de vias de biossíntese de vitaminas, mas também há GEs que contribuem para o desenvolvimento embrionário destas espécies. Como virtualmente todos os ensaios para a predição de GEs envolve o uso de técnicas de varredura *in vitro* em larga escala, a maior parcela dos organismos que possuem informação em escala genômica sobre essencialidade gênica compreendem organismos procarióticos (CHEN et al., 2017; LUO et al., 2021, p. 15; PENG et al., 2017).

Embora existam alguns estudos de AM buscando GEs com eucariotos multicelulares, eles carecem de capacidade de generalização. Um exemplo é o de Lloyd et al. (2015), que gerou modelos de AM para predizer GEs na planta *Arabidopsis thaliana* usando como rótulos de genes a informação de genes letais quando há perda de função. Entretanto, esse estudo virtualmente não utiliza atributos intrínsecos às sequências. Estes autores utilizaram seis atributos para avaliar a presença e a frequência de eventos de duplicação gênica (eg. tamanho da família gênica, pseudogene presente), cinco atributos baseados na expressão gênica (eg. mediana da expressão, correlação da expressão de GEs com GNEs), 12 atributos para avaliar o perfil conservação evolutiva dos genes (eg. homólogo não encontrado em arroz, percentual de identidade em plantas, percentual de identidade em metazoários), três atributos de interação via redes (rede de co-expressão, conexões gênicas, interação proteína-proteína), e quatro atributos gerais (gene metilado, porcentagem de identidade com o parálogo, tamanho da proteína, quantidade de domínios proteicos).

Os atributos intrínsecos das sequências de nucleotídeos e de aminoácidos podem ser obtidos para virtualmente qualquer sequência, independente do fato de apresentarem homólogos conhecidos ou não. Assim, essas informações podem ser utilizadas para, eventualmente, treinar classificadores que identifiquem GEs linhagem-específicos em insetos,

um grupo negligenciado em estudos dessa natureza. Isso permite que genes sem similaridade com genes previamente caracterizados ou sem experimentos que tenham definido suas funções biológicas possam ser eventualmente preditos como essenciais, e consequentemente priorizados para caracterização fenotípica e eventuais usos biotecnológicos.

O trabalho publicado por Nigatu et al. (2017) compreende uma tentativa em utilizar somente atributos intrínsecos às sequência de nucleotídeos para se treinar modelos de florestas aleatórias capazes de distinguir GEs de GNEs para 12 espécies de bactérias e um eucarioto unicelular (a levedura *Schizosaccharomyces pombe)*. Estes autores obtiveram valores de AUC-ROC acima de 0,75 para diversos testes do tipo *leave-one-organism-out*, onde um modelo é treinado em um conjunto de espécie e testado em outras, demostrando a viabilidade de se usar atributos independentes de homologia e independentes dados experimentais (e.g. expressão gênica) para encontrar GEs.

Utilizamos como estudo de caso a predição de GEs em organismos procarióticos já avaliados por Nigatu et al. (2017), de modo a avaliar como a incorporação de atributos computados a partir de proteínas poderia melhorar a predição destes. Os resultados compreendendo a implementação da rotina de cálculo de atributos para nucleotídeos e proteínas, treinamento de modelos e validação em organismos procarióticos compreende o capítulo 1 do presente documento. Posteriormente, obtivemos informações sobre GEs e GNEs para os dois insetos que possuem informações estruturadas e em larga escala sobre letalidade gênica: *D. melanogaster* (banco de dados FlyBase) e *T. castaneum* (banco de dados iBeetle). De posse dessa informação, utilizamos a estratégia descrita no capítulo 1 para treinar e validar modelos capazes de predizer genes essenciais em insetos. Os resultados dessa análise compreendem o capítulo 2 do presente documento.

2. Objetivos

2.1. Objetivo geral

Desenvolvimento e validação de preditores de GEs através de AM usando atributos intrínsecos das sequências de nucleotídeos e aminoácidos, possibilitando a predição de GEs codificadores de proteínas em organismos não-modelos e em genes sem anotação conhecida (*biological dark matter*).

2.2. Objetivos específicos

- Desenvolvimento de rotinas computacionais para o cálculo de atributos intrínsecos para os genes e para o treino dos modelos
- Verificação da integridade de bancos de dados de GEs/GNEs para o treinamento dos modelos
- Desenvolvimento e validação de preditores de GEs em procariotos
- Desenvolvimento e validação de preditores de GEs em insetos

Capítulo 1:

Predição de genes essenciais utilizando atributos intrínsecos às sequências em procariotos: onde estamos realmente?

3. Introdução

Genes essenciais (GEs) são definidos como aqueles onde a interrupção da funcionalidade do produto gênico é incompatível com a viabilidade do organismo, ou geram a sua incapacidade reprodutiva, enquanto genes não-essenciais (GNEs) são aqueles onde a ausência completa de função biológica do produto gênico é compatível com a viabilidade e reprodução. O estudo sistemático desses genes fornece informações importantes para o estabelecimento de genomas mínimos para a manutenção da vida em condições experimentais específicas. Adicionalmente, a identificação de GEs permite a descoberta de novas funções genicas, e de novos componentes moleculares de vias já caracterizadas (WAINBERG et al., 2021). Finalmente, a caracterização de GEs também permite a identificação de alvos potenciais para a intervenção molecular em campos tão diversos como o controle de pragas agrícolas, patógenos de interesse médico e veterinário, e até o tratamento do câncer (BIRHANU et al., 2018; CHANG et al., 2021; QURESHI et al., 2021; ULRICH et al., 2015).

Atualmente, informações sobre a essencialidade gênica através de ensaios em larga escala de inativação gênica e silenciamento transcricional encontra-se disponível para diversos organismos (BELLEN et al., 2011; DÖNITZ et al., 2015; GLASS et al., 2006; PRICE et al., 2018; SPRADLING et al., 1999; VISWANATHA et al., 2018). Os resultados de alguns desses experimentos também estão disponíveis de maneira estruturada em bancos de dados com tal finalidade, tais como o DEG (*Database of essential genes*) (LUO et al., 2014, 2021) e o *Online GEne Essentiality* (OGEE) (CHEN et al., 2012, 2017).

A disponibilidade de informação estruturada sobre o status de essencialidade para uma larga coleção de genes e um número razoável de espécies permitiu o desenvolvimento de diversas estratégias computacionais de aprendizado de máquina para a predição de genes essenciais (LIU et al., 2017; PENG et al., 2017). Em relação aos atributos dos genes utilizados para treinar preditores, os mesmos podem ser classificados em duas grandes classes: extrínsecos e intrínsecos. Atributos extrínsecos compreendem um conjunto heterogêneo de camadas de informação biológica que são atribuídas aos genes a partir de dados externos à sua própria sequência. Como exemplos, mencionamos o perfil de expressão gênica, a presença de parálogos no genoma em questão, redes de interação proteína-proteína, conservação filogenética, data de origem do gene e presença de homólogos essenciais já caracterizados em outras espécies, dentre outros (CAMPOS et al., 2019; COUTINHO; FRANCO; LOBO, 2015; NAGY et al., 2020).

Atributos intrínsecos são aqueles que, por definição, podem ser computados a partir da

informação contida na sequência do gene, somente. Estes atributos compreendem desde valores como o tamanho do gene e a frequência relativa de k-mers de tamanhos distintos até parâmetros estatísticos, como a quantidade de informação codificada no gene. No caso de genes codificadores de proteínas, diversos outros descritores que capturam diferentes propriedades físico-químicas das proteínas, tais como carga, polaridade e estrutura secundária, dentre outros, podem ser calculados ou estimados a partir da sequência de aminoácidos (XIAO et al., 2015).

Recentemente, o trabalho de Nigatu et al. (2017) reportou o desenvolvimento de classificadores com alta performance para 12 espécies de procariotos e um organismo eucariótico utilizando somente atributos intrínsecos calculados a partir de sequências nucleotídicas obtidas a partir do banco de dados DEG. Posteriormente, outros autores utilizaram este banco de dados para realizar trabalhos semelhantes, novamente utilizando apenas sequências nucleotídicas (YU et al., 2017; ZHOU; QI; REN, 2021).

Em nosso trabalho, inicialmente pretendíamos avaliar o desempenho de classificadores de genes essenciais ao incorporar também atributos intrínsecos que possam ser computados para sequências proteicas. Para tal, utilizamos o trabalho de Nigatu et al. (2017) como referência a título de comparação de desempenho de classificadores treinados somente com atributos intrínsecos de sequências nucleotídicas versus classificadores treinados com atributos intrínsecos obtidos de sequências nucleotídicas e proteicas.

Entretanto, observamos inconsistências nos dados de GNE obtidos do banco de dados DEG que revelam que aproximadamente um terço dos GNEs disponíveis nesse banco não codificam a sequência proteica esperada em nenhuma fase de leitura. Como evidências adicionais, verificamos que estas regiões não possuem as marcas canônicas de regiões codificadoras, tais como a presença de códons de início e de término. Adicionalmente, buscas via alinhamento de sequência visando localizar estas sequências em seus genomas de origem se mostraram infrutíferas. Entretanto, encontramos evidências de que estas sequências foram produzidas por algum erro computacional durante sua manipulação. Isso porque elas correspondem ao complemento da região codificadora que deveriam representar.

Também encontramos evidências de que estas sequências de GNEs com inconsistências foram sistematicamente utilizadas para o desenvolvimento e validação de preditores de genes essenciais em procariotos, os quais possuem desempenho artificialmente inflado em função do erro no DEG. Especificamente, demonstramos como os valores das curvas ROC dos classificadores obtidos por Nigatu et al. (2017) são compatíveis com o cenário de contaminação do banco de dados de GNEs por sequências não-codificadoras que reportamos. Sendo assim, a eficiência real de classificadores de GEs em organismos procarióticos utilizando somente

atributos intrínsecos às sequências pode ser significativamente menor do que o reportado até o momento.

Nosso trabalho revela que mesmo a informação advinda de bancos de dados tradicionais e amplamente utilizados pela comunidade científica podem conter erros substanciais, e tais erros podem inflar de maneira artificial o estado-da-arte da predição de GEs em procariotos. Adicionalmente, ressaltamos que rotinas internas simples de verificação de erros, tais como a conferência da presença de regiões canônicas de início e fim de tradução, devem ser utilizadas sempre que possível para verificar a consistência interna dos resultados esperados. Finalmente, reportamos o que acreditamos ser o cenário mais próximo do estado-da-arte da predição de GEs em organismos procarióticos e avaliamos que, embora a predição em organismos filogeneticamente distantes seja consideravelmente mais eficiente que um modelo do tipo ZR, os valores que obtivemos são consideravelmente inferiores aos já reportados por trabalhos que devem ter usado os dados enviesados.

4. Métodos

4.1. Levantamento bibliográfico

Para a avaliação de nosso algoritmo de predição de GEs a partir de atributos intrínsecos nucleotídicos e proteicos, selecionamos um estudo que desenvolveu tais preditores utilizando somente atributos intrínsecos nucleotídicos, para 12 espécies filogeneticamente diversas de procariotos e um organismo eucariótico unicelular (a levedura *Schizosaccharomyces pombe*) (NIGATU et al., 2017). Além desse trabalho, outros artigos foram publicados que também usam atributos intrínsecos, (AROMOLARAN et al., 2020; CAMPOS et al., 2019, 2020), os quais foram avaliados e usados como referência para integrar atributos adicionais não usados pelo trabalho de Nigatu et al. (2017). Somente genes codificadores de proteínas foram utilizados, dada a vasta quantidade de informação já existente para estes em contraste aos genes não-codificadores. Além disso, genes codificadores permitem que mais atributos sejam extraídos a partir da sequência de aminoácidos.

A base de dados do DEG foi escolhida para validar o nosso classificador, uma vez que possuía tanto as sequências nucleotídicas de GEs quanto de GNEs disponíveis para download (versão 15.2, de 2017). Acreditamos que essa também tenha sido a fonte original de dados de sequências utilizada por Nigatu et al. (2017), conforme pudemos deduzir a partir do relatado na seção de metodologia em seu artigo.

4.2. Alinhamento de sequências

A fim de realizar o alinhamento par-a-par entre cada CDS codificadora de um gene condido no DEG e a sequência protéica armazenada no mesmo banco de dados, utilizamos o software tBLASTn com o parâmetro de código genético 11 (bactérias) e reportando a melhor HSP[2] por alinhamento. A identidade de um alinhamento par a par reflete o quanto deste é composto por exatamente os mesmos aminoácidos em ambas as sequências. No caso de alinhamentos perfeitos, espera-se um valor de identidade de 100% ao longo de toda a região alinhada. O outro parâmetro, a cobertura da *query* (parâmetro *qcovs* do BLAST+), reflete o quanto da *query* foi alinhada pela sequência do banco de dados (*subject*). Novamente, é esperado que uma CDS completa alinhe por todo o comprimento da proteína que ela codifica, logo sendo 100% da proteína coberta pelo alinhamento. Assim, utilizamos estes dois valores, obtidos a partir do resultado do tBLASTn, para avaliar as sequências nucleotídicas e proteicas do banco de dados de GEs e GNEs obtidos a partir do DEG.

No entanto, consideramos que alguns erros podem eventualmente interferir nestes valores, tais como a presença de bases degeneradas nos arquivos de nucleotídeos em função de erros de sequenciamento, ajustes nos parâmetros para o alinhamento, ou sequências de nucleotídeos que eventualmente não contivessem somente a CDS. Por esses motivos, utilizamos como ponto de corte o valor de 90% para ambos os parâmetros de identidade e cobertura da *query*. Conjuntos onde as sequências possuíam valores acima de 90% para ambos foram consideradas CDS verdadeiras, enquanto as que não passarem no filtro foram consideradas CDSs falsas.

4.3. Desenvolvimento de rotinas computacionais para o cálculo de atributos homologia-independentes para os genes

Os atributos que utilizamos foram escolhidos para que o modelo tenha a menor dependência possível de dados experimentais ou de informação transferida via alinhamento de sequências e inferência de relações de homologia, de modo a permitir que virtualmente todos os genes codificadores de proteínas possam ser analisados, independentemente do organismo. Adicionalmente, diversos desses atributos já foram utilizados para a obtenção e validação computacional de classificadores de GEs em procariotos (NIGATU et al., 2017). Os atributos

[2] HSP: Maiores pontuações entre os pares de segmentos (em inglês: HSP - High-scoring Segment Pair). O algoritmo do Blast+ inicia os alinhamentos com pares de sequencias da *query* exatamente idênticas a da base de dados (subject), as extensões desses alinhamentos que tem as maiores pontuações são chamadas de HSPs. Como é esperado um alinhamento completo entre uma CDS e sua própria proteína, apenas 1 HSP é suficiente, facilitando o processamento do arquivo resultante.

atualmente calculados a partir da sequência de nucleotídeos e de aminoácidos estão resumidos na Tabela 2, e são descritos a seguir. O código para o cálculo de atributos, treinamento e validação dos modelos está disponível em https://github.com/g1o/GeneEssentiality. Os dados foram processados no servidor do LAB (Laboratório de Algoritmos em Biologia) da UFMG. Este servidor tem 128GB de RAM (118 atualmente, devido a um pente de 8GB queimado) e 64 threads.

4.3.1. Informação Mútua (*Mutual Information* – MI)

A informação mútua mede a associação entre duas variáveis, x e y, comparando a probabilidade de serem observadas juntas com a probabilidade de serem observadas separadas. No nosso caso, as variáveis são, para sequências nucleotídicas, as bases, e para sequências proteicas, os resíduos consecutivos. A fórmula usada para calcular a informação mútua foi a de Nigatu et al. (2017) e implementada nesse trabalho, adaptando a definição para o caso de sequencias proteicas, e pode ser vista a seguir:

$$I(X,Y) = \sum_{x\in\Omega}\sum_{y\in\Omega} P\,(x,y) log_2 \frac{P(x,y)}{P(x)P(y)}$$

Onde Ω representa bases (A, C, G, T) para sequências de nucleotídeos, ou resíduos (G, P, A, V, L, I, M, C, F, Y, W, H, K, R, Q, N, E, D, S, T) para sequências proteicas. P(x,y) é a probabilidade de encontrar x seguido de y, e P(x) e P(y) as probabilidades marginais. As probabilidades são estimadas a partir das frequências relativas, ao longo de toda a sequência, sendo similar a "taxa de associação" (CHURCH; HANKS, 1990). Ex: "ACACTC" teria os valores P(a,c)=2/6 com P(a)=2/6 e P(c)=3/6. Assim como feito por Nigatu et. al. (2017), para cada dupla é calculada a quantidade $P(x,y) log_2 \frac{P(x,y)}{P(x)P(y)}$ e usada como um atributo, resultando em 17 atributos para nucleotídeos e 401 para aminoácidos

4.3.2. Informação Condicional Mútua (*Conditional Mutual Information* – CMI)

A CMI é a informação mútua de duas variáveis condicionada a uma terceira variável, sendo usada a mesma fórmula de Nigatu et al (2017):

$$I(X;Y|Z) = \sum_{x\in\Omega}\sum_{y\in\Omega}\sum_{z\in\Omega} P\,(x,y,z) log_2 \frac{P(x,y,z)P(z)}{P(x,z)P(y,z)}$$

Onde P(X,Y,Z), P(X,Z) e P(Y,Z) são as probabilidades conjuntas. A informação mútua entre as bases/resíduos na posição inicial e final é condicionada pela posição central e calculada como um atributo, sendo a probabilidade estimada da frequência relativa da trinca. Ex:

P(X,Y,Z) e P(X,Z) teriam as frequências relativas de “XZY” e de “XZ” respectivamente ao longo de toda a sequência. Assim como descrito por Nigatu et. al. (2017), para cada trinca o valor $P(x,y,z)log_2\frac{P(x,y,z)P(z)}{P(x,z)P(y,z)}$ é calculado e usado como atributo, resultando em 65 atributos para nucleotídeos e 8001 para aminoácidos.

4.3.3. Entropia de Shannon (H)

A entropia de Shannon (SHANNON, 1948) calcula a quantidade de informação digital, quanto maior a desordem no sistema (maior diversidade de bases ou resíduos), maior será a quantidade de informação. No caso da sequência de nucleotídeos ou aminoácidos ela é calculada a partir da distribuição das letras que codificam os nucleotídeos ou aminoácidos. Para uma palavra de tamanho *T* é dada por*:*

$$H_T = -\sum_n P_T\,(n) log_2 P_T(n)$$

Onde $P_T(n)$ é a probabilidade de se encontrar a enésima palavra de tamanho *T*, que neste caso é a frequência relativa de uma palavra de tamanho 2 (e.g. “GG”) ou 3 (e.g. “KKK”) ao longo de toda a sequência. Tendo a entropia de Shannon calculada para palavras de tamanho 2 e 3, tanto para nucleotídeos como para aminoácidos, ficamos com 2 atributos para nucleotídeos e 2 para aminoácidos.

4.3.4. Entropia de Gibbs

A estabilidade termodinâmica entre GEs e GNEs podem ter diferenças e, devido a essa possibilidade, foram incluídas como atributos as entalpias e as entropias das sequências nucleotídicas. Para predizer a entropia e entalpia, cada sequência nucleotídica foi usada como entrada para o programa VarGibbs (WEBER, 2015), usando o modelo D-CMB, e tendo como resultado os 2 atributos.

4.3.5. Covariância Auto-cruzada baseada em Dinucleotídeo e Trinucleotídeo (*Dinucleotide-based Auto-cross Covariance*; *Trinucleotide-based Auto-cross Covariance*)

A covariância auto-cruzada é a combinação da auto-covariância, que incorpora a correlação de mesmas propriedades entre dois di/tri nucleotídeos, e da covariância cruzada, que incorpora a correlação entre diferentes propriedades entre dois di/tri nucleotídeos. Essas propriedades refletem a correlação entre dois di/tri nucleotídeos, separados por uma distância, em termos de suas propriedades físico-químicas (LIU et al., 2015). Ao mensurar a correlação

entre as propriedades, as sequências que possuem tamanhos diferentes são transformadas em vetores de tamanhos fixos (DONG; ZHOU; GUAN, 2009).

As funções usadas para obter os vetores foram *extrDACC* e *extrTACC,* para dínucleotídeos e trinucleotídeos respectivamente, do pacote rDNAse (ZHU; DONG; CAO, 2016). Os parâmetros para a função *extrDACC* foram *allprop=T e nlag=2,* nesta função o parâmetro *allprop=T* usa todos os 38 índices físico-químicos disponíveis no pacote, e o *nlag* representa a distância entre os dínucleotídeos, valores maiores que 2 para *nlag* geram atributos vazios (na). Os parâmetros para a função *extrTACC* foram *allprop=T e nlag=150,* nesta função o parâmetro *allprop=T* usa todos os 12 índices físico-químicos disponíveis no pacote, e o *nlag* representa a distância entre os trinucleotídeos. A covariância auto-cruzada baseada em dinucleotídeo produz 2888 atributos, enquanto a covariância auto-cruzada baseada em trinucleotídeo produz 2100 atributos.

4.3.6. Composição de Pseudo Dínucleotídeos (*Pseudo Dinucleotide Composition*)

Está composição representa sequencias de DNA incorporando efeitos de ordem da sequência. Foi desenvolvido para melhorar a qualidade da predição de spots de recombinação (CHEN et al., 2013). Por ter sido desenvolvido seguindo a ideia da composição de pseudo aminoácidos (CHOU, 2001), a **Figura 2** também pode ser usada para exemplificar o cálculo, substituindo aminoácidos por dinucleotídeos. A função usada foi a extrPseDNC do pacote rDNAse. O parâmetro *lambda* foi mantido em 3, resultando em 19 atributos.

4.3.7. *Conjoint triad*

Inicialmente usado para predizer interações proteína-proteína, este atributo classifica os 20 aminoácidos em 7 classes de acordo com suas propriedades físico-químicas. O cálculo é feito contando a frequência de cada trinca de aminoácidos de acordo com suas classes, resultando no total de 343 atributos (SHEN et al., 2007). A função usada foi *extractCTriad* do pacote *protR*.

4.3.8. Autocorrelação: *Moreau-Broto, Moran* e *Geary*

Os três cálculos usam as mesmas propriedades dos aminoácidos, que por padrão são 8 obtidos do índice de aminoácidos (KAWASHIMA et al., 2008): CIDH920105 Média normalizada da escala de hidrofobicidade (CID et al., 1992), BHAR880101 média dos índices de flexibilidade (BHASKARAN; PONNUSWAMY, 1988), CHAM820102 - energia livre na

solução de água (CHARTON; CHARTON, 1982), CHAM820101 - Parâmetro de Polarizabilidade (CHARTON; CHARTON, 1982), CHOC760101 - área de superfície acessível no tripeptídeo (CHOTHIA, 1976), BIGC670101 - volume do resíduo (BIGELOW, 1967), CHAM810101 - parâmetro estérico (CHARTON, 1981), DAYM780201 - mutabilidade relativa (DAYHOFF; SCHWARTZ; ORCUTT, 1978).

A autocorrelação *Moreau-Broto* usa os valores das propriedades como bases para as medições. A autocorrelação *Moran* utiliza os desvios das médias das propriedades. E *Geary* utiliza a o quadrado da diferença dos valores das propriedades. Essas autocorrelações mensuram autocorrelações espaciais, que é a correlação da variável com ela mesma pelo espaço (ONG et al., 2007), isto é, ao longo da sequência. Os cálculos foram feitos pelas funções do pacote protR (XIAO et al., 2015). Os parâmetros padrões foram mantidos, ou seja, *nlag* ficou como 30, e as 8 propriedades descritas acima foram usadas, assim, ficam 240 atributos (*nlag* multiplicado pela quantidade de propriedades, 8x30) para cada um dos 3 cálculos, totalizando 720 atributos.

4.3.9. Descritores CTD

Os descritores CTD (Composição, Transição, Distribuição) foram inicialmente desenvolvidos para auxiliar a predição da classe de enovelamento da proteína (DUBCHAK et al., 1995, 1999). Os aminoácidos são categorizados em três grupos de acordo com seus atributos, sendo cada aminoácido é codificado por um de seus índices de acordo com o grupo a que pertence (Tabela 1). A composição, que é definida pelo percentual de cada grupo por atributo, foi usada com a função *extractCTDC*, resultando em 21 atributos. A transição, que consiste na frequência relativa em que um aminoácido de um grupo é seguido por um de outro grupo, indicando assim a transição entre estados, foi obtida usando a função *extractCTDT*, resultando em 21 atributos.

A distribuição é calculada, para cada atributo, obtendo 5 percentuais de suas posições na sequência, sendo estes 5 percentuais os valores para a posição do primeiro resíduo do grupo na sequência, e para os quantis 25% (25% * número de resíduos do grupo), 50% (50% * número de resíduos do grupo), 75% (75% * número de resíduos do grupo), e para 100% (100% * número de resíduos do grupo). Uma vez obtidas as posições desses resíduos, cada posição é dividida pelo número total de aminoácidos da sequência e multiplicada por 100 (ex: O descritor da distribuição de um resíduo na posição 2 de uma sequência de 20 aminoácidos: 2/20 * 100 = 10). Novamente, a função *extractCTDD* do pacote protr foi utilizada para obter os valores de

distribuição, resultando em 105 atributos.

Tabela 1: Atributos dos aminoácidos, e a classificação em 3 grupos para os 20 aminoácidos para cada atributo.

Atributo	**Grupo 1**	**Grupo 2**	**Grupo 3**
Hidrofobicidade	Polar	Neutra	Hidrofóbica
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, W
Volume de van der Waals normalizado	0-2.78	2.95-4.0	4.03-8.08
	G, A, S, T, P, D, C	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polaridade	4.9-6.2	8.0-9.2	10.4-13.0
	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
polarizabilidade	0-1.08	0.128-0.186	0.219-0.409
	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Carga	Positiva	Neutra	Negativa
	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
Estrutura secundária	Hélice (*Helix*)	Fita (*Strand*)	Desordenada (*Coil*)
	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Acessibilidade de solvente	"Enterrado"	Exposto	Intermediário
	A, L, F, C, G, I, V, W	R, K, Q, E, N, D	M, S, P, T, H, Y

4.3.10. Composição físico-química

Os aminoácidos podem ser classificados em grupos com propriedade físico-química semelhantes. Neste caso foram usadas as frequências relativas das seguintes classes obtidas pela função AAstat do pacote SeqinR (CHARIF; LOBRY, 2007): aromático, ácido, básico, alifático, apolar, polar, carregado, pequeno (*small*), minúsculo (*tiny*)

4.3.11. Ponto isoelétrico

O valor de pH onde há equilíbrio das cargas positivas e negativas na proteína é chamado de ponto isoelétrico. O valor foi obtido pela função AAstat do pacote SeqinR.

4.3.12. Composição de pseudo-aminoácidos

O uso somente da frequência de aminoácidos não leva em consideração que pode haver

interações entre eles, perdendo informação da ordem dos aminoácidos. O conceito de pseudo-aminoácido foi originalmente introduzido por Chou (2001) para melhorar a representação da proteína em forma numérica, auxiliando na predição da localização celular e de proteínas de membrana.

O cálculo leva em consideração dois parâmetros, *w* (*weigth*) que é um fator de peso (padrão de 0,05) e um grau de ordem (λ). Este cálculo também leva em consideração fatores adicionais, e aqui foram usados o valor de hidrofobicidade, hidrofilia, e massa da cadeia lateral igual ao descrito no trabalho original (CHOU, 2001). Assim, o cálculo considera não apenas a composição dos aminoácidos e sua ordem, como também características que melhor refletem a proteína num modelo discreto.

Na **Figura 2** é possível verificar que há um limite para o valor de λ, sendo o valor máximo possível igual ao tamanho da proteína menos um. O λ limita a quantidade de atributos que podem ser extraídos desse cálculo, sendo uma troca entre o corte pelo tamanho das proteínas e a quantidade de atributos extraídos. No nosso caso foi usado λ=50, e proteínas com tamanhos menores que 50 foram removidas. O cálculo da composição de pseudo-aminoácido foi feito usando a função *extractPAAC* do pacote protr.

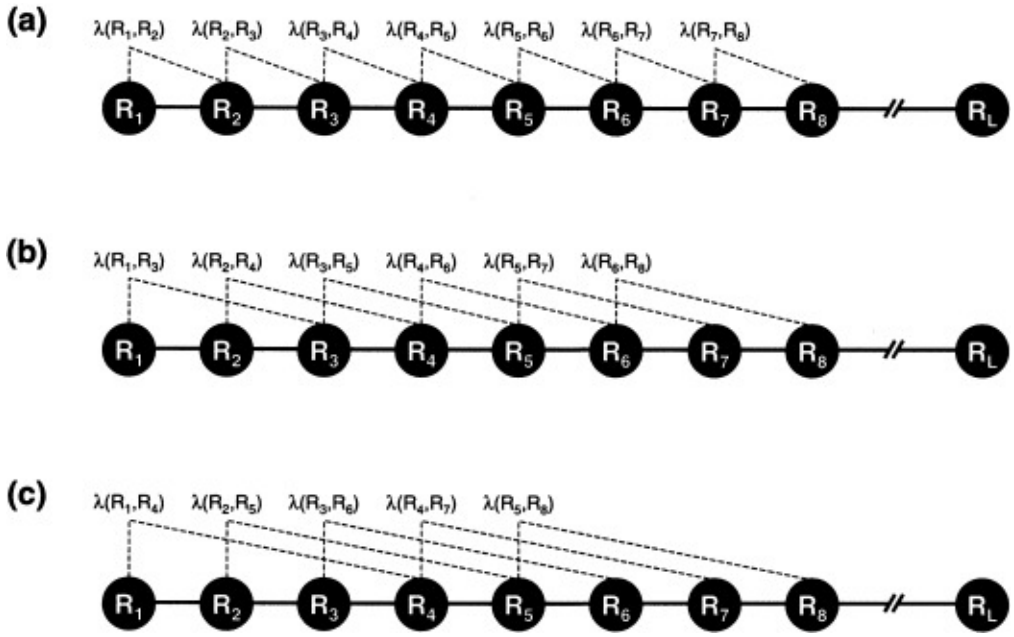


Figura 2: Esquema mostrando a correlação da ordem da sequência para (a) 1ª ordem, (b) 2ª ordem e (c) 3ª ordem ao longo da proteína. *R_1 representa o primeiro resíduo de aminoácido, R_2 o segundo e assim por diante. (a) reflete a correlação entre resíduos mais próximos, (b) entre todo segundo resíduo mais próximo e (c) entre todo terceiro resíduo mais próximo (CHOU, 2001).*

4.3.13. Composição de pseudo-aminoácido anfílico

Esta composição é similar a anterior, com a diferença de que dois fatores de correlação são usados para refletirem diferentes distribuições de hidrofobicidade e hidrofilicidade ao longo de uma proteína, ao invés de um único. Essa diferença é melhor visualizada na Figura 3, tendo H^1 e H^2 como funções de correlação para hidrofobicidade e hidrofilicidade (CHOU, 2005). Os parâmetros foram mantidos iguais aos anteriores, com λ=50, e w=0,05. O cálculo da composição de pseudo-aminoácido anfílico foi feito usando a função *extractAPAAC* do pacote protr, resultando em 120 atributos.

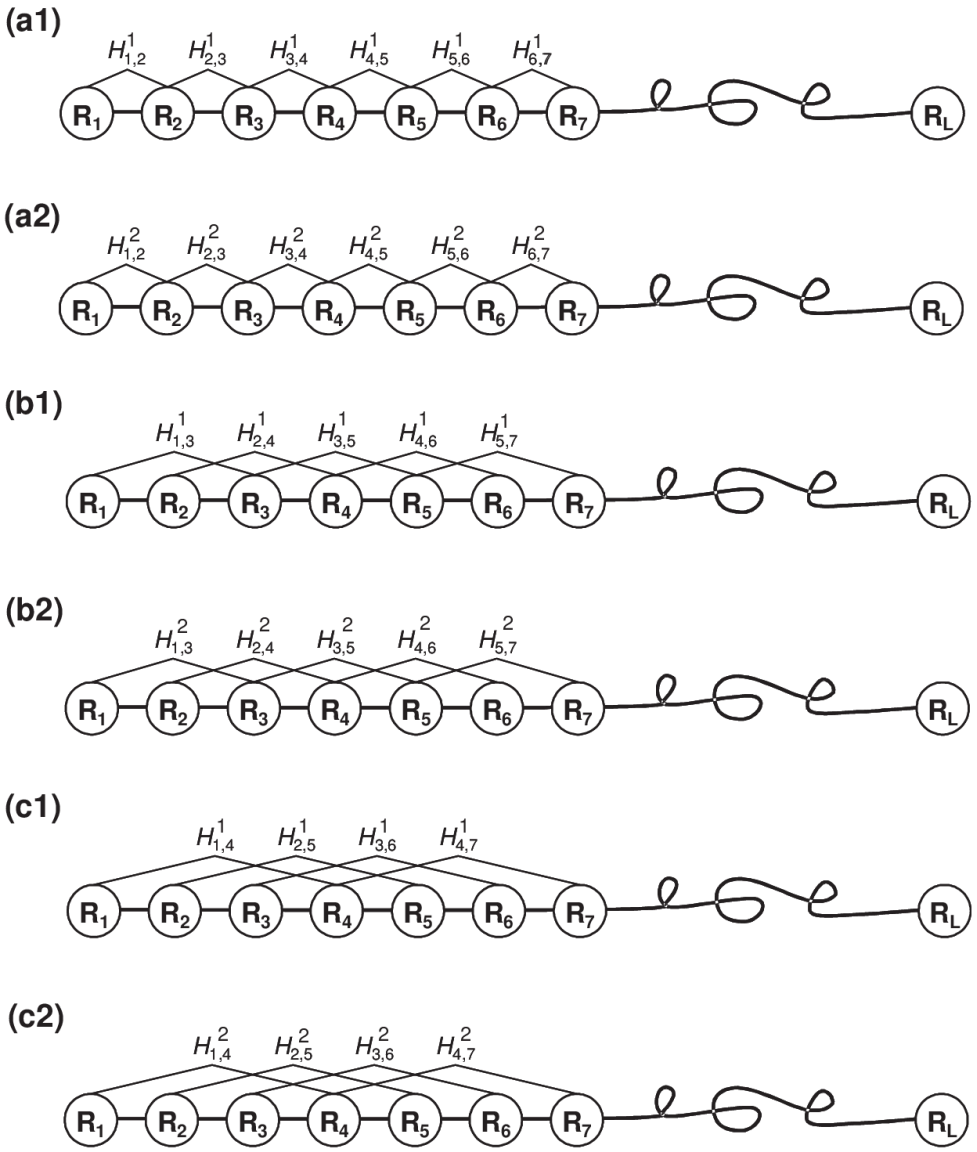


Figura 3: Esquema mostrando a correlação da ordem da sequência. *Para (a1/a2) 1ª ordem, (b1/b2) 2ª ordem e (c1/c2) 3ª ordem ao longo da proteína (CHOU, 2005).*

4.3.14. Tamanho da proteína

O tamanho da proteína é um atributo simples, que já foi utilizado em outros algoritmos de AM e associado à essencialidade (BRANDES; OFER; LINIAL, 2016; HWANG et al., 2009). Ele nada mais é do que o número de resíduos da proteína.

Tabela 2: Atributos, quantidades e programas usados para aminoácidos e nucleotídeos.

Atributos dos nucleotídeos	**Número de atributos**	**Software/pacote**
Informação mútua	17	Este estudo
Informação condicional mútua	65	Este estudo
Entropia de Shannon	2	Este estudo
Covariância auto-cruzada baseada em dinucleotídeo	2888	rDNAse
Covariância auto-cruzada baseada em trinucleotídeo	2100	rDNAse
Composição de pseudo dínucleotídeos	19	rDNAse
Entropia e entalpia de Gibbs	2	VarGibbs
Atributos das proteínas		
Composição físico-química e ponto isoelétrico	10	Seqinr
Informação mútua	401	Este estudo
Informação condicional mútua	8001	Este estudo
Entropia de Shannon	2	Este estudo
Tamanho da proteína	1	Este estudo
Conjoint triad	343	Protr
Moreau-Broto	240	Protr
Moran	240	Protr
Geary	240	Protr
Descritores CTD - Composição (CTDC)	21	Protr
Descritores CTD - Distribuição (CTDD)	105	Protr
Descritores CTD - Transição (CTDT)	21	Protr
Composição de pseudo-aminoácido	70	Protr
Composição de pseudo-aminoácido anfílico	120	Protr

4.4. Desenvolvimento e validação de preditores de genes essenciais em bactérias

O programa para o cálculo dos atributos e treino dos modelos de AM foi escrito na linguagem de programação R. O modelo escolhido e testado foi com o algoritmo implementado

no pacote ranger (WRIGHT; ZIEGLER, 2017) encapsulado no pacote CARET (KUHN, 2018) do R como o método ‘ranger’.

Para acelerar os processos de extração de atributos e treino, ambos foram paralelizados. Enquanto os cálculos para um gene são feitos sequencialmente, cada sequência é independente das outras. Então, para a extração de atributos, cada gene é processado independentemente dos outros em paralelo, e limitado pelo número de threads disponibilizados. Depois, o resultado de cada gene é unido em um objeto do tipo *data frame*, onde as linhas dos genes que não passaram nos filtros (e.g. tamanhos menores que 50 aminoácidos) ou tiveram erros (e.g. bases ambíguas, falha em programas externos), são removidas. O processo de paralelização do treino do algoritmo de AM é realizado internamente pelo pacote caret.

Os testes foram feitos com os dados de duas bactérias filogeneticamente diversas para a finalidade de testar o modelo e compará-lo com resultados de outros trabalhos já publicados (LIU et al., 2017; NIGATU et al., 2017): *Acinetobacter baylyi* ADP1 (Filo Proteobacteria) (DE BERARDINIS et al., 2008) e *Staphylococcus aureus* NCTC 8325 (Filo Firmicutes) (CHAUDHURI et al., 2009). Os GEs e GNEs destas bactérias foram obtidos da base de dados do DEG v15.2. Especificamente, para cada modelo, realizamos a etapa de treinamento em um organismo e a etapa de avaliação final de desempenho do melhor modelo na outra espécie, de modo a avaliarmos como seria o desempenho de nosso modelo em uma situação que emula o uso real de um modelo predizendo dados em organismos nunca utilizados em seu treinamento.

Para o treinamento de modelos de florestas aleatórias (*Random Forest*), alguns parâmetros foram ajustados. Um destes compreende o *mtry*, que é a quantidade de atributos que serão escolhidos aleatoriamente do conjunto total de atributos a cada nó da árvore de decisão. O valor padrão de *mtry* é rotineiramente usado como a raiz quadrada do total de atributos (JAMES et al., 2013). A fim de permitir um ajuste no parâmetro, possibilitando treinar modelos melhores, além do valor padrão também foi incluído o dobro da raiz quadrada do total de atributos. Assim, os valores testados para *mtry* foram os números inteiros mais próximos da raiz quadrada da quantidade de atributos e o dobro deste valor (122 e 244, respectivamente). O segundo parâmetro (*ntree*) compreende a quantidade de árvores de decisões que vão ser combinadas para se gerar o modelo, onde utilizamos 1000 árvores para tal fim. Por fim, a área sob a curva ROC (AUC-ROC, AUC: do inglês *Area Under the Curve*) foi escolhida como variável a ser maximizada, uma vez que a acurácia, que é o padrão do programa, não é adequada

ao problema de classes desbalanceadas[3]. Além disso, esse também foi o valor utilizado por outros trabalhos nesta área (CAMPOS et al., 2019; NIGATU et al., 2017). Assim, para cada valor de *mtry* (122 e 244), realizamos a validação cruzada de 10 vezes repetida 3 vezes (*repeatedcv*), com 1000 árvores cada, e selecionamos o modelo que obtiver o maior valor da AUC- ROC.

4.5. Avaliação do impacto das sequências nucleotídicas de GNEs nos classificadores

Para a averiguação do impacto do erro que observamos nas sequências de GNEs obtidas a partir do banco de dados DEG nos modelos de aprendizado de máquina, construímos diferentes conjuntos de dados onde utilizamos sempre todos os dados dos GEs como exemplos positivos, mas dividimos os GNEs em três conjuntos: A) GNEs que passaram no filtro de identidade e cobertura (valores maiores que 90% para ambos ao compararmos a sequência nucleotídica e proteica via tBLASTN, alinhamentos "verdadeiros") B) GNEs que não passaram no filtro de identidade e cobertura (valores menores que 90% para identidade e/ou cobertura, alinhamentos "falsos") C) Todos os GNEs obtidos do DEG independentemente do status de alinhamento da região codificadora e da sequência proteica (o que deve ser o conjunto que melhor reflete a análise realizada por outros trabalhos (NIGATU et al., 2017; YU et al., 2017)).

Para cada bactéria, treinamos três modelos, cada um utilizando um dos conjuntos de dados de GNEs (A, B e C). Posteriormente, os 3 modelos de *A. baylyi* foram utilizados para predizer os conjuntos de dados A, B e C de *S. aureus*. Igualmente, os 3 modelos de *S. aureus* foram usados para predizer os conjuntos de dados A, B e C de *A. baylyi.* Com os resultados destas predições, produzimos curvas ROCs e obtivemos as respectivas AUCs. A AUC obtida a partir do modelo treinado utilizando o conjunto de dados C de uma bactéria e testada com o conjunto de dados C da outra bactéria foram comparadas com o resultado de Nigatu *et. al* (2017), uma vez que este possivelmente é o cenário análogo a esse estudo.

Um último modelo, usualmente denominado "classificador de Regra Zero" (*zero-rule classifier,* ZR), foi utilizado para comparar os resultados obtidos com uma linha base. Nosso classificador ZR prediz todos os elementos como GEs com a mesma probabilidade, e consequentemente, os resultados possuem uma AUC para a curva ROC de 0,5. Assim, se um resultado é significativamente diferente do obtido por classificadores ZR, ele pode ser

[3] A acurácia é a fração de quantos elementos foram corretamente classificados. Supondo que temos 10% de GEs, se simplesmente classificarmos todos os genes como GNEs, vamos ter 90% de acurácia, afinal, classificamos 90% dos genes corretamente como GNEs. Um resultado muito bom visto por essa métrica, mas que é inútil para nosso objetivo.

considerado melhor que uma classificação aleatória. Para as comparações com as outras curvas, o parâmetro *direction=">"* foi acrescentado ao produzir as curvas ZR através da função *roc*. As comparações estatísticas entre as AUCs das curvas ROCs foram realizadas par a par usando a função *roc.test* do pacote pROC (ROBIN et al., 2011) do R, utilizando o método DeLong (DELONG; DELONG; CLARKE-PEARSON, 1988).

5. Resultados

5.1. Desenvolvimento de rotinas computacionais para o cálculo de atributos homologia-independentes para os genes

Após revisão da literatura científica, escolhemos diversos atributos homologia-independentes que podem eventualmente contribuir para a obtenção de classificadores capazes de distinguir GEs de GNEs. Especificamente, nossa rotina computacional calcula 5093 atributos a serem computados a partir de sequências nucleotídicas e 9815 atributos que podem ser computados a partir de sequências proteicas, totalizando 14908 atributos. O código para importar as sequências nucleotídicas, realizar a sua tradução *in silico* e computar os atributos, bem como para o treinamento e teste do modelo, está disponível como um pacote do R em https://github.com/g1o/GeneEssentiality.

5.2. O banco de dados DEG (v15.2) possui um erro sistêmico nas sequências de nucleotídeos de GNEs

Após revisão da literatura científica, identificamos artigos que fizeram uso de AM e de atributos homologia-independentes para a predição de GEs (CAMPOS et al., 2019; LIU et al., 2017; NIGATU et al., 2017; YU et al., 2017). Embora outros autores também façam uso parcial de atributos homologia-independentes para a predição de GEs, tais modelos foram criados utilizando uma combinação de atributos homologia-independentes com atributos dependentes de homologia e, consequentemente, carecem de capacidade de generalização, uma vez que não podem ser utilizados para organismos não-modelo onde informações extrínsecas não estão disponíveis (DENG et al., 2011; LLOYD et al., 2015, p. 2; SERINGHAUS et al., 2006; SONG; TONG; WU, 2014).

Entretanto, ao iniciarmos nossas rotinas de controle de qualidade das traduções *in silico* de genes codificadores de proteínas disponíveis no banco de dados DEG (versão 15.2, de dezembro de 2017, disponível em http://tubic.org/deg_bak/), observamos que uma fração

considerável de GNEs não apresentam uma região codificadora (ORF) contendo a informação necessária para codificar a proteína também descrita no DEG como sendo o produto do gene em questão. Essa observação foi confirmada após realizarmos o alinhamento par a par via tBLASTn das sequências de nucleotídeos dos GNEs do DEG com a sua sequência de aminoácidos, também extraída do DEG (Figura 4).

Observamos que, dentre os 126.886 GNEs descritos para todas as espécies de procariotos contidas no banco de dados DEG, 84.409 (66,53%) genes possuem uma sequência de nucleotídeos que se alinha corretamente à proteína supostamente codificada (> 90% para identidade e cobertura) enquanto 42477 (33,48%) das sequências apresentaram alinhamento insatisfatório. Destas, 31178 (24,57%), possuíam algum alinhamento com E-values menores que 10, e encontram-se representadas na Figura 4, enquanto 11299 (8,90%) possuem E-value maior que 10 ou não apresentaram alinhamento algum, e não estão representadas na Figura 4. Em contraste, dentre os 18.835 GEs reportados no DEG para organismos procarióticos, 18058 (95,87%) se alinham corretamente às sequências protéicas correspondentes obtidas a partir do DEG (identidade e cobertura > 90%), evidenciando que o problema reportado está restrito aos GNEs.

Esse resultado demonstra que aproximadamente 33,3% das sequências de nucleotídeos dos GNEs do DEG não contêm uma CDS capaz de codificar a proteína descrita para o próprio gene. Adicionalmente, estes alinhamentos concentram-se majoritariamente na região contendo alinhamentos com identidade e cobertura abaixo de 50%, uma região que compreende possivelmente alinhamentos espúrios, enquanto as 84409 sequências selecionadas pelos filtros de identidade e cobertura estão virtualmente todas contidas na coordenada esperada para alinhamentos perfeitos (100% de identidade e cobertura). A partir desse ponto, iremos nos referir às sequências nucleotídicas de GNEs que não alinham com as proteínas supostamente codificadas por essas sequências e às sequências que se alinham como GNEs falsos e verdadeiros, respectivamente.

Para investigarmos os possíveis fenômenos responsáveis pela ausência de alinhamento entre GNEs falsos e suas respectivas proteínas, investigamos a presença de sequências canônicas presentes em regiões codificadoras (códons de início e término, Figura 5). Observamos que as sequências de GEs possuem, em sua vasta maioria, códons canônicos de início e fim de tradução (Figura 5A, “Essential genes”, frequência de 95,6% para códons de início [ATG/GTG/TTG] e de 95,8% de para códons de término [TAA/TAG/TGA]).

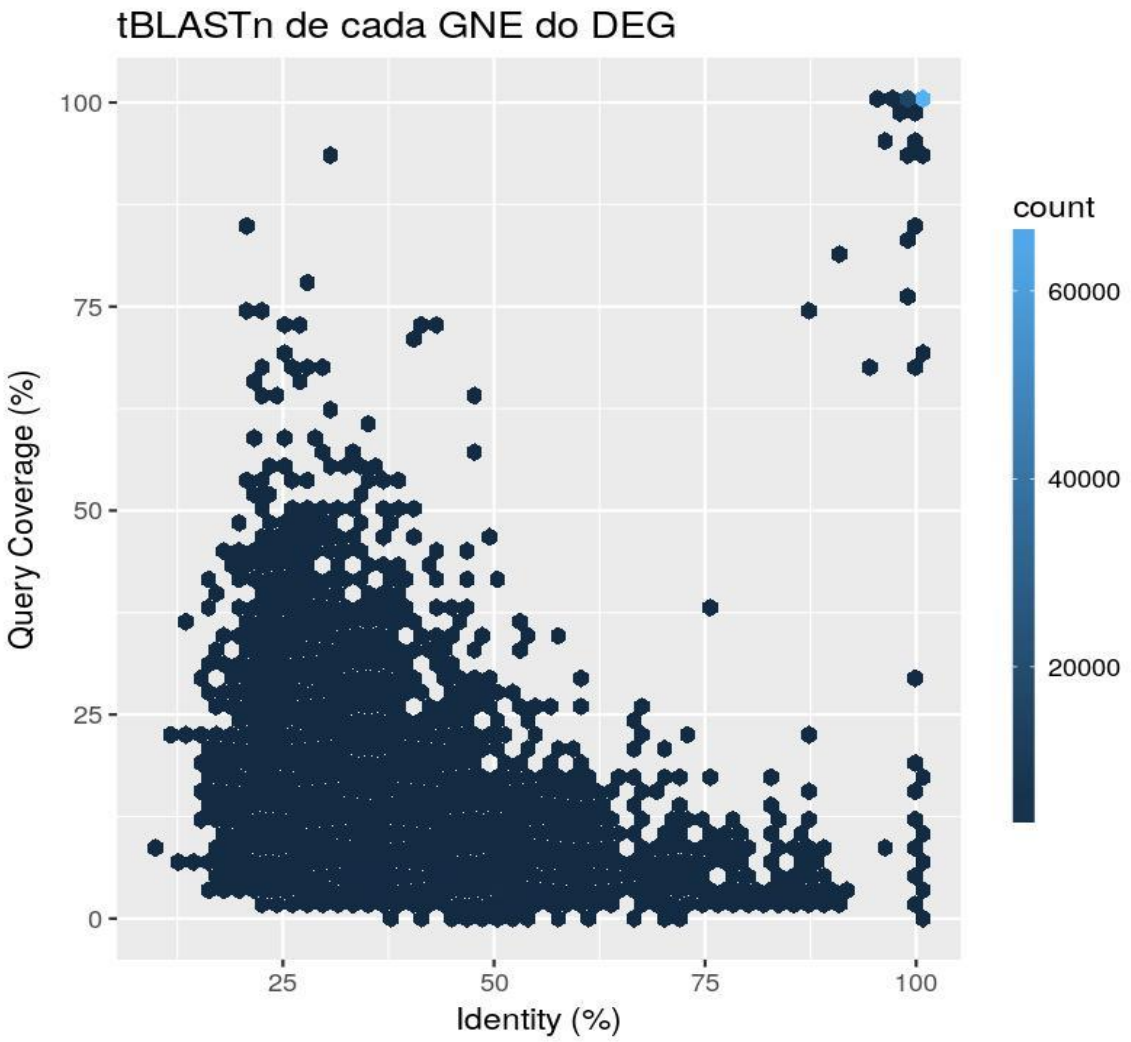
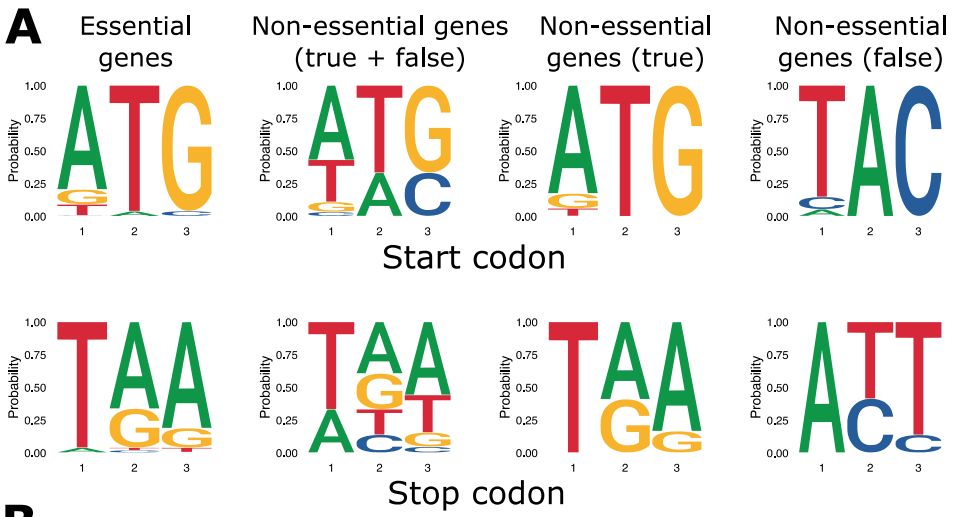


Figura 4: Valores de identidade e cobertura do alinhamento tBLASTn de GNEs do DEG. *Alinhamento feito par a par entre as sequências de nucleotídeo que codificam GNEs e suas próprias proteínas para todas as bactérias presentes no banco de dados DEG.*

.

Em contraste, o conjunto total de GNEs disponíveis no DEG apresentam um perfil claramente distinto de códons de início e de término, onde observamos frequências de 66,3% e 66,6% para os códons canônicos de início e término, respectivamente (Figura 5, "Non-essential genes (true + false)"). Nestas sequências, os nucleotídeos com a segunda maior frequência em cada posição dos códons correspondem ao nucleotídeo complementar do nucleotídeo canônico esperado (e.g. "TAC" para o códon de início "ATG").

Ao avaliarmos os códons de início e término dos conjuntos verdadeiros e falsos de GNEs, observamos que as CDSs verdadeiras apresentam os sinais canônicos similares aos dos genes essenciais, enquanto os nucleotídeos das sequências falsas correspondem ao seu complemento (Figura 5A, "Non-essential genes (true)" e "Non-essential genes (false)").



B

Score	Expect	Method	Identities	Positives	Gaps	Frame
655 bits(1691)	0.0	Compositional matrix adjust.	315/315(100%)	315/315(100%)	0/315(0%)	+1

```
Query  1    MTYHEWKDLALFYSVESTQKFLEKVYILNGINDAKKNSFKNSERFIYFLKHAESFYKQAA  60
            MTYHEWKDLALFYSVESTQKFLEKVYILNGINDAKKNSFKNSERFIYFLKHAESFYKQAA
Sbjct  1    MTYHEWKDLALFYSVESTQKFLEKVYILNGINDAKKNSFKNSERFIYFLKHAESFYKQAA  180

Query  61   YSPLEIKPILLFYGMAQLIKACLITRDPHYPSHTSVLAHGVTTRKRKKQNYCFSDDEVKI  120
            YSPLEIKPILLFYGMAQLIKACLITRDPHYPSHTSVLAHGVTTRKRKKQNYCFSDDEVKI
Sbjct  181  YSPLEIKPILLFYGMAQLIKACLITRDPHYPSHTSVLAHGVTTRKRKKQNYCFSDDEVKI  360

Query  121  QRNGLCVHFMKHLFGQSDIVDERYTMKKLLMAIPELSDIFYFQQKERFMTKVEKDKNEIF  180
            QRNGLCVHFMKHLFGQSDIVDERYTMKKLLMAIPELSDIFYFQQKERFMTKVEKDKNEIF
Sbjct  361  QRNGLCVHFMKHLFGQSDIVDERYTMKKLLMAIPELSDIFYFQQKERFMTKVEKDKNEIF  540

Query  181  VPEEVVINYKMSDSRFAEYMSHHYQWSFTKKNEHGLLFEISPQDKEPWTSTSLLFDMEKN  240
            VPEEVVINYKMSDSRFAEYMSHHYQWSFTKKNEHGLLFEISPQDKEPWTSTSLLFDMEKN
Sbjct  541  VPEEVVINYKMSDSRFAEYMSHHYQWSFTKKNEHGLLFEISPQDKEPWTSTSLLFDMEKN  720

Query  241  QYYIPSQREQFLRLPEMTIHYLILYNVGMIARYETEWWYELLTQHISDDYVLIQQFLLVS  300
            QYYIPSQREQFLRLPEMTIHYLILYNVGMIARYETEWWYELLTQHISDDYVLIQQFLLVS
Sbjct  721  QYYIPSQREQFLRLPEMTIHYLILYNVGMIARYETEWWYELLTQHISDDYVLIQQFLLVS  900

Query  301  EKKFPKYASQFLLHF  315
            EKKFPKYASQFLLHF
Sbjct  901  EKKFPKYASQFLLHF  945
```

Figura 5: Avaliação das regiões codificadoras de genes essenciais e não-essenciais armazenadas no banco de dados DEG. *A) Composição nucleotídica dos códons de início de genes essenciais, genes não-essenciais (todos) e dos conjuntos verdadeiros e falsos de genes não-essenciais. B) Alinhamento par-a-par via tBLASTn entre a sequência proteica obtida a partir do complemento de um gene não-essencial falso (query) e da sequência proteica do mesmo obtida no banco de dados DEG.*

Selecionamos um GNE falso que não apresentou nenhum alinhamento via tBLASTn (gene DNEG10010004, cuja sequência foi salva no *internet archive* e está disponível em: https://web.archive.org/web/20210603144839/http://tubic.tju.edu.cn/deg/information.php?ac=DNEG10010004) e realizamos a operação necessária para a obtenção de sua sequência complementar. O alinhamento via tBLASTn desta sequência complementar com a proteína supostamente codificada por este locus resulta em um alinhamento perfeito em termos de identidade e cobertura (Figura 5B).

Considerando todos os fatos (um erro sistemático observado em aproximadamente um terço das sequências de GNEs de procariotos e que pode ser corrigido através da obtenção de

sequências complementares às sequências nucleotídicas), hipotetizamos que a inconsistência observada se deva a um erro computacional do banco de dados DEG ao extrair essas informações dos genomas destes organismos para GNEs, possivelmente durante a obtenção de sequências do tipo complemento reverso para representar informações da fita de DNA não representada na sequência genômica.

Embora não saibamos exatamente desde quando esse erro existe no DEG, ele possivelmente foi introduzido após a versão 10 do DEG, uma vez que essa é a versão a partir da qual as sequências de GNEs foram adicionadas (LUO et al., 2014). Assim, trabalhos que usam os dados do DEG, e que são anteriores a 2014, não fazem uso dos dados equivocados para CDS de GNEs do DEG. Entretanto, trabalhos que utilizam versões posteriores do DEG, e que fazem uso de GNEs fornecidos por esse banco de dados, possivelmente contém um terço das sequências de GNEs contendo o erro aqui descrito. Na versão mais atual do banco de dados DEG, não há a disponibilização das sequências de GNEs, o que sugere, embora não tenham relatado o erro, que os administradores do banco também encontraram essa inconsistência e optaram por remover o conjunto de dados de GNEs (LUO et al., 2021).

5.3. Predição de genes essenciais em organismos procarióticos utilizando atributos intrínsecos: qual é o estado-da-arte real?

Para verificarmos a eficiência das rotinas computacionais desenvolvidas nesse projeto, bem como para avaliarmos como as inconsistências observadas nos conjuntos de GNEs do DEG podem ter influenciado o resultado de preditores de GEs via AM previamente publicados, utilizamos os dados de sequência do DEG de GEs e GNEs disponíveis para duas espécies filogeneticamente distantes e originalmente presentes nos trabalhos de (NIGATU et al., 2017; YU et al., 2017): *Acinetobacter baylyi* ADP1 (Filo Proteobacteria) e *Staphylococcus aureus* NCTC 8325 (Filo Firmicutes). Especificamente, utilizamos todas as sequências de GEs de uma determinada espécie como classe “GE” e desenvolvemos, para cada espécie, três modelos utilizando diferentes conjuntos de GNEs como classe “NGE”:

Modelo A (estado-da-arte, CDSs verdadeiras): nesse modelo, utilizamos como exemplos de GNEs todos aqueles genes cujas CDS apresentaram alinhamento com identidade e cobertura maior que 90% com a sua proteína correspondente. Este modelo visa estimar qual seria o estado-da-arte da predição de GEs ao utilizar conjuntos de GNEs que apresentam alta similaridade com a proteína codificada pelo gene e que, portanto, compreendem as verdadeiras CDSs dos GNEs.

Modelo B (pior cenário, CDSs falsas): nesse modelo, utilizamos como exemplos de GNEs todos aqueles cujas CDS não passaram no filtro do alinhamento par-a-par via tBLASTn (identidade e/ou cobertura < 90%), sendo assim considerados como CDSs falsas. Este modelo visa observar os efeitos extremos de se treinar um modelo para a predição de GEs utilizando somente sequências que não correspondem às CDSs verdadeiras dos GNEs.

Modelo C (análogo à literatura, CDSs totais): Nesse modelo, utilizamos todos os GNEs sem qualquer tipo de filtro, assim como estavam disponíveis no DEG. Este classificador foi desenvolvido visando permitir a comparação entre o modelo que consideramos estado-da-arte e os classificadores já desenvolvidos, e possivelmente corresponde a uma análise equivalente à feita por Nigatu et al. (2017).

Como validação de cada um dos modelos desenvolvidos, utilizamos os conjuntos de dados utilizados para treinar os modelos "A", "B" e "C" na outra espécie (espécie-teste). Quando utilizamos os classificadores treinados utilizando somente CDSs verdadeiras de GNEs da espécie-teste (Figura 6), todas as curvas apresentaram diferenças significativas entre os valores de suas AUCs (comparações par-a-par, função *roc.test*, pacote pROC, todos p-valores < 0,01). Adicionalmente, observamos que os modelos treinados com GNEs verdadeiros e validados com conjuntos verdadeiros (curvas pretas) apresentam desempenho significativamente superior às validações utilizando os conjuntos de sequências falsas (curva vermelha) ou todas as sequências descritas como GNEs (curva azul). Adicionalmente, os modelos apresentam desempenho intermediário quando validados utilizando o conjunto total de sequências (curva azul).

Considerados em conjunto, estes resultados demonstram que é possível realizar a predição satisfatória de GEs em procariotos em um cenário de uso em organismos filogeneticamente distantes aos utilizados para treinar o modelo. Observamos também que a presença das regiões GNEs falsas, compostas majoritariamente de regiões não-codificadoras, piora significativamente o desempenho dos modelos treinados somente com as regiões verdadeiras, o que fornece evidências de que os resultados de Nigatu et al. (2017) possivelmente apresentam desempenho inferior aos preditores treinados utilizando CDSs verdadeiras, somente.

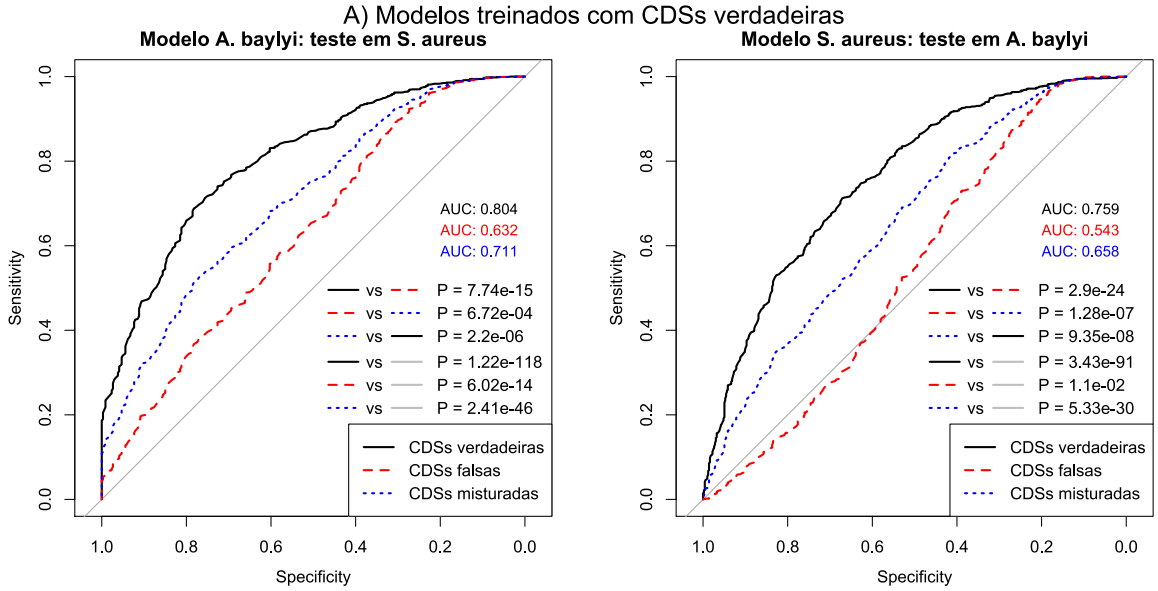


Figura 6: Desempenho dos modelos treinados com GNEs contendo CDSs verdadeiras de uma espécie (modelo A, estado-da-arte) e validados na espécie-teste. *Linha preta: validação utilizando CDSs verdadeiras.* ***Linha vermelha:*** *validação utilizando CDSs falsas. Linha azul: validação utilizando CDS sem filtro (verdadeiras + falsas). A linha cinza representa as predições de um modelo aleatório, no caso foi usado a Regra Zero para comparação.*

Ao compararmos as AUCs com a linha base produzida por um modelo ZR, apenas o modelo treinado em *S. aureus* e testado com CDSs falsas de *A. baylyi* não teve diferença significativa (p-valor=0.011), embora este modelo apresente desempenho significativo quando avaliado com GNEs verdadeiros. Este resultado fornece evidência adicional de que este modelo, embora seja capaz de diferenciar GEs de GNEs verdadeiros, tem um sucesso limitado nas sequências de GNEs falsas.

Como avaliação final do desempenho dos classificadores obtidos para o conjunto de dados "A", que teve AUCs de 0,804 e 0,759 (Figura 6, curvas pretas), comparamos nossos classificadores com um trabalho que menciona explicitamente a utilização de sequências genômicas extraídas do NCBI, e que fez uso somente da informação de rótulos, apenas, obtidos a partir do DEG (LIU et al., 2017).

Liu et. al, (2017). obtiveram uma média de AUCs para 31 bactérias de 0,794, e as comparações par-a-par foram de 0,68 para o treinamento em *A. baylyi* e teste em e *S. aureus* e 0,73 para o treinamento em *S. aureus* e teste em *A. baylyi*. Assim, consideramos que nosso classificador utilizando sequências protéicas e nucleotídicas consegue predizer GEs e GNEs tão bem quanto, ou até melhor, do que classificadores treinados utilizando somente informações nucleotídicas.

Prosseguimos realizando o treinamento de um modelo capaz de distinguir entre GEs e um conjunto de GNEs composto exclusivamente de sequências falsas (Figura 7). Nesse cenário extremo, buscamos destacar as consequências da utilização de regiões de baixa qualidade na etapa de treinamento de modelos e de validação deles.

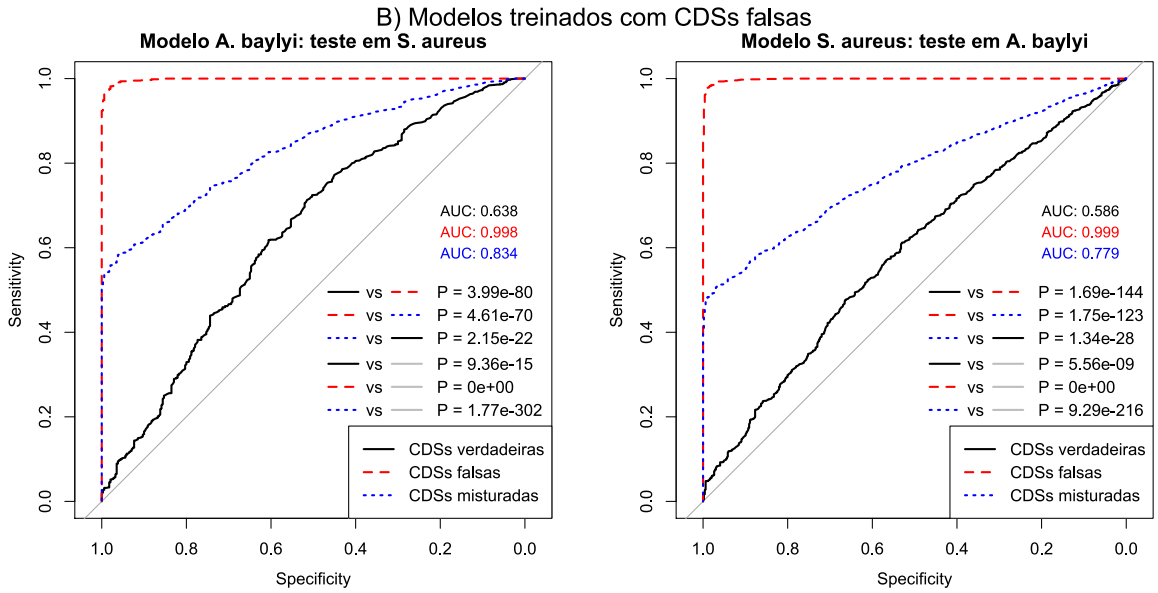


Figura 7: Desempenho dos modelos treinados com GNEs contendo CDSs falsas de uma espécie (modelo B, pior cenário) e validados na espécie-teste. *Linha preta: validação utilizando CDSs verdadeiras.* ***Linha vermelha:*** *validação utilizando CDSs falsas. Linha azul: validação utilizando CDS sem filtro (verdadeiras + falsas). A linha cinza representa as predições de um modelo aleatório, no caso foi usado a Regra Zero para comparação.*

Observamos que os modelos aqui treinados apresentam o melhor desempenho global dentre todos os modelos quando avaliados utilizando as sequências falsas de GNEs do organismo teste (Figura 7, curvas vermelhas). Adicionalmente, observamos também que o modelo apresenta desempenho intermediário ao avaliar o conjunto contendo todas as sequências de GNEs (Figura 7, curvas azuis), e apresenta os piores desempenhos ao ser utilizado para distinguir GEs de GNEs verdadeiros (Figura 7, curvas pretas).

Assim, concluímos com esse experimento que a presença de sequências de GNEs falsas, embora permita a produção dos classificadores mais eficientes dentre todos os que desenvolvemos nessa análise, são treinados em um cenário que obviamente não corresponde ao desejável. Embora não tenhamos investigado essa questão mais profundamente, acreditamos que, uma vez que as regiões GNEs falsas não descrevem de maneira adequada as informações contidas em regiões codificadoras, os modelos produzidos aqui estejam na verdade aprendendo

a distinguir regiões codificadoras de regiões não-codificadoras. Essa hipótese, se confirmada, explicaria o excelente desempenho destes modelos, uma vez que a predição de ORFs em genomas procarióticos é uma tarefa consideravelmente mais simples se comparada à predição de GEs.

Os modelos treinados utilizando todas as CDSs de GNEs disponibilizadas pelo banco de dados DEG, independentemente da qualidade de seu alinhamento com as proteínas codificadas, consistem na simulação mais fidedigna às análises realizadas anteriormente visando a predição de GEs em procariotos (Figura 7) (NIGATU et al., 2017; YU et al., 2017). Estes trabalhos mencionam explicitamente a utilização das sequências de GEs e GNEs conforme fornecido pelo banco de dados DEG, embora não mencionem nenhum tipo de controle de qualidade interno como o realizado em nossa análise. Além disso, estes trabalhos utilizaram somente sequências nucleotídicas para a construção de seus classificadores, o que dificulta a detecção do erro no DEG. Assim, assumimos que o erro que detectamos não foi levado em consideração nos mesmos.

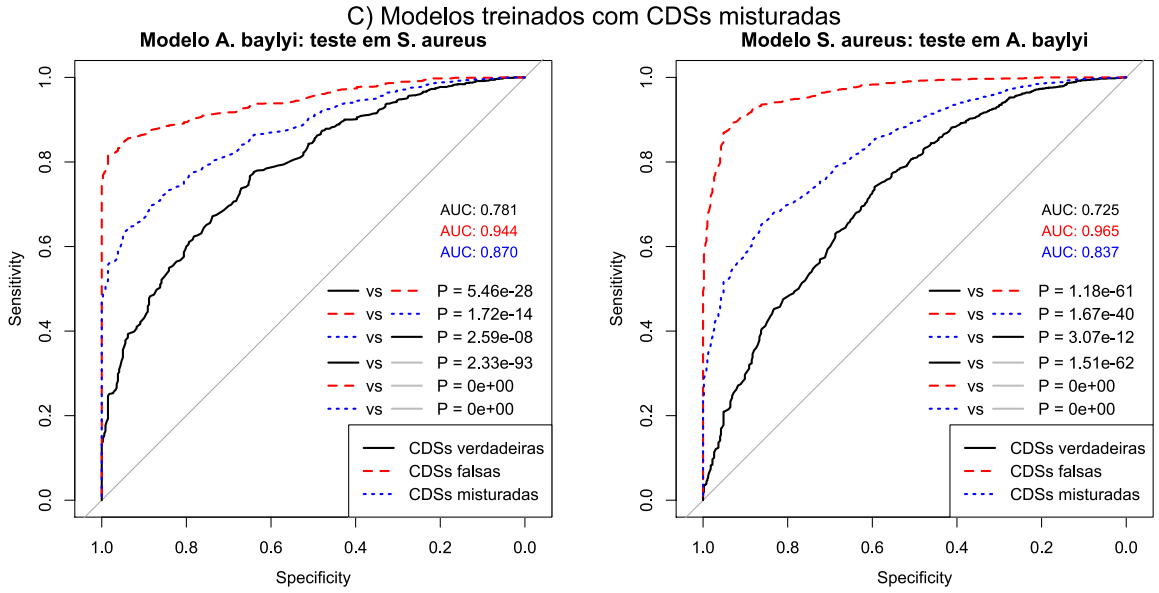


Figura 8: Desempenho dos modelos treinados com GNEs contendo CDSs sem filtro (modelo C, verdadeiras + falsas) de uma espécie e validados na espécie-teste. Linha preta: validação utilizando CDSs verdadeiras. *Linha vermelha: validação utilizando CDSs falsas. Linha azul: validação utilizando CDS sem filtro (verdadeiras + falsas). A linha cinza representa as predições de um modelo aleatório, no caso foi usado a Regra Zero para comparação.*

Novamente, observamos que todas as curvas tiveram suas AUCs melhores que o modelo ZR, sendo também significativamente diferentes entre si (p-valor<0.01). Similarmente aos modelos treinados somente com CDSs falsas de GNEs, os modelos treinados com todas as

CDSs de GNEs também apresentam o melhor desempenho no conjunto de dados de teste que compreende somente CDSs falsas (Figura 8, curvas vermelhas, AUCs de 0,944, para o modelo *A. baylyi* testado em *S. aureus* e de 0,965 para o modelo *S. aureus* testado em *A. baylyi*), o que indica que estes modelos derivam regras eficientes para a tarefa trivial de separar CDSs falsas de GNEs de CDSs de GEs.

Novamente de maneira similar aos modelos treinados somente em CDSs falsas de GNEs, observamos um desempenho intermediário do modelo treinado com todas as CDSs de GNE quando utilizado para avaliar todas as CDSs de GNEs (Figura 8, curvas azuis). Estas curvas representariam um cenário análogo ao utilizado por Nigatu et al. (2017) em seu estudo e, interessantemente, os valores das curvas ROC obtidos em nossas análises são bastante próximos aos reportados por estes autores. Especificamente, obtivemos valores de AUCs de 0.87 e 0.84 para os modelos treinados em *A. baylyi* e validados em *S. aureus* e vice-versa, respectivamente, ao passo que Nigatu reporta AUCs de 0.84 e 0.83, respectivamente. Ressaltamos também que os modelos treinados utilizando todas as CDSs de GNEs apresentam o seu pior desempenho quando validados utilizando somente as CDSs verdadeiras de GNEs (Figura 8, curvas pretas). Avaliados em conjunto, os resultados que obtivemos indicam que o desempenho real de preditores desenvolvidos utilizando todos os dados do de GNEs do DEG encontra-se inflado artificialmente tanto em função do treinamento utilizando CDSs falsas de GNEs como em função da utilização de sequências de GNEs de espécies-teste contendo CDSs falsas de GNEs como conjunto de validação.

Como experimento final, avaliamos o desempenho dos modelos desenvolvidos em nosso trabalho e que julgamos refletir o desempenho real de preditores de GEs em procariotos utilizando somente atributos intrínsecos quando treinados e validados com dados verdadeiros de GNEs comparado aos modelos análogos publicados como, por exemplo, por Nigatu et al., (2017), os quais utilizam todos os dados de GNEs disponíveis no DEG, incluindo as sequências contendo erros.

Inicialmente, avaliamos o desempenho dos modelos que consideramos corretos (Figura 6, curva preta) e dos modelos análogos à literatura científica para classificar o conjunto verdadeiro de GNEs (Figura 8, curva preta), de modo a avaliar o desempenho destes modelos em um cenário contendo somente dados corretos (Figura 9a e Figura 9b). Nesse experimento, observamos que os modelos que treinamos com dados corretos apresentam desempenho significativamente superior ao apresentado pelos modelos treinados utilizando todas as sequências de GNEs. Assim, concluímos que o treinamento utilizando somente os conjuntos de dados de GNEs corretos produz modelos melhores do que os modelos análogos aos já

publicados e treinados utilizando toda as GNEs.

Posteriormente, avaliamos o desempenho dos modelos corretos que treinamos quando utilizados para predizer GNEs verdadeiros versus o desempenho dos modelos análogos à literatura (treinados com GNEs verdadeiros mais falsos) quando utilizados para predizer todos os GNEs (verdadeiros mais falsos). Esse experimento visa comparar os resultados que consideramos o estado-da-arte verdadeiro da predição de GEs em organismos procarióticos versus os resultados já publicados, os quais foram treinados e validados utilizando todas as sequências de GNEs.

De maneira preocupante, observamos que o experimento que realizamos para simular os modelos já publicados apresentam desempenho significativamente melhor do que os modelos que utilizam os dados adequados, o que reforça a hipótese de que a predição de GEs em procariotos já reportada em estudos científicos anteriores encontra-se possivelmente inflada artificialmente em função dos vieses descritos no conjunto total sequências de GNEs do DEG.

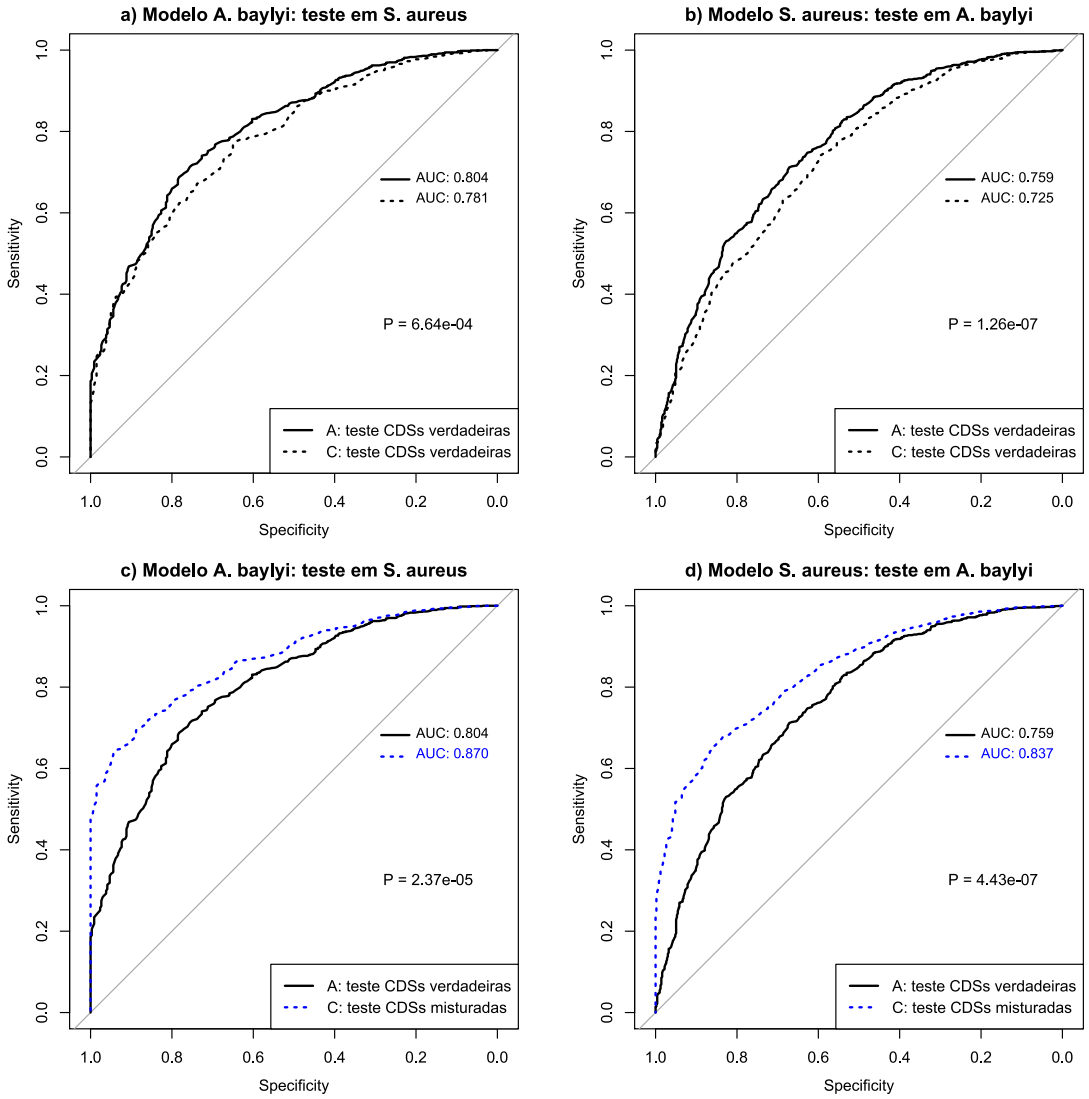


Figura 9: Avaliação do desempenho de classificadores de genes essenciais treinados com conjuntos verdadeiros e falsos de genes não-essenciais. *a) Modelos treinados com CDSs verdadeiras (curva preta) ou CDSs misturadas (curva preta pontilhada) de A. baylyi e validado utilizando CDSs verdadeiras de S. aureus*. b) *Modelos treinados com CDSs verdadeiras (curva preta) ou CDSs misturadas (curva preta pontilhada) de S. aureus e validado utilizando CDSs verdadeiras de A. baylyi. c) Modelos treinados com CDSs verdadeiras de GNEs (curva preta) ou todas as CDS de GNEs (curva azul) de A. baylyi e validados utilizando CDSs verdadeiras ou todas as CDSs de GNEs de S. aureus, respectivamente. d) Modelos treinados com CDSs verdadeiras de GNEs (curva preta) ou todas as CDS de GNEs (curva azul) de S. aureus e validados utilizando CDSs verdadeiras ou todas as CDSs de GNEs de A. baylyi, respectivamente.*

6. Discussão

A detecção de GEs é de fundamental importância para diversos campos do conhecimento, permitindo o desenvolvimento de áreas tão diversas que vão desde a síntese completas de genomas mínimos celulares até o desenvolvimento de drogas para parasitos de interesse médico e agropecuário (BIRHANU et al., 2018; CHANG et al., 2021; QURESHI et al., 2021; ULRICH et al., 2015). Não surpreendentemente, diversos grupos de pesquisa já reportaram análises em larga escala de ensaios para a avaliação de essencialidade em diferentes espécies de procariotos e eucariotos (CHEN et al., 2017; LUO et al., 2021).

A integração da informação de essencialidade e não-essencialidade para diversos organismos é de fundamental importância para o desenvolvimento e validação de preditores de genes essenciais, os quais podem ser utilizados para acelerar o processo de descoberta de tais genes sem a necessidade de ensaios em larga escala. O banco de dados DEG provê um serviço de fundamental importância para a comunidade científica ao integrar informações sobre essencialidade para diversos organismos, fornecendo subsídio para a produção de diversos classificadores.

Embora a versão atual desse banco não provenha dados de sequências para GNEs, versões anteriores desse banco disponibilizavam acesso aos dados de sequência de conjuntos de GEs e também de GNEs. Identificamos um de um erro na especificação das regiões codificadoras de aproximadamente um terço dos GNEs de bactérias na versão 15.7 do DEG, as quais representam o complemento das sequências codificadoras reais. Acreditamos que esse erro provavelmente decorre da produção de sequências complemento reverso de maneira automatizada. Rotinas simples de avaliação de auto consistência de dados, como o alinhamento das sequências proteicas obtidas via tradução *in silico* de regiões codificadoras e sequências proteicas obtidas diretamente do DEG, bem como a busca por sinais canônicos de regiões codificadoras que realizamos nas sequências de GNEs, deveriam ser utilizadas sempre que possível como protocolos internos e externos de controle de qualidade, de maneira a minimizar possíveis erros dessa natureza.

Além dos trabalhos de (NIGATU et al., 2017; YU et al., 2017), nos quais há menção explícita no uso de dados de GNEs do DEG para o treinamento de classificadores de GEs, detectamos outros trabalhos publicados na literatura científica que também afirmam fazer uso da informação de sequência de GNEs do DEG em versões anteriores do banco de dados (ZHOU; QI; REN, 2021; ZHOU; YU, 2014), embora não haja menção explícita do uso de dados de sequência. Portanto, sugerimos que os mesmos devem ser avaliados com cautela.

Aqui ressaltamos também um problema de reprodutibilidade, uma vez que diversos desses trabalhos listam também os números de acesso dos genomas como fonte primária dos dados de sequência, o que traz dúvida quanto à origem real da informação das sequências utilizadas. Como exceção, mencionamos o único artigo que disponibiliza, além do código-fonte utilizado na análise, uma cópia dos dados de sequência do DEG utilizado em sua página do Github (HASAN; LONARDI, 2020). Neste caso, foi possível constatar o uso das sequências com erro do DEG por parte desses autores de maneira objetiva.

Outros artigos só utilizam as informações do DEG para rotular GEs e GNEs, mas mencionam explicitamente que os dados genômicos foram obtidos de outras fontes, como do NCBI (DILUCCA; CIMINI; GIANSANTI, 2018; LIU et al., 2017; XU; GUO; LIU, 2020). Assim, consideramos que estes trabalhos aparentemente não utilizaram os dados incorretos, e seus resultados devem refletir com mais fidedignidade o estado-da-arte da predição de GEs em procariotos.

7. Conclusão

Após a detecção do erro sistemático nas sequências de GNEs do DEG, realizamos diversos experimentos controlados *in silico* que demonstram objetivamente que, mesmo após a remoção do erro sistemático das sequências de GNEs do banco de dados DEG, é possível produzir preditores de GEs para procariotos com desempenho satisfatório em espécies filogeneticamente distantes.

Também demonstramos que o treinamento e a validação de preditores de GEs utilizando todas as sequências de GNEs do DEG, incluindo as sequências inconsistentes representando o complemento das regiões codificadoras, produzem resultados artificialmente inflados em função da presença do erro nas sequências de GNEs.

Finalizamos constatando que protocolos básicos para análises de bioinformática devem ser usados, assim como os controles internos de qualidade utilizados para rotinas de laboratórios convencionais. Um desses protocolos deve incluir a verificação de sequências codificadoras de proteínas, que são muito bem estabelecidas para a vasta maioria dos casos.

8. Referências do capítulo 1

ACENCIO, M. L.; LEMKE, N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. **BMC Bioinformatics**, v. 10, n. 1, p. 290, 2009.

AIRS, P.; BARTHOLOMAY, L. RNA Interference for Mosquito and Mosquito-Borne Disease Control. **Insects**, v. 8, n. 1, p. 4, 5 jan. 2017.

AROMOLARAN, O. et al. Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features. **Computational and Structural Biotechnology Journal**, v. 18, p. 612–621, 2020.

BACHMAN, P. M. et al. Characterization of the spectrum of insecticidal activity of a double-stranded RNA with targeted activity against Western Corn Rootworm (<Emphasis Type="Italic">Diabrotica virgifera virgifera</Emphasis> LeConte). **Transgenic Research**, v. 22, n. 6, p. 1207–1222, 1 dez. 2013.

BALDI, P.; SADOWSKI, P.; WHITESON, D. Searching for exotic particles in high-energy physics with deep learning. **Nature Communications**, v. 5, n. 1, dez. 2014.

BARUTCUOGLU, Z.; SCHAPIRE, R. E.; TROYANSKAYA, O. G. Hierarchical multi-label prediction of gene function. **Bioinformatics**, v. 22, n. 7, p. 830–836, 1 abr. 2006.

BAUM, J. A. et al. Control of coleopteran insect pests through RNA interference. **Nature Biotechnology**, v. 25, n. 11, p. 1322–1326, nov. 2007.

BELLEN, H. J. et al. The Drosophila Gene Disruption Project: Progress Using Transposons With Distinctive Site Specificities. **Genetics**, v. 188, n. 3, p. 731–743, 1 jul. 2011.

BENGIO, Y.; GRANDVALET, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. **J. Mach. Learn. Res.**, v. 5, p. 1089–1105, Dezembro 2004.

BHASKARAN, R.; PONNUSWAMY, P. K. Positional flexibilities of amino acid residues in globular proteins. **International Journal of Peptide and Protein Research**, v. 32, n. 4, p. 241–255, Outubro 1988.

BIGELOW, C. C. On the average hydrophobicity of proteins and the relation between it and protein structure. **Journal of Theoretical Biology**, v. 16, n. 2, p. 187–211, Agosto 1967.

BIRHANU, B. T. et al. *In silico* analysis of putative drug and vaccine targets of the metabolic pathways of *Actinobacillus pleuropneumoniae* using a subtractive/comparative genomics approach. **Journal of Veterinary Science**, v. 19, n. 2, p. 188, 2018.

BRANDES, N.; OFER, D.; LINIAL, M. ASAP: a machine learning framework for local protein properties. **Database: The Journal of Biological Databases and Curation**, v. 2016,

2016.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

CAMPOS, T. L. et al. An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-Derived Features. **Computational and Structural Biotechnology Journal**, v. 17, p. 785–796, 8 jun. 2019.

CAMPOS, T. L. et al. Combined use of feature engineering and machine-learning to predict essential genes in Drosophila melanogaster. **NAR Genomics and Bioinformatics**, v. 2, n. 3, 1 set. 2020.

CHANG, L. et al. Targeting pan-essential genes in cancer: Challenges and opportunities. **Cancer Cell**, v. 39, n. 4, p. 466–479, 12 abr. 2021.

CHAPMAN, A. D. **Numbers of living species in Australia and the world**. 2nd. ed. Australia: Department of the Environment, Water, Heritage and the Arts Canberra, 2009.

CHARIF, D.; LOBRY, J. R. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. Em: **Structural approaches to sequence evolution**. [s.l.] Springer, 2007. p. 207–232.

CHARTON, M. Electrical Effect Substituent Constants for Correlation Analysis. Em: **Progress in Physical Organic Chemistry**. [s.l.] John Wiley & Sons, Ltd, 1981. p. 119–251.

CHARTON, M.; CHARTON, B. I. The structural dependence of amino acid hydrophobicity parameters. **Journal of Theoretical Biology**, v. 99, n. 4, p. 629–644, Dezembro 1982.

CHAUDHURI, R. R. et al. Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). **BMC Genomics**, v. 10, n. 1, p. 291, 1 jul. 2009.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**... Em: KDD '16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. San Francisco California USA: ACM, 13 ago. 2016. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 12 dez. 2020

CHEN, W. et al. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. **Nucleic Acids Research**, v. 41, n. 6, p. e68–e68, abr. 2013.

CHEN, W.-H. et al. OGEE: an online gene essentiality database. **Nucleic Acids Research**, v. 40, n. D1, p. D901–D906, 1 jan. 2012.

CHEN, W.-H. et al. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. **Nucleic Acids Research**, v.

45, n. D1, p. D940–D944, 4 jan. 2017.

CHENG, J. et al. A new computational strategy for predicting essential genes. **BMC Genomics**, v. 14, n. 1, p. 910, 2013.

CHENG, J. et al. Training Set Selection for the Prediction of Essential Genes. **PLoS ONE**, v. 9, n. 1, p. e86805, 22 jan. 2014.

CHOTHIA, C. The nature of the accessible and buried surfaces in proteins. **Journal of Molecular Biology**, v. 105, n. 1, p. 1–12, 25 jul. 1976.

CHOU, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. **Proteins: Structure, Function, and Genetics**, v. 43, n. 3, p. 246–255, 15 maio 2001.

CHOU, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. **Bioinformatics**, v. 21, n. 1, p. 10–19, 1 jan. 2005.

CHURCH, K. W.; HANKS, P. Word Association Norms, Mutual Information, and Lexicography. **Computational Linguistics**, v. 16, n. 1, p. 22–29, 1990.

CID, H. et al. Hydrophobicity and structural classes in proteins. **Protein Engineering, Design and Selection**, v. 5, n. 5, p. 373–375, 1992.

COUTINHO, T. J. D.; FRANCO, G. R.; LOBO, F. P. Homology-independent metrics for comparative genomics. **Computational and structural biotechnology journal**, v. 13, p. 352–357, 2015.

DAYHOFF, M.; SCHWARTZ, R.; ORCUTT, B. 22 a model of evolutionary change in proteins. **Atlas of protein sequence and structure**, v. 5, p. 345–352, 1978.

DE BERARDINIS, V. et al. A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. **Molecular Systems Biology**, v. 4, 4 mar. 2008.

DELONG, E. R.; DELONG, D. M.; CLARKE-PEARSON, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. **Biometrics**, v. 44, n. 3, p. 837–845, set. 1988.

DENG, J. et al. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. **Nucleic Acids Research**, v. 39, n. 3, p. 795–807, fev. 2011.

DILUCCA, M.; CIMINI, G.; GIANSANTI, A. Essentiality, conservation, evolutionary pressure and codon bias in bacterial genomes. **Gene**, v. 663, p. 178–188, jul. 2018.

DOBZHANSKY, T. Genetics of natural populations; recombination and variability in populations of Drosophila pseudoobscura. **Genetics**, v. 31, p. 269–290, maio 1946.

DONG, Q.; ZHOU, S.; GUAN, J. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. **Bioinformatics**, v. 25, n. 20, p. 2655–2662, 15

out. 2009.

DÖNITZ, J. et al. iBeetle-Base: a database for RNAi phenotypes in the red flour beetle Tribolium castaneum. **Nucleic Acids Research**, v. 43, n. D1, p. D720–D725, 28 jan. 2015.

DUBCHAK, I. et al. Prediction of protein folding class using global description of amino acid sequence. **Proceedings of the National Academy of Sciences of the United States of America**, v. 92, n. 19, p. 8700–8704, Setembro 1995.

DUBCHAK, I. et al. Recognition of a protein fold in the context of the SCOP classification. **Proteins: Structure, Function, and Bioinformatics**, v. 35, n. 4, p. 401–407, 1 jun. 1999.

FIRE, A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. **Nature**, v. 391, p. 806, Fevereiro 1998.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

GERDES, S. Y. et al. Experimental Determination and System Level Analysis of Essential Genes in *Escherichia coli* MG1655. **Journal of Bacteriology**, v. 185, n. 19, p. 5673–5684, 1 out. 2003.

GLASS, J. I. et al. Essential genes of a minimal bacterium. **Proceedings of the National Academy of Sciences**, v. 103, n. 2, p. 425–430, 10 jan. 2006.

GRISHOK, A. RNAi mechanisms in *Caenorhabditis elegans*. **FEBS Letters**, v. 579, n. 26, p. 5932–5939, 31 out. 2005.

HALL, M. et al. The WEKA Data Mining Software: An Update. **SIGKDD Explor. Newsl.**, v. 11, n. 1, p. 10–18, nov. 2009.

HANSEN, L. M. Economic damage threshold model for pollen beetles (Meligethes aeneus F.) in spring oilseed rape (Brassica napus L.) crops. **Crop Protection**, v. 23, n. 1, p. 43–46, jan. 2004.

HASAN, M. A.; LONARDI, S. DeeplyEssential: a deep neural network for predicting essential genes in microbes. **BMC Bioinformatics**, v. 21, n. 14, p. 367, 30 set. 2020.

HINTON, G. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. **IEEE Signal Processing Magazine**, v. 29, n. 6, p. 82–97, nov. 2012.

HORN, T.; SANDMANN, T.; BOUTROS, M. Design and evaluation of genome-wide libraries for RNA interference screens. **Genome Biology**, v. 11, n. 6, p. R61, 2010.

HU, W. et al. Essential Gene Identification and Drug Target Prioritization in Aspergillus fumigatus. **PLoS Pathogens**, v. 3, n. 3, p. e24, 2007.

HUERTA-CEPAS, J.; DOPAZO, J.; GABALDÓN, T. ETE: a python Environment for Tree

Exploration. **BMC Bioinformatics**, v. 11, n. 1, p. 24, 2010.

HUTCHISON, C. A. et al. Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. **Science**, v. 286, n. 5447, p. 2165, Dezembro 1999.

HUTCHISON, C. A. et al. Design and synthesis of a minimal bacterial genome. **Science**, v. 351, n. 6280, p. aad6253–aad6253, 25 mar. 2016.

HWANG, Y.-C. et al. Predicting essential genes based on network and sequence analysis. **Molecular BioSystems**, v. 5, n. 12, p. 1672, 2009.

JAMES, G. et al. **An Introduction to Statistical Learning**. 1. ed. New York, NY: Springer New York, 2013. v. 103

JEROME FRIEDMAN; TREVOR HASTIE; ROBERT TIBSHIRANI. Additive logistic regression: a statistical view of boosting (Withdiscussion and a rejoinder by the authors). **The Annals of Statistics**, v. 28, n. 2, p. 337–407, Abril 2000.

JO, T. et al. Improving Protein Fold Recognition by Deep Learning Networks. **Scientific Reports**, v. 5, p. 17573, 4 dez. 2015.

KANAKALA, S.; GHANIM, M. RNA Interference in Insect Vectors for Plant Viruses. **Viruses**, v. 8, n. 12, 12 2016.

KARR, J. R. et al. A Whole-Cell Computational Model Predicts Phenotype from Genotype. **Cell**, v. 150, n. 2, p. 389–401, jul. 2012.

KAWASHIMA, S. et al. AAindex: amino acid index database, progress report 2008. **Nucleic Acids Research**, v. 36, n. Database issue, p. D202-205, jan. 2008.

KIM, S.-K. et al. miTarget: microRNA target gene prediction using a support vector machine. **BMC Bioinformatics**, v. 7, n. 1, p. 411, Setembro 2006.

KNORR, E. et al. Gene silencing in *Tribolium castaneum* as a tool for the targeted identification of candidate RNAi targets in crop pests. **Scientific Reports**, v. 8, n. 1, p. 2061, Fevereiro 2018.

KUBAT, M.; HOLTE, R. C.; MATWIN, S. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. **Machine Learning**, v. 30, n. 2, p. 195–215, Fevereiro 1998.

KUHN, M. **caret: Classification and Regression Training**. [s.l: s.n.].

LI, L. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. **Genome Research**, v. 13, n. 9, p. 2178–2189, 1 set. 2003.

LI, Y. et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. **Bioinformatics**, v. 34, n. 5, p. 760–769, 1 mar. 2018.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, v. 16, n. 6, p. 321–332, jun. 2015.

LIBERATI, N. T. et al. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. **Proceedings of the National Academy of Sciences**, v. 103, n. 8, p. 2833–2838, 21 fev. 2006.

LIU, B. et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. **Nucleic Acids Research**, v. 43, n. W1, p. W65–W71, 1 jul. 2015.

LIU, X. et al. Selection of key sequence-based features for prediction of essential genes in 31 diverse bacterial species. **PLOS ONE**, v. 12, n. 3, p. e0174638, 30 mar. 2017.

LLOYD, J. P. et al. Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. **The Plant Cell**, v. 27, n. 8, p. 2133–2147, ago. 2015.

LU, Y. et al. Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus. **Computational Biology and Chemistry**, v. 50, p. 29–40, jun. 2014.

LUO, H. et al. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements: Table 1. **Nucleic Acids Research**, v. 42, n. D1, p. D574–D580, jan. 2014.

LUO, H. et al. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. **Nucleic Acids Research**, v. 49, n. D1, p. D677–D686, 8 jan. 2021.

MARTÍN ABADI et al. **TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems**. [s.l: s.n.].

MARTÍNEZ-CARRANZA, E. et al. Variability of Bacterial Essential Genes Among Closely Related Bacteria: The Case of Escherichia coli. **Frontiers in Microbiology**, v. 9, p. 1059, 2018.

MIKOLOV, T. et al. **Strategies for training large scale neural network language models**. IEEE, dez. 2011. Disponível em: <http://ieeexplore.ieee.org/document/6163930/>. Acesso em: 5 jul. 2018

MOLINA-HENARES, M. A. et al. Identification of conditionally essential genes for growth of *Pseudomonas putida* KT2440 on minimal medium through the screening of a genome-wide mutant library. **Environmental Microbiology**, mar. 2010.

NAGY, L. G. et al. Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. **Nucleic Acids Research**, v. 48, n. 5, p. 2209–2219, 18 mar. 2020.

NIGATU, D. et al. Sequence-based information-theoretic features for gene essentiality prediction. **BMC bioinformatics**, v. 18, n. 1, p. 473, 9 nov. 2017.

NING, L. W. et al. Predicting bacterial essential genes using only sequence composition

information. **Genetics and Molecular Research**, v. 13, n. 2, p. 4564–4572, 2014.

OHLER, U. et al. Computational analysis of core promoters in the Drosophila genome. **Genome Biology**, v. 3, n. 12, p. RESEARCH0087, 2002.

ONG, S. A. et al. Efficacy of different protein descriptors in predicting protein functional families. **BMC Bioinformatics**, v. 8, n. 1, p. 300, 2007.

PENG, C. et al. A Comprehensive Overview of Online Resources to Identify and Predict Bacterial Essential Genes. **Frontiers in Microbiology**, v. 8, p. 2331, 27 nov. 2017.

PHADNIS, N. et al. An essential cell cycle regulation gene causes hybrid inviability in Drosophila. **Science (New York, N.Y.)**, v. 350, n. 6267, p. 1552–1555, 18 dez. 2015.

PLAIMAS, K.; EILS, R.; KÖNIG, R. Identifying essential genes in bacterial metabolic networks with machine learning methods. **BMC Systems Biology**, v. 4, n. 1, p. 56, 2010.

PORT, F. et al. A large-scale resource for tissue-specific CRISPR mutagenesis in Drosophila. **eLife**, v. 9, p. e53865, 13 fev. 2020.

PRICE, M. N. et al. Mutant phenotypes for thousands of bacterial genes of unknown function. **Nature**, v. 557, n. 7706, p. 503–509, Maio 2018.

QURESHI, N. A. et al. Genome-Based Drug Target Identification in Human Pathogen Streptococcus gallolyticus. **Frontiers in Genetics**, v. 12, p. 564056, 25 mar. 2021.

ROBIN, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. **BMC Bioinformatics**, v. 12, p. 77, 2011.

RUBIN, B. E. et al. The essential gene set of a photosynthetic organism. **Proceedings of the National Academy of Sciences**, v. 112, n. 48, p. E6634–E6643, 1 dez. 2015.

SASSETTI CHRISTOPHER M.; BOYD DANA H.; RUBIN ERIC J. Genes required for mycobacterial growth defined by high density mutagenesis. **Molecular Microbiology**, v. 48, n. 1, p. 77–84, 25 mar. 2003.

SERINGHAUS, M. et al. Predicting essential genes in fungal genomes. **Genome Research**, v. 16, n. 9, p. 1126–1135, set. 2006.

SHABALINA, S.; KOONIN, E. Origins and evolution of eukaryotic RNA interference. **Trends in Ecology & Evolution**, v. 23, n. 10, p. 578–587, out. 2008.

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379–423, jul. 1948.

SHEN, J. et al. Predicting protein-protein interactions based only on sequences information. **Proceedings of the National Academy of Sciences**, v. 104, n. 11, p. 4337–4341, 13 mar. 2007.

SONG, K.; TONG, T.; WU, F. Predicting essential genes in prokaryotic genomes using a linear method: ZUPLS. **Integr. Biol.**, v. 6, n. 4, p. 460–469, 2014.

SPRADLING, A. C. et al. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. **Genetics**, v. 153, n. 1, p. 135–177, set. 1999.

ULRICH, J. et al. Large scale RNAi screen in *Tribolium* reveals novel target genes for pest control and the proteasome as prime target. **BMC Genomics**, v. 16, p. 674, 2015.

VALLETTA, J. J. et al. Applications of machine learning in animal behaviour studies. **Animal Behaviour**, v. 124, p. 203–220, fev. 2017.

VERVIER, K. et al. Large-scale machine learning for metagenomics sequence classification. **Bioinformatics**, v. 32, n. 7, p. 1023–1032, 1 abr. 2016.

VIOLA, P.; JONES, M. **Rapid object detection using a boosted cascade of simple features**. IEEE Comput. Soc, 2001. Disponível em: <http://ieeexplore.ieee.org/document/990517/>. Acesso em: 5 jul. 2018

VISWANATHA, R. et al. Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in Drosophila cells. **eLife**, v. 7, p. e36333, 27 jul. 2018.

WAINBERG, M. et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. **Nature Genetics**, v. 53, n. 5, p. 638–649, 1 maio 2021.

WANG, N. et al. Genome-wide identification of Acinetobacter baumannii genes necessary for persistence in the lung. **mBio**, v. 5, n. 3, p. e01163- 01114, 3 jun. 2014.

WANG, S. et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. **Scientific Reports**, v. 6, n. 1, maio 2016.

WEBER, G. Optimization method for obtaining nearest-neighbour DNA entropies and enthalpies directly from melting temperatures. **Bioinformatics**, v. 31, n. 6, p. 871–877, 15 mar. 2015.

WESSELER, J.; FALL, E. H. Potential damage costs of Diabrotica virgifera virgifera infestation in Europe - the ‘no control’ scenario: Potential damage costs of Dvv. in Europe. **Journal of Applied Entomology**, v. 134, n. 5, p. 385–394, 11 mar. 2010.

WHYARD, S.; SINGH, A. D.; WONG, S. Ingested double-stranded RNAs can act as species-specific insecticides. **Insect Biochemistry and Molecular Biology**, v. 39, n. 11, p. 824–832, nov. 2009.

WRIGHT, M. N.; ZIEGLER, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software, Articles**, v. 77, n. 1, p. 1–17, 2017.

XIAO, N. et al. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. **Bioinformatics**, v. 31, n. 11, p. 1857–1859, 1

jun. 2015.

XING, X.; LIU, J. S.; ZHONG, W. MetaGen: reference-free learning with multiple metagenomic samples. **Genome Biology**, v. 18, n. 1, dez. 2017.

XU, L.; GUO, Z.; LIU, X. Prediction of essential genes in prokaryote based on artificial neural network. **Genes & Genomics**, v. 42, n. 1, p. 97–106, jan. 2020.

YANG, Y. et al. Genome-scale characterization of RNA tertiary structures and their functional impact by RNA solvent accessibility prediction. **RNA (New York, N.Y.)**, v. 23, n. 1, p. 14–22, 2017.

YU, Y. et al. Gene essentiality prediction based on fractal features and machine learning. **Molecular BioSystems**, v. 13, n. 3, p. 577–584, 2017.

ZHOU, Q.; QI, S.; REN, C. Gene essentiality prediction based on chaos game representation and spiking neural networks. **Chaos, Solitons & Fractals**, v. 144, p. 110649, mar. 2021.

ZHOU, Q.; YU, Y. Information dimension analysis of bacterial essential and nonessential genes based on chaos game representation. **Journal of Physics D: Applied Physics**, v. 47, n. 46, p. 465401, 19 nov. 2014.

ZHU, M.; DONG, J.; CAO, D.-S. rDNAse: R package for generating various numerical representation schemes of DNA sequences. 2016.

ZOTTI, M. et al. RNA interference technology in crop protection against arthropod pests, pathogens and nematodes: RNA interference technology in crop protection against arthropod pests, pathogens and nematodes. **Pest Management Science**, v. 74, n. 6, p. 1239–1250, jun. 2018.

Capítulo 2:

Predição de genes essenciais em insetos

1. Introdução

O conjunto de genes onde mutações de perda de função (LOF – do inglês, *Loss Of Function*) causam a morte de um organismo antes da reprodução ou sua falta de capacidade reprodutiva é definido como genes essenciais (GEs) (RANCATI et al., 2018). Na biologia celular, a detecção do conjunto mínimo de genes para uma espécie prosperar em ambientes específicos define os genomas mínimos para pares organismo/ambiente específicos, permitindo tanto a definição de processos moleculares centrais para a vida quanto o desenvolvimento de organismos modificados para uma ampla gama de aplicações biotecnológicas (HUTCHISON et al., 2016). Tais genes também são alvos moleculares interessantes para intervenção farmacológica em bactérias patogênicas (WANG et al., 2014), manejo de pragas de insetos (KNORR et al., 2018) e terapias contra o câncer (CHEN et al., 2017).

A maioria dos modelos estatísticos para prever a essencialidade dos genes foram construídos para procariontes e usaram uma combinação de características intrínsecas e extrínsecas ao treinar modelos (DONG et al., 2018). O viés para organismos procarióticos é devido provavelmente à viabilidade de ensaios de triagem em larga escala para procurar GEs destes organismos cultivados em condições distintas, e também devido à disponibilidade dessas informações como bancos de dados estruturados, fornecendo, portanto, uma riqueza de essencialidade gênica dados para várias espécies bacterianas (LUO et al., 2014).

A descoberta de GEs é muito mais desafiadora em organismos multicelulares, pois testar diferentes condições *in vivo* pode não ser viável devido ao custo e a questões experimentais. Além disso, linhagens eucarióticas multicelulares complexas requerem vários programas de expressão gênica de desenvolvimento para produzir um organismo viável, e esses GEs podem ser descobertos apenas ao avaliar essas espécies no nível do organismo. Assim, a maioria dos métodos computacionais para prever GEs em eucariotos usa como modelo de treinamento a levedura unicelular *Saccharomyces cerevisiae*, que pode ser estudada de maneira semelhante a uma espécie procariótica para procurar GEs em nível celular (DONG et al., 2018). Em metazoários, a maioria dos modelos estatísticos para prever a essencialidade dos genes foram treinados usando dados produzidos a partir de ensaios em linhagens celulares de organismos modelo *Homo sapiens* e *Mus musculus* (GUO et al., 2017; PHILIPS; WU; LI, 2017; YANG et al., 2014), com menos estudos visando prever GEs associados ao estilo de vida multicelular em outros animais (TIAN et al., 2018).

Os insetos são um grupo altamente diversificado de metazoários com uma vasta diversidade taxonômica e mais de um milhão de espécies estimadas. Este táxon também é altamente diversificado em termos de seus papéis ecológicos e importância para as sociedades humanas, com várias espécies observadas como polinizadoras de culturas e espécies silvestres, vetores de doenças humanas e animais, parasitas, organismos modelo para pesquisa básica e aplicada, componentes importantes de ciclos de nutrientes e cadeias alimentares e produtores de substâncias economicamente relevantes (CRESPO-PÉREZ et al., 2020; RUST; SU, 2012; STORK, 2018), por exemplo, alguns mosquitos são vetores de doenças e as abelhas que ajudam na polinização. Portanto, é altamente desejável poder identificar GEs para grupos específicos de insetos que possam ser usados tanto para entender o *modus operandi molecular* desse táxon quanto para desenvolver inseticidas de linhagem restrita para controle de pragas com pegada ecológica potencialmente menor.

A mosca da fruta *Drosophila melanogaster,* indiscutivelmente um dos organismos modelo multicelulares mais bem caracterizados, tem aproximadamente 52% de seus genes codificadores de proteínas anotados com fenótipos LOF (EWEN-CAMPEN et al., 2017). Um recurso útil para fornecer informações adicionais de LOF para *D. melanogaster* é o Flybase[4], um banco de dados contendo dados organizados e estruturados sobre loci conhecidos na mosca-das-frutas, como alelos, genótipos e seus fenótipos associados, permitindo o levantamento sistemático dessas informações para alelos com perda de função em configurações genotípicas específicas que resultam em fenótipos incompatíveis com a vida ou viáveis e, consequentemente, obter lista de GEs e de genes não essenciais (NEGs), respectivamente (LARKIN et al., 2021). Dados de essencialidade gênica em *D. melanogaster* também estão disponíveis em estudos genômicos feitos em linhagens celulares; indiscutivelmente, essa informação não captura a essencialidade dos GEs de desenvolvimento necessários para espécies multicelulares, pois dependem principalmente de pesquisas *in vitro* em larga escala de GEs no nível celular (VISWANATHA et al., 2018).

O segundo inseto em que a informação sobre a essencialidade do gene em escala genômica está disponível, incluindo GEs no nível do organismo, é o Besouro-Castanho (*Tribolium castaneum*), um organismo modelo emergente com uma quantidade crescente de

[4] Além do Flybase, existem esforços para listar genes essenciais em bases de dados especificas, como a base DEG discutida no capítulo 1, e a OGEE, que contém dados de *D. melanogaster*, mas não foram escolhidas para este estudo por motivos descritos na sessão de anexos.

informações genômicas e fenotípicas disponíveis. Essas informações estão disponíveis no iBeetle-base, um banco de dados estruturado contendo informações curadas sobre o silenciamento de genes em *T. castaneum* usando RNAi em diferentes estágios de desenvolvimento, incluindo dados de letalidade (DÖNITZ et al., 2015).

As ordens de insetos Diptera e Coleoptera divergiram há aproximadamente 300 milhões de anos, enquanto os insetos evoluíram há cerca de 600 milhões de anos (KUMAR et al., 2017). Consequentemente, embora se espere que essas duas ordens compartilhem uma fração considerável de genes homólogos, conjuntos de genes linhagem-específicos provavelmente também evoluíram. É razoável supor que alguns desses genes linhagem-específicos codificam funções biológicas necessárias para processos linhagem-específicos essenciais, incluindo os de desenvolvimento (CHEN; ZHANG; LONG, 2010). Portanto, *D. melanogaster* e *T. castaneum* constituem um cenário interessante para desenvolver e avaliar um fluxo de trabalho estatístico geral para predizer GEs em organismos filogeneticamente distantes, incluindo a avaliação do desempenho de classificação em GEs linhagem-específicos.

O uso combinado de recursos extrínsecos e intrínsecos demonstrou melhorar o desempenho da previsão ao procurar por GEs. Recentemente, dois estudos demonstraram um desempenho considerável para predizer GEs em *D. melanogaster* usando informações extrínsecas e intrínsecas (AROMOLARAN et al., 2020; CAMPOS et al., 2020). No entanto, como exposto anteriormente, as características extrínsecas têm grandes desvantagens que podem impedir que sejam usadas em estratégias gerais para prever GEs em insetos. Embora se possa evitar o uso de recursos extrínsecos e ainda obter sucesso ao desenvolver estratégias de aprendizado de máquina para prever GEs em espécies unicelulares, não foi objetivamente avaliado como esses modelos preditivos se comportam ao avaliar organismos eucarióticos complexos ou genes restritos a táxons (CAMPOS et al., 2019; GUO et al., 2017; LIU et al., 2017; NIGATU et al., 2017).

Neste capítulo é descrito um conjunto de dados completo para desenvolver preditores gerais para GEs em organismos multicelulares, usando insetos como estudo de caso. Começamos descrevendo métodos para coletar dados de genes codificadores de proteínas essenciais e não essenciais para *D. melanogaster* e para *T. castaneum* da base Flybase e da base iBeetle, respectivamente, compreendendo tanto o nível celular quanto os exemplos de desenvolvimento de GEs e GNEs. Também desenvolvemos rotinas computacionais para reunir, para cada gene codificador de proteína, um conjunto de características intrínsecas (baseadas em genes e proteínas) e extrínsecas (expressão gênica e predição de localização subcelular) que foram usadas para treinar modelos estatísticos para a predição de GEs nestes dois organismos.

Para a validação de nossos modelos, empregamos uma validação entre espécies onde os classificadores foram treinados em um organismo e validados no outro, usando o conjunto completo de genes com status de essencialidade ou um subconjunto de genes específicos da linhagem. Especificamente, estávamos interessados em avaliar a influência dos seguintes parâmetros no desempenho da classificação: (i) importância relativa dos atributos intrínsecos e extrínsecos e, para estes últimos, sua disponibilidade para insetos com genomas completos disponíveis; (ii) fontes distintas de rotulagem de dados para GEs e GNEs sem *D. melanogaster* (listas já disponíveis de estudos anteriores versus a nossa estratégia de busca FlyBase); (iii) pré-processamento e seleção de atributos; (iv) método de aprendizagem para desenvolvimento de modelos. Demonstramos nossa estratégia de rotulagem de genes para superar significativamente os rótulos de genes atuais para a mosca quando usados para desenvolver classificadores para prever GEs no besouro. Prosseguimos descobrindo que, para a maioria dos insetos com dados genômicos já disponíveis, há menos de três experimentos de RNA-Seq disponíveis, e a maioria dos genomas não possui dados de expressão gênica disponíveis. Prosseguimos demonstrando que é possível prever GEs em insetos usando apenas atributos de sequência intrínsecos, embora os dados de RNA-Seq aumentem significativamente o desempenho do modelo, com a previsão de localização subcelular tendo uma contribuição mais modesta. Por fim, desenvolvemos vários modelos que podem ser usados em uma ampla gama de cenários para prever GEs em outras espécies de insetos, a partir de sequências de codificação e seus atributos intrínsecos sozinhos, ou incorporando dados de expressão gênica de adultos e/ou estágios larvais. Todo o código, dados brutos e modelos finais estão disponíveis em https://github.com/g1o/GeneEssentiality/.

2. Métodos

A **Figura 10** resume o pipeline geral de previsão de GEs desenvolvido para este estudo e será usado como um guia nesta seção.

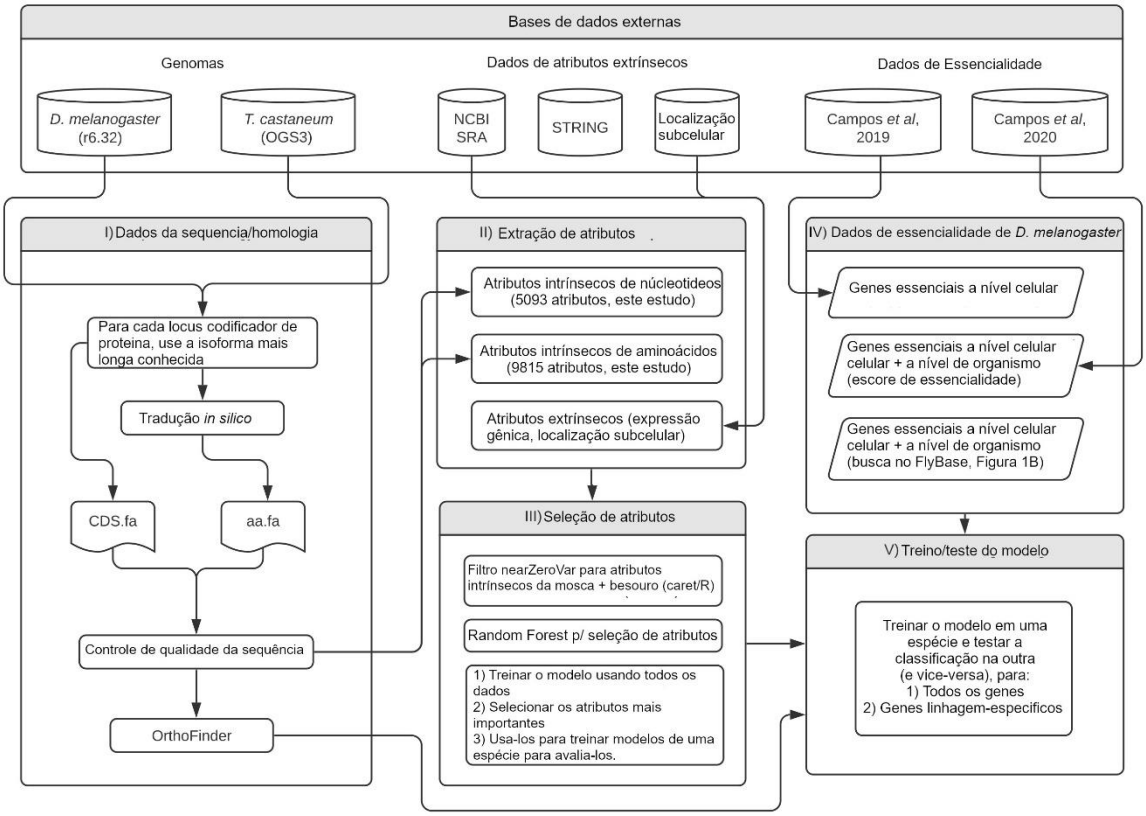


Figura 10 - Fluxograma conceitual deste capítulo.

2.1 Dados de sequência/homologia

Os dados de sequência para genes codificadores de proteínas para as espécies *D. melanogaster* e *T. castaneum* foram obtidos das bases de dados FlyBase e iBeetle-base, respectivamente, e procedemos por (i) remoção de sequências com nucleotídeos não padronizados; (ii) seleção da CDS and sequência proteica mais longa; (iii) remoção de sequências menores que 50 aminoácidos; e (iii) inferência de homologia para as sequências de proteínas restantes usando OrthoFinder com parâmetros padrão (EMMS; KELLY, 2019) **Figura 10**, “I) Dados de sequência e homologia”).

2.2 Extração/computação/coleta de recursos

Os atributos intrínsecos de cada gene foram extraídos das sequências codificadoras e de proteínas usando uma combinação de software externo, código interno e pacotes R (R CORE TEAM, 2016)(Tabela 2 no capítulo 1) (**Figura 10**, "II) Extração de atributos") sendo computados como descrito no capítulo 1, com a exceção dos casos descritos a seguir. Dentre os genes de *D. melanogaster,* haviam 350 genes com essencialidade conhecida com pelo menos uma transleitura (*readthrough*) do códon de parada, o que resultava em erros durante a extração de características, pois a proteína mais longa tinha o aminoácido X no lugar de um códon de parada, e 3 outros genes tinham os aminoácidos modificados selenocisteína (U) e pirrolisina (O), o que também resultava em erros, pois as ferramentas para calcular as características das proteínas consideram apenas os aminoácidos mais comuns. Para evitar a remoção desses genes, cada aminoácido X que representa um códon de parada foi removido, a selenocisteína foi substituída por cisteína e a pirrolisina foi substituída por lisina. Nenhuma das proteínas mais longas de *T. castaneum* tinham aminoácidos fora do padrão, o que mostra que há uma anotação muito mais minuciosa em *D. melanogaster* devido a maior quantidade de estudos realizados.

Também coletamos a disponibilidade de atributos extrínsecos a nível de gene para os insetos com genomas completos disponíveis no NCBI em agosto de 2021 por meio de pesquisas no NBI Short Read Archives para dados de RNA-Seq (SAYERS et al., 2021) e na base de dados STRING para interação proteína-proteína (SZKLARCZYK et al., 2021), para avaliar a viabilidade de usar esses dados para insetos com genomas completos disponíveis. Como demonstrado abaixo, uma vez que os dados de interação proteína-proteína estão amplamente ausentes para a grande maioria das espécies de insetos (Figura 11A), procedemos usando apenas dados de expressão gênica como atributos extrínsecos. Especificamente, calculamos os valores de TPM para cada gene (representado por sua isoforma mais longa[5]) de dois experimentos de RNA-Seq representando estágios larvais e indivíduos adultos para ambas as espécies usando seus respectivos dados de CDS como transcriptomas de referência e o programa Salmon com parâmetros padrão (PATRO et al., 2017) (Número de acesso dos experimentos de RNA-Seq: *D. melanogaster*: SRR15663867 (adulto) e SRR7866341 (larva); *T. castaneum*: SRR1048129 (adulto) e SRR15082131 (larva)).

[5] Usado como uma simplificação de um problema mais complexo. Como mostrado na Figura 11, Das espécies que possem dados de RNA-Seq disponíveis, grande parte tem apenas um experimento. Logo, é esperado que a uma parte considerável das isoformas alternativas não seja observada usando apenas um experimento como base.

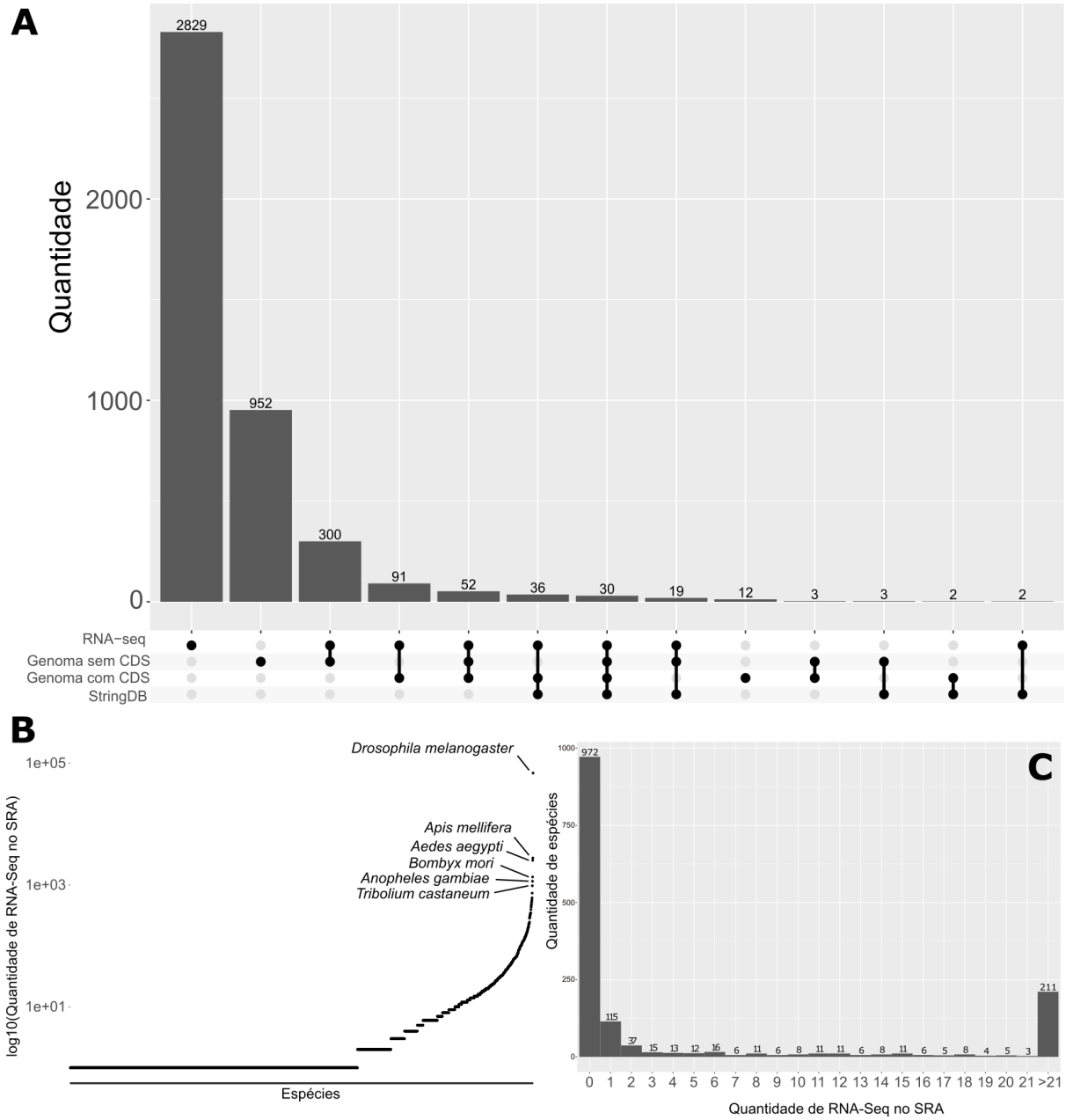


Figura 11: Disponibilidade de atributos extrínsecos para espécies de insetos. *A) Upset plot da disponibilidade de RNA-Seq, informações genômicas e dados do StringDB para todas as espécies de insetos disponíveis. B) Disponibilidade de experimentos de RNA-Seq por espécie de inseto. C) Distribuição dos dados de RNA-Seq por número de espécies de insetos.*

2.3 Rótulos de genes (essenciais/não essenciais)

Nossa busca por GEs no Flybase foi feita usando a ferramenta *query-builder*. Duas consultas lógicas seguidas por um passo de união e filtro foram construídas para classificar inequivocamente os genes como essenciais, não essenciais ou inconclusivos (Figura 12, "Dados de essencialidade dos genes de *D. melanogaster*"). A primeira consulta teve como

objetivo encontrar GEs, definidos como aqueles que possuem pelo menos um alelo LOF amórfico ou hipomórfico em homozigose ou em heterozigose associado a um fenótipo letal. A segunda consulta teve como objetivo encontrar GNEs, definidos como os genes que possuem ao menos um alelo LOF, mas que nenhum alelo LOF tenha fenótipo letal. Complementamos nossa lista de GEs realizando uma revisão da literatura para GEs essenciais e GNEs (Figura 12, "Revisão da literatura")

2.4 Busca no FlyBase para D. melanogaster

Em detalhe, a busca de genes essenciais (EGS - Essential Genes Search), realizada em Maio/2021, usada para consultar o FlyBase tinha os seguintes parâmetros: A classe do alelo tinha que ser "alelo de perda de função" ou "alelo amorfo" ou "alelo amorfo - evidência genética" ou "alelo amorfo - evidência molecular" ou " alelo hipomórfico " ou "alelo hipomórfico – evidência genética" ou " alelo hipomórfico – evidência molecular" e a classe fenotípica tinha que ser "letal". Alelos amorfos são aqueles em que nenhuma função gênica é relatada, enquanto os alelos hipomórficos têm suas funções reduzidas em comparação aos alelos do tipo selvagem. Se uma expressão mais baixa de um alelo é suficiente para causar letalidade, então argumentamos que é lógico classificar esse gene como essencial. Nenhum dado de ensaios de RNAi foi selecionado, pois estes não fornecem estatus da classe de alelos. Com a EGS construída e executada, usamos a interface web do Flybase para selecionar tais alelos e procedemos convertendo-os em IDs de genes.

Quanto a busca de genes não-essenciais (NEGS - Non-Essential Gene Search), os parâmetros foram os seguintes: a classe de alelos deveria ser "alelo de perda de função" ou "alelo amorfo" ou "alelo amorfo - evidência genética" ou "alelo amorfo - evidência molecular ", mas não " alelo hipomórfico ", " alelo hipomórfico – evidência genética", ou " alelo hipomórfico – evidência molecular", e a classe fenotípica tinha que ser "não letal". Excluímos os alelos hipomórficos porque eles retêm alguma atividade biológica por definição e, consequentemente, não fornecem informações para avaliar objetivamente se um evento LOF para esse locus é compatível com a sobrevivência do organismo, que é a definição estrita de genes não essenciais. Em ambas as buscas, foram excluídos os alelos que não continham descrição de classe fenotípica, pois não é possível inferir seu status de essencialidade (Tabela 3). Os alelos resultantes também foram convertidos em seus IDs de genes.

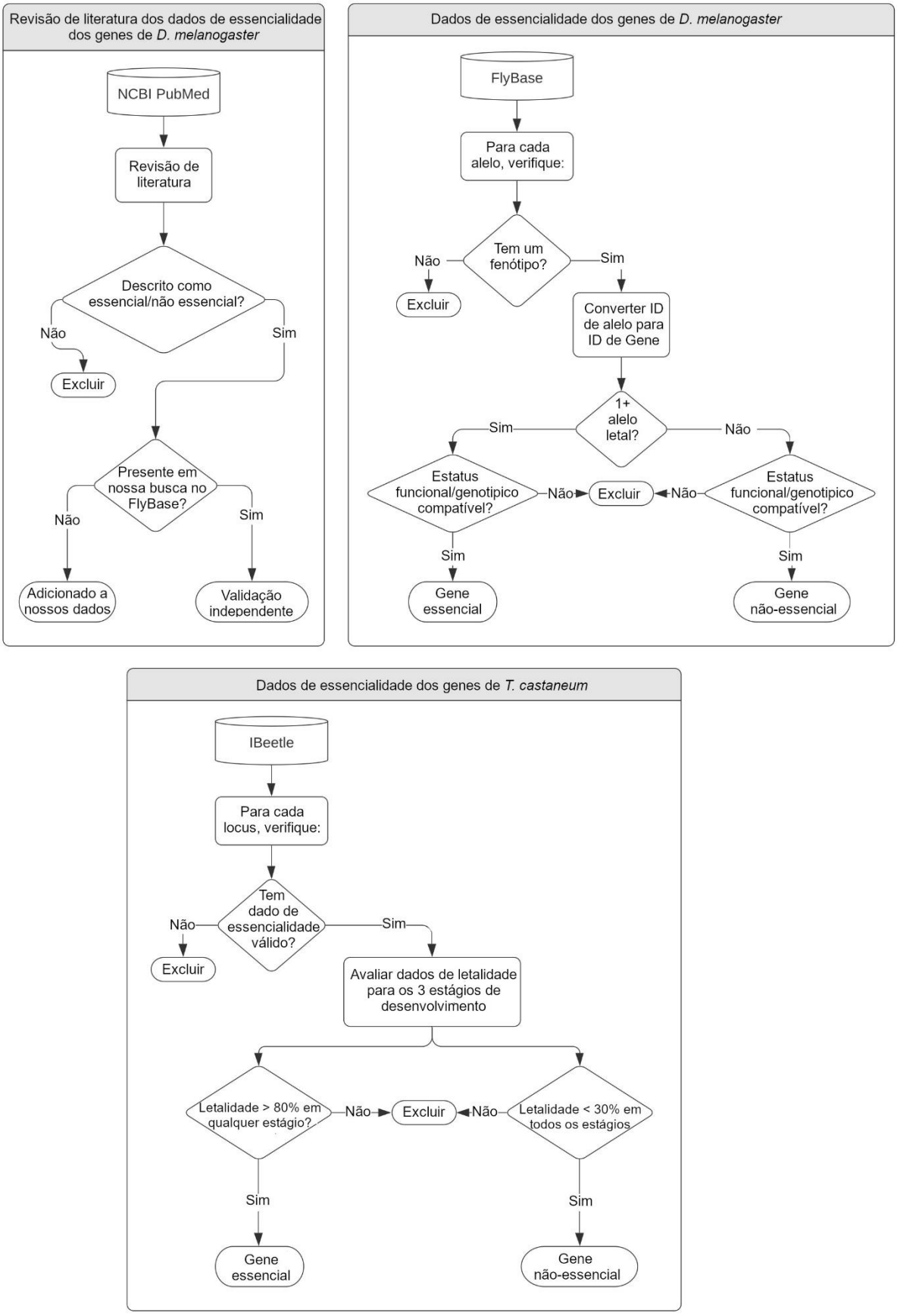


Figura 12: Fluxograma da obtenção e revisão de genes essenciais.

Para considerar um gene como essencial, buscamos alelos observados em heterozigosidade ou homozigose, pois o LOF de um único alelo é suficiente para detectar alelos letais dominantes. Os genes não essenciais, no entanto, podem ser classificados com segurança apenas quando um fenótipo viável é observado em indivíduos homozigotos, pois os alelos letais recessivos com LOF são conhecidos há mais de um século. Como exemplo, em *D. melanogaster*, o gene Duox é um exemplo clássico de gene com alelos letais recessivos (Xie, et al., 2010). Portanto, consideramos como genes não essenciais apenas aqueles em que indivíduos com alelos amorfos em homozigose são viáveis, pois indivíduos viáveis heterozigotos podem representar genes não essenciais verdadeiros ou genes letais recessivos, logo, são inconclusivos.

Os genes encontrados tanto em EGS quanto em NEGS foram considerados genes essenciais com base na premissa de que um gene com pelo menos um alelo letal LOF é um gene essencial em pelo menos uma condição (genes essenciais condicionais). Avaliamos nossa consulta automática para recuperar genes de moscas essenciais e não essenciais por meio de uma extensa revisão de literatura para genes já descritos como essenciais ou não essenciais. Ao ler a literatura, também encontramos genes essenciais e não essenciais anteriormente não categorizados por nossas consultas que foram usados para expandir nosso conjunto de dados. O conjunto de dados combinado da pesquisa do Flybase e da revisão da literatura será referido a partir de agora como conjunto de dados DMEL.

2.5 *D. melanogaster* em OGEE, DEG e Flybase

A qualidade dos dados usados no aprendizado de máquina é de extrema importância, pois dados ausentes e rotulados incorretamente para o conjunto de treinamento podem introduzir vieses sistemáticos nos modelos que provavelmente diminuirão o desempenho do modelo (Batista, 2003). Para desenvolver um preditor de genes essenciais para insetos, inicialmente buscamos experimentos em larga escala em *D. melanogaster* ou bancos de dados descrevendo genes essenciais e não essenciais.

Encontramos essa informação disponível em dois bancos de dados especializados contendo listas de genes essenciais de *D. melanogaster* diretamente disponíveis – OGEE (Chen, et al., 2017) e DEG (Gao, et al., 2015) . OGEE contém dois conjuntos de dados para *D. melanogaster* : 1) 13781 genes pesquisados quanto à essencialidade no nível celular através de knockdown de RNAi *in vitro* (Boutros, et al., 2004) , 2) 437 genes silenciados com knockdown de RNAi transgênico em moscas em diferentes estágios de desenvolvimento (Chen, et al.,

2010). No DEG, encontramos um conjunto de dados contendo 339 genes essenciais que foram interrompidos por inserções de elementos P (Spradling, et al., 1999) . O banco de dados Flybase usa descrições controladas para descrever dados centrados em alelos, como tipos de mutação, suas consequências funcionais e os fenótipos associados para genótipos distintos (Larkin et al., 2021). Essas informações estruturadas permitem a seleção de genes essenciais e não essenciais por meio de consultas específicas para selecionar tais conjuntos de forma inequívoca (consulte "Métodos" para nossa estratégia de consulta).

Argumentamos que os conjuntos de dados OGEE e DEG sofrem de várias desvantagens que prejudicam sua eficácia para produzir conjuntos confiáveis de genes essenciais e não essenciais. Primeiro, nenhum desses conjuntos de dados relata dados suficientes para nos permitir selecionar inequivocamente conjuntos de genes essenciais e não essenciais, pois eles não possuem os dados detalhados centrados em alelos necessários para selecionar esses conjuntos de genes. Além disso, o conjunto de dados OGEE 1 provavelmente não relata genes essenciais ao desenvolvimento, classificando-os falsamente como genes não essenciais, enquanto o conjunto de dados OGEE 2 mostrou conter resultados que podem não ser reproduzíveis (Kondo et al., 2017).

Uma análise mais profunda dos alelos representados no conjunto de dados DEG e no FlyBase descobriu que a maioria deles não possui informações de classe de alelos (86,7%, 1183 de 1364). Quanto aos 181 alelos onde a informação da classe de alelos está disponível, encontramos 90 alelos hipomórficos, 2 com ganho de função e 1 alelo hipermórfico, o que nos impede de utilizá-lo como fonte confiável de LOF e alelos hipomórficos, que são fundamentais para definir alelos essenciais e genes não essenciais.

Também investigamos se o conjunto de dados de genes essenciais que obtivemos com nossa busca no Flybase concorda com o maior disponível no OGEE, contendo 267 genes essenciais de um total de 13781 genes testados (Boutros, et al., 2004). Este conjunto de dados altamente desequilibrado, contendo 98,1% dos genes classificados como não essenciais, provavelmente conterá apenas genes essenciais de manutenção detectados por meio de triagem *in vivo de* alto rendimento. Em comparação, nossa busca no Flybase resultou em 1.393 genes essenciais e 899 genes não essenciais. A intersecção entre genes não essenciais do OGEE e genes essenciais da nossa pesquisa revelou que mais de 1000 genes essenciais do Flybase seriam considerados não essenciais por este conjunto de dados OGEE.

2.6 Informação de essencialidade gênica para *Drosophila melanogaster*

Uma diferença que pode explicar o maior número de genes disponíveis exclusivamente no conjunto de dados do Enrichment Score (ES) produzido por Campos e colaboradores em 2020 em relação aos nossos dados é o fato de usarmos informações de fenótipo e classe de alelos, juntamente com o status do genótipo por locus, para classificar objetivamente os genes como EG ou NEG. Uma fração considerável de loci é inconclusiva depois de levar em conta esses fatores biológicos de confusão conhecidos, e esses dados são removidos das análises a jusante. Como exemplo, mencionamos a alta fração de alelos sem informações de classe de alelos nos genes de *D. melanogaster* (

Tabela 3).

Tabela 3: Distribuição da classe de alelos do FlyBase. Somente os 10 mais frequentes estão mostrados.

Most frequent terms	Related records	Frequency
[empty field - no data available]	1267	0,1274
viable	949	0,0954
recessive	787	0,0791
visible	698	0,0702
adult stage	690	0,0694
neuroanatomy defective	547	0,055
fertile	292	0,0294
third instar larval stage	269	0,027
neurophysiology defective	256	0,0257
somatic clone	243	0,0244

Campos e colaboradores utilizaram todos os genes onde alelos com status fenotípico identificados como “letal” ou “viável” estão disponíveis para calcular o ES *ad-hoc* para cada um desses genes, definido como o número total de alelos ligados a termos essenciais/letais ao quadrado dividido pelo número total de experimentos ligados a termos essenciais/letais mais não essenciais/viáveis ao quadrado. A pontuação ES foi então usada para classificar cada gene em uma das três categorias: essencial (ES > 0,9), essencial condicional (0/9 > ES > 0,1) ou não essencial (ES < 0,1). Portanto, o próprio conceito de essencialidade e as estratégias utilizadas para defini-lo são substancialmente distintos em nossos estudos, que produziram conjuntos gênicos distintos desde o início.

Nossa definição de genes essenciais (genes em que um LOF completo ou um alelo hipomórfico causam fenótipos letais, seja em heterozigose ou em homozigose, independentemente das condições ambientais) tenta traduzir a definição lógica mais ampla de genes essenciais, levando em consideração genotípica, fenotípica e dados funcionais, e certamente inclui genes essenciais condicionais. Este fato é coerente com a distribuição de nossos genes essenciais de acordo com o ES de Campos (a maioria de nossos genes essenciais é classificada como condicional essencial por eles).

A definição de genes não essenciais em nosso trabalho – loci onde alelos LOF em homozigose resultam em indivíduos viáveis – também leva em consideração dados genotípicos, fenotípicos e funcionais para excluir dados que possam impedir a classificação inequívoca de um gene como não essencial, como como a ocorrência de alelos letais e hipomórficos recessivos. Argumentamos que essas diferenças podem ser pelo menos parcialmente responsáveis pelo número muito maior de genes classificados exclusivamente como não essenciais por Campos *et al.* em comparação com nossos dados, pois sua estratégia baseada em pontuação aceitaria como não essenciais, por exemplo, os loci onde uma pequena fração de alelos causa letalidade, que pode até incluir genes essenciais com menor penetrância (Schmitt-Engel, et al., 2015).

Depois de considerar todos os fatos mencionados, decidimos nossas consultas no Flybase para integrar dados genéticos, funcionais, genotípicos e fenotípicos, incluindo dados de várias fontes experimentais e estágios de desenvolvimento, permitindo definir sem ambiguidade conjuntos de genes essenciais e não essenciais. Portanto, decidimos continuar usando este banco de dados como nossa fonte de dados de essencialidade gênica para *D. melanogaster*.

2.7 Seleção de GEs e GNEs em *D. melanogaster* e *T. castaneum*

Os genes descritos como essenciais/não essenciais na literatura e presentes em nossa busca no FlyBase foram usados para avaliar nossa estratégia de busca, enquanto os encontrados apenas na revisão de literatura foram adicionados ao nosso conjunto de dados de genes de *D. melanogaster* com status de essencialidade conhecido. Este conjunto de GEs e GNEs de *D. melanogaster* será referido a partir de agora como o conjunto de dados DMEL.

Dois estudos recentes também relatam o desenvolvimento de classificadores para GEs em *D. melanogaster.* Um deles baseou-se exclusivamente em dados do OGEE, que foram gerados a partir de linhagens celulares e, consequentemente, espera-se que capturem genes

essenciais ao nível celular, e serão referidos daqui em diante como “nível celular” (CAMPOS et al., 2019). O segundo estudo baseia-se em um escore de essencialidade *ad hoc* (ES) calculado a partir do FlyBase para classificar genes como GNEs (ES < 0,1), GEs condicionais (0,1 ≤ ES ≤ 0,9) e GEs (ES > 0,9). Espera-se que esse escore capture GEs celulares e de desenvolvimento, embora essa estratégia possa classificar como GEs apenas os genes em que uma alta fração de alelos produz um fenótipo não viável (Campos, et al., 2020). Esta lista é, a partir de agora, referida como “ES”. Usamos o WebGestalt para procurar possíveis vias enriquecidas nessas listas de genes para avaliar ainda mais os possíveis vieses para GEs em nível de célula e organismo usando os seguintes parâmetros: KEGG como banco de dados funcional e FDR < 0,05 para significância. As 10 vias com as maiores taxas de enriquecimento foram selecionadas para inspeção adicional.

O banco de dados iBeetle contém resultados de experimentos de milhares de genes silenciados em *T. castaneum* através da injeção de dsRNA durante a fase de pupa e larval de 5º/6º instar, incluindo dados de letalidade 11 dias após a injeção (dpi) para injeção de pupa e 11 e 22 dpi para injeção larval (DÖNITZ et al., 2015). Consultamos o banco de dados do iBeetle para obter as informações de letalidade do gene para os três estágios de desenvolvimento e computamos a distribuição da letalidade de forma semelhante à descrita na literatura (SCHMITT-ENGEL et al., 2015) (Figura 12, “Dados de essencialidade dos genes de T. castaneum”). Especificamente, exigimos que os GEs tivessem mais de 80% de letalidade em 11 ou 22 dias após a injeção larval ou 11 dias após a injeção de dsRNA em pupa. Quanto aos NEGs, foram definidos como aqueles com no máximo 30% de letalidade em todos os momentos, após a injeção de larvas e pupas. Este conjunto de essenciais e GNEs de *T. castaneum* será referido a partir de agora como o conjunto de dados TRIB.

2.8 Seleção de atributos (Feature *selection*)

Após a extração de atributos intrínsecos conforme feita no capítulo anterior, foi realizada uma etapa de pré-processamento usando a função nearZeroVar() do pacote Caret do R, com os parâmetros padrão para remover atributos intrínsecos de baixa variância, que foram computados a partir de todas CDSs e proteínas dos dados mesclados de mosca mais besouro (**Figura 10**, “III) Seleção de atributos”). Em seguida, estimamos a importância relativa do recurso para todos os nossos dados de recurso treinando um modelo de RF usando os atributos de DMEL + TRIB conforme descrito na seção 2.5 e obtendo a importância dos atributos usando o varImp() do pacote R caret. Em seguida, estimamos o melhor ponto de corte para a

importância do recurso computando AUCs para ROC e PR para modelos treinados com DMEL e validados em TRIB (e vice-versa), variando o ponto de corte de importância dos atributos.

2.9 Treinamento e validação do modelo

Usamos combinações distintas de atributos e rótulos de genes para treinar e avaliar modelos estatísticos específicos capazes de distinguir entre GEs e NEGs, visando desenvolver a ferramenta mais útil para prever tais genes em novas espécies de insetos. Especificamente, estávamos interessados em avaliar a importância dos seguintes procedimentos: (i) pré-processamento de atributos e seleção de atributos (computados usando as funções nearZerVar() e varImp() do caret, respectivamente); (ii) classe de modelo preditivo (SVM vs. RF vs. XGBT); (iii) classe de atributo (intrínseco vs. extrínseco vs. ambos); (iv) rotulagem de dados de *D. melanogaster* (nossa estratégia vs. GEs em nível celular (CAMPOS et al., 2019) vs pontuação de essencialidade (CAMPOS et al., 2020).

Usamos os recursos computados como entrada para treinar modelos de *Support Vector Machines* (SVM) usando os métodos “ svmRadial ” e “ svmPoly ” (KARATZOGLOU et al., 2004), *Extreme Gradient Boosting Trees* (XGBT) usando o pacote ' xgbTree ' (CHEN; GUESTRIN, 2016) e *Random Forest* (RF) usando o pacote 'ranger' (WRIGHT; ZIEGLER, 2017) (todos disponíveis no pacote caret (KUHN, 2018)). Os parâmetros para os modelos foram os seguintes: (i) SVM: kernel polinomial (grau = (2:4), C = 2^c(0, 1, 2), escala = c(0.1,0.2)) e kernel radial (sigma = 2^c(-30,-25, -20), C = 2^c(0, 0,5, 0,75, 1, 1,5)); (ii) RF: 1000 árvores e ajuste selecionando o *mtry* entre o quadrado do número de atributos e duas vezes ele mesmo; (iii) XGBT: os parâmetros mantidos constantes foram *eta* = 0,1, *gamma* = 1, *colsample* _bytree = 1, *min_child_weight* = 1, *subsample* = 1 e, para ajuste de parâmetros, *nrounds* foi 100, 200 e 500 e *max_depth* foi 4 e 10.

A validação cruzada de dez vezes usada no treinamento dos modelos foi repetida 3 vezes, usando a AUC máxima da ROC (AUC-ROC) como a métrica de desempenho a ser maximizada. Comparamos o desempenho de modelos distintos usando o teste de DeLong implementado na função roc.test () do pacote pROC (SUN; XU, 2014). Para avaliar o desempenho de modelos específicos de insetos em relação a uma linha de base, usamos um classificador Zero Rule (ZR) que classifica cada proteína como um EG. Embora não tenhamos um conjunto de dados de treinamento altamente desequilibrado, esse não é o caso dos GEs da pontuação de essencialidade e dos a nível de célula, portanto, também relatamos AUCs de precisão e revocação (PR-AUC: *Precision Recall-Area Under Curve*) para modelos treinados

e ZR para avaliar o desempenho do classificador em termos de taxas de falsos positivos e falsos negativos.

Para avaliar se os modelos treinados usando dados de *D. melanogaster* ou *T. castaneum* são classificadores gerais que podem prever GEs em outras espécies de insetos, validamos cada modelo para prever GEs usando a outra espécie de inseto como conjunto de dados de teste (**Figura 10**, "V) Treino/teste do modelo"). Para verificar se o modelo de *D. melanogaster* pode predizer essencialidade em genes restritos a *T. castaneum* (genes que não possuem homólogos em *D. melanogaster*) e vice-versa, usamos o modelo treinado em uma espécie para predizer GEs no conjunto de genes sem homólogos, obtidos do resultado do OrthoFinder, no outro.

3. Resultados

3.1 Avaliação da disponibilidade de recursos extrínsecos para Insecta

De um total de 4.331 espécies de insetos representadas em pelo menos um banco de dados externo, encontramos 3.359 (77,6%), 1.500 (34,6%) e 92 (2,1%) espécies que possuem dados disponíveis como RNA-Seq, genomas montados e dados de interação proteína-proteína, respectivamente (Figura 11A). A maioria das espécies (2.829, 65,3%) está representada apenas pelo transcriptoma. Tanto a segunda quanto a terceira maior categoria, com 952 (22,0%) e 300 (6,9%) espécies, são aquelas com conjuntos genômicos que não possuem dados de anotação ou predição gênica, sendo que esta última também contém dados transcriptômicos. Um total de 226 (5,22%) espécies possuem genomas com informações de modelo gênico, sendo que 211 (4,9%) delas contêm pelo menos uma fonte de atributos extrínsecos. Neste grupo, a grande maioria (209, 4,8%) compreende espécies com dados de RNA-Seq, enquanto os dados de interação proteína-proteína correspondem a 41 (0,9%) espécies.

Exploramos ainda mais a distribuição de espécies com dados de RNA-Seq, descobrindo que ela é altamente tendenciosa para organismos modelo de importância médica, científica e agrícola (Figura 11B), e com aproximadamente 90% das espécies com genomas disponíveis sendo representadas por dois ou menos experimentos de RNA-Seq (Figura 11C). Com base nesses achados, decidimos usar duas bibliotecas de RNA-Seq como fonte de atributos extrínsecos para emular o cenário para a grande maioria dos insetos com dados genômicos e transcriptômicos disponíveis. Além disso, selecionamos bibliotecas compreendendo um estado larval e um adulto, pois essas bibliotecas desses estágios específicos são uma escolha comum para aumentar o desempenho de previsão do modelo genético (BRŮNA et al., 2021).

Embora, dois grupos tenham relatado o desenvolvimento bem sucedido de preditores de GEs para *D. melanogaster* (Aromolaran, et al., 2020; Campos, et al., 2020), O uso dessas ferramentas como uma abordagem geral para prever GEs em outras espécies de insetos é altamente limitado. Ambas as estratégias dependem fortemente de propriedades extrínsecas ao nível do gene, como predição do domínio da proteína e perfis de expressão gênica e redes de interação proteína-proteína, para citar alguns, ao desenvolver seus preditores. Esta informação está prontamente disponível para os organismos modelo *D. melanogaster* e *T. castaneum,* mas a coleta desta informação para outros insetos é uma atividade trabalhosa e pode não estar disponível para a maioria dos genomas de insetos disponíveis (por exemplo, dados STRINGdb, como demonstrado neste trabalho). Não surpreendentemente, e em contraste com o nosso trabalho, ambos os estudos carecem da demonstração formal de que seus modelos podem ser usados para prever GEs em outras espécies de insetos.

3.2 Dados de sequência/homologia e extração/computação de recursos

Começamos nossa análise com 13.968 e 16.576 sequências de DNA e proteínas válidas para cada locus codificador de proteínas para os genomas da mosca e do besouro, respectivamente, com 5.126 (mosca) e 7.423 (besouro) genes sem homólogos nas outras espécies. Ao considerar os dados de essencialidade, encontramos 397 genes de besouro sem homólogos de mosca (76EGs e 321 NEGs) e 369 genes de mosca sem homólogos de besouro (144 GEs e 225 NEGs) (usando TRIB e DMEL para determinar a essencialidade do gene de besouro e mosca, veja as seções 3.4 e 3.5).

Integramos vários softwares que, juntos, computaram um total de 15.388 atributos intrínsecos para cada gene válido (Tabela 2), juntamente com os seguintes atributos extrínsecos: (i) o TPM para cada biblioteca, que produz duas características para cada espécie, correspondendo a expressão em adultos e uma fase larval; (ii) predição da localização subcelular fornecida pelo DeepLoc 1.0, que inclui 11 características (ALMAGRO ARMENTEROS et al., 2017). A etapa de pre-processessamento usando o filtro nearZeroVar removeu aproximadamente 45% (6.914 dos 15.388 atributos intrínsecos calculados pelo nosso pipeline). Todos os atributos removidos neste ponto são baseados em proteínas, e a grande maioria (6.780) são da classe CMI (informação mútua condicional de peptídeos de comprimento 3), o que se espera produzir uma matriz muito esparsa, com maior ruido e consequentemente, baixo poder preditivo e alta demanda de recursos computacionais.

3.3 Definindo o conjunto de genes essenciais e não essenciais em *T. castaneum*

A base do iBeetle contém informações sobre a letalidade de 4.084 genes avaliados aos 11 dpi para injeção nos estágios de pupa (P11) e larval (L11) e aos 22 dpi para o estágio larval (L22), enquanto 4.174 genes com dados faltantes no campo de letalidade e foram excluídos a jusante análise. Ao avaliar a frequência relativa dos genes em função da letalidade nos três estágios de desenvolvimento, descobrimos que ela é tendenciosa para genes com valores de letalidade altos ou baixos (Figura 13 A). Após uma inspeção visual de nossos histogramas, e conforme feito anteriormente por Schmitt-Engel et al. (2015), definimos GEs como aqueles com um valor de letalidade superior a 80% em pelo menos um estágio de desenvolvimento (Figura 13 A, círculos) e como GNEs aquelas com valor de letalidade menor ou igual a 30% nos três estágios de desenvolvimento (Figura 13 A, triângulos). Esses são valores razoáveis, devido a técnica de silenciamento gênico não ter garantia de eficiência na inibição de 100% e, portanto, a letalidade nem sempre ser observada, e que no outro extremo a própria técnica de injetar RNAi no pequeno inseto possa levar a letalidade, entre outros possíveis ruídos.

Observou-se uma ampla distribuição dos valores de letalidade para GEs nos estágios de desenvolvimento larval e pupal 11 dpi (L11 e P11) com discreto aumento nos valores de letalidade em torno de 100%. Isso contrasta com a mortalidade muito maior de GEs no único estágio de desenvolvimento em que os dados de letalidade 22 dpi estão disponíveis (larva 22 dpi, L22). Essa distribuição sugere que pode haver relativamente poucos GEs domésticos com 100% de mortalidade em estágios distintos de desenvolvimento antes de 11 dias. Por outro lado, a maioria dos GEs necessitou de pelo menos 12 dias para ser detectado na fase larval, um fenômeno provavelmente causado por fenótipos letais de baixa penetrância (SCHMITT-ENGEL et al., 2015). Após excluir dois genes com informações de letalidade conflitantes, selecionamos 1.073 e 1.077 GEs e GNEs para *T. castaneum,* respectivamente (Figura 12, "*T. castaneum*, conjunto de dados TRIB").

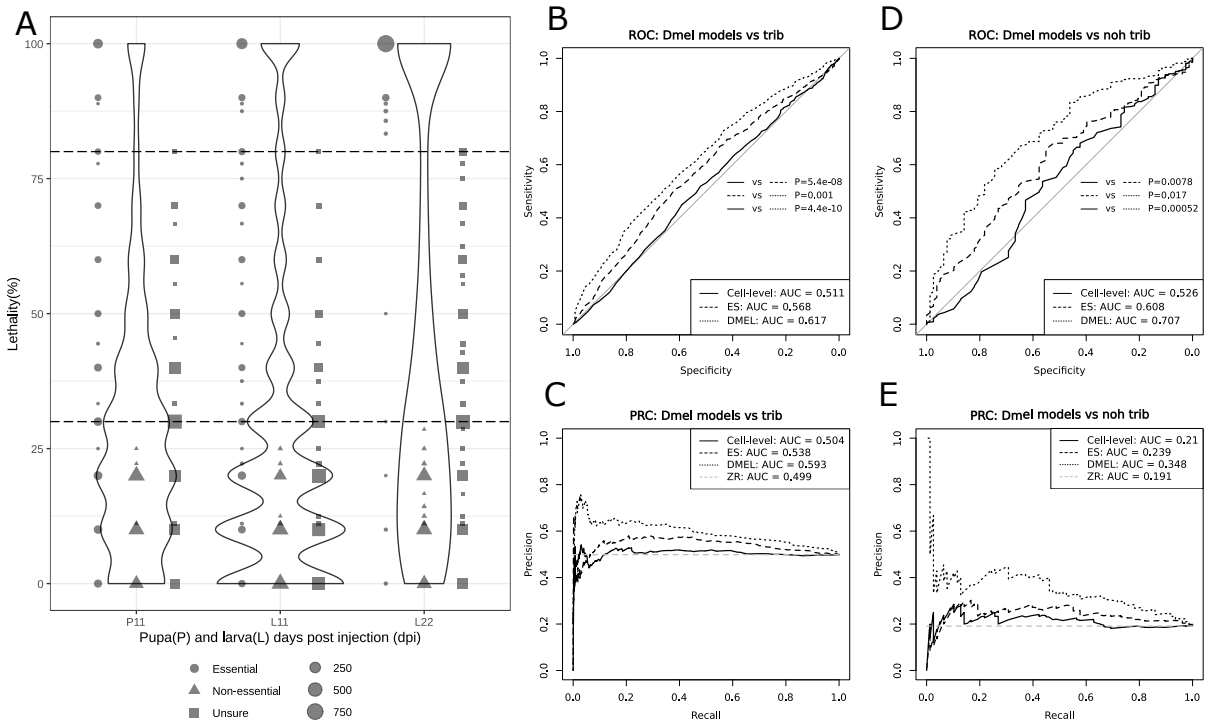


Figura 13: Dados de essencialidade genica. *A) Distribuição dos dados de letalidade dos genes de T. castaneum por estágios de desenvolvimento. Círculo: genes essenciais (GEs); triângulo: genes não essenciais (GNEs); quadrado: não classificável. As linhas tracejadas são os pontos de corte para definição da classe (GEs: letalidade >= 80% em qualquer estágio; GNEs: letalidade < 30% em todos os estágios). B) AUC-ROCs para modelos treinados usando estratégias de rótulos dos genes em nível de célula, ES e DMEL em todas as moscas e testados em genes de besouros. C) AUC-PRCs para modelos treinados usando esquemas de rotulagem de genes em nível de célula, ES e DMEL em todas as moscas e testados em genes de besouros. D) AUC-ROCs para modelos treinados usando esquemas de rotulagem de genes em nível de célula, ES e DMEL em todas as moscas e testados em genes de besouros específicos de linhagem. E) AUC-PRCs para modelos treinados usando esquemas de rotulagem de genes em nível de célula, ES e DMEL em todas as moscas e testados em genes de besouros específicos de linhagem.*

3.4 Definindo o conjunto de genes essenciais e não essenciais em *D. melanogaster*

Construímos duas consultas para pesquisar o banco de dados Flybase e selecionar conjuntos de alelos onde combinações específicas de informações genotípicas e funcionais sobre alelos em um locus resultaram inequivocamente em um fenótipo letal (GEs) ou viável (GNEs), após a exclusão de casos inconclusivos (Figura 10; veja também a seção “

Busca no FlyBase para D. melanogaster”). Em nossa busca, observamos 1.267 alelos sem classe fenotípica descrita que foram excluídos da análise a jusante, pois não puderam ser

avaliados automaticamente quanto ao seu fenótipo de essencialidade (Apêndice–Tabela complementar 1: Seleção manual de genes essenciais e não essenciais de D. melanogaster da revisão de literatura.). Após mapear e filtrar os alelos do GEs e GNEs para os IDs dos genes, obtivemos 1.393 GEs e 899 GNEs.

Nossa revisão de literatura, juntamente com uma recente revisão de literatura independente para genes essenciais na mosca (PORT et al., 2020), encontrou 181 e 99 GEs e GNEs, respectivamente, também presentes em nossos resultados de consulta do Flybase. Destes, 164 (91%) e 80 (81%) genes foram corretamente classificados como GEs e GNEs por nossa consulta, respectivamente, enquanto 17 (9%) GEs e 19 (19%) GNEs eram falsos positivos e negativos de nossa consulta, respectivamente (precisão de 0,90, recall de 0,91 e F-measure de 0,90). Portanto, concluímos que nossa consulta é capaz de selecionar automaticamente conjuntos de genes essenciais e não essenciais.

Para considerar um gene essencial, procuramos alelos que tinham a classe fenotípica anotada como letal, que estavam em heterozigose ou homozigose, pois um único alelo LOF é suficiente para detectar alelos letais dominantes. Os genes não essenciais, no entanto, podem ser classificados com segurança apenas quando um fenótipo viável é observado em indivíduos homozigotos, pois os alelos letais recessivos com LOF são conhecidos há mais de um século.

O conjunto de dados DMEL produzido em nossa análise compartilha algumas semelhanças, mas também apresenta diferenças consideráveis, em relação aos obtidos da literatura (conjuntos de dados em nível de célula e ES). As duas listas da literatura têm muito mais entradas do que o conjunto de dados DMEL (DMEL: 2690 entradas, versus 13640 e 7283 para nível de célula e ES, respectivamente). Os conjuntos de dados da literatura também são fortemente desequilibrados, com uma considerável fração de genes é marcada como GNEs (Figura 14A). Como esperado, esses dois conjuntos de dados compartilham uma fração considerável de GNEs que estão ausentes do conjunto de dados DMEL (5818 genes, o maior conjunto). Os GNEs encontrados exclusivamente no conjunto de dados em nível de célula compreendem a segunda maior categoria (4981 genes), e é razoável supor que alguns destes podem ser essenciais ao nível do organismo. Esta hipótese é ainda apoiada pela observação de que a terceira maior categoria compreende genes rotulados como essenciais no conjunto de dados DMEL, mas como não-essenciais nos dados de nível celular (1200 genes, ou 77,6% GEs de DMEL).

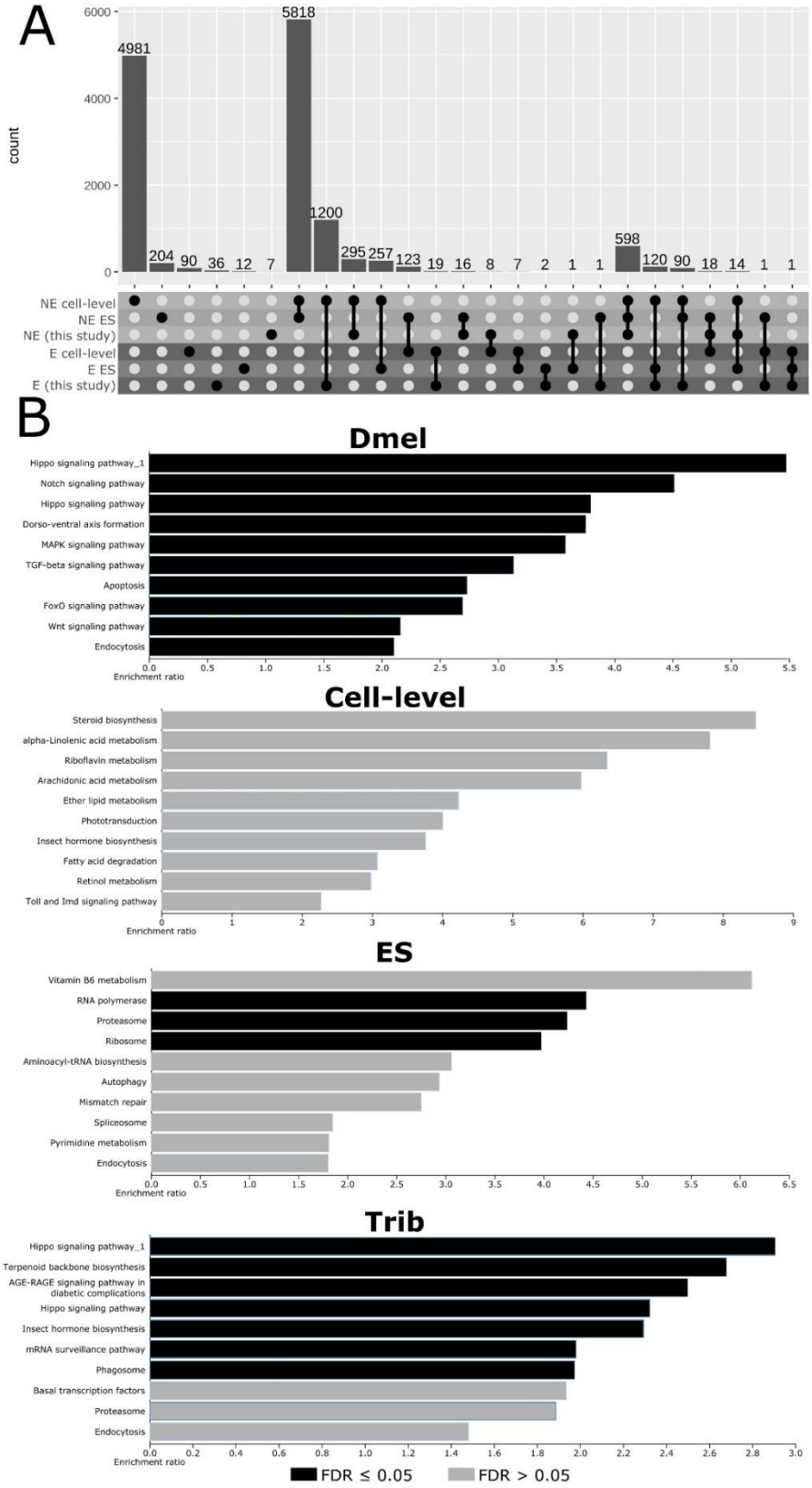


Figura 14: Comparação de diferentes esquemas de rótulos dos genes para *D. melanogaster*. A) Distribuição de genes rotulados como essenciais ou não essenciais em nível celular (Cell-level), por *enrichment score* (ES), e pelo esquema de rotulagem deste estudo (NE- Não essencial, E – Essencial). B) Enriquecimento de vias biológicas nos esquemas de rotulagem em nível de célula, ES e DMEL.

Curiosamente, DMEL compartilha uma fração considerável de GNEs com esses estudos (as próximas maiores categorias, compreendendo 598 genes compartilhados com ambos e 295 compartilhados apenas com o conjunto de dados em nível de célula), o que pode sugerir uma robustez para identificar GNEs em nível de célula em todas as estratégias. No entanto, a lista DMEL de GEs é amplamente discrepante das fornecidas por esses estudos, pois a maioria deles está ausente da lista de ES e rotulada como GNEs na lista de nível de célula. Dos GEs restantes no DMEL, 120 (7,8% dos GEs do DMEL) também são rotulados como GEs nos dados de ES, mas como GNEs nos dados em nível de célula, o que pode sugerir que ambos os conjuntos de dados podem capturar uma fração de GEs de desenvolvimento que não são essenciais a nível celular.

Também avaliamos os datasets DMEL, TRIB, nível celular e ES para possível enriquecimento de categorias funcionais para destacar possíveis vias biológicas super-representadas nessas listas (Figura 14B). Encontramos várias vias de desenvolvimento significativamente enriquecidas em DMEL (por exemplo, Hippo, Notch, MAPK, TGF-beta e Wnt) (SONOSHITA; CAGAN, 2017). Nenhuma categoria enriquecida foi encontrada no conjunto de dados em nível de célula e apenas três categorias representando processos fundamentais em nível celular (RNA polimerase, Proteassoma e Ribossomo) foram encontrados enriquecidos no banco de dados ES, resultado semelhante à análise de enriquecimento feita pelos autores desta lista de genes (CAMPOS et al., 2020). Como a estratégia ES depende de uma pontuação que provavelmente selecionará como GEs os genes onde a grande maioria dos alelos é letal, e à luz dos resultados da análise de enriquecimento, argumentamos que a lista ES parece conter uma fração substancial de células genes de limpeza de nível. Mapeamos os 1643 GEs do besouro em 1466 genes da mosca usando a ferramenta de consulta iBeetle, e estes foram usados para pesquisar o conjunto TRIB de GEs para enriquecimento das vias KEGG. Também observamos um enriquecimento de genes pertencentes a vias de desenvolvimento (por exemplo, biossíntese de hormônios de insetos, Hippo, AGE-RAGE), uma semelhança funcional compartilhada com o conjunto de dados DMEL e uma evidência importante de que nossa seleção de GEs para esses insetos capturou uma fração de processos essenciais conhecidos ao nível do organismo. Em menor grau, o conjunto de dados TRIB também compartilhou semelhanças funcionais com a lista de genes ES (por exemplo, via Proteassoma).

Para comparar o desempenho relativo dos rótulos distintos da lista de genes de mosca para prever GEs no besouro, treinamos modelos individuais usando-os como fonte de rótulos de dados e avaliamos seu desempenho no conjunto de dados TRIB (todos os genes e conjunto

de dados NOH). Especificamente, nosso experimento foi o seguinte: (i) uso apenas de atributos intrínsecos, para refletir o pior cenário de desempenho; (ii) nearZeroVar() para pré-filtragem de atributos de baixa variância; e (iii) RF como modelo, pois é conhecido por ter um bom desempenho em uma ampla gama de tarefas de classificação (Smith and Frank, 2016).

Vimos que ambos os conjuntos DMEL e ES produzem modelos capazes de prever GEs no besouro significativamente melhor do que os modelos ZR (P <0,05, Figura 13 B e C, valores AUC). O modelo treinado usando o conjunto DMEL também superou significativamente o ES. É importante ressaltar que o conjunto de dados DMEL também foi mais bem-sucedido do que os dados em nível de célula e ES para desenvolver modelos capazes de prever GEs específicos de besouro (Figura 13 D e E, valores de AUC maiores do que os obtidos nas Figura 13 C e D). Levando em conta todas as propriedades observadas no conjunto de dados DMEL quando comparados com os esquemas de rotulagem em nível de célula e ES - uma distribuição de classes mais equilibrada, o enriquecimento de vias de desenvolvimento e o desenvolvimento de melhores modelos preditivos para genes essenciais no besouro - usamos o conjunto de dados DMEL como nossa fonte de informações de essencialidade dos genes de *D. melanogaster* para todas as análises a jusante.

3.5 Avaliando a importância relativa dos atributos intrínsecos

Até onde sabemos, todos os preditores para GEs codificadores de proteínas que dependem de características intrínsecas usam apenas dados de proteínas com base no raciocínio de que a essencialidade de tais genes é determinada pela ausência do papel funcional desempenhado pelas proteínas codificadas por eles. Como nosso pipeline computa características intrínsecas baseadas em nucleotídeos e aminoácidos, avaliamos se é possível prever GEs em insetos usando apenas nucleotídeos, proteínas ou a combinação de ambas as fontes de características intrínsecas. Especificamente, usamos modelos RF e XGBT treinados usando todos os recursos intrínsecos disponíveis após o filtro nearZeroVar em uma espécie de inseto e avaliamos seu desempenho nas outras espécies enquanto variamos os tipos de recursos intrínsecos (Figura 15 A-D). Todos os modelos tiveram um desempenho melhor do que os modelos ZR ($p < 0,05$). O algoritmo de RF é equivalente ou melhor que o XGBT para todos os cenários, com as maiores diferenças observadas no conjunto de dados NOH, especialmente para o conjunto treinado em tempo real. Curiosamente, os modelos treinados com atributos baseados em nucleotídeos também são capazes de prever GEs sozinhos, embora geralmente tenham um desempenho um pouco pior do que os modelos que usam dados baseados em

proteínas, novamente com exceção dos dados NOH treinados na mosca. Considerando que as características baseadas em nucleotídeos são pouco exploradas e, como demonstramos, podem gerar modelos razoavelmente eficientes para prever EGs, decidimos usar características intrínsecas derivadas de nucleotídeos e aminoácidos em nossos próximos experimentos.

Nós estimamos a melhor compensação entre números de recursos e importância. Para isso, inicialmente treinamos um modelo de RF usando todas as características intrínsecas após o filtro nearZeroVar para os dados da mosca e do besouro juntos (DMEL+TRIB) para estimar a importância relativa das características para todos os atributos intrínsecos para ambas as espécies simultaneamente. De posse dos valores de importância relativa para todos os atributos intrínsecos, treinamos novos modelos em uma única espécie enquanto validamos na outra, usando apenas subconjuntos de atributos com base em sua importância relativa. Então, avaliamos o desempenho do modelo - definido como valores de AUC para curvas ROC e PR - enquanto se variava o corte de importância do atributo (Figura 15 E-F). Além disso, para avaliar se os atributos selecionados usando RF aumentam o desempenho de outros modelos estatísticos, usamos os atributos selecionados para treinar também modelos XGBT.

Descobrimos que a distribuição de AUCs tem uma distribuição em forma de sino, com classificadores exibindo um desempenho inferior quando treinados usando todos os atributos independentemente de suas importâncias, ou quando treinados com apenas alguns poucos atributos dos que tiveram as maiores importâncias relativas (Figura 15 E e F). Para fins de comparação, o modelo de RF treinado na mosca com importância de recurso igual a zero (AUC-ROC e PRC de 0,617 e 0,593, respectivamente) é o mesmo mostrado nas Figura 15 B e C. Nenhum ponto de corte único forneceu valores máximos simultaneamente para as AUCs de ROC e PRC em todas as combinações de conjuntos de validação e abordagens de aprendizado de máquina. Mas o intervalo de importância variando de 11 a 13 exibiu os melhores resultados para a maioria das combinações. Portanto, decidimos usar 12 como nosso ponto de corte para a importância do atributo, pois parece representar o melhor desempenho geral para os diferentes modelos.

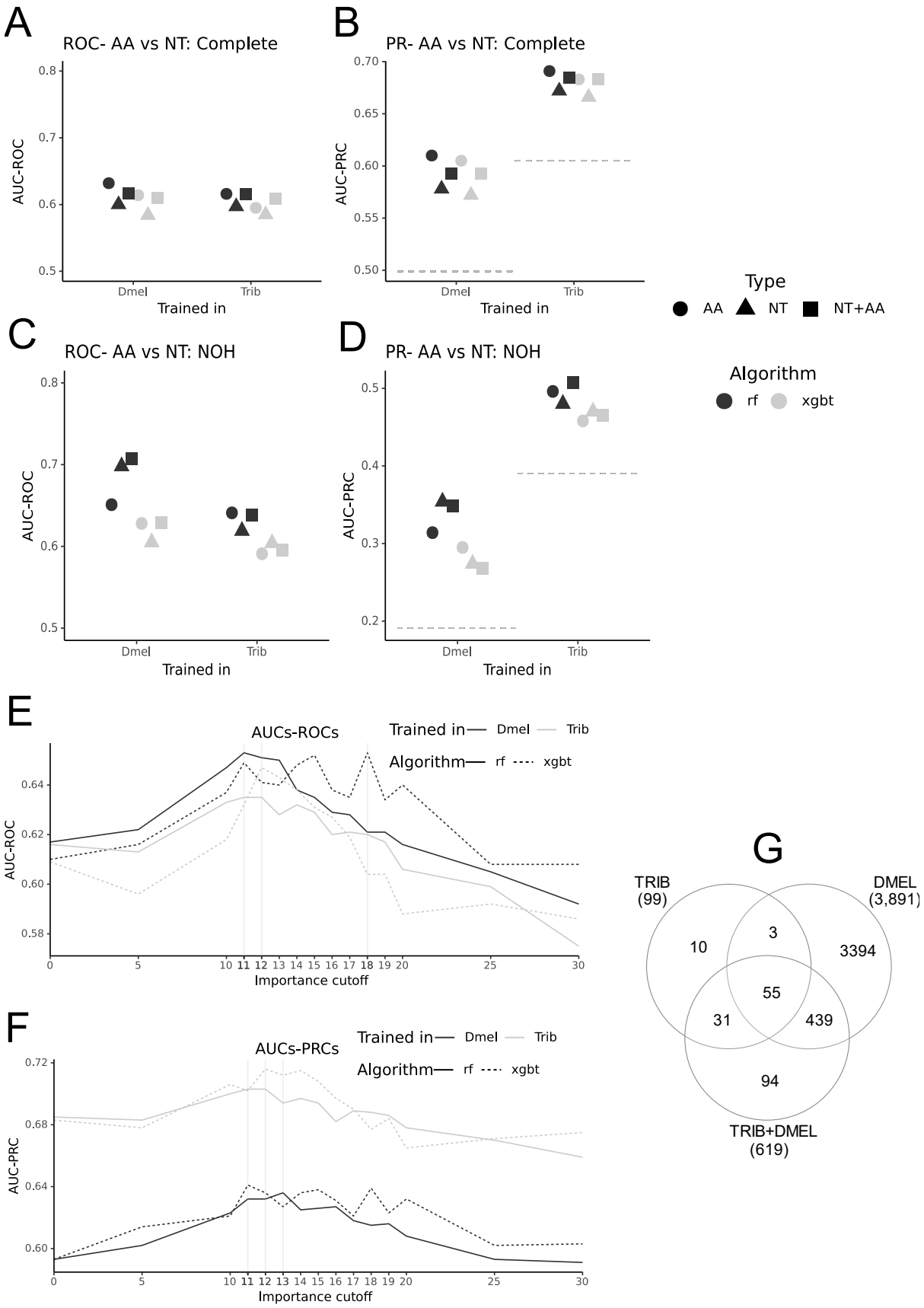


Figura 15: Estratégias de seleção de recursos para melhoria de desempenho do modelo. *A) AUC-ROCs para modelos treinados usando RF (em preto) e XGBT (em cinza) e categorias distintas de atributos intrínsecos (AA-círculos, somente aminoácidos; NT-Triângulos: somente*

nucleotídeos; NT+AA-quadrados: ambos) em todos os genes de uma espécie e validados em todos os genes da outra. B) AUC-PRCs para modelos treinados usando RF e XGBT e categorias distintas de atributos intrínsecos em todos os genes de uma espécie e validados em todos os genes da outra. C) AUC-ROCs para modelos treinados usando RF e XGBT e categorias distintas de atributos intrínsecos em todos os genes de uma espécie e validados nos genes linhagem-específicos da outra. D) AUC-PRCs para modelos treinados usando RF e XGBT e categorias distintas de atributos intrínsecos em todos os genes de uma espécie e validados nos genes linhagem-específicos da outra. E) Distribuição de AUC-ROCs para modelos RF e XGBT, variando a importância do atributo. F) Distribuição de AUC-PRCs para modelos RF e XGBT, variando a importância do atributo. G) Atributos compartilhados e exclusivos com importância > 12 em modelos treinados usando apenas dados DMEL, dados TRIB ou dados TRIB+DMEL.

Do total de 8.472 atributos selecionados após o filtro nearZeroVar, 5.093 e 3.379 são calculados a partir de sequências de nucleotídeos e proteínas, e 619 atributos intrínsecos têm importância relativa maior que 12 no conjunto de dados DMEL+TRIB (337 e 282 são baseados em nucleotídeos e aminoácidos, respectivamente). Comparamos esta lista de atributos com os obtidos dos modelos treinados em apenas uma espécie para avaliar melhor quais atributos foram considerados importantes nesses três experimentos (definidos como variáveis com importância relativa > 12 obtidas dos modelos de RF treinados em uma única espécie usando Características NT+AA, Figura 15 A e B). Descobrimos que os modelos treinados usando apenas dados DMEL ou TRIB têm um grau considerável de variabilidade em relação às variáveis selecionadas como importantes (Figura 15 G). Com o TRIB foi selecionado 99 variáveis (88 nucleotídeos e 11 à base de proteínas), enquanto com o DMEL foi selecionado 3.891 variáveis (2.256 nucleotídeos e 1.635 à base de proteínas). Curiosamente, encontramos um núcleo de 55 atributos selecionados independentemente pelos três modelos (46 nucleotídeos e 9 baseados em proteínas). Os atributos importantes do DMEL+TRIB também compartilham uma sobreposição considerável com o DMEL (439 recursos) e o TRIB (31 recursos), indicando que ele captura independentemente propriedades importantes inferidas por meio desses experimentos.

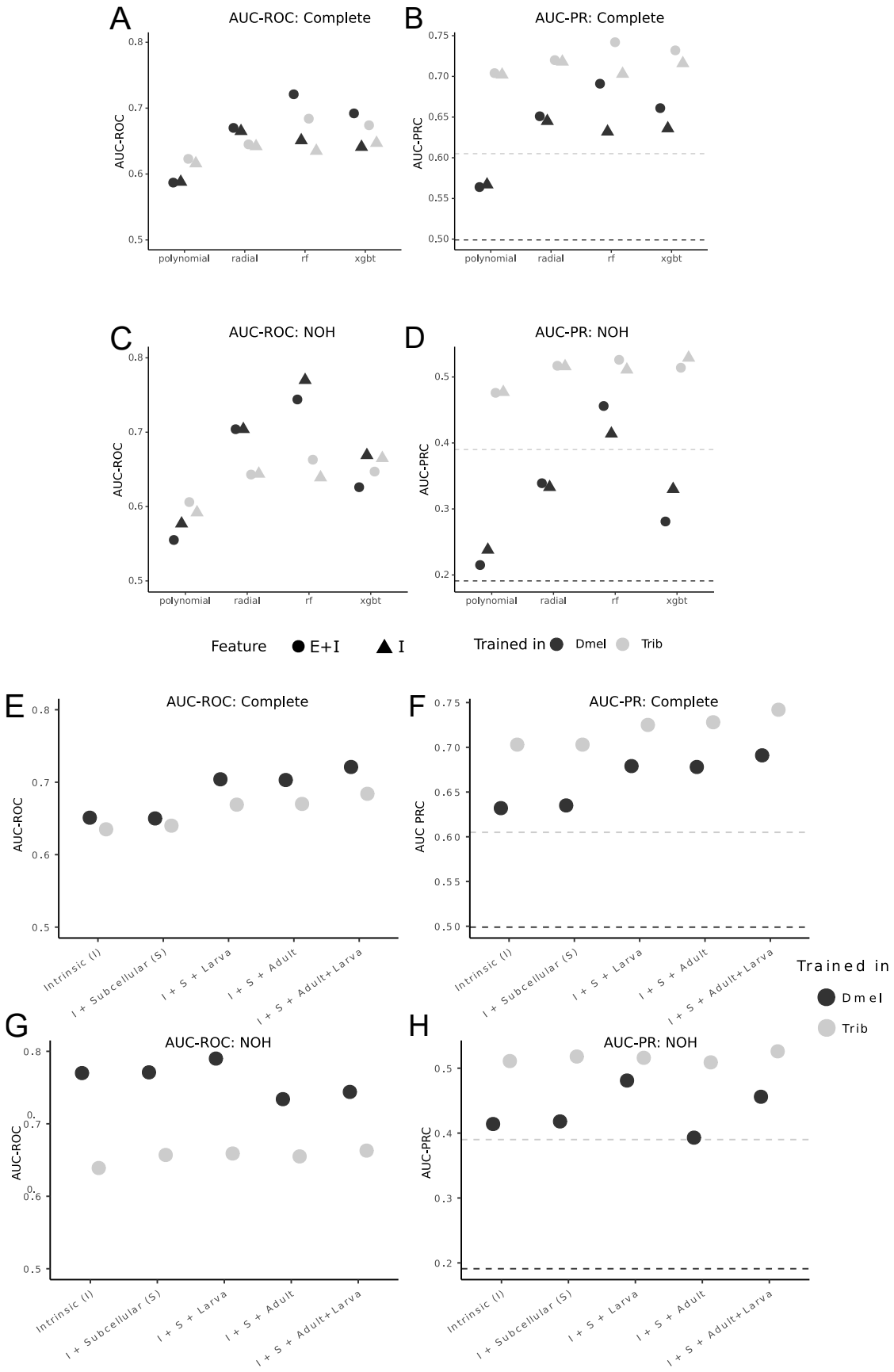


Figura 16: Avaliação de estratégias distintas de modelos e classes de atributos no desempenho do classificador. *A) AUC-ROCs para modelos treinados usando Support Vector*

Machines (SVM) com kernels polinomiais (POLY) ou radiais (RAD), Random Forests (RF) e Extreme Gradient Boosting (XGBT), com apenas atributos intrínsecos ou uma combinação de atributos intrínsecos e extrínsecos em todos os genes de uma espécie e validados em todos os genes de outra espécie. B) AUC-PRs para os modelos treinados em 'A' e validados em todos os genes das outras espécies. C) AUC-ROCs para os modelos treinados em 'A' e validados nos genes específicos da linhagem das outras espécies. D) AUC-PRs para os modelos treinados em 'A' e validados nos genes específicos da linhagem das outras espécies. E) AUC-ROCs para modelos treinados usando Random Forests e combinações de atributos intrínsecos e extrínsecos em todos os genes de uma espécie e validados em todos os genes da outra (I - Intrínseco; S - Localização subcelular; L - RNA-seq do estágio Larva; A - RNA-Seq do Estágio Adulto). F) AUC-PRs para os modelos treinados em 'E' e validados em todos os genes das outras espécies. G) AUC-ROCs para os modelos treinados em 'E' e validados nos genes específicos da linhagem das outras espécies. H) AUC-PRs para modelos treinados em 'E' e validados nos genes específicos de linhagem de outras espécies

Entre as 9 características baseadas em proteínas principais encontramos várias propriedades já associadas a genes essenciais em outros estudos, como a proporção de aminoácidos aromáticos (GONG et al., 2008) e a polaridade de aminoácidos (SERINGHAUS et al., 2006). Quanto às características baseadas em nucleotídeos, que, em contraste, e surpreendentemente, compreendem os atributos mais abundantes e importantes para todos os modelos, encontramos um enriquecimento de mais de quatro vezes de características de covariância auto-cruzada baseadas em dínucleotídeos, que compreendiam 2.888 dos 15.388 atributos intrínsecos (~18%), mas correspondeu a 43 dos 55 atributos principais (~78%). Essas 43 feições representam diversas propriedades locais heterogêneas, variando de características físico-químicas a composicionais, e sua interpretação biológica ainda precisa ser avaliada. Como os 619 atributos encontrados no conjunto de dados DMEL+TRIB parecem capturar a maioria dos atributos mais importantes compartilhados pelos modelos treinados em uma única espécie, bem como características únicas observadas exclusivamente nos modelos de espécie única, procedemos usando esses 619 atributos intrínsecos para desenvolver ainda mais nosso preditor.

3.6 Construindo nossos modelos finais: integrando atributos extrínsecos e avaliando abordagens distintas de aprendizado de máquina

O passo final para o desenvolvimento do nosso preditor de EGs em insetos consistiu na avaliação da influência de duas classes de atributos extrínsecos que podem ser obtidos para a maioria dos insetos com dados genômicos disponíveis (RNA-Seq) ou computados para

virtualmente qualquer sequência de proteínas (previsão de localização celular) no desempenho de nossos classificadores. Além disso, como realizamos a seleção de recursos, agora também é possível ter uma comparação justa do desempenho do SVM com RF e XGBT, pois o primeiro não possui seleção de recursos intrínsecos como componente de seu funcionamento interno. Treinamos e testamos nossos modelos da seguinte forma: (i) pré-filtragem de recursos [nearZeroVar()]; (ii) filtragem de recursos (importância de recursos < 12 no conjunto de dados DMEL+TRIB, abrangendo 619 atributos); (iii) adição opcional de atributos extrínsecos (2 RNA-Seq dos estágios adulto e larval + 11 probabilidades de predição da localização celular).

Todos os métodos (SVM polinomial, SVM radial, RF e XGBT) tiveram um desempenho significativamente melhor do que os modelos ZR, considerando apenas características intrínsecas ou a combinação de atributos intrínsecos e extrínsecos (Figura 16 A e B para todos os genes, Figura 16 C e D para genes linhagem-específicos, $P < 0,05$). O SVM polinomial teve o pior desempenho, enquanto os demais métodos possuem perfis mais semelhantes, sendo o RF o melhor modelo na maioria dos cenários. A adição de atributos extrínsecos aumentou ainda mais o desempenho do classificador para modelos RF e XGBT na maioria dos cenários, mas não apresentou grandes mudanças no modelo radial SVM. Nossa análise final consistiu em avaliar a importância relativa de classes distintas de atributos para o treinamento do modelo de RF (Figura 16 E e F para todos os genes, Figura 16 G e H para genes específicos de linhagem). Descobrimos que, para a maioria dos cenários, os dados de localização de células subcelulares aumentam apenas marginalmente o desempenho do classificador. Os dados de RNA-Seq, por outro lado, desempenharam um papel muito mais proeminente na melhoria do desempenho do classificador, especialmente ao considerar todos os genes.

4. Discussão

Aqui relatamos, até onde sabemos, o desenvolvimento e validação do primeiro preditor geral de GEs usando atributos intrínsecos e extrínsecos que podem ser obtidos para a maioria dos genomas de insetos com informações de previsão de genes disponíveis. Também relatamos a compilação de conjuntos de genes essenciais e não essenciais para os dois organismos modelo de insetos onde esta informação está disponível – a mosca da fruta *D. melanogaster* e o besouro vermelho da farinha *T. castaneum*. Além disso, demonstramos que nossa lista de genes de mosca para GEs e GNEs é mais bem sucedida para o desenvolvimento de preditores capazes de predizer GEs de besouros quando comparados com os atualmente disponíveis.

Nossos conjuntos de dados GEs e GNEs para as duas espécies de insetos foram obtidos usando estratégias experimentais distintas com ontologias curadas, o que diminui a possibilidade de vieses sistemáticos introduzidos por variáveis experimentais (por exemplo, a detecção de GEs em nível celular sendo causada apenas pelo uso exclusivo de experimentos *in vitro* para procurar GEs). Especificamente, os procedimentos experimentais usados para gerar dados de essencialidade para ambos os organismos compreendem ensaios *in vivo* em vários estágios de desenvolvimento, permitindo assim treinar modelos eventualmente capazes de detectar GEs de desenvolvimento além dos GEs correspondentes a processos em nível celular. As informações de essencialidade gênica em larga escala disponíveis para as espécies *D. melanogaster* (Diptera) e *T. castaneum* (Coleoptera) e compiladas neste trabalho são um recurso útil para desenvolver e validar preditores gerais de GEs de insetos, incluindo candidatos específicos de linhagem.

Descobrimos que os modelos de RF treinados em dados DMEL ou TRIB compartilham um conjunto de recursos principais entre os mais relevantes, bem como vários atributos não sobrepostos, demonstrando assim que combinações distintas de recursos estão sendo selecionadas durante o treinamento do modelo. No entanto, esses conjuntos de características distintas ainda são capazes de prever GEs em outras espécies de insetos. Esse fato, juntamente com os valores de desempenho observados para nossos modelos, sugere que podemos melhorar ainda mais o desempenho do classificador por meio da engenharia de recursos e explorando outros esquemas de modelo de classificação disponíveis.

Nosso esquema de validação com duas espécies filogeneticamente distantes simula um cenário de caso real onde se deseja prever GEs para uma fração considerável das espécies de insetos restantes. Nossas ferramentas podem prever GEs mesmo no pior cenário (usando apenas atributos intrínsecos e em genes específicos de linhagem). Esse fato sugere que GEs ainda mais jovens, com restrição de linhagem, compartilham propriedades comuns suficientes incorporadas nas sequências biológicas de GEs de uma espécie de inseto filogeneticamente distante para permitir a previsão entre espécies.

5. Conclusão

Consideramos que nosso trabalho fornece um rico conjunto de possibilidades para desenvolvimento futuro na predição de genes essenciais em insetos, pois fornecemos (i) conjuntos de dados de GEs e GNEs para dois insetos filogeneticamente distantes; (ii) dados de atributos e homologia para esses genes e (iii) uma ampla variedade de modelos já pesquisados

para desempenho de classificação. Em termos práticos, demonstramos que nossos melhores modelos podem prever genes essenciais melhor do que os modelos de regra-zero (ZR) em uma ampla gama de cenários e com combinações distintas de características intrínsecas e extrínsecas. Também fornecemos modelos treinados que podem ser usados para prever GEs em espécies de insetos, desde modelos que usam atributos apenas intrínsecos até aqueles que integram atributos extrínsecos (dados de expressão e localização subcelular). Finalmente, como nosso pipeline calcula, até onde sabemos, o maior conjunto de recursos intrínsecos baseados em nucleotídeos e proteínas, podendo ser um recurso útil para o desenvolvimento de abordagens de aprendizado de máquina para prever outras classes de genes.

6. Referências do capítulo 2

ALMAGRO ARMENTEROS, J. J. et al. DeepLoc: prediction of protein subcellular localization using deep learning. **Bioinformatics**, v. 33, n. 21, p. 3387–3395, 1 nov. 2017.

AROMOLARAN, O. et al. Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features. **Computational and Structural Biotechnology Journal**, v. 18, p. 612–621, 2020.

BRŮNA, T. et al. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. **NAR Genomics and Bioinformatics**, v. 3, n. 1, p. lqaa108, 6 jan. 2021.

CAMPOS, T. L. et al. An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-Derived Features. **Computational and Structural Biotechnology Journal**, v. 17, p. 785–796, 8 jun. 2019.

CAMPOS, T. L. et al. Combined use of feature engineering and machine-learning to predict essential genes in Drosophila melanogaster. **NAR Genomics and Bioinformatics**, v. 2, n. 3, 1 set. 2020.

CHEN, S.; ZHANG, Y. E.; LONG, M. New Genes in *Drosophila* Quickly Become Essential. **Science**, v. 330, n. 6011, p. 1682–1685, 17 dez. 2010.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**... Em: KDD '16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. San Francisco California USA: ACM, 13 ago. 2016. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 21 set. 2022

CHEN, W.-H. et al. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. **Nucleic Acids Research**, v. 45, n. D1, p. D940–D944, 4 jan. 2017.

CRESPO-PÉREZ, V. et al. The importance of insects on land and in water: a tropical view. **Vectors and medical and veterinary entomology • Special Section on Insects and the UN sustainable development goals**, v. 40, p. 31–38, 1 ago. 2020.

DONG, C. et al. Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment. **Briefings in Bioinformatics**, 29 nov. 2018.

DÖNITZ, J. et al. iBeetle-Base: a database for RNAi phenotypes in the red flour beetle Tribolium castaneum. **Nucleic Acids Research**, v. 43, n. D1, p. D720–D725, 28 jan. 2015.

EMMS, D. M.; KELLY, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. **Genome Biology**, v. 20, n. 1, p. 238, dez. 2019.

EWEN-CAMPEN, B. et al. Accessing the Phenotype Gap: Enabling Systematic Investigation of Paralog Functional Complexity with CRISPR. **Developmental Cell**, v. 43, n. 1, p. 6–9, out. 2017.

GONG, X. et al. Comparative analysis of essential genes and nonessential genes in Escherichia coli K12. **Molecular Genetics and Genomics**, v. 279, n. 1, p. 87–94, jan. 2008.

GUO, F.-B. et al. Accurate prediction of human essential genes using only nucleotide composition and association information. **Bioinformatics**, v. 33, n. 12, p. 1758–1764, 15 jun. 2017.

HUTCHISON, C. A. et al. Design and synthesis of a minimal bacterial genome. **Science**, v. 351, n. 6280, p. aad6253–aad6253, 25 mar. 2016.

KARATZOGLOU, A. et al. **kernlab** - An *S4* Package for Kernel Methods in *R*. **Journal of Statistical Software**, v. 11, n. 9, 2004.

KNORR, E. et al. Gene silencing in *Tribolium castaneum* as a tool for the targeted identification of candidate RNAi targets in crop pests. **Scientific Reports**, v. 8, n. 1, p. 2061, Fevereiro 2018.

KUHN, M. **caret: Classification and Regression Training**. [s.l: s.n.].

KUMAR, S. et al. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. **Molecular Biology and Evolution**, v. 34, n. 7, p. 1812–1819, jul. 2017.

LARKIN, A. et al. FlyBase: updates to the *Drosophila melanogaster* knowledge base. **Nucleic Acids Research**, v. 49, n. D1, p. D899–D907, 8 jan. 2021.

LIU, X. et al. Selection of key sequence-based features for prediction of essential genes in 31

diverse bacterial species. **PLOS ONE**, v. 12, n. 3, p. e0174638, 30 mar. 2017.

LUO, H. et al. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements: Table 1. **Nucleic Acids Research**, v. 42, n. D1, p. D574–D580, jan. 2014.

NIGATU, D. et al. Sequence-based information-theoretic features for gene essentiality prediction. **BMC bioinformatics**, v. 18, n. 1, p. 473, 9 nov. 2017.

PATRO, R. et al. Salmon provides fast and bias-aware quantification of transcript expression. **Nature Methods**, v. 14, n. 4, p. 417–419, abr. 2017.

PHILIPS, S.; WU, H.-Y.; LI, L. Using machine learning algorithms to identify genes essential for cell survival. **BMC Bioinformatics**, v. 18, n. S11, p. 397, out. 2017.

PORT, F. et al. A large-scale resource for tissue-specific CRISPR mutagenesis in Drosophila. **eLife**, v. 9, p. e53865, 13 fev. 2020.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2016.

RANCATI, G. et al. Emerging and evolving concepts in gene essentiality. **Nature Reviews Genetics**, v. 19, n. 1, p. 34–49, jan. 2018.

RUST, M. K.; SU, N.-Y. Managing Social Insects of Urban Importance. **Annual Review of Entomology**, v. 57, n. 1, p. 355–375, 7 jan. 2012.

SAYERS, E. W. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 49, n. D1, p. D10–D17, 8 jan. 2021.

SCHMITT-ENGEL, C. et al. The iBeetle large-scale RNAi screen reveals gene functions for insect development and physiology. **Nature Communications**, v. 6, p. 7822, 28 jul. 2015.

SERINGHAUS, M. et al. Predicting essential genes in fungal genomes. **Genome Research**, v. 16, n. 9, p. 1126–1135, set. 2006.

SONOSHITA, M.; CAGAN, R. L. Modeling Human Cancers in Drosophila. Em: **Current Topics in Developmental Biology**. [s.l.] Elsevier, 2017. v. 121p. 287–309.

STORK, N. E. How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? **Annual Review of Entomology**, v. 63, n. 1, p. 31–45, 7 jan. 2018.

SUN, X.; XU, W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. **IEEE Signal Processing Letters**, v. 21, n. 11, p. 1389–1393, nov. 2014.

SZKLARCZYK, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. **Nucleic Acids Research**, v. 49, n. D1, p. D605–D612, 8 jan. 2021.

TIAN, D. et al. Identifying mouse developmental essential genes using machine learning. **Disease Models & Mechanisms**, v. 11, n. 12, p. dmm034546, 1 dez. 2018.

VISWANATHA, R. et al. Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in Drosophila cells. **eLife**, v. 7, p. e36333, 27 jul. 2018.

WANG, N. et al. Genome-wide identification of Acinetobacter baumannii genes necessary for persistence in the lung. **mBio**, v. 5, n. 3, p. e01163- 01114, 3 jun. 2014.

WRIGHT, M. N.; ZIEGLER, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software, Articles**, v. 77, n. 1, p. 1–17, 2017.

YANG, L. et al. Analysis and identification of essential genes in humans using topological properties and biological information. **Gene**, v. 551, n. 2, p. 138–151, nov. 2014.

Conclusão Geral

Nesse trabalho, detectamos um viés sistemático em um banco de dados amplamente utilizado para estudos que visam produzir classificadores de GEs para organismos procarióticos, e demonstramos objetivamente como tais vieses possivelmente inflaram diversos resultados de predição de GEs publicados. Nós também demonstramos como rotinas simples, como alinhamentos de sequência e verificação de sinais canônicos de regiões codificadoras, poderiam ter sido utilizados para detectar estes erros.

Diversos dos trabalhos que consultamos não permitam a conclusão final de que usaram as sequências erradas do banco de dados DEG, embora haja evidências consideráveis de que que houve o uso delas. Esse fato demonstra, mais uma vez, como a impossibilidade de reprodução de resultados contribui para a crise de credibilidade na ciência.

De maneira louvável, o banco de dados DEG mantém um registro das versões anteriores do banco, o que possibilitou a recuperação da informação de sequência dos GNEs e a eventual detecção do erro sistemático mencionado no capítulo 1. Também demonstrando boas práticas, o trabalho de Hasan e Lonardi (2020), ao disponibilizar os dados brutos de análise, permitiu que pudéssemos concluir que os dados inconsistentes do DEG foram utilizados em publicações recentes para o treinamento de preditores de GEs.

O pacote R desenvolvido nesse projeto contém todos os dados utilizados em sua validação em procariotos e eucariotos, além do código utilizado na produção de todas as figuras, o que permite a auditoria completa da análise de maneira a garantir a reprodutibilidade dos nossos resultados, a sua eventual melhoria, bem como algum erro que eventualmente tenhamos cometido. E embora nosso estudo de caso consista na predição de GEs, os atributos aqui calculados podem, em teoria, ser utilizados para quaisquer tarefas que visem o desenvolvimento de classificadores de dados de sequências biológicas.

Como validação nos organismos procarióticos, utilizamos experimentos *in silico* controlados para avaliar objetivamente como os diferentes conjuntos de sequências de GNEs influenciaram o treinamento e a validação de diversas classes de preditores de GEs, incluindo modelos análogos aos já publicados e que contém o erro do banco DEG e modelos que consideramos ser o real estado-da-arte para a predição de GEs em procariotos.

Embora tenhamos demonstrado que nossos preditores possuem desempenho equivalente, e possivelmente superior, à estudos que não fazem uso de dados de sequência do DEG, demonstramos também que diversos dos modelos já publicados estão possivelmente inflados artificialmente em função da presença do erro sistemático do DEG.

A nossa validação em insetos demonstrou que o protocolo aqui desenvolvido também pode ser utilizado para a predição de GEs em organismos eucarióticos complexos. Com o uso de duas espécies de insetos e validações do tipo *leave-one-organism-out*, foi demonstrado que nossa ferramenta potencialmente funcionará em novas espécies de insetos ainda não caracterizadas, sendo inclusive capaz de detectar GEs linhagem-específicos.

Até onde pudemos verificar, o protocolo de análise aqui apresentado compreende o mais amplo projeto de desenvolvimento e validação de algoritmos de aprendizado de máquina que fazem uso de atributos intrínsecos às sequências nucleotídicas, proteicas e com a opção de adicionar os extrínsecos com bibliotecas de RNA-Seq, para a predição de GEs em organismos procarióticos e eucarióticos, aumentando de maneira substancial a disponibilidade de dados brutos, código e protocolos de validação para tal.

Apêndice

Tabela complementar 1: Seleção manual de genes essenciais e não essenciais de D. melanogaster da revisão de literatura.

GeneID	0 = Não essencial 1 = Essencial	Referência
FBgn0042178	0	[1]
FBgn0004475	0	[1]
FBgn0264707	0	[2]
FBgn0004513	0	[3]
FBgn0004512	0	[3]
FBgn0010241	0	[3]
FBgn0262782	0	[4]
FBgn0264953	0	[5]
FBgn0035154	0	[6]
FBgn0263934	0	[7]
FBgn0260747	0	[8]
FBgn0030706	0	[9]
FBgn0035610	0	[10]
FBgn0034590	0	[11]
FBgn0024992	0	[12]
FBgn0011676	0	[13]
FBgn0036919	0	[14]
FBgn0261787	0	[15]
FBgn0038303	0	[15]
FBgn0032879	0	[16]
FBgn0052732	0	[17]
FBgn0040475	0	[18]
FBgn0037780	0	[19]
FBgn0040752	0	[20]
FBgn0030313	0	[21]
FBgn0034617	0	[22]
FBgn0036411	0	[23]
FBgn0033051	0	[24]
FBgn0040070	0	[25]
FBgn0004108	0	[26]
FBgn0028717	0	[27]
FBgn0033483	0	[28]
FBgn0035207	0	[29]
FBgn0035111	0	[30]
FBgn0037470	0	[30]
FBgn0030600	0	[31]
FBgn0051217	0	[32]
FBgn0039055	0	[33]
FBgn0053052	0	[34]
FBgn0034739	0	[35]
FBgn0002940	0	[36]
FBgn0036260	0	[37]
FBgn0039862	0	[38]
FBgn0035379	0	[38]
FBgn0259683	0	[39]
FBgn0025382	0	[40]
FBgn0260986	0	[41]
FBgn0263199	0	[42]
FBgn0003741	0	[43]
FBgn0037659	0	[44]
FBgn0037703	0	[44]
FBgn0033233	0	[44]
FBgn0053182	0	[44]
FBgn0266570	0	[44]
FBgn0036366	0	[44]
FBgn0035166	0	[44]
FBgn0263025	0	[44]

FBgn0038948	0	[44]
FBgn0032671	0	[44]
FBgn0033238	0	[45]
FBgn0004050	0	[46]
FBgn0264272	0	[47]
FBgn0262369	0	[48]
FBgn0036125	0	[49]
FBgn0027528	0	[50]
FBgn0015575	0	[51]
FBgn0034135	0	[52]
FBgn0023517	0	[53]
FBgn0004575	0	[54]
FBgn0029976	0	[55]
FBgn0033744	0	[56]
FBgn0039666	0	[57]
FBgn0035847	0	[58]
FBgn0046885	0	[58]
FBgn0051438	0	[58]
FBgn0032439	0	[58]
FBgn0051882	0	[58]
FBgn0032754	0	[58]
FBgn0036970	0	[58]
FBgn0032585	0	[58]
FBgn0051406	0	[58]
FBgn0011832	0	[58]
FBgn0034156	0	[58]
FBgn0037974	0	[58]
FBgn0052301	0	[58]
FBgn0052282	0	[58]
FBgn0028987	0	[58]
FBgn0034427	0	[58]
FBgn0053462	0	[58]
FBgn0038299	0	[58]
FBgn0039739	0	[58]
FBgn0038888	0	[58]
FBgn0034870	0	[58]
FBgn0286516	1	[59]
FBgn0028418	1	[10]
FBgn0035416	1	[15]
FBgn0260859	1	[15]
FBgn0266722	1	[15]
FBgn0266724	1	[15]
FBgn0260655	1	[15]
FBgn0037551	1	[60]
FBgn0000723	1	[61]
FBgn0000463	1	[62]
FBgn0263864	1	[63]
FBgn0005654	1	[64]
FBgn0267975	1	[65]
FBgn0032407	1	[66]
FBgn0266418	1	[67]
FBgn0027053	1	[68]
FBgn0000063	1	[41]
FBgn0000449	1	[69]
FBgn0260855	1	[70]
FBgn0260749	1	[44]
FBgn0036003	1	[44]
FBgn0000575	1	[71]
FBgn0025641	1	[72]
FBgn0003870	1	[73]
FBgn0260635	1	[74]
FBgn0002781	1	[75]
FBgn0264307	1	[76]
FBgn0263396	1	[77]
FBgn0031359	1	[78]
FBgn0015795	1	[79]

FBgn0020309	1	[58]
FBgn0020305	1	[58]
FBgn0021796	1	[58]
FBgn0004374	1	[58]
FBgn0267828	1	[58]
FBgn0031604	1	[58]
FBgn0003145	1	[58]
FBgn0004914	1	[58]
FBgn0000299	1	[58]
FBgn0001981	1	[58]
FBgn0032683	1	[58]
FBgn0261983	1	[58]
FBgn0040232	1	[58]
FBgn0023388	1	[58]
FBgn0000422	1	[58]
FBgn0263933	1	[58]
FBgn0040228	1	[58]
FBgn0014127	1	[58]
FBgn0086444	1	[58]
FBgn0002524	1	[58]
FBgn0000546	1	[80]
FBgn0003964	1	[81]

Referências da tabela complementar 1

1. VanKuren NW, Long M. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. Nat Ecol Evol. 2018;2:705–12.

2. Couturier L, Mazouni K, Bernard F, Besson C, Reynaud E, Schweisguth F. Regulation of cortical stability by RhoGEF3 in mitotic Sensory Organ Precursor cells in *Drosophila*. Biology Open. 2017;6:1851–60.

3. Denecke S, Fusetto R, Batterham P. Describing the role of Drosophila melanogaster ABC transporters in insecticide biology using CRISPR-Cas9 knockouts. Insect Biochemistry and Molecular Biology. 2017;91:1–9.

4. Eanes WF, Merritt TJS, Flowers JM, Kumagai S, Zhu C-T. Direct Evidence That Genetic Variation in Glycerol-3-Phosphate and Malate Dehydrogenase Genes (*Gpdh* and *Mdh1*) Affects Adult Ethanol Tolerance in *Drosophila melanogaster*. Genetics. 2009;181:607–14.

5. He L, Si G, Huang J, Samuel ADT, Perrimon N. Mechanical regulation of stem-cell differentiation by the stretch-activated Piezo channel. Nature. 2018;555:103–6.

6. Huang H-W, Brown B, Chung J, Domingos PM, Ryoo HD. highroad Is a Carboxypetidase Induced by Retinoids to Clear Mutant Rhodopsin-1 in Drosophila Retinitis Pigmentosa Models. Cell Reports. 2018;22:1384–91.

7. Matsubara D, Horiuchi S-Y, Shimono K, Usui T, Uemura T. The seven-pass transmembrane cadherin Flamingo controls dendritic self-avoidance via its binding to a LIM domain protein, Espinas, in Drosophila sensory neurons. Genes Dev. 2011;25:1982–96.

8. Meng H, Yamashita C, Shiba-Fukushima K, Inoshita T, Funayama M, Sato S, et al. Loss of Parkinson's disease-associated protein CHCHD2 affects mitochondrial crista structure and destabilizes cytochrome c. Nature Communications. 2017;8:15500.

9. Mosca TJ, Luginbuhl DJ, Wang IE, Luo L. Presynaptic LRP4 promotes synapse number and function of excitatory CNS neurons. eLife. 2017;6. doi:10.7554/eLife.27347.

10. Ohashi H, Sakai T. Leucokinin signaling regulates hunger–driven reduction of behavioral responses to noxious heat in Drosophila. Biochemical and Biophysical Research Communications. 2018;499:221–6.

11. Padash Barmchi M, Samarasekera G, Gilbert M, Auld VJ, Zhang B. Magi Is Associated with the Par Complex and Functions Antagonistically with Bazooka to Regulate the Apical Polarity Complex. PLOS ONE. 2016;11:e0153259.

12. Pareek G, Thomas RE, Pallanck LJ. Loss of the Drosophila m-AAA mitochondrial protease paraplegin results in mitochondrial dysfunction, shortened lifespan, and neuronal and muscular degeneration. Cell Death & Disease. 2018;9. doi:10.1038/s41419-018-0365-8.

13. Rabinovich D, Yaniv SP, Alyagor I, Schuldiner O. Nitric Oxide as a Switching Mechanism between Axon Degeneration and Regrowth during Developmental Remodeling. Cell. 2016;164:170–82.

14. Rajan A, Housden BE, Wirtz-Peitz F, Holderbaum L, Perrimon N. A Mechanism Coupling Systemic

Energy Sensing to Adipokine Secretion. Developmental Cell. 2017;43:83-98.e6.

15. Riedel F, Galindo A, Muschalik N, Munro S. The two TRAPP complexes of metazoans have distinct roles and act on different Rab GTPases. The Journal of Cell Biology. 2018;217:601–17.

16. Stenesen D, Moehlman AT, Krämer H. The carcinine transporter CarT is required in Drosophila photoreceptor neurons to sustain histamine recycling. eLife. 2015;4. doi:10.7554/eLife.10972.

17. Tiebe M, Lutz M, Levy D, Teleman AA. Phenotypic characterization of SETD3 knockout Drosophila. PLOS ONE. 2018;13:e0201609.

18. Ukken FP, Bruckner JJ, Weir KL, Hope SJ, Sison SL, Birschbach RM, et al. BAR-SH3 sorting nexins are conserved interacting proteins of Nervous wreck that organize synapses and promote neurotransmission. Journal of Cell Science. 2016;129:166–77.

19. Wakabayashi S, Sawamura N, Voelzmann A, Broemer M, Asahi T, Hoch M. Ohgata, the Single *Drosophila* Ortholog of Human Cereblon, Regulates Insulin Signaling-dependent Organismic Growth. Journal of Biological Chemistry. 2016;291:25120–32.

20. Wu S, Gan G, Zhang Z, Sun J, Wang Q, Gao Z, et al. A Presynaptic Function of Shank Protein in *Drosophila*. The Journal of Neuroscience. 2017;37:11592–604.

21. Yang SY, Chang Y-C, Wan YH, Whitworth C, Baxter EM, Primus S, et al. Control of a Novel Spermatocyte-Promoting Factor by the Male Germline Sex Determination Factor PHF7 of *Drosophila melanogaster*. Genetics. 2017;206:1939–49.

22. Yu Y, Gu J, Jin Y, Luo Y, Preall JB, Ma J, et al. Panoramix enforces piRNA-dependent cotranscriptional silencing. Science. 2015;350:339–42.

23. Zhai Z, Kondo S, Ha N, Boquete J-P, Brunner M, Ueda R, et al. Accumulation of differentiating intestinal stem cell progenies drives tumorigenesis. Nature Communications. 2015;6. doi:10.1038/ncomms10219.

24. Baum JS, Arama E, Steller H, McCall K. The Drosophila caspases Strica and Dronc function redundantly in programmed cell death during oogenesis. Cell Death Differ. 2007;14:1508–17.

25. Tsuda M, Ootaka R, Ohkura C, Kishita Y, Seong K-H, Matsuo T, et al. Loss of *Trx-2* enhances oxidative stress-dependent phenotypes in *Drosophila*. FEBS Letters. 2010;584:3398–401.

26. Speicher S, García-Alonso L, Carmena A, Martín-Bermudo MD, de la Escalera S, Jiménez F. Neurotactin Functions in Concert with Other Identified CAMs in Growth Cone Guidance in Drosophila. Neuron. 1998;20:221–33.

27. Slack C, Werz C, Wieser D, Alic N, Foley A, Stocker H, et al. Regulation of Lifespan, Metabolism, and Stress Responses by the Drosophila SH2B Protein, Lnk. PLoS Genet. 2010;6:e1000881.

28. Geuking P, Narasimamurthy R, Lemaitre B, Basler K, Leulier F. A Non-Redundant Role for Drosophila Mkk4 and Hemipterous/Mkk7 in TAK1-Mediated Activation of JNK. PLoS ONE. 2009;4:e7709.

29. Aerne BL, Gailite I, Sims D, Tapon N. Hippo Stabilises Its Adaptor Salvador by Antagonising the HECT Ubiquitin Ligase Herc4. PLoS ONE. 2015;10:e0131113.

30. Lin C-J, Wen J, Bejarano F, Hu F, Bortolamiol-Becet D, Kan L, et al. Characterization of a TUTase/RNase complex required for *Drosophila* gametogenesis. RNA. 2017;23:284–96.

31. Wan HI, DiAntonio A, Fetter RD, Bergstrom K, Strauss R, Goodman CS. Highwire Regulates Synaptic Growth in Drosophila. Neuron. 2000;26:313–29.

32. Buchon N, Poidevin M, Kwon H-M, Guillou A, Sottas V, Lee B-L, et al. A single modular serine protease integrates signals from pattern-recognition receptors upstream of the Drosophila Toll pathway. Proceedings of the National Academy of Sciences. 2009;106:12442–7.

33. Polesello C, Huelsmann S, Brown NH, Tapon N. The Drosophila RASSF Homolog Antagonizes the Hippo Pathway. Current Biology. 2006;16:2459–65.

34. Kovacs L, Chao-Chu J, Schneider S, Gottardo M, Tzolovsky G, Dzhindzhev NS, et al. Gorab is a Golgi protein required for structure and duplication of Drosophila centrioles. Nat Genet. 2018;50:1021–31.

35. Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, Zhao R, et al. A Young Drosophila Duplicate Gene Plays Essential Roles in Spermatogenesis by Regulating Several Y-Linked Male Fertility Genes. PLoS Genet. 2010;6:e1001255.

36. Kumar JP, Ready DF. Rhodopsin plays an essential structural role in Drosophila photoreceptor development. Development. 1995;121:4359.

37. Ni JD, Baik LS, Holmes TC, Montell C. A rhodopsin in the brain functions in circadian photoentrainment in Drosophila. Nature. 2017;545:340–4.

38. Ulian-Benitez S, Bishop S, Foldi I, Wentzell J, Okenwa C, Forero MG, et al. Kek-6: A truncated-Trk-like

receptor for Drosophila neurotrophin 2 regulates structural synaptic plasticity. PLoS Genet. 2017;13:e1006968.

39. Knecht ZA, Silbering AF, Ni L, Klein M, Budelli G, Bell R, et al. Distinct combinations of variant ionotropic glutamate receptors mediate thermosensation and hygrosensation in Drosophila. eLife. 2016;5:e17879.

40. Chan C-C, Scoggin S, Wang D, Cherry S, Dembo T, Greenberg B, et al. Systematic Discovery of Rab GTPases with Synaptic Functions in Drosophila. Current Biology. 2011;21:1704–15.

41. Wu C, Singaram V, McKim KS. *mei-38* Is Required for Chromosome Segregation During Meiosis in Drosophila Females. Genetics. 2008;180:61–72.

42. Daenzer JMI, Jumbo-Lucioni PP, Hopson ML, Garza KR, Ryan EL, Fridovich-Keil JL. Acute and long-term outcomes in a *Drosophila melanogaster* model of classic galactosemia occur independently of galactose-1-phosphate accumulation. Dis Model Mech. 2016;9:1375–82.

43. Kimura K, Hachiya T, Koganezawa M, Tazawa T, Yamamoto D. Fruitless and Doublesex Coordinate to Generate Male-Specific Neurons that Can Initiate Courtship. Neuron. 2008;59:759–69.

44. Shalaby NA, Sayed R, Zhang Q, Scoggin S, Eliazer S, Rothenfluh A, et al. Systematic discovery of genetic modulation by Jumonji histone demethylases in Drosophila. Sci Rep. 2017;7:5240.

45. Merino MM, Rhiner C, Lopez-Gay JM, Buechel D, Hauert B, Moreno E. Elimination of Unfit Cells Maintains Tissue Health and Prolongs Lifespan. Cell. 2015;160:461–76.

46. Judd BH. Mutations of zeste that mediate transvection are recessive enhancers of position-effect variegation in Drosophila melanogaster. Genetics. 1995;141:245–53.

47. Yan J, Huen D, Morely T, Johnson G, Gubb D, Roote J, et al. The *multiple-wing-hairs* Gene Encodes a Novel GBD–FH3 Domain-Containing Protein That Functions Both Prior to and After Wing Hair Initiation. Genetics. 2008;180:219–28.

48. Chen Y-W, Song S, Weng R, Verma P, Kugler J-M, Buescher M, et al. Systematic Study of Drosophila MicroRNA Functions Using a Collection of Targeted Knockout Mutations. Developmental Cell. 2014;31:784–800.

49. Phatarphekar A, Su Q, Eun SH, Chen X, Rokita SE. The importance of a halotyrosine dehalogenase for *Drosophila* fertility. J Biol Chem. 2018;293:10314–21.

50. Matsuoka S, Gupta S, Suzuki E, Hiromi Y, Asaoka M. gone early, a Novel Germline Factor, Ensures the Proper Size of the Stem Cell Precursor Pool in the Drosophila Ovary. PLoS ONE. 2014;9:e113423.

51. Birner-Gruenberger R, Bickmeyer I, Lange J, Hehlert P, Hermetter A, Kollroser M, et al. Functional fat body proteomics and gene targeting reveal in vivo functions of Drosophila melanogaster α-Esterase-7. Insect Biochemistry and Molecular Biology. 2012;42:220–9.

52. Nagai R, Hashimoto R, Yamaguchi M. Drosophila Syntrophins are involved in locomotion and regulation of synaptic morphology. Experimental Cell Research. 2010;316:2313–21.

53. Ishida Y, Sekine Y, Oguchi H, Chihara T, Miura M, Ichijo H, et al. Prevention of Apoptosis by Mitochondrial Phosphatase PGAM5 in the Mushroom Body Is Crucial for Heat Shock Resistance in Drosophila melanogaster. PLoS ONE. 2012;7:e30265.

54. Godenschwege TA, Reisch D, Diegelmann S, Eberle K, Funk N, Heisenberg M, et al. Flies lacking all synapsins are unexpectedly healthy but are impaired in complex behaviour. Eur J Neurosci. 2004;20:611–22.

55. Ugrankar R, Bowerman J, Hariri H, Chandra M, Chen K, Bossanyi M-F, et al. Drosophila Snazarus Regulates a Lipid Droplet Population at Plasma Membrane-Droplet Contacts in Adipocytes. Developmental Cell. 2019;50:557-572.e5.

56. Yang Z, Huang R, Fu X, Wang G, Qi W, Mao D, et al. A post-ingestive amino acid sensor promotes food consumption in Drosophila. Cell Res. 2018;28:1013–25.

57. Lamiable O, Kellenberger C, Kemp C, Troxler L, Pelte N, Boutros M, et al. Cytokine Diedel and a viral homologue suppress the IMD pathway in *Drosophila*. Proc Natl Acad Sci USA. 2016;113:698–703.

58. Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan L, Pang N, et al. New genes often acquire male-specific functions but rarely become essential in *Drosophila*. Genes & Development. 2017;31:1841–6.

59. Mendoza-Ortíz MA, Murillo-Maldonado JM, Riesgo-Escovar JR. *aaquetzalli* is required for epithelial cell polarity and neural tissue formation in *Drosophila*. PeerJ. 2018;6:e5042.

60. Rosa-Ferreira C, Sweeney ST, Munro S. The small G protein Arl8 contributes to lysosomal function and long-range axonal transport in *Drosophila*. Biology Open. 2018;7:bio035964.

61. Murray MJ. The Fes/Fer non-receptor tyrosine kinase cooperates with Src42A to regulate dorsal closure

in Drosophila. Development. 2006;133:3063–73.

62. Kopczynski CC, Alton AK, Fechtel K, Kooh PJ, Muskavitch MA. Delta, a Drosophila neurogenic gene, is transcriptionally complex and encodes a protein related to blood coagulation factors and epidermal growth factor of vertebrates. Genes & Development. 1988;2:1723–35.

63. Akdemir F. Autophagy occurs upstream or parallel to the apoptosome during histolytic cell death. Development. 2006;133:1457–65.

64. Pinto S, Quintana DG, Smith P, Mihalek RM, Hou Z-H, Boynton S, et al. latheo Encodes a Subunit of the Origin Recognition Complex and Disrupts Neuronal Proliferation and Adult Olfactory Memory When Mutant. Neuron. 1999;23:45–54.

65. Giansanti MG, Bonaccorsi S, Kurek R, Farkas RM, Dimitri P, Fuller MT, et al. The Class I PITP Giotto Is Required for Drosophila Cytokinesis. Current Biology. 2006;16:195–201.

66. Bülow MH, Wingen C, Senyilmaz D, Gosejacob D, Sociale M, Bauer R, et al. Unbalanced lipolysis results in lipotoxicity and mitochondrial damage in peroxisome-deficient *Pex19* mutants. MBoC. 2018;29:396–407.

67. Zhang S, Ross KD, Seidner GA, Gorman MR, Poon TH, Wang X, et al. Nmf9 Encodes a Highly Conserved Protein Important to Neurological Function in Mice and Flies. PLoS Genet. 2015;11:e1005344.

68. Oron E, Mannervik M, Rencus S, Harari-Steinberg O, Neuman-Silberberg S, Segal D, et al. COP9 signalosome subunits 4 and 5 regulate multiple pleiotropic pathways in Drosophila melanogaster. Development. 2002;129:4399.

69. Chavez VM, Marques G, Delbecque JP, Kobayashi K, Hollingsworth M, Burr J, et al. The Drosophila disembodied gene controls late embryonic morphogenesis and codes for a cytochrome P450 enzyme that regulates embryonic ecdysone levels. Development. 2000;127:4115.

70. Zhao X, Yang H, Liu W, Duan X, Shang W, Xia D, et al. Sec22 Regulates Endoplasmic Reticulum Morphology but Not Autophagy and Is Required for Eye Development in *Drosophila*. J Biol Chem. 2015;290:7943–51.

71. Li K, Baker NE. Regulation of the Drosophila ID protein Extra macrochaetae by proneural dimerization partners. eLife. 2018;7:e33967.

72. Matusek T. The Drosophila formin DAAM regulates the tracheal cuticle pattern through organizing the actin cytoskeleton. Development. 2006;133:957–66.

73. Guo M, Bier E, Jan LY, Jan YN. tramtrack acts downstream of numb to specify distinct daughter cell fates during asymmetric cell divisions in the drosophila PNS. Neuron. 1995;14:913–25.

74. Lisi S, Mazzon I, White K. Diverse Domains of THREAD/DIAP1 Are Required to Inhibit Apoptosis Induced by REAPER and HID in Drosophila. Genetics. 2000;154:669.

75. Soltani-Bejnood M, Thomas SE, Villeneuve L, Schwartz K, Hong C, McKee BD. Role of the *mod(mdg4)* Common Region in Homolog Segregation in Drosophila Male Meiosis. Genetics. 2007;176:161–80.

76. Krüttner S, Stepien B, Noordermeer JN, Mommaas MA, Mechtler K, Dickson BJ, et al. Drosophila CPEB Orb2A Mediates Memory Independent of Its RNA-Binding Domain. Neuron. 2012;76:383–95.

77. Kelley RL. Initial organization of the Drosophila dorsoventral axis depends on an RNA-binding protein encoded by the squid gene. Genes & Development. 1993;7:948–60.

78. Da-Rè C, Franzolin E, Biscontin A, Piazzesi A, Pacchioni B, Gagliani MC, et al. Functional Characterization of d *rim2* , the *Drosophila melanogaster* Homolog of the Yeast Mitochondrial Deoxynucleotide Transporter. J Biol Chem. 2014;289:7448–59.

79. Cherry S, Jin EJ, Özel MN, Lu Z, Agi E, Wang D, et al. Charcot-Marie-Tooth 2B mutations in rab7 cause dosage-dependent neurodegeneration due to partial loss of function. eLife. 2013;2:e01064.

80. Li T, Bender M. A conditional rescue system reveals essential functions for the ecdysone receptor (EcR) gene during molting and metamorphosis in Drosophila. Development. 2000;127:2897–905.

81. Jones G, Teal P, Henrich VC, Krzywonos A, Sapa A, Wozniak M, et al. Ligand binding pocket function of Drosophila USP is necessary for metamorphosis. General and Comparative Endocrinology. 2013;182:73-82.