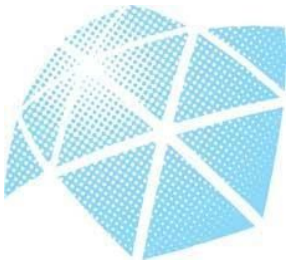
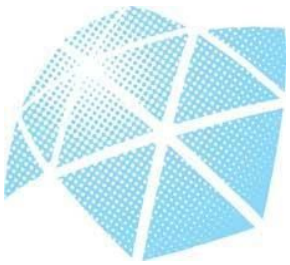


Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências de Botucatu
Programa de Pós-Graduação em Biotecnologia

Predição de rotas metabólicas de enzimas utilizando aprendizado de máquina

Rodrigo de Oliveira Almeida

Botucatu - SP
2018



Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências de Botucatu
Programa de Pós-Graduação em Biotecnologia

Predição de rotas metabólicas de enzimas utilizando aprendizado de máquina

Doutorando: Rodrigo de Oliveira Almeida
Orientador: Dr. Guilherme Targino Valente

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia do Instituto de Biociências de Botucatu da Universidade Estadual Paulista “Júlio de Mesquita Filho”, para obtenção do título de doutor.

Botucatu - SP
2018

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Almeida, Rodrigo de Oliveira.
Predição de rotas metabólicas de enzimas utilizando aprendizado de máquina / Rodrigo de Oliveira Almeida. - Botucatu, 2018

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Guilherme Targino Valente
Capes: 90400003

1. Aprendizado do computador. 2. Bioinformática. 3. Enzimas. 4. Proteínas - Metabolismo.

Palavras-chave: Aprendizado de máquina; Bioinformática; Enzimas; Rotas metabólicas.

"Quanto mais nos elevamos, menores parecemos aos olhos daqueles que não sabem voar".

Friedrich Wilhelm Nietzsche

Dedico este trabalho à minha família e amigos que, mesmo à distância, contribuíram para minha formação pessoal e profissional.

Agradecimentos

Ao Programa de Pós-Graduação em Biotecnologia da Universidade Estadual Paulista, pela oportunidade de atuar no curso de doutorado e suporte aos trabalhos.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro.

Ao meu orientador, Dr. Guilherme Targino Valente, pela amizade e confiança, além das contribuições importantes para o desenvolvimento do trabalho e para minha formação.

Ao Dr. Rafael Plana Simões e Dr. Ney Lemke, pela amizade e auxílio no trabalho.

Ao Dr. Henrik Stotz, da University of Hertfordshire – UK, que me recebeu em seu laboratório para a realização de meu estágio no exterior, além da confiança em meu trabalho.

Aos amigos do SBGL (Systems Biology and Genomic Laboratory – UNESP) e LGI (Laboratório de Genômica Integrativa - UNESP), pelo convívio e aprendizado durante estes anos.

Aos colegas de pós-graduação e professores da UNESP, pela oportunidade de participar e contribuir em suas pesquisas e trabalhos.

Resumo

Enzimas são uma classe de proteínas responsáveis por catalisar diversos tipos de reações químicas presentes em diferentes rotas metabólicas, sendo assim o principal foco de estudo nas áreas de engenharia metabólica e biologia sintética. Contudo, a anotação de enzimas e a identificação da rota metabólica em que atuam, são frequentemente baseados na similaridade de sequências previamente descritas. A falta e dificuldade de anotação das enzimas se devem pela diversidade funcional em sequências similares de famílias proteicas, sequências espécie-específicas e a dificuldade na definição de homologia em larga escala. De modo a auxiliar a superar tais problemas, o presente trabalho objetivou criar um classificador de rotas metabólicas de enzimas baseado inteiramente nas características da estrutura primária de enzimas e utilizando aprendizado de máquina. A ferramenta computacional criada (mAppLe - Metabolic Pathway Prediction of Enzymes) é composta por 11 preditores de rotas metabólicas de fungos, podendo assim auxiliar nas anotações dos bancos de dados e em trabalhos nas diferentes áreas de pesquisa, como biologia sintética e engenharia metabólica. As performances médias de predição foram de 94% de acurácia, 44% de taxa de falsa descoberta, 67% de F-*score*, 98% de sensitividade, 93% de especificidade e 0,69 para coeficiente de correlação de Matthews. Com base no desempenho dos preditores criados, constata-se que a ferramenta computacional criada pode ser aplicada com grande sucesso na predição de rotas metabólicas de enzimas de fungos, independente da similaridade das sequências.

Palavras-chave: Aprendizado de máquina, Enzimas, Rotas metabólicas.

Abstract

Enzymes are a class of proteins that are responsible for catalyzing chemical reactions in numerous metabolic pathways and are often "main targets" in metabolic engineering and synthetic biology. However, enzyme annotation and metabolic pathway identifications are often based on sequence similarities to previously well-described enzymes. Functional diversity in similar sequences of protein families, species-specificity, and difficult-to-define large-scale homologies results in difficulties and a lack of annotation. Here, we present the mAppLe (Metabolic Pathway Prediction of Enzymes), the first metabolic pathway classifier for enzymes based only on primary structure features and a machine learning approach, surpassing limitations imposed by sequence similarities. This tool is composed of 11 pathways predictors for fungi, that can help databank annotations and several type of researches like synthetic biology and metabolic engineering. Results show an average performance of 94% to accuracy, 44% false discovery rate, 67% F-*score*, 98% sensitivity, 93% specificity and 0.69 to Matthews coefficient correlation. Based on the performance of this predictors, the computational tool created (mAppLe) can be applied successfully to predict pathways of enzymes of the fungi, independent of sequence similarity.

key-words: Machine learning, Enzymes, Metabolic pathways.

Sumário

1. INTRODUÇÃO

1.1. Enzimas

Enzimas são as proteínas mais notáveis e altamente especializadas, ponto central nos processos bioquímicos. Catalisam inúmeras reações intra e extracelulares, com alta velocidade e especificidade, degradando macromoléculas para precursores mais simples, e transformando e conservando energia (LODISH et al., 2004 ; NELSON e COX, 2011). Realizam as mais diversas reações bioquímicas, com pH e temperaturas específicas para o correto funcionamento. Suas concentrações e atividades podem ser reguladas, de forma a permitir suas ações dentro das oscilações do meio na qual se encontram. Essas regulações podem ser via inibição por *feedback*, regulação alostérica, fosforilação, compartimentalização, cofatores, entre outros.

O sítio ativo de uma enzima (local onde o substrato se liga para conversão em um produto) contém resíduos de aminoácidos que se ligam ao substrato e agem na substituição de grupos específicos, realizando assim a transformação química (NELSON e COX, 2011). A transformação e conservação da energia acontecem com uma série de reações interconectadas, formando longas rotas que permitem a sobrevivência, crescimento e reprodução celular (ALBERTS, 2015), gerando, contudo, uma extensa rede metabólica (ORTH et al., 2011).

Atualmente, são conhecidos diversos tipos de reações bioquímicas/metabólicas, genes que regulam cada tipo de enzima, e substratos e produtos referentes a uma enzima específica, possibilitando assim calcular o fluxo de metabólitos de uma determinada rota metabólica. Com este tipo de abordagem (análise de fluxo de metabólitos), pode-se reconstruir redes metabólicas e possibilitar a predição da taxa de crescimento de um organismo, ou até mesmo a taxa de produção de um metabólito específico de interesse (ORTH et al., 2011). Um determinado conjunto de enzimas que catalisam reações bioquímicas específicas em um organismo, transformando um composto inicial até ao composto final necessário, é chamado de rota metabólica (PLANES e BEASLEY, 2009).

Sendo assim, uma rota metabólica é uma parte da extensa e complexa rede metabólica (SCHREIBER, 2003).

A atividade biológica de uma enzima é tipicamente determinada por uma parte da cadeia polipeptídica conhecida como domínio (TIAN et al., 2004, NELSON e COX, 2011). Estes domínios são regiões com funções bem definidas e uma proteína pode possuir diferentes domínios em sua estrutura. Embora a função de um domínio seja conservado, ele pode ser alterado por mutações, deleções e inserções, podendo gerar um novo domínio e até mesmo uma nova função (BULJAN e BATEMAN, 2009). Sendo assim, os domínios definem a função a ser exercida pela enzima e seu local de atuação de uma rota bioquímica. Logo, a informação da função enzimática (*EC number*, discutido logo em seguida) das enzimas de um determinado genoma abre a possibilidade de reconstrução de rotas bioquímicas/metabólicas completas, sendo que um único *gap* em alguma rota pode indicar uma anotação equivocada ou um gene ainda não anotado (GINSBURG, 2009).

Extensos estudos ainda são realizados com enzimas na busca de uma classificação funcional (FREILICH et al., 2005), bem como suas participações em uma ou mais rotas metabólicas e conservação da sequência (PEREGRIN-ALVAREZ et al., 2003). Analisando a distribuição filogenética de enzimas de *Escherichia coli*, Peregrin-Alvarez et al. (2003) relatam que, embora as enzimas sejam amplamente distribuídas e altamente conservadas durante a evolução, sua participação nas rotas metabólicas podem variar significativamente.

Contudo, em 1956, o presidente da União Internacional de Bioquímica estabeleceu uma Comissão Internacional sobre Enzimas com o objetivo de resolver problemas relacionados à classificação e nomenclatura das mesmas. Foi então aprovado em 1961 um relatório com unidades, símbolos e nomenclatura para enzimas, no qual cada classe de enzima é subdividida e cada enzima contém um código único de quatro dígitos, chamado de "*Enzyme Commission number*" ou "*EC number*" (TIPTON e BOYCE, 2000).

A respeito do *EC number*, sendo as enzimas classificadas de acordo com o tipo de reação que realizam, o primeiro dígito (classe) define o tipo de reação geral catalisada, com valores de 1 a 6, constituindo as reações de oxidoredutase, transferases, hidrolases, liases,

isomerases e ligases, respectivamente. O segundo e terceiro dígitos indicam a subclasse e sub-subclasse, respectivamente; nessas subclasses e sub-subclasses, geralmente, encontra-se diversas especificações como grupos químicos de atuação da enzima e o produto a ser formado, respectivamente. O quarto dígito é um número identificador da enzima dentro de uma determinada sub-subclasse.

Atribuir um código *EC number* a uma enzima está longe de ser uma tarefa trivial, tanto computacionalmente quanto experimentalmente. Por vias computacionais, realiza-se uma análise de similaridade das sequências de aminoácidos da enzima desconhecida contra um banco de dados. Ao obter uma alta taxa de similaridade com alguma sequência deste banco de dados, as anotações da função enzimática (*EC number*) desta sequência serão transferidas para a nova enzima analisada (sendo este procedimento discutido mais à frente, na seção 1.3). Por vias experimentais (moroso e muito mais complexo, porém com maior exatidão), a enzima deve ser purificada e uma série de análises bioquímicas devem ser realizadas (determinação do pH e da temperatura de maior atividade, necessidade de cofatores, velocidade da reação, entre outros) a fim de definir o tipo de substrato (ou substratos) que tal enzima consegue degradar além de determinar o produto formado. Vários bancos de dados como KEGG (KANEHISA et al., 2002), BRENDA (SCHOMBURG et al., 2002), ExplorEnz (McDONALD et al., 2009), Uniprot (The UniProt Consortium) e EcoCyc (KARP et al., 2000) fornecem informações sobre rota metabólica e a reação enzimática nos processos celulares. Tais informações são primordiais para desenvolvimento de novos produtos biotecnológicos e para pesquisas mais detalhadas, como análise de fluxo de metabólitos, engenharia metabólica e biologia sintética.

Ressalta-se que uma boa anotação do genoma de uma espécie é de extrema importância para toda a comunidade científica, pois permite realizar e fundamentar pesquisas na área computacional (possibilitando a melhoria de sistemas de predição), assim como pesquisas com foco em novos produtos biotecnológicos (engenharia metabólica e biologia sintética). Enzimas bem anotadas formam um grupo funcional ideal para estudos de mudanças fenotípicas e divergência/redundância funcional nas espécies (FREILICH et al., 2005).

Além disso, estudos em evolução de enzimas podem fornecer diversas contribuições para um melhor entendimento da biologia, tais como um melhor entendimento das funções enzimáticas, identificação de enzimas com funções ainda desconhecidas, caracterização de novos domínios, famílias e superfamílias de proteínas, e melhorar os métodos de anotação de genomas (GLASNER et al., 2006). Como relatado por Freilich et al. (2005), “espécies mais complexas” contém menor quantidade de enzimas, indicando uma maior redundância funcional, entretanto, possuem uma maior quantidade de enzimas envolvidas em processos de sinalização de degradação.

1.2. Engenharia metabólica e biologia sintética: a importância de determinar uma rota metabólica

Enormes avanços biotecnológicos têm sido realizados desde 1980, tais como melhorias nos sequenciadores de ácidos nucleicos e métodos de quantificar e identificar proteínas, descoberta e disponibilização de novas enzimas de restrição e modificação de plasmídeos para expressão de genes. Isso vem permitindo o desenvolvimento de novos e sustentáveis bioprocessos de produção de combustíveis, compostos químicos e materiais diversos (NIELSEN et al., 2014), levando assim ao surgimento de duas importantes áreas de estudo: a engenharia metabólica e a biologia sintética.

A engenharia metabólica pode ser definida como o desenvolvimento de métodos e conceitos para análises e modificações de redes metabólicas, geralmente com o objetivo de encontrar alvos para projetar biofábricas de modo a aumentar a produção de um determinado bioproduto (NIELSEN et al., 2014, STEPHANOPOULOS et al., 2012). Já a biologia sintética pode ser definida como o desenho e construção de novos sistemas biológicos ainda não existentes na natureza, com foco em uma montagem bem caracterizada, padronizada e com aproveitamento dos componentes, realizando modularização e ajustes na expressão e na estrutura dos genes (NIELSEN et al., 2014, STEPHANOPOULOS et al., 2012).

Tais áreas de estudo (engenharia metabólica e biologia sintética) são complementares, pois se auxiliam na busca por biofábricas mais eficientes para produção de compostos químicos especiais, combustíveis e materiais renováveis diversos, além de aplicações na área alimentar e farmacêutica (NIELSEN et al., 2014, WU et al., 2016). Diferenças e atuação em comum destas duas áreas são mostradas na tabela 1.

Tabela 1. Diferenças e sobreposição de atuação entre engenharia metabólica e biologia sintética.

	Engenharia metabólica	Biologia sintética
Área de atuação comum	rotas metabólicas sintéticas	
Ferramentas em comum	biologia molecular, modelagem matemática, bioinformática, modelagem molecular	
Área de domínio	química, enzimas, genes, rotas metabólicas, redes biológicas, células	genes, circuitos de expressão, células
Objetivo	engenharia de fenótipo celular	engenharia de partes biológicas
Ferramentas especializadas	quantificação de fluxo e análises de rede	síntese de componentes biológicos
Aplicação	indústria biotecnológica	diversas

Fonte: Adaptado de Nielsen et al., 2014, p. 321.

Diferentes plataformas foram desenvolvidas para desenhar rotas metabólicas com reações enzimáticas múltiplas *in vitro*, como a biotransformação por rotas enzimáticas sintéticas, reações em cascata minimizada e sistema bioquímico sintético (TANIGUCHI et al., 2017). Biotransformação (ou biocatálise) refere-se a enzimas (ou microorganismos) que aceleram reações bioquímicas, enquanto rota sintética é a montagem de uma rota metabólica baseada em uma rota metabólica natural, porém com devidas modificações

(ZHANG et al., 2010). A nova rota bioquímica a ser montada precisa ser cuidadosamente planejada, pois uma determinada reação bioquímica pode ser realizada por diferentes rotas. Com isso, se faz necessário considerar diversos fatores que podem influenciar esta nova rota, como o balanço de ATP/NADP, equilíbrio de reação, termodinâmica e separação de produtos. Em seguida, seleciona-se as enzimas com base em suas reações catalíticas e suas especificidades com substratos, de preferência enzimas com alta seletividade de substrato, alta eficiência catalítica, baixa inibição pelo produto e alta estabilidade. Entretanto, algumas enzimas devem ser produzidas com algumas modificações, de modo a aumentar sua eficiência para aplicações industriais. Por último, a parte de engenharia de processos deve ser realizada para definir os meios necessários para a imobilização de enzimas, estabilização de cofatores, separação do produto *in situ* e montagem do específico biorreator, assim como a reutilização de enzimas, reciclagem de cofatores e remoção do produto *in situ* para evitar a inibição das reações por *feedback* (ZHANG et al., 2010). Em resumo, a biotransformação por rotas enzimáticas sintéticas consiste na montagem de um biorreator, utilizando um conjunto de enzimas para realizar as reações bioquímicas planejadas para a obtenção de um produto desejado, baseado em uma rota metabólica modificada, seleção de enzimas, produção e modificações de algumas enzimas especiais (ZHANG et al., 2010).

Reações em cascatas minimizadas consiste em utilizar biocatalisadores (enzimas) purificados, de modo a superar barreiras associadas aos mecanismos moleculares da célula, como restrição para acessar o substrato, produção de metabólito tóxico e redirecionamento de rota metabólica induzida por substrato. Essa abordagem também apresenta maior facilidade para ser empregado na indústria. Entretanto, este sistema precisa de adicionar uma concentração equilibrada de cofatores, assim como pode ser restringido por limitações termodinâmicas e variação de performance das enzimas ao longo do tempo. Sendo assim, reações em cascata minimizadas consiste na utilização de enzimas (purificadas) necessária para realizar somente as reações bioquímicas desejadas, de modo a evitar ou eliminar rotas metabólicas não desejáveis (presentes nos mecanismos moleculares de uma célula como rotas alternativas) e obter o produto final desejado (GUTERL et al., 2012).

Semelhante ao sistema de reações em cascatas minimizadas, a bioquímica sintética se diferencia pelo fato de que o sistema enzimático deve ser projetado de forma que consiga se auto sustentar por longo tempo. Isso significa que durante o processo de biocatálise, não há necessidade de adição extra de cofatores e/ou enzimas, pois estes se encontram em equilíbrio no sistema (KORMAN et al., 2017).

Atualmente, a maior parte das pesquisas em engenharia metabólica trabalham com biocatalizadores purificados *in vitro*, uma vez que barreiras associadas à célula (como substrato, toxicidade de produto, redirecionamento de metabolismo induzido por substrato) podem ser eliminados (GUTERL et al., 2012). As principais diferenças entre engenharia metabólica *in vitro* e a convencional (por fermentação) são mostradas na tabela 2.

Tabela 2. Comparação entre engenharia metabólica convencional e *in vitro*.

	Convencional (Fermentação)	*In vitro*
Controle	- níveis de expressão ajustados por modificação genética. - regulação do crescimento celular ajustando a disponibilidade de nutrientes e oxigênio.	- concentração enzimática pode ser ajustada precisamente. - regulação simples para as reações enzimáticas.
Desenho da rota	- limitação do produto metabólico. - efeitos colaterais, devido à modificação da rota, são difíceis de serem preditos. - otimização das rotas podem ser adquirida através da evolução e seleção dos organismos.	- rotas podem ser desenhadas com alta flexibilidade. - possibilidade de desenhar rotas artificiais. - possibilidade de produção de compostos citotóxicos.
Custo de produção	- os microorganismos se reproduzem (consequentemente, produzindo enzimas e cofatores). - Alto custo na purificação do produto final, quando a concentração deste é baixa.	- necessidade de preparo de enzimas. - necessário a adição de cofatores. - purificação do produto final de forma fácil e com baixo custo.
Produção	- altas produções podem ser alcançadas com fornecimento suficiente de substrato. - substrato também é utilizado no crescimento celular.	- necessidade da regeneração dos cofatores para total conversão do substrato. - substrato é totalmente convertido em produto, sem formação de co-produto.
Permeação da membrana	- necessário.	- não necessário.
Implementação em escala comercial	- já implementado e com vários exemplos.	- ainda em fase de estudo.

Fonte: adaptado de Taniguchi et al., 2017, p. 66

A prospecção de enzimas (biocatalisadores) com alto potencial de aplicação na área industrial, tem estabelecido vários empreendimentos biotecnológicos e ligando diversos

estudos à processos industriais (FILHO, 2011). A topologia e diversificação funcional de rotas bioquímicas ainda vem sendo exploradas, uma vez que a maioria das informações disponíveis advindas de forma experimental, são restritas a pouco modelos de espécies (PEREGRIN-ALVAREz et al., 2003). As espécies mais estudadas e com maior quantidade e qualidade de informações são *Arabidopsis thaliana*, *Bos taurus*, *Caenorhabditis elegans*, *Chlamydomonas reinhardtii*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Escherichia coli*, *Hepatitis C virus*, *Homo sapiens*, *Mus musculus*, *Mycoplasma pneumoniae*, *Oryza sativa*, *Plasmodium falciparum*, *Pneumocystis carinii*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Takifugu rubripes*, *Xenopus laevis* e *Zea mays*. Destas, somente *Pneumocystis carinii*, *Saccharomyces cerevisiae* e *Schizosaccharomyces pombe* representam o reino fungi, foco deste trabalho. Logo, é extremamente comum a falta deste tipo de informação (rotas bioquímicas e metabólicas) nos diversos proteomas de fungos presentes nos bancos de dados.

1.3. Anotações nos bancos de dados

Com o surgimento dos sequenciadores de nova geração (NGS – *Next Generation Sequencing*), grande quantidade de informações vem sendo depositadas nos mais diversos bancos de dados públicos, tais como NCBI, EMBL e UniProt. Apesar destes dados poderem ter potencial uso em pesquisas biológicas e médicas, a caracterização experimental foi realizada em apenas uma pequena parte deste grande conjunto de sequências (DNA, RNA e proteínas) disponíveis (SCHNOES et al., 2009).

Ferramentas de bioinformática utilizadas para anotações de sequências, em geral, são baseadas na busca por similaridade de sequências em um determinado banco de dados, constituindo assim uma transferência de anotação baseada em similaridade (FRIEDBERG, 2006). Logo, se duas sequências têm alto grau de similaridade, então muito provavelmente têm um ancestral comum e, consequentemente, a mesma função (ROST, 2002). Entretanto, esta abordagem pode não ser tão confiável para anotações funcionais, mesmo obtendo altas porcentagens de identidade nos alinhamentos das sequências (FRIEDBERG, 2006). Como

exemplo, para enzimas, menos de 30% das sequências com taxa de similaridade maior que 50% realmente compartilham a mesma função enzimática (ROST, 2002).

Além disso, erros contidos em anotações podem se acumular em diferentes estágios, do sequenciamento à anotação de rotas metabólicas, podendo se propagar para novos genomas sequenciados (POPTSOVA e GOGARTEN, 2010). Logo, anotações de rotas metabólicas podem ser transferidas erroneamente para algumas sequências de uma determinada espécie, podendo causar falhas no processo de biotransformação (referente aos sistemas anteriormente citados) utilizado na área de engenharia metabólica e biologia sintética, assim como ser uma fonte causadora de viés em uma análise computacional.

Investigando erros de anotação em bancos de dados mais utilizados, como NCBI, UniProtKB/TrEMBL, UniProtKB/Swiss-Prot e KEGG, Schnoes et al. (2009) compararam as anotações de famílias de enzimas destes bancos de dados com informações de famílias de enzimas curadas e bem caracterizadas (sequências, estrutura, mecanismos de ação) proveniente do banco de dados Structure-Function Linkage Database (SFLD) e da literatura. Eles demonstraram altos níveis de anotações equivocadas nessas bases de dados, com exceção do UniProtKB/Swiss-Prot.

Devido ao aumento de novas e diferentes sequências proteicas, a transferência de anotação por similaridade pode se tornar menos efetiva, pois como o banco de dados não terá um cluster de proteínas para uma devida comparação, um cluster mais próximo será utilizado, comprometendo assim a anotação desta nova proteína. Além disso, ao se realizar esta análise e inferir tal anotação como verdadeira (devido a um alto grau de similaridade), corre-se o risco de aumentar a propagação destas anotações equivocadas (FRIEDBERG, 2006).

1.4. Bioinformática aplicada a estudo de enzimas

Atualmente, vários bancos de dados biológicos disponibilizam anotações genômicas de uma ampla variedade de organismos. Contudo, a curadoria manual dessas anotações

estão disponíveis apenas para organismos modelos bem investigados (QUESTER e SCHOMBURG, 2011).

Acuradas anotações funcionais geradas computacionalmente para macromoléculas biológicas permitem aos pesquisadores identificar determinadas proteínas e suas rotas metabólicas. O aumento da acurácia dessas anotações funcionais via bioinformática está atrelado ao aumento das anotações realizadas após comprovação experimental e melhoria dos métodos e técnicas de predição de funções. Entretanto, a predição da função de uma proteína não é nada trivial e, ainda, não há clareza sobre quais as melhores ferramentas de predição (JIANG et al., 2016). Neste contexto, diversos programas de bioinformática têm sido desenvolvidos para predição de funções de enzimas, no qual a maioria utiliza análise de similaridade de sequências, tais como abaixo exemplificado.

O programa EFFICAZ (TIAN et al., 2004) foi criado para predizer funções enzimáticas combinando diferentes tipos de análises (detecção de funcionalidade de resíduos, informações sobre os domínios do banco de dados Pfam e PROSITE). Posteriormente, foi incorporado ao programa mais duas análises utilizando *Support Vector Machine* (aprendizado de máquina) sobre informações do banco de dados Pfam (ARAKAKI et al., 2009; KUMAR and SKOLNICK, 2012). Utilizando sequências genômicas, EFFICAZ conseguiu predizer 8.886 funções enzimáticas (*EC numbers* com 4 e 3 dígitos) dentre 50.475 sequências (traduzidas pelo TrEMBL) do proteoma humano. Entretanto, este programa não fornece a informação da rota metabólica que a enzima classificada atua.

De modo a predizer funções enzimáticas, o PRIAM (CLAUDEL-RENARD et al., 2003) baseia-se em uma coleção de enzimas específicas provenientes do banco de dados ENZYME, caracterizando os módulos de cada coleção, criando conjunto de regras baseado nos módulos para inferir uma enzima e, por fim, obtendo uma matriz de pontuação de posição específica de cada módulo. Testado no genoma de *Sinorhizobium meliloti*, foram preditas 1.460 enzimas em 6.204 proteínas. Assim como o programa EFFICAZ, o PRIAM não fornece a informação sobre a rota metabólica de atuação da enzima classificada.

A ferramenta *on line* Pathway Analyst (PIREDDU et al., 2006), disponibiliza 10 modelos preditores de rotas bioquímicas, fornecendo também a informação sobre a função enzimática (*EC number*). Ele utiliza ferramentas como o BLAST, HMMs (*hidden Markov models*) e *Support Vector Machine*. Utiliza também informações de reações enzimáticas e alinhamento de sequências, de modo a inferir a rota bioquímica e a função enzimática. Os testes utilizaram 125 rotas de um total de 1.759 reações, no qual o preditor mais acurado alcançou a média de precisão de 78,3%. Contudo, o programa não fornece a informação da rota metabólica que a enzima classificada atua. Somando a isso, a ferramenta (que deveria ser acessada pelo link http://path-a.cs.ualberta.ca) não está mais disponível.

Outra ferramenta *on line*, o ComPath (CHOI e KIM, 2008), utiliza primariamente informações do banco de dados KEGG (PATHWAY, GENES, LIGAND e BRITE) tais como rotas, sequências, compostos/reações e classificação funcional. Informações sobre *motif* de domínio estrutural são fornecidos pelos bancos de dados Pfam, PROSITE, SCOP, SCOPEC, SUPERFAMILY e PDB. Informações pertinentes à sequência genômica foram adquiridas nos bancos de dados Swiss-Prot and KEGG/GENES. Com a integração dessas informações, o ComPath verifica a predição dos componentes das rotas bioquímicas utilizando ferramentas de análise de sequências, *motif* e filogenia. Assim como o programa Pathway Analyst, o ComPath não fornece a informação da rota metabólica que a enzima classificada atua, além de não mais estar disponível na *world wide web* (que deveria ser acessado pelo link http://www.compath.org/).

Criado para realizar modelagem e predição de estrutura de proteínas e sua função, o I-TASSER (YANG et al., 2014) se baseia na simulação iterativa de montagem de estruturas, seleção de modelo, refinamento do modelo e anotação funcional baseado em estrutura. Além disso, utiliza algoritmos complementares para melhorar a inferência de função baseado em *Support Vector Machine*. Como os demais programas citados, o I-TASSER não fornece informação da rota metabólica da enzima classificada.

Utilizando atributos calculados baseados em estrutura secundária, propensão de aminoácidos e propriedades de ligantes e de superfície, Dobson e Doig (2004) combinaram predições feitas por *multi-class Support Vector Machine* (aprendizado de máquina) e

conseguiram até 60% de acurácia na predição de funções enzimáticas (*EC number*). Apesar desta abordagem não fornecer também a informação da rota metabólica da enzima classificada, difere dos demais programas por não utilizar como base uma análise de similaridade das sequências.

Como observado, a maioria dos programas se baseiam em análise de similaridade de sequências, com o objetivo de inferir a classificação funcional da enzima (*EC number*) e/ou fornecer informação sobre a rota bioquímica (não a rota metabólica) de atuação desta enzima classificada. Entretanto, como abordado na seção anterior, tal procedimento pode levar a anotações equivocadas e consequentemente propagar o erro de anotação para futuras novas sequências proteicas. Logo, surge a necessidade de outra abordagem computacional que não utilize análise de similaridade de sequências e que forneça a informação (predição) em qual rota metabólica uma enzima (ainda com suas funções enzimáticas desconhecidas) está atuando.

1.5. Aprendizado de máquina

A cada ano, a quantidade de dados biológicos disponíveis vem crescendo de forma exponencial, dificultando a extração de informações. Sendo este o maior e atual desafio da biologia computacional, o desenvolvimento de ferramentas e métodos capazes de explorar estes dados se tornam cada vez mais necessários e, nesse contexto, o aprendizado de máquina vem se destacando ao longo dos anos (LARRAÑAGA et al., 2006).

Aprendizado de máquina consiste em programar computadores para otimizar o critério de performance de extração de padrões usando dados de exemplo e informações prévias. Isso pode possibilitar a criação de um modelo preditivo, com experimentos relativamente rápidos, os quais podem resultar na sugestão de experimentos mais promissores a serem executados em laboratório (LARRAÑAGA et al., 2006; FABRIS et al., 2017).

Há diversas áreas biológicas onde técnicas de aprendizado de máquina vêm sendo aplicadas para extração de padrões e informações, como no melhoramento genético (HECKMANN et al., 2017), farmacêutica (AZUAJE, 2016), genômica (localização e estrutura

de genes, elementos regulatórios e genes não codificantes de RNA), proteômica (predição de estrutura e de função) (MALHIS et al., 2015), microarrays (identificação de padrão de expressão/classificação de redes genéticas), biologia de sistemas (modelagem de redes genéticas, redes de sinais de transdução e de rotas metabólicas) (KANDOI et al., 2015), evolução (reconstrução de árvores filogenéticas) e mineração de texto (anotações de genes e proteínas) (LARRAÑAGA et al., 2006).

Os dados utilizados no aprendizado de máquina devem ser primeiramente estruturados, formando assim um conjunto de dados de treinamento e de teste. Estes conjuntos contém instâncias, geralmente representadas por um conjunto de tamanho fixo e com variáveis numéricas ou nominais (características associadas a cada instância), os quais são chamados de atributos (FABRIS et al., 2017). Além disso, o aprendizado de máquina pode ser dividido em aprendizado supervisionado, não supervisionado e semi-supervisionado.

O método supervisionado utiliza dados rotulados (por exemplo, positivo e negativo) vinculados a cada instância, os quais são utilizados para a construção de modelos capazes de realizar predições em dados não rotulados. São utilizados para predições de dados contínuos (regressão) ou discretos (classificação). No método não supervisionado não há utilização de dados rotulados, permitindo identificar potenciais elementos no conjunto de dados; são requeridos passos adicionais no processo de rotulagem ao final do processo. O aprendizado não supervisionado pode também ser utilizado para reduzir a dimensão dos dados formando assim *clusters* com base nos atributos analisados. O método semi-supervisionado é um método intermediário, no qual o algoritmo realiza o treinamento com um conjunto de dados rotulados (conjunto menor) para gerar um modelo capaz de rotular os dados de um conjunto não rotulado (conjunto maior). Em seguida, os novos dados rotulados são adicionados iterativamente ao conjunto de treinamento, melhorando assim o modelo a cada ciclo realizado (LIBBRECHT e NOBLE, 2015). Logo, no caso de explorar, classificar e predizer rotas metabólicas que uma enzima pode atuar, o aprendizado supervisionado é a melhor opção, pois todos os atributos desse conjunto de dados estarão vinculados a esses rótulos (neste trabalho as rotas metabólica), realizando assim a predição

se uma enzima atua ou não (neste trabalho é uma classificação positiva ou negativa) nessas determinadas rotas metabólicas.

É de suma importância realizar uma seleção *a priori* dos atributos (características vinculadas a cada instância) dos dados a serem utilizados em qualquer um dos métodos de aprendizado de máquina. O objetivo dessa etapa é reduzir o conjunto de atributos iniciais para àqueles que possibilitem uma melhor performance do classificador em termos de acurácia (VARSHAVSKY et al., 2006; LIBBRECHT e NOBLE, 2015). Além disso, a seleção de atributos são importantes para prevenir *overfitting* e fornecer modelos rápidos e de maior custo efetivo computacional (GUYON e ELISSEEFF, 2003; SAEYS et al., 2007). O termo *overfiting*, muito utilizado na área estatística, descreve uma situação no qual um modelo gerado se ajusta somente na base de dados que gerou o modelo, obtendo baixa performance quando aplicado em uma base de dados desconhecida.

Além da seleção dos atributos, deve-se também filtrar e ajustar devidamente as instâncias a serem utilizadas, removendo dados redundantes, retirando *outliers* (dados muito discrepantes dos demais), normalizando os dados (manter proporcionalidade), aleatorizando as instâncias (para prevenir a inserção de viés), balanceando os dados (técnica conhecida como *undersampling*) e separação das instâncias (conjunto de treinamento, teste e validação). Outro fator extremamente importante é a definição da quantidade de instâncias a ser utilizada, pois caso tenha poucas instâncias, o modelo gerado poderá ter baixa performance. No entanto, instâncias em demasia levam a um maior custo computacional (podendo levar semanas) para gerar um modelo sem a garantia de ter uma alta performance preditiva.

Os procedimentos acima mencionados são necessários para evitar algum tipo de viés no conjunto de dados, aumentar a performance de treinamento do algoritmo de aprendizagem, e consequentemente gerar modelos de predição de ampla capacidade preditiva (modelo generalista).

Diversos tipos de algoritmos de aprendizado de máquina têm sido utilizados para distinguir propriedades específicas de duas ou mais classes funcionais. Vários algoritmos

utilizados são os baseados em SVM (*Support Vector Machine*), árvores de decisão, regressão logística, redes Bayesianas, combinação de classificadores, entre outros (KANDOI et al., 2015).

Os algoritmos baseados em SVM são conjuntos de modelos que mapeiam os dados para, em seguida, construir um hiperplano (separando assim os dados) para ser utilizado na classificação, ou até mesmo selecionar hiperplanos em uma alta dimensão ou espaço dimensional infinito (KANDOI et al., 2015; KARIMPOUR-FARD et al., 2015). Como vantagens, o SVM consegue lidar bem com grandes conjuntos de dados, alta dimensionalidade e com processo rápido de classificação. Por outro lado, a definição do *kernel* precisa ser bem definida e o tempo de treinamento pode ser longo.

As árvores de decisão são algoritmos mais simples mas com grande poder de classificação. São muito utilizados para a construção de modelos preditivos para classificação de dados, sendo muito popular o algoritmo Random Forest, pela habilidade de construir classificadores robustos e por selecionar variáveis discriminantes (KANDOI et al., 2015; KARIMPOUR-FARD et al., 2015). A utilização de algoritmos do tipo árvore de decisão oferece a vantagem de facilidade de interpretação dos dados e na construção das regras de classificação, tendo como desvantagem a possibilidade de gerar árvores extremamente grandes.

Os algoritmos de Redes Neurais Artificiais (algoritmos que simulam funções cerebrais) são uma poderosa ferramenta de modelagem e vem sendo amplamente utilizadas. Baseado na topologia estrutural, essa abordagem pode ser dividida em redes *forward*, *backward*, *random* e *self-organized*, sendo a rede neural *back propagation* (rede do tipo *foward*) uma das mais populares e utilizadas (ZHANG et al., 2017). Redes Neurais Profundas (também conhecida como *Deep learning*) utiliza dados brutos (como entrada de dados) na menor camada e transforma esses dados em representações abstratas por combinação sucessiva dos dados de saída de cada camada anterior, gerenciando esses dados e encapsulando funções extremamente complexas durante o processo. O *deep learning* vem sendo atualmente o campo mais ativo da área de aprendizado de máquina e mostrando melhorias nas performances em biologia computacional (ANGERMUELLER et al., 2016). A vantagem

das redes neurais são a flexibilidade em trabalhar com os mais diversos tipo de dados. Quanto às desvantagens, destacam-se a difícil determinação do número de camadas e neurônios a serem utilizados, e os modelos gerados são extremamente complexos. Além disso, exige maior consumo de *hardware* (processadores e memória RAM) e tempo de máquina.

Independente do tipo de algoritmo utilizado para gerar o modelo de predição, uma importante etapa é a avaliação da performance preditiva, principalmente na aplicação do modelo no conjunto de teste e/ou de validação, pois estes são conjuntos de dados desbalanceados (geralmente com a classe negativa superior à quantidade de classe positiva) e desconhecidos para o modelo. Nesse contexto, várias métricas podem ser calculadas tais como acurácia, sensitividade (*recall*), especificidade, precisão, F-*score*, ROC (*Receiver Operating Characteristic*), taxa de falsa descoberta, coeficiente de correlação de Matthews, entre outras. As mais utilizadas são sensitividade, especificidade, ROC e acurácia. Entretanto, utilizar somente uma ou duas métricas para avaliar o desempenho do modelo nos dados "desconhecidos" (conjunto de teste e validação) pode levar a uma conclusão equivocada dos resultados. Por exemplo, um modelo cuja a performance apresenta alta acurácia, pode ter um baixo valor de sensitividade e um alto valor de especificidade.

2. OBJETIVOS

2.1. Geral

Criar um classificador/preditor de rotas metabólicas de enzimas baseado em aprendizagem de máquina, utilizando unicamente sequências de aminoácidos (características da estrutura primária) de enzimas de fungos.

2.1. Específicos

- Gerar e selecionar atributos específicos para cada rota metabólica.

- Gerar modelos específicos para cada rota metabólica.
- Criar ferramenta computacional que utiliza os modelos gerados para aplicar em dados desconhecidos (enzimas sem anotação de rota metabólica).

3. JUSTIFICATIVA

Devido à diversidade funcional em sequências similares de famílias proteicas e sequências espécie-específicas, a anotação de enzimas e a identificação da rota metabólica em que atuam se torna um trabalho extremamente difícil, pois tais tarefas são baseadas frequentemente na similaridade de sequências previamente descritas nos bancos de dados.

4. MATERIAL E MÉTODOS

A ferramenta computacional mAppLe (*Metabolic Pathway Prediction of Enzymes*) foi desenvolvida utilizando a linguagem de programação R (R Core Team, 2017) e o software Weka-3.9.1 (Hall et al., 2009). Foi implementado o conceito de aprendizado de máquina supervisionado, no qual as enzimas de uma determinada rota metabólica são identificadas como sendo da classe "positivo" e todas as outras como classe "negativo". Em outras palavras, quando as enzimas de uma determinada rota metabólica estão em um conjunto de treinamento/teste ou validação, são instâncias “positivo”, e todas as outras enzimas de diferentes rotas metabólicas são marcadas como instâncias “negativo”. Já as características da estrutura primária destas enzimas são assumidas como atributos dessas instâncias.
Todas as análises foram feitas de modo independente, gerando modelos específicos para cada metabolismo.

4.1. Seleção das espécies, instâncias e rotas metabólicas

Foram selecionadas 604 proteomas de fungos (e suas diferentes cepas) presentes no banco de dados Uniprot para compor o banco de dados inicial

(ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/) (Material Suplementar T1). O *download* das proteínas sequenciadas das diferentes espécies, em formato de tabelas, contêm as informações "*Entry*", "*Organism*", "*EC number*" , "*Pathway*", "*Annotation score*" e "*Sequence*".

Um processo de filtragem foi realizado de modo a selecionar somente as devidas sequências de aminoácidos pertencentes à enzimas, para assim formar o conjunto de dados de treinamento e validação. Sendo assim, foram utilizados as seguintes informações:

- *EC numbe*r: Conforme descrito anteriormente, esta informação remete à função enzimática exercida, podendo uma enzima atuar em mais de um tipo de reação bioquímica (pode assim conter um ou mais códigos *EC number*). Este código pode ser completo (4 dígitos) ou incompleto (um a três dígitos). Logo, somente as sequências contendo apenas um código *EC number*, completo ou não, foram selecionadas.
- *Annotation Score*: Esta informação fornecida pelo banco de dados Uniprot refere-se o quão estudada, explorada e/ou citada é uma determinada sequência. Em uma escala de 1 a 5, sequências pouco estudadas recebem nota 1, enquanto as sequências mais estudadas recebem nota 5. No presente trabalho, esta informação foi utilizada como um fator de "qualidade" da informação, utilizando no mínimo um valor igual a 2.
- *Pathway*: Este campo contém a anotação da rota metabólica de atuação da enzima. Podendo uma enzima atuar em mais de uma rota metabólica, este campo pode conter várias anotações de rotas metabólicas (em ordem crescente de detalhamento). A primeira anotação, remete a uma informação simplificada da rota metabólica, enquanto as demais anotações fornecem informações mais detalhadas de onde e como estas enzimas atuam na referida rota metabólica. No entanto, as anotações poderiam gerar duplicação dos dados e, sendo assim, neste trabalho utilizou-se somente a primeira anotação.
- Aminoácidos especiais: Algumas enzimas contém aminoácidos diferente dos 20 aminoácidos usuais. Devido ao R *package protr* trabalhar somente com os 20 aminoácidos usuais, enzimas que continham aminoácidos especiais foram removidas do conjunto de dados.
- Sequências redundantes: De modo a criar um conjunto de dados com menor redundância nas informações contidas e consequentemente evitar algum tipo de viés na geração dos modelos de predição, sequências com similaridade ≥99% foram removidas utilizando o programa CD-Hit (LI e GODZIK, 2006).
- Pontuação de corte: De modo a selecionar somente as espécies que contenham maiores quantidades de informações, espécies que continham menos de 50 sequências foram removidas, não participando assim da base de dados que deverá formar os conjuntos de treinamento/teste e validação.

Após esse processamento, as espécies que continham 50 ou mais sequências foram selecionadas para compor a base de dados deste trabalho (Figura 1).

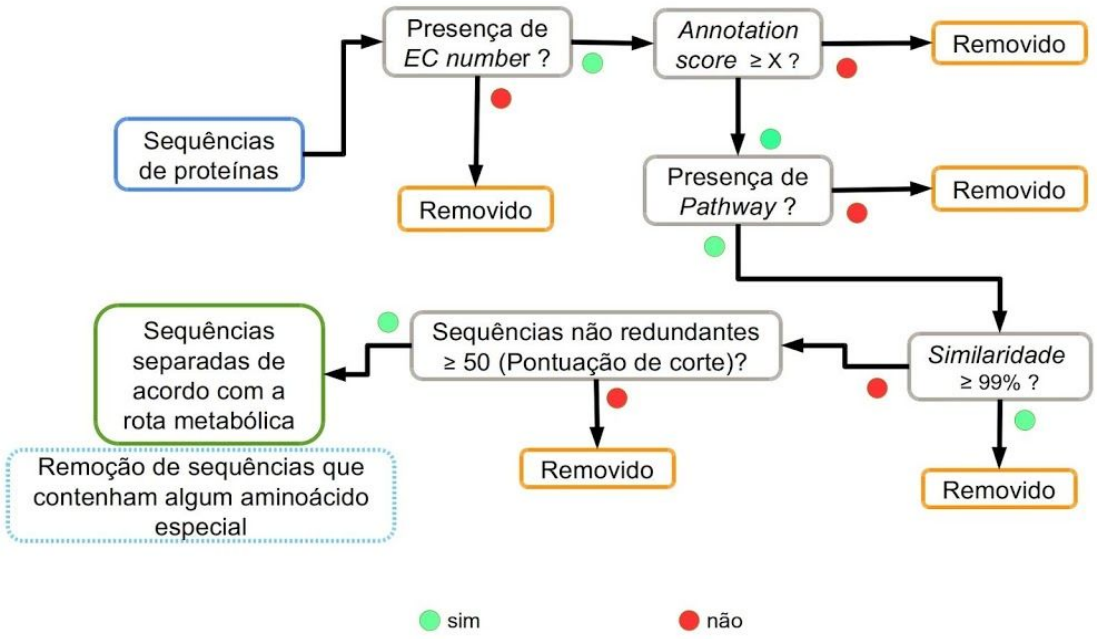


Figura 1. Processo de seleção das instâncias e suas respectivas rotas metabólicas para compor os conjuntos de treinamento/teste e validação, de acordo com o valor do parâmetro *Annotation score* ($2 \leq X < 5$).

Com base somente nas espécies selecionadas, verificou-se quais rotas metabólicas estavam presentes no campo *Pathway*, fazendo destas as rotas metabólicas a serem estudadas.

Inicialmente contando com 604 organismos (espécies de fungos e suas diferentes cepas), aplicou-se todo o processo de filtragem, utilizando o parâmetro *Annotation score* igual a um, para utilizar somente as espécies/cepas que continham mais de 50 enzimas não redundantes. Apenas 178 espécies/cepas foram selecionadas (Material Suplementar T1). Com esse procedimento, houve uma remoção de aproximadamente ⅓ dos dados (de

3.345.198 para 2.220.779 sequências), diminuindo assim o tempo necessário de uso de *hardware* e priorizando somente as espécies com um mínimo de quantidade de informação.

Em seguida, aplicou-se novamente o processo de filtragem, utilizando valores de 2 a 5 para o parâmetro *Annotation Score*. Em cada passo (valor utilizado para esse parâmetro), foi aplicado uma pontuação de corte referente à quantidade de enzimas não redundantes, fornecendo assim diferentes quantidades de espécies a serem selecionadas, de enzimas não redundantes e de anotações de rotas metabólicas (Tabela 3A-C e Material Suplementar T2-10).

Tabela 3. Quantidades de espécies selecionadas (A), de enzimas não redundantes (B) e de anotações de rotas metabólicas (C), referente a cada pontuação de corte em relação ao número de enzimas não redundantes e a cada valor de *Annotation score* utilizado.

A

	Pontuação de corte						
Annotation score	50	75	100	125	150	175	200
2	177	172	147	97	36	16	8
3	39	16	9	5	5	4	2
4	4	2	2	2	2	2	2
5	2	2	1	1	1	1	1

B

	Pontuação de corte						
Annotation score	50	75	100	125	150	175	200
2	23.564	23.280	21.031	15.373	7.146	3.919	2.418

3	3.694	2.407	1.797	1.368	1.368	1.207	821
4	720	591	591	591	591	591	591
5	321	321	242	242	242	242	242

C

	Pontuação de corte						
Annotation score	50	75	100	125	150	175	200
2	29	29	28	28	23	18	14
3	16	13	11	7	7	6	4
4	6	4	4	4	4	4	4
5	4	4	3	3	3	3	3

De modo a selecionar o maior número de espécies, mas que contenham somente 10 a 20 tipos de anotações de rotas metabólicas (com um mínimo de 50 instâncias por espécie), utilizou-se a matriz A e B (espécies e tipos de rotas metabólicas selecionadas de acordo com *Annotation Score* x Pontuação de corte, respectivamente tabela 3A e 3C), para em seguida, gerar a matriz C:

Matriz C_{ij}: a_{ij} x b_{ij} se $10 \leq b_{ij} \leq 20$; caso contrário, C_{ij}: a_{ij} x 1

A $\begin{bmatrix} 177 & 171 & 147 & 97 & 36 & 16 & 8 \\ 39 & 16 & 9 & 5 & 5 & 4 & 2 \\ 4 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$ B $\begin{bmatrix} 29 & 29 & 28 & 28 & 23 & 18 & 14 \\ 16 & 13 & 11 & 7 & 7 & 6 & 4 \\ 6 & 4 & 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 3 & 3 & 3 & 3 & 3 \end{bmatrix}$ C $\begin{bmatrix} 177 & 172 & 147 & 97 & 36 & 288 & 112 \\ 624 & 208 & 99 & 5 & 5 & 4 & 4 \\ 4 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$

A matriz C é uma matriz de pontuação para selecionar a quantidade de espécies (maior pontuação) a compor o banco de dados, assim como definir os valores a serem utilizados de *Annotation Score* e pontuação de corte da quantidade de sequências não

redundantes de enzimas. Selecionando a maior pontuação da matriz C (624, alocado linha 2, coluna 1), encontramos a seleção correspondente na matriz A e B (linha 2, coluna 1). Com esse procedimento, foram selecionadas 39 espécies e as instâncias selecionadas agrupadas em 16 rotas metabólicas (utilizando o parâmetro *Annotation Score* igual a 3 e pontuação de corte igual a 50) de modo a compor os conjuntos de treinamento/teste e validação (Figura 2 e Material Suplementar T10-11).

	Pontuação de corte						
	50	75	100	125	150	175	200
amino ácido	551	355	267	204	204	185	146
carboidrato	510	305	211	148	148	123	86
cofator	362	198	137	102	102	90	75
glicano	152	125	86	65	65	65	6
glicolipídeo	78	71	55	41	41	34	28
lipídeo	146	87	62	45	45	36	28
metabólico intermediário	66	43	36	32	32	27	23
micotoxina	123	56	30	23	23	23	0
açúcar nucleotídeo	107	66	44	32	32	26	18
fosfolipídeo	187	96	61	45	45	38	27
modificação de proteína	469	359	286	228	228	195	170
purina	192	102	70	53	53	45	36
pirimidina	82	64	51	35	35	30	28
metabolismo secundário	148	126	114	88	88	88	6
enxofre	66	38	26	20	20	18	14
modificação de trna	64	32	19	13	13	12	10

Figura 2. Quantidade de sequências não redundantes de enzimas e suas respectivas rotas metabólicas, utilizando *Annotation Score* igual a 3.

4.2. Conjunto de treinamento, teste, validação, avaliador 1 e avaliador 2

Após o processo de filtragem dos dados, a etapa de seleção e separação dos dados em diferentes conjuntos (treinamento, teste e validação) foram executados. Sendo 16 rotas metabólicas selecionadas para este estudo, das 39 espécies selecionadas. As espécies *Emericella nidulans* (*Aspergillus nidulans*) e *Neosartorya fumigata* (*Aspergillus fumigatus*)

foram imediatamente alocadas para formar o conjunto de treinamento/teste, pois continham pelo menos uma sequência com a anotação das referidas 16 rotas metabólicas. As 37 espécies/cepas restantes (que não continham pelo menos uma sequência com anotação de cada rota metabólica), foram combinadas em pares, de modo que essa combinação obtivesse pelo menos uma instância para cada um dos 16 grupos de rotas metabólicas (amino ácido, carboidrato, cofator, glicano, glicolipídeo, lipídeo, metabólico intermediário, micotoxina, açúcar nucleotídeo, fosfolipídeo, modificação de proteína, purina, pirimidina, metabólito secundário, enxofre e modificação de trna). De forma aleatória, foi selecionado apenas um par dentro das combinações acima mencionadas para compor as espécies do conjunto de validação, selecionando assim as espécies *Aspergillus clavatus* (strain ATCC 1007 CBS 513.65 DSM 816 NCTC) e *Debaryomyces hansenii* (strain ATCC 36239 CBS 767 JCM 1990 NBRC) para essa finalidade. Este conjunto contém apenas instâncias que não participarão da fase de treinamento, sendo assim, instâncias completamente desconhecidas pelos algoritmos. As demais demais espécies/cepas não selecionadas foram inseridas no conjunto de treinamento (Figura 3 e Material Suplementar T12).

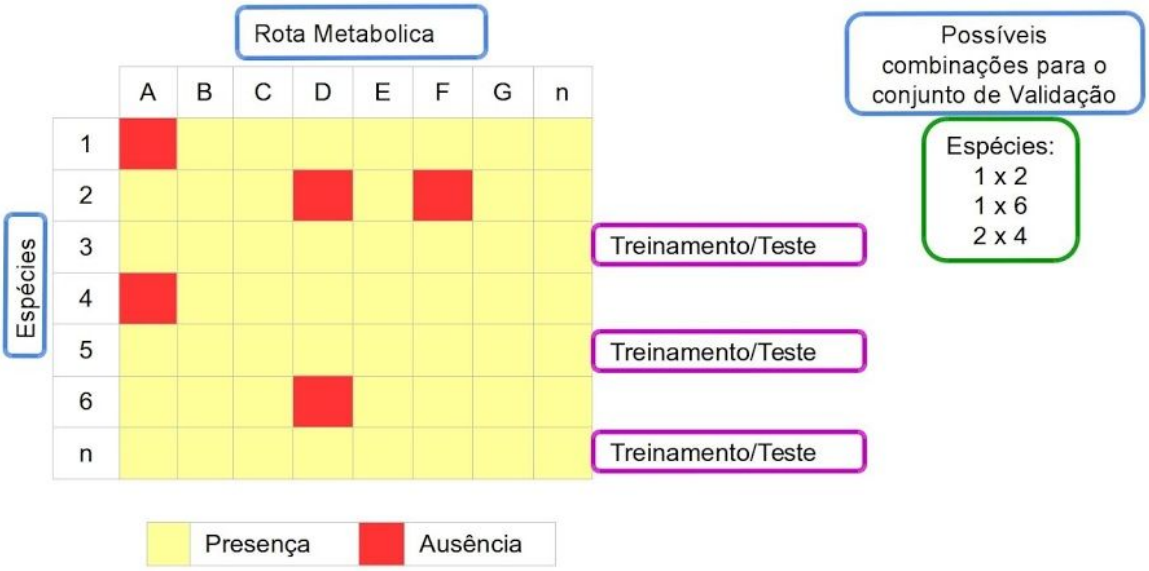


Figure 3. Processo de seleção de instâncias/espécies para o conjunto de validação.

Após selecionar os algoritmos devidamente otimizados e gerar os modelos para cada rota metabólica, os melhores modelos serão aplicados sobre as instâncias positivas de cada conjunto de treinamento, de modo a selecionar sequências de aminoácidos representativas para cada tipo de rota metabólica (conjunto avaliador 1). Para gerar o conjunto avaliador 1, selecionou-se as “sequências controle” cuja a probabilidade de predição positiva para uma determinada rota metabólica foi maior que 0,70, ao mesmo tempo que esta mesma instância tenha altos valores probabilísticas de predição negativa para outras rotas metabólicas (Figura 4). O conjunto avaliador 1 é utilizado na ferramenta mAppLe como instâncias controle, de modo a comparar com os resultados obtidos na classificação de instâncias desconhecidas.

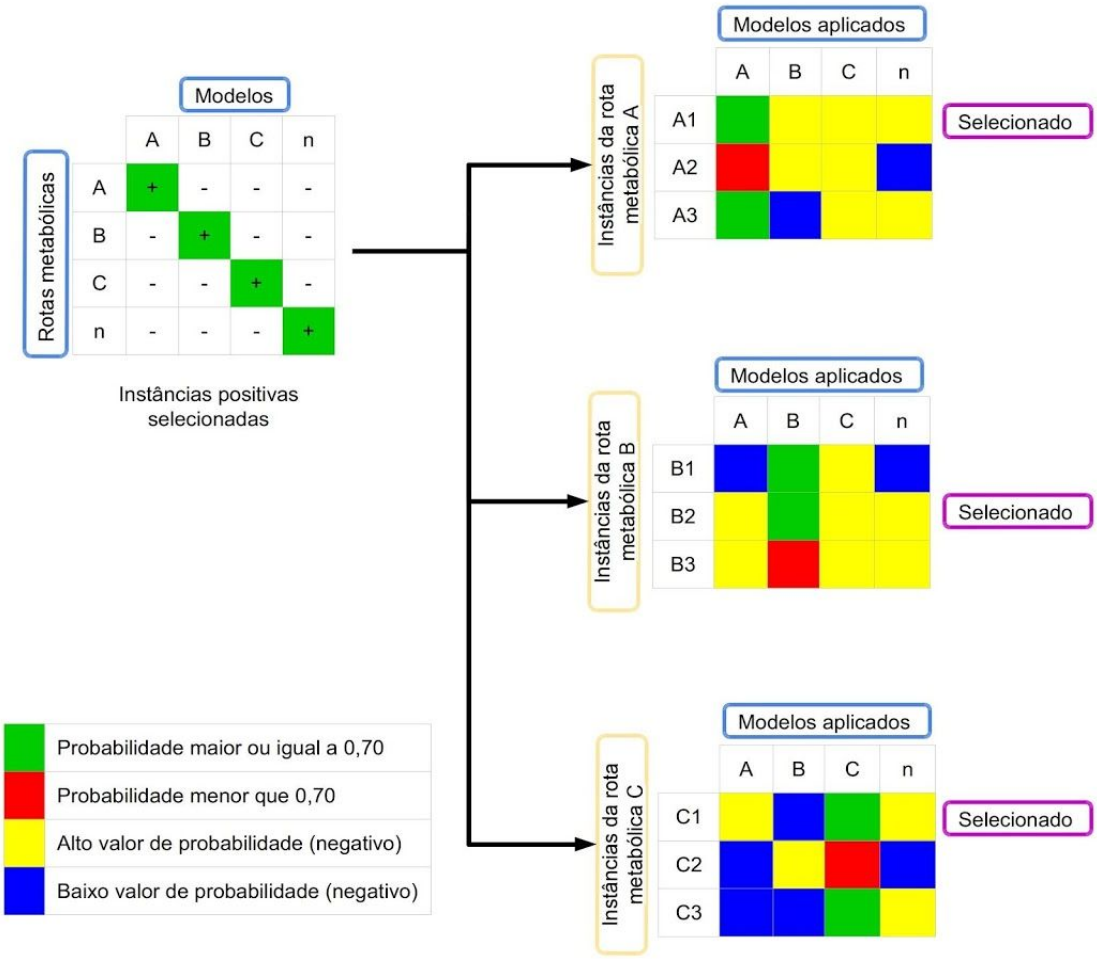


Figura 4. Seleção de sequências de aminoácidos controle para cada tipo de rota metabólica para a ferramenta mAppLe. Aplicação de todos os modelos sobre as instâncias positivas de cada conjunto de treinamento.

Para avaliar a performance preditiva da ferramenta mAppLe, o conjunto avaliador 2 foi criado utilizando as sequências de aminoácidos das espécies não selecionadas para compor a base de dados do treinamento e modelagem (conjuntos de treinamento, teste e validação), sendo assim um conjunto de dados completamente novo e desconhecido para os modelos gerados. Para esse conjunto de dados, o processo de filtragem também foi aplicado (enzimas somente com um *EC number*, *annotation score* ≥ 3, somente sequências não redundantes e sem aminoácidos especiais) de modo a selecionar enzimas (sequência de aminoácidos) com alta qualidade de informação.

4.3. Geração dos atributos

Utilizando o R package Peptides (OSORIO et al, 2015) e protr (XIAO et al.,2015), foram gerados um total de 1.024 atributos para todas as instâncias de cada rota metabólica dos conjuntos de treinamento, teste e validação. Os atributos calculados foram Composição/Transição/Distribuição, Auto Correlação de Moreau-Broto, Conjoint Triad, *Quasi Sequence Order*, Composição de Pseudo Aminoácidos, Composição de Pseudo Aminoácidos Anfifílicos, Composição de Monopeptídeos, Composição de Dípeptideos e Características Fisico-Químicas (Tabela 4).

Tabela 4. Características calculadas (atributos) para cada sequência e quantidade de atributos gerados.

Características calculadas	Quantidade de atributos
Composição/Transição/Distribuição	147
Auto Correlação de Moreau Broto	8
Conjoint Triad	343
Quase Sequence Order	42

Composição de Pseudo Aminoácidos	21
Composição de Pseudo Aminoácidos Anfilícos	22
Composição de Monopeptideos	20
Composição de Dipeptideos	400
Características Físico-Químicas	21
Total de atributos	1.024

4.4. Normalização dos dados

Antes dos conjuntos de dados serem devidamente separados (conjunto de treinamento, teste e validação), os atributos foram normalizados. Para isso, todos os valores de uma dada coluna são subtraídos pelo menor valor desta coluna e posteriormente dividido pela amplitude. Após a normalização, os conjuntos foram separados em treinamento/teste e validação.

4.5. Identificação de classes

Para todos os conjunto de dados (treinamento/teste e validação), as instâncias (sequências de aminoácido das enzimas selecionadas) foram rotuladas como “positivo” ou “negativo”, baseando-se na anotação da rota metabólica. Quando enzimas da rota metabólica X_1 são rotuladas como “positivo”, as demais enzimas de outras rotas são rotuladas como “negativos”. Logo, quando as enzimas da rota X_2 são rotuladas como “positivo”, as demais rotas são rotuladas como “negativos”, incluindo as enzimas da rota X_1. Isso se repete para todas as rotas metabólicas existentes no conjunto de dados. Após esse procedimento, 90% das instâncias do conjunto de treinamento/teste foram selecionadas aleatoriamente para compor o conjunto de dados de treinamento, sendo os 10% restantes alocados como o conjunto de teste (Figura 5).

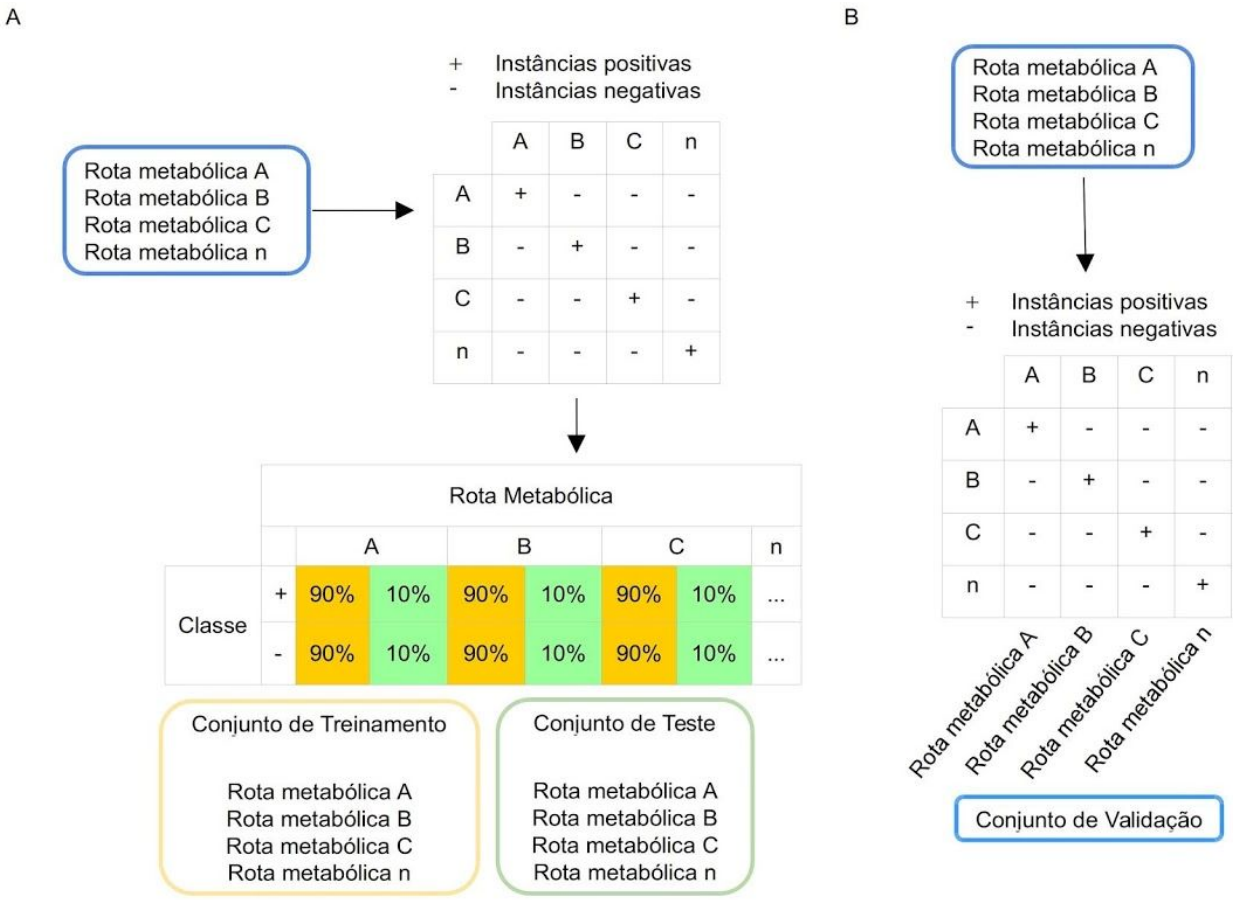


Figura 5. Construção dos conjuntos de dados de treinamento, teste e validação. A – Separação dos dados de treinamento/teste (90% para o conjunto de treinamento e 10% para o conjunto de teste), passando por aleatorização e *undersampling*. B – Montagem do conjunto de validação. Todos os conjuntos (treinamento, teste e validação) passaram pelo processo de normalização e somente o conjunto de treinamento foi balanceado.

Sendo assim, forma-se uma matriz quadrada A (quantidade de linhas é igual à quantidade de colunas), onde a_{ij} é “positivo” quando i=j, e “negativo” quando i≠j.

$$A = \begin{bmatrix} + & - & - & - \\ - & + & - & - \\ - & - & + & - \\ - & - & - & + \end{bmatrix}$$

Com isso, a quantidade de instâncias sem mantêm fixas dentro de cada conjunto de dados (para treinamento, teste e validação).

4.6. Adequação do conjunto de treinamento

Sendo a quantidade de instâncias “negativas” muito superior às “positivas”, o processo de *undersampling* foi realizado no conjunto de treinamento. Esta técnica é aplicada em um conjunto de dados no qual a distribuição das classes ou categorias estão desbalanceadas, com o objetivo final de ajustar esta distribuição, gerando diversos subconjuntos balanceados. Assim sendo, as instâncias negativas foram aleatorizadas e divididas de acordo com a quantidade de instâncias positivas, formando X subconjuntos. Em cada subconjunto X, as instâncias positivas são as mesmas e todos os subconjuntos X, a qual têm suas instâncias totais (positivos e negativos) randomizadas de modo a evitar algum tipo de viés no processo de *cross-validation*. Este processo é uma técnica que separa o conjunto de dados em diversos subconjuntos (geralmente, é dividido aleatoriamente em 10 subconjuntos - 10 *fold cross-validation*), contendo uma proporção aproximadamente igual das classes (positivo e negativo). Cada subconjunto é isolado por vez, de modo que o algoritmo de aprendizagem é aplicado em cada 9 dos subconjuntos gerando um submodelo a ser aplicado no subconjunto previamente isolado, para no final (após esse processo ser executado 10 vezes) obter uma estimativa de erro do algoritmo de aprendizagem.

4.7. Redução da dimensionalidade

Com o conjunto de treinamento devidamente separado, normalizado e com todos os atributos calculados, realizou-se a etapa de seleção dos melhores atributos de cada rota metabólica. As ferramentas "*principal components*" e "*attribute selection ranker*" do software Weka-3.9.1 (todos os parâmetros em *default*) foram aplicadas em todos os conjuntos de treinamento, de modo a selecionar 10, 20, 30, 40, 50, 100, 150 e 200 melhores atributos ranqueados (dentro dos 1.024 atributos disponíveis), sendo este procedimento realizado para todos os metabolismos de forma independente. Além disso, este procedimento faz com que

os menores grupos de atributos selecionados sempre estejam inseridos nos grupos maiores, ou seja, o aumento do número de atributos é iterativo (Figura 6).

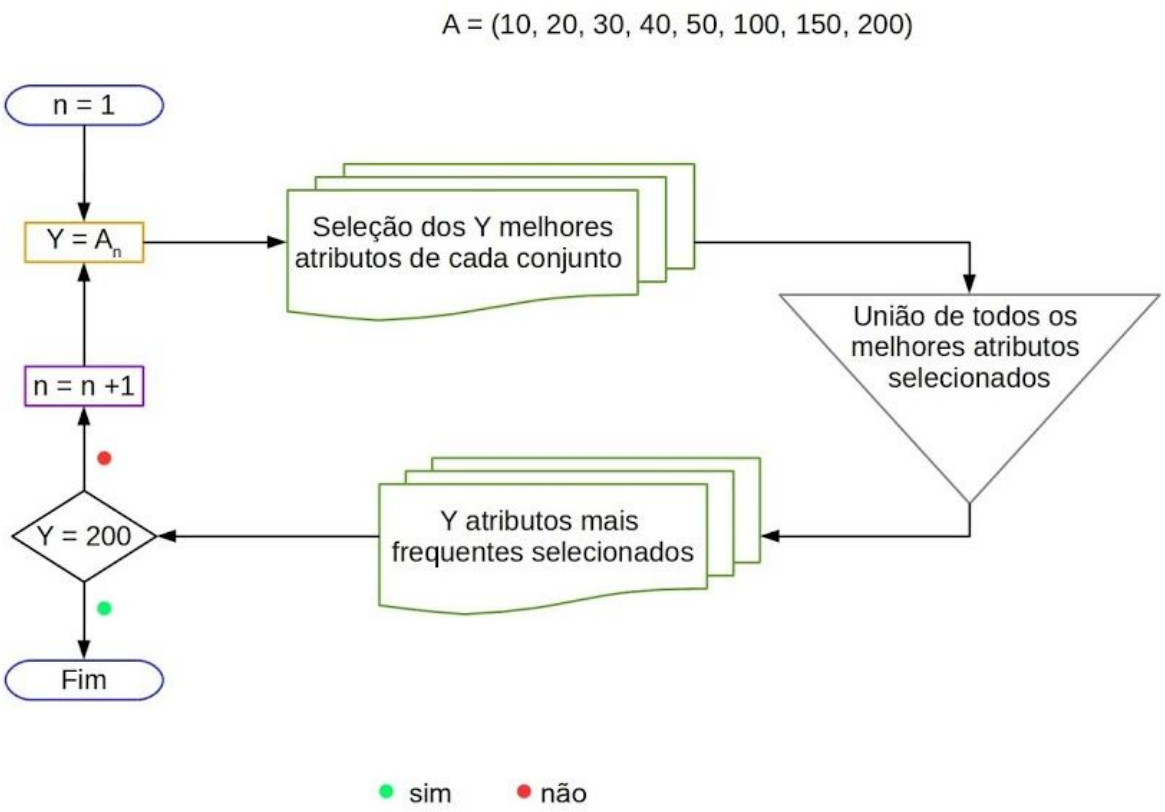


Figura 6. Procedimento para criação dos conjuntos de dados de treinamento com base na quantidade de melhores atributos selecionados (de 10 a 200), para cada metabolismo.

Para cada metabolismo, os mesmos atributos selecionados no conjunto de treinamento foram selecionados nos conjuntos de teste, validação, avaliador 1 e 2. Após esse processo, os atributos selecionados (10 a 200) nos conjuntos de treinamento foram testados independentemente no aprendizado supervisionado de modo a selecionar o melhor grupo de atributos para criar o modelo classificador final para cada metabolismo.

4.8. Algoritmos classificadores e seleção dos melhores parâmetros

Foram selecionados 47 algoritmos (Material Suplementar T13) foram aplicados (em configuração *default*) em 5 subconjuntos de treinamento de cada uma das 16 rotas metabólicas para cada *N* melhores atributos selecionados, para assim obter uma média da performance de treinamento (Material Suplementar T14). Para essa medição de performance, utilizou-se as métricas fornecidas pelo software tais como porcentagem de instâncias corretamente classificadas (ICC), verdadeiros positivo (VP), falsos positivos (FP), precisão, recall, F-*score* e ROC. Os algoritmos com menores valores de FP e maiores valores de F-*score* foram selecionados e o teste estatístico de Kruskal-Wallis (seguido do teste *post hoc* Diferença Mínima Significativa de Fisher, com $p \leq 0.01$) foi aplicado sobre os valores de ICC desses algoritmos selecionados, possibilitando assim a seleção de até três algoritmos classificadores para cada rota metabólica (Figura 7, Tabela 5 e Material Suplementar T15).

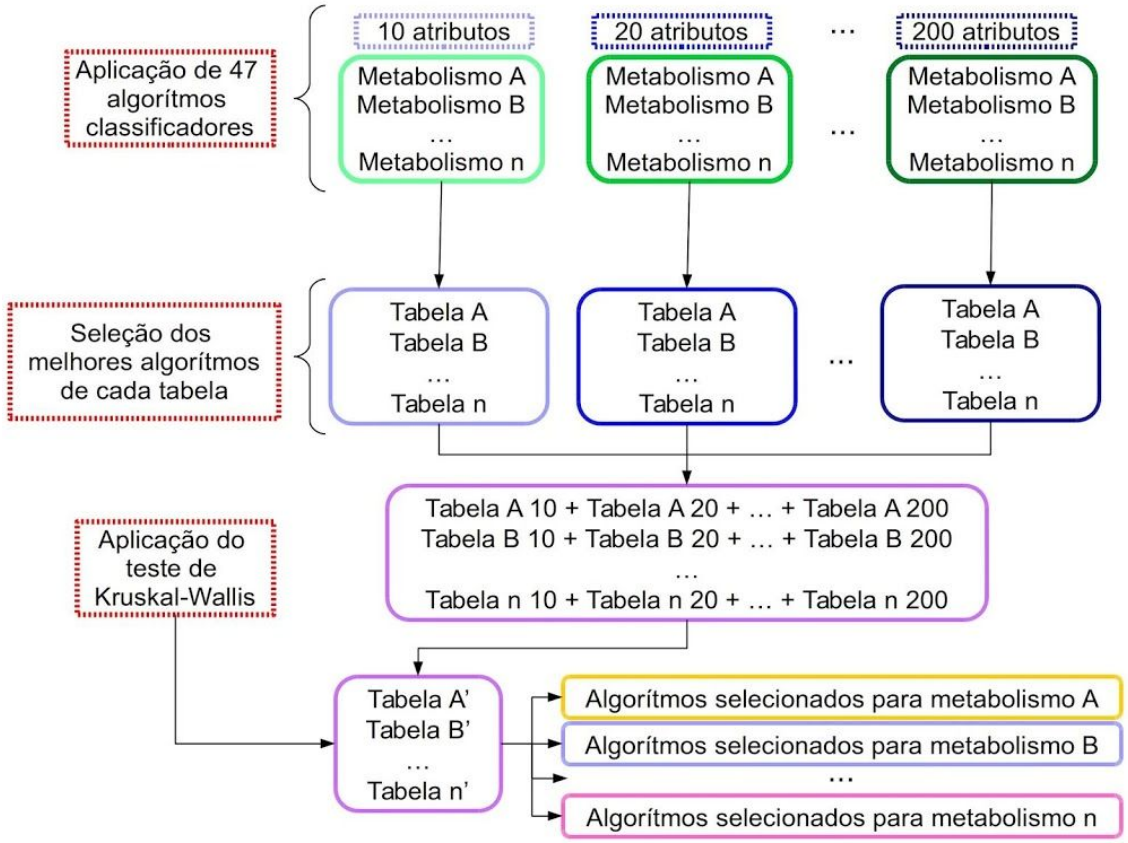


Figura 7. Procedimento de seleção de até três algoritmos classificadores a serem aplicados em cada tipo de rota metabólica para a geração dos modelos.

Tabela 5. Algoritmos selecionados para cada tipo de rota metabólica, após teste de Kruskal-Wallis (seguido do teste *post hoc* Diferença Mínima Significativa de Fisher, com $p \leq 0.01$). ICC - Instâncias corretamente classificadas. DP - Desvio padrão.

Rota Metabólica	Média ICC	Média DP	Algoritmo selecionado
Amino Ácido	74.61	7.33	IB1, Ibk, RseslibKnn
Carboidrato	81.30	6.41	IB1, Ibk, Random Forest
Cofator	73.38	5.79	IB1, RandomForest, RseslibKnn
Glicano	83.97	5.29	RandomForest, RBFClassifier

Glicolipídeo	85.81	1.82	RandomForest
Lipídeo	72.90	6.24	LibLINEAR, RandomForest, RBFClassifier
Metabólico Intermediário	77.39	8.47	RseslibKnn
Micotoxina	61.62	3.89	RandomForest, RBFClassifier, RseslibKnn
Açúcar-Nucleotídeo	80.84	9.73	RseslibKnn
Fosfolipídeo	82.76	6.89	RandomForest
Modificação de Proteína	74.92	7.00	RandomForest
Purina	83.18	4.38	RandomForest
Pirimidina	73.40	6.25	Random Forest, RseslibKnn
Metabólito Secundário	61.46	3.99	Random Forest
Enxofre	82.58	5.55	LibLINEAR, RBFClassifier
Modificação de Trna	82.84	10.33	LMT

Os algoritmos selecionados foram aplicados 8 vezes (em cada *N* melhores atributos selecionados, conforme seção 4.7), assim como cada algoritmo é aplicado juntamente com o algoritmo Bagging, que melhora a performance reduzindo a variância dos dados.

Uma vez selecionados os algoritmos classificadores para cada rota metabólica, realizou-se o processo de testes dos parâmetros de cada um desses algoritmos, a fim de obter as maiores performances para o treinamento. Para esse procedimento, foram utilizados 5 subconjuntos de treinamento de cada rota metabólica (conjunto de dados onde se aplicará o algoritmo classificador a ser otimizado) e diferentes combinações possíveis dos parâmetros de cada algoritmo classificador foram testadas um a um (Figura 8).

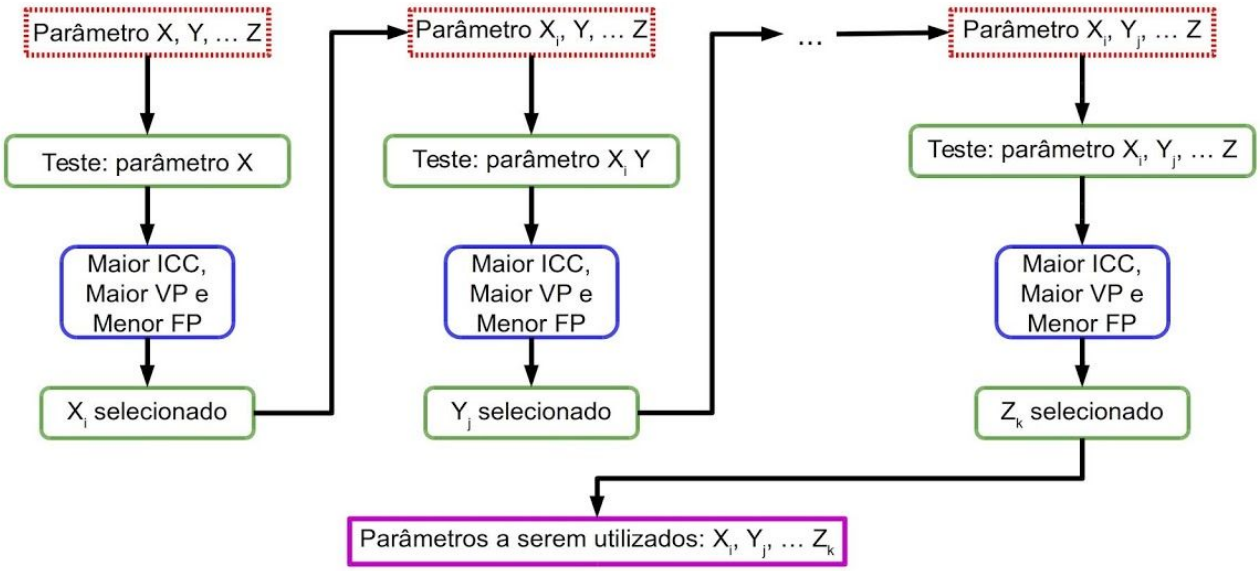


Figura 8. Processo de seleção dos melhores parâmetros de cada algoritmo classificador para cada rota metabólica baseado-se nos melhores resultados das performances preditivas. ICC - Instâncias corretamente classificadas. VP - Taxa de verdadeiros positivos. FP - Taxa de falsos positivos.

Assim, se um algoritmo possui X e Y como parâmetros que podem ser modificados, primeiro testa-se X utilizando 5 (ou 10) valores, e com base na performance preditiva, seleciona-se o valor de X que forneceu o melhor resultado. Em seguida, o mesmo é feito para o parâmetro Y e assim por diante, caso o algoritmo tenha diversos parâmetro que podem ser otimizados. A melhor performance preditiva é selecionada obedecendo, em ordem, os critérios de maior valor de ICC (instâncias corretamente classificadas), maior valor de VP (verdadeiros positivos) e menor valor de FP (falsos positivos). Somando a isso, a otimização dos parâmetros dos algoritmos Bagging e Vote também são realizados. Após obter os dados de otimização de parâmetros para cada algoritmo, estes foram devidamente melhorados, combinados (utilizando os algoritmos Bagging e Vote, respectivamente) e aplicados em todos os subconjuntos de treinamento para cada rota metabólica (juntamente com a aplicação de *10-fold cross-validation*), gerando assim, ao final, os modelos finais específicos.

4.9. Ferramenta mAppLe (Metabolic Pathway Prediction of Enzymes)

Sequências controle (conjunto avaliador 1, conforme descrito na seção 4.2) selecionadas para compor o grupo controle da ferramenta mAppLe (Tabela 6 e Material Suplementar T17), serão utilizadas pela ferramenta mAppLe no processo de predição de rotas metabólicas de enzimas. Com isso, três modelos foram removidos (rotas metabólicas de glicolipídeo, micotoxina e metabólito secundário, respectivamente) devido ao baixo valor de predição das instâncias controle (Tabela 6), ficando somente 11 rotas metabólicas a serem exploradas por esta ferramenta.

Tabela 6. Seleção de sequências controle para a ferramenta mAppLE. Sequências que obtiveram predições positivas menores que 0,70 foram removidas.

Rota metabólica	Predição	Entry – mAppLe controle	Removido ?
Amino Ácido	0.900 – 0.919	Q59QC4, A0A168PEW4, A7E4S9	Não
Carboidrato	0.922 – 0.923	P17819, J9VRH1, A3LQ70	Não
Cofator	0.934 – 0.952	A0A061AEW9, A0A1D9PW33, G2XYG7	Não
Glicano	0.799 – 0.817	Q4WLV2, B0XXF3, B8NKE9	Não
Glcolipídeo	0.522	Q5AMR5	Sim
Metabólico Intermediário	0.921	E6RG60	Não
Micotoxina	0.538 – 0.546	Q6Q884, N4WR35	Sim
Açúcar-Nucleotídeo	1	Q75AB5, A0A0D1BUK2, M1W8V6, A5DNZ9, A7TSJ1, E5A4Z2, M2U6W0, M2U9I8, G8ZZX5	Não
Fosfolipídeo	0.710 – 0.750	C0NQR3, A6QX77, Q2UC55	Não
Modificação de Proteína	0.831 – 0.857	P0CH67, Q6FR76, A0A0W4ZKM2	Não
Purina	0.760 – 0.844	Q7S604, Q7SFX7, J9VS03	Não

Pirimidina	0.904 – 0.936	P33317, O13867, Q6FKQ6	Não
Metabólito Secundário	0.581 – 0.631	Q5AUW8, Q4WLD0, Q4WQY6	Sim
Modificação de Trna	0.954 – 0.963	Q7RZC1, A0A1Q2ZTN3, A0A1Q3A7Y7	Não

A ferramenta mAppLe, formada pelos modelos que obtiveram melhores resultados nas performances de predições sobre os conjuntos de teste e validação, tem seu funcionamento simples e automatizado, bastando apenas a adição dos arquivos (exemplos no Material Suplementar T18 - subconjuntos do conjunto avaliador 2) a serem analisados (contendo sequências de aminoácidos, nos formatos .fasta, .tab ou .csv), dos modelos disponíveis para análise (total de 11 modelos) e execução de um *script* em linguagem R (somente para máquinas com sistema operacional Linux). Ao final, a ferramenta gera uma tabela com a probabilidade de cada instância para cada rota metabólica. De forma simplificada, a figura 9 mostra as etapas necessárias para a criação dos modelos de predição e o seu devido uso na ferramenta mAppLe.

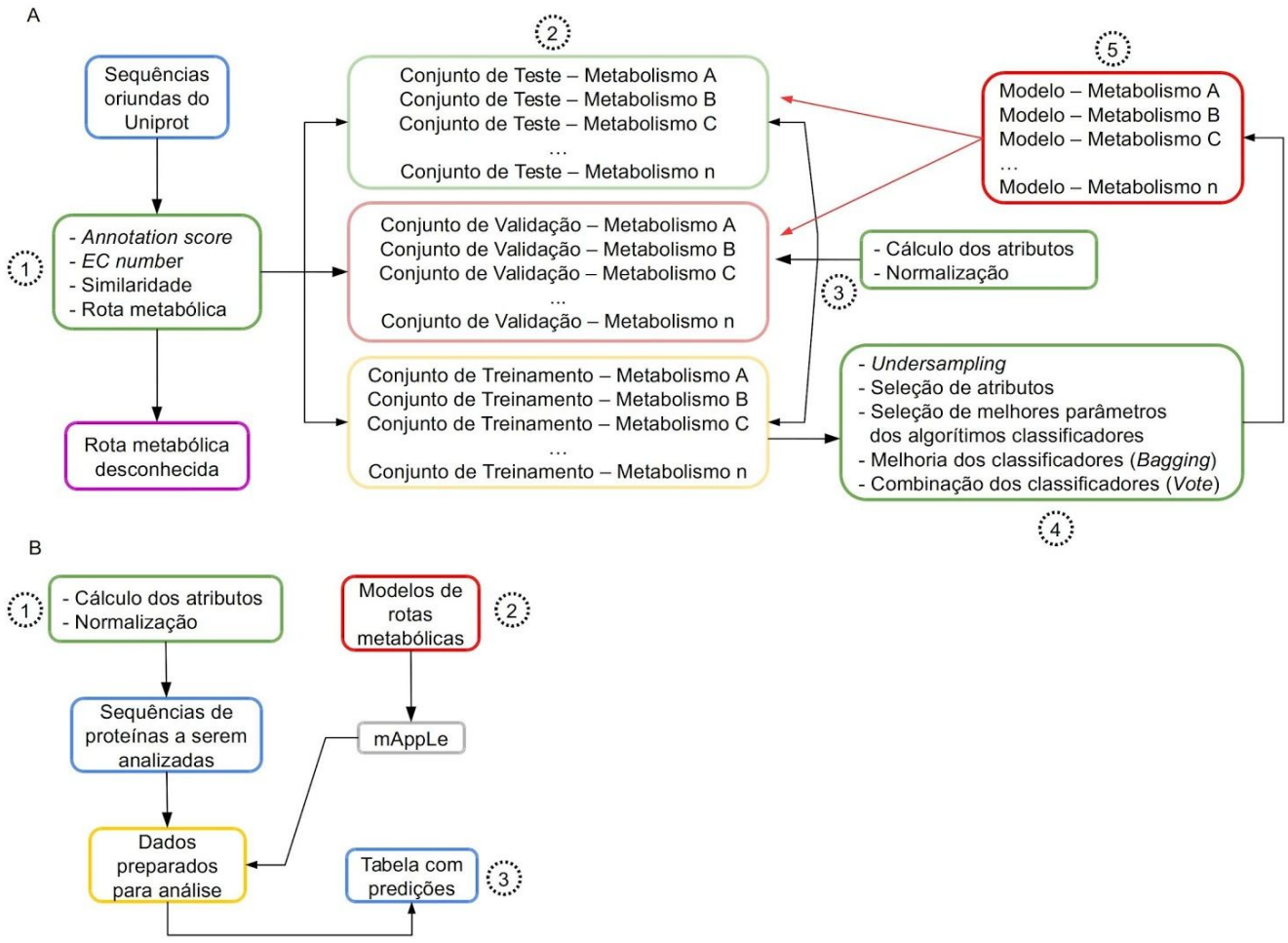


Figura 9 – Procedimentos utilizados para obtenção dos modelos preditivos e sua aplicação. A – Processo de filtragem (1), criação dos conjuntos de dados (2), preparação dos dados (3), processo de obtenção dos modelos preditivos (4), aplicação dos modelos nos conjuntos de dados (5); B – Preparação dos dados (1), aplicação dos modelos (2), resultado das predições (3).

5. RESULTADOS E DISCUSSÃO

5.1. Seleção das espécies e instâncias

Utilizando 2.220.779 sequências de aminoácidos de 177 espécies/cepas pré selecionadas (sendo 131.710 sequências referentes a enzimas) e utilizando o processo de filtragem conforme anteriormente descrito, 3.144 instâncias foram selecionadas para compor

o conjunto de treinamento/teste, contendo 37 espécies/cepas. Além disso, 159 instâncias foram selecionadas para compor o conjunto de validação, contendo 2 espécies/cepas (Material suplementar T12).

Isso nos mostra que, das enzimas selecionadas (sequências não redundantes, não contendo aminoácidos especiais, exercendo somente uma função enzimática e contendo anotações sobre a rota metabólica de atuação), apenas 2,5% contém informações mais detalhadas. Além disso, fica bem evidente que os estudos enzimáticos são direcionados para poucas espécies, com um maior destaque para as espécies *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Aspergillus nidulans* (*Emericella nidulans*), *Aspergillus fumigatus* (*Neosartorya fumigata*) e *Candida albicans* (com 7.159, 1.460, 1.275, 3.659 e 1.290 sequências, respectivamente), representando 32,3% dos dados. De fato, estas espécies ainda são estudadas exaustivamente e sendo os principais modelos (dentro da biologia molecular e celular) de organismos do reino fungi, existem bancos de dados próprios tais como o SGD (Saccharomyces Genome Database), PomBase, AspGD (Aspergillus Genome Database) e CGD (Candida Genome Database). Contudo, essas espécies formam a principal base de dados para comparação de novas sequências, o que pode levar a inferir uma anotação equivocada simplesmente por levar em consideração a porcentagem de identidade.

5.2. Conjunto de treinamento, teste e validação

Como abordado anteriormente, as 5 espécies com maior quantidade de informação (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Emericella nidulans* (*Aspergillus nidulans*), *Neosartorya fumigata* (*Aspergillus fumigatus*) e *Candida albicans*) representam 37,3% do conjunto de treinamento/teste (1174 instâncias de 3144), sendo 12,5, 9,8, 5,6, 5,0 e 4,4% dos dados, respectivamente. Entretanto, utilizar somente estas espécies para criar os modelos preditivos poderia ocasionar o problema de *overfiting*, um modelo com alta performance preditiva no treinamento e com baixa taxa preditiva em dados desconhecidos. Logo, as 32 espécies/cepas restantes, que representam 62,7% do conjunto de

treinamento/teste, promovem uma maior diversificação nos dados, evitando assim o referido problema e podendo colaborar para a geração de um modelo generalista.

Após selecionar as espécies/instâncias para seus devidos conjuntos de dados (treinamento/teste e validação), foram determinados 1.024 atributos para cada instância, utilizando os R *packages* protr e Peptide (Tabela 4). Em seguida, os dados foram normalizados, conforme descrito na seção 4.4. Este processo (cálculo de atributos e normalização) precisa ser realizado antes da separação dos dados, pois caso realize essa etapa separadamente em cada conjunto de dados, a amplitude de cada atributo poderá ser diferente, dificultando o algoritmo de aprendizagem e/ou gerando modelos de baixa performance (testes preliminares realizados - dados não mostrados).

Em seguida, as instâncias foram rotuladas como “positivo” ou “negativo”, baseando-se na anotação da rota metabólica (Figura 5A-B), conforme descrito na seção 4.5 (Identificação de classes). Após essa etapa, foram selecionados 10% das instâncias de cada rota metabólica para assim gerar o conjunto de teste (Tabela 7A). Os 90% restantes das instâncias (de cada rota metabólica) foram agrupados para formar o conjunto de dados de treinamento (Tabela 7B). Tal processo não foi aplicado nos dados de validação, pois estes não participam no processo de treinamento (Tabela 7C). Destes três conjuntos de dados, somente o conjunto de treinamento teve os dados balanceados (*undersampling*).

Tabela 7. Quantidade de instâncias pertencentes aos conjuntos de dados. A - Teste; B-Treinamento; C - Validação.

A

Rotas metabólicas	Instâncias Positivas	Instâncias Negativas	Total de instâncias
amino ácido	53	262	315
carboidrato	48	266	314
cofator	35	280	315
glicano	14	301	315
glicolipídeo	7	307	314
lipídeo	14	301	315

metabólico intermediário	7	308	315
micotoxina	11	303	314
açúcar nucleotídeo	10	304	315
fosfolipídeo	18	297	315
modificação de proteína	45	270	315
purina	18	296	314
pirimidina	8	306	314
metabolismo secundário	15	300	315
enxofre	6	309	315
modificação de trna	6	309	315

B

Rotas metabólicas	Instâncias Positivas	Instâncias Negativas	Subconjuntos criados após *undersampling*	Total de instâncias
amino ácido	475	2.354	5	2.829
carboidrato	437	2.393	6	2.830
cofator	311	2.518	8	2.829
glicano	122	2.707	23	2.829
glicolipídeo	66	2.764	42	2.830
lipídeo	125	2.705	22	2.830
metabólico intermediário	59	2.771	50	2.830
micotoxina	99	2.731	28	2.830
açúcar nucleotídeo	92	2.738	30	2.830
fosfolipídeo	161	2.669	17	2.830
modificação de proteína	401	2.428	6	2.829
purina	165	2.665	17	2.830
pirimidina	72	2.758	39	2.830
metabolismo secundário	132	2.697	21	2.829
enxofre	58	2.772	48	2.830
modificação de trna	55	2.775	51	2.830

C

Rotas metabólicas	Instâncias Positivas	Instâncias Negativas	Total de instâncias
amino ácido	23	136	159
carboidrato	25	134	159
cofator	16	143	159

glicano	16	143	159
glicolipídeo	5	154	159
lipídeo	7	152	159
metabólico intermediário	1	158	159
micotoxina	13	146	159
açúcar nucleotídeo	5	154	159
fosfolipídeo	8	151	159
modificação de proteína	23	136	159
purina	9	150	159
pirimidina	2	157	159
metabolismo secundário	1	158	159
enxofre	2	157	159
modificação de trna	3	156	159

Como pode ser observado nas tabelas 7A-C, tal procedimento faz com que a quantidade de instâncias para cada rota metabólica sejam praticamente iguais, variando apenas a proporção das classes “positivas” (sempre em menor quantidade) e “negativas”. Para balancear as classes (somente no conjunto de treinamento, conforme descrito na seção 4.6), aplicou-se a técnica de *undersampling* (randomização das instâncias negativas e dividindo-as de acordo com a quantidade de instâncias positivas). Após inserir instâncias positivas em cada pequeno conjunto das instâncias negativas, os dados foram novamente aleatorizados para evitar qualquer tipo de viés no processo de *cross-validation*, gerando assim diversos conjuntos de treinamento balanceados para cada tipo de metabolismo. Como esperado, cada rota metabólica do conjunto de treinamento, passa assim a ter diferentes quantidades de subconjuntos de treinamento (Tabela 7B). Apesar das instâncias positivas serem as mesmas dentro de cada subconjunto de treinamento de cada rota metabólica, as instâncias negativas são sempre diferentes. Além disso, a ordem de todas as instâncias (positivas e negativas) são diferentes em cada subconjunto, sempre com o objetivo de evitar (ou minimizar) algum tipo de viés. É importante chamar a atenção para os conjuntos de teste e validação, conjuntos que não passam pela técnica de *undersampling*. Isso se deve ao fato de serem dados utilizados para a avaliação da performance do modelo criado, no qual as classes positivo e negativo são “mascaradas” para os algoritmos (substituindo os rótulos pelo carácter “?”).

5.3. Seleção dos atributos

Esse procedimento visa reduzir a dimensão dos conjuntos de dados, podendo assim gerar um melhor desempenho dos modelos a serem criados. Entretanto, não há uma regra que defina um número mínimo ou máximo de atributos a serem utilizados em um aprendizado de máquina, pois isso varia de acordo com os dados analisados. Sendo assim, optou-se por testar 8 possibilidades de redução dimensional dos dados, selecionando 10, 20, 30, 40, 50, 100, 150 e 200 atributos dentre os 1.024 iniciais; esse procedimento foi feito independentemente para cada rota metabólica. Importante frisar que, para essa redução de dimensão, seleciona-se apenas os atributos de maior relevância para a análise.

Deve-se destacar que, após realizar essa etapa, todos os dados (treinamento, teste e validação) foram testados e analisados 8 vezes nas etapas subsequentes (aprendizado supervisionado), no qual cada análise utiliza seus devidos conjuntos de dados de *N* melhores atributos. Isso se faz necessário para que, ao final de todo processo (geração dos modelos de predição e aplicação deste nos conjuntos de dados de teste e validação, de forma independente para cada rota metabólica), seja possível comparar as performances de todos os modelos criados. Outro detalhe importante foi a predominância dos atributos selecionados fornecidos pelo descritor *Conjoint Triad*. Mesmo sendo estes atributos selecionados em diferentes ordens de relevância em cada rota metabólica, é notável seu destaque nesta etapa de seleção de atributos mais relevantes.

5.4. Seleção dos algoritmos classificadores

Diferentemente de outros trabalhos onde seleciona-se *a priori* os algoritmos de aprendizagem a serem aplicados, no presente trabalho optou-se por explorar os mais diversos algoritmos implementados no software weka-3-9-1, possibilitando assim o uso de algoritmos pouco explorados e com alta performance preditiva. Pode-se constatar que os algoritmos classificadores selecionados podem ser os mesmos nas diferentes rotas metabólicas. A combinação de mais de três algoritmos classificadores para cada rota

metabólica não resultava em melhoras expressivas dos modelos gerados. . Sendo assim, combinar quantidade excessiva de classificadores com Vote e Bagging gera modelos muito complexos que, na maioria das vezes, não apresentam ganho significativo de performance na predição (de acordo com testes realizados - dados não mostrados). Além disso, o tempo de uso de *hardware* se eleva consideravelmente, inviabilizando a análise e criação dos modelos preditivos.

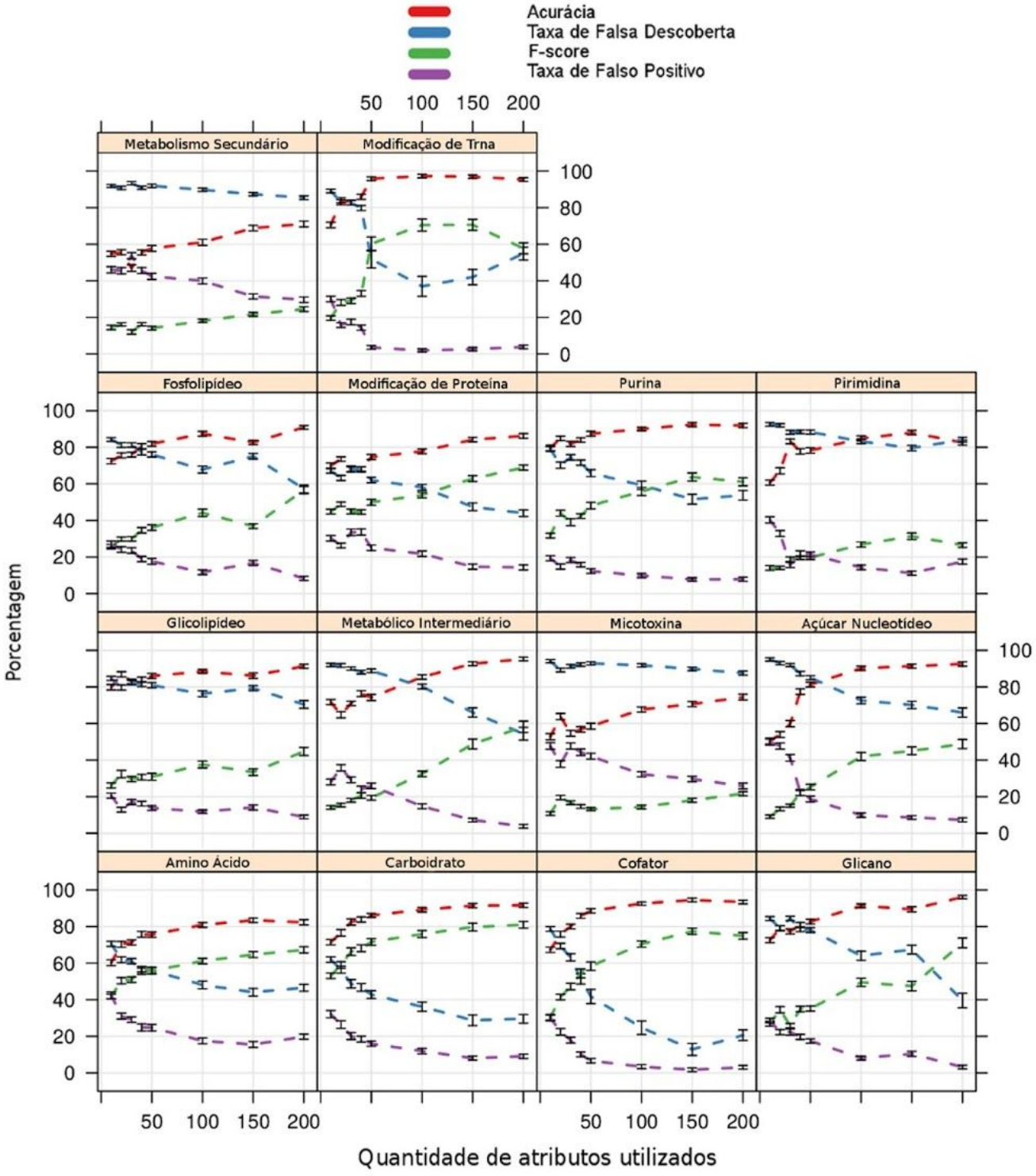
5.5. Aplicação dos modelos nos conjuntos de teste e de validação

Esta etapa representa 90% do tempo de utilização de *hardware*, exigindo o máximo da capacidade da máquina tais como elevado uso de processadores (*multi-core*) e memória RAM. Visando diminuir o tempo de uso do *hardware*, foi utilizado também o processamento em paralelo (utilização de vários núcleos simultaneamente para processamento dos dados). Notavelmente, esta etapa é a mais importante de todo o trabalho, pois uma vez que o modelo final gerado não seja satisfatório (avaliado nas etapas seguintes), as etapas de seleção de algoritmos precisam serem refeitas para assim fazer a devida modificação no modelo (trocar/retirar/inserir de algoritmo classificador, modificar parâmetro de otimização, etc).

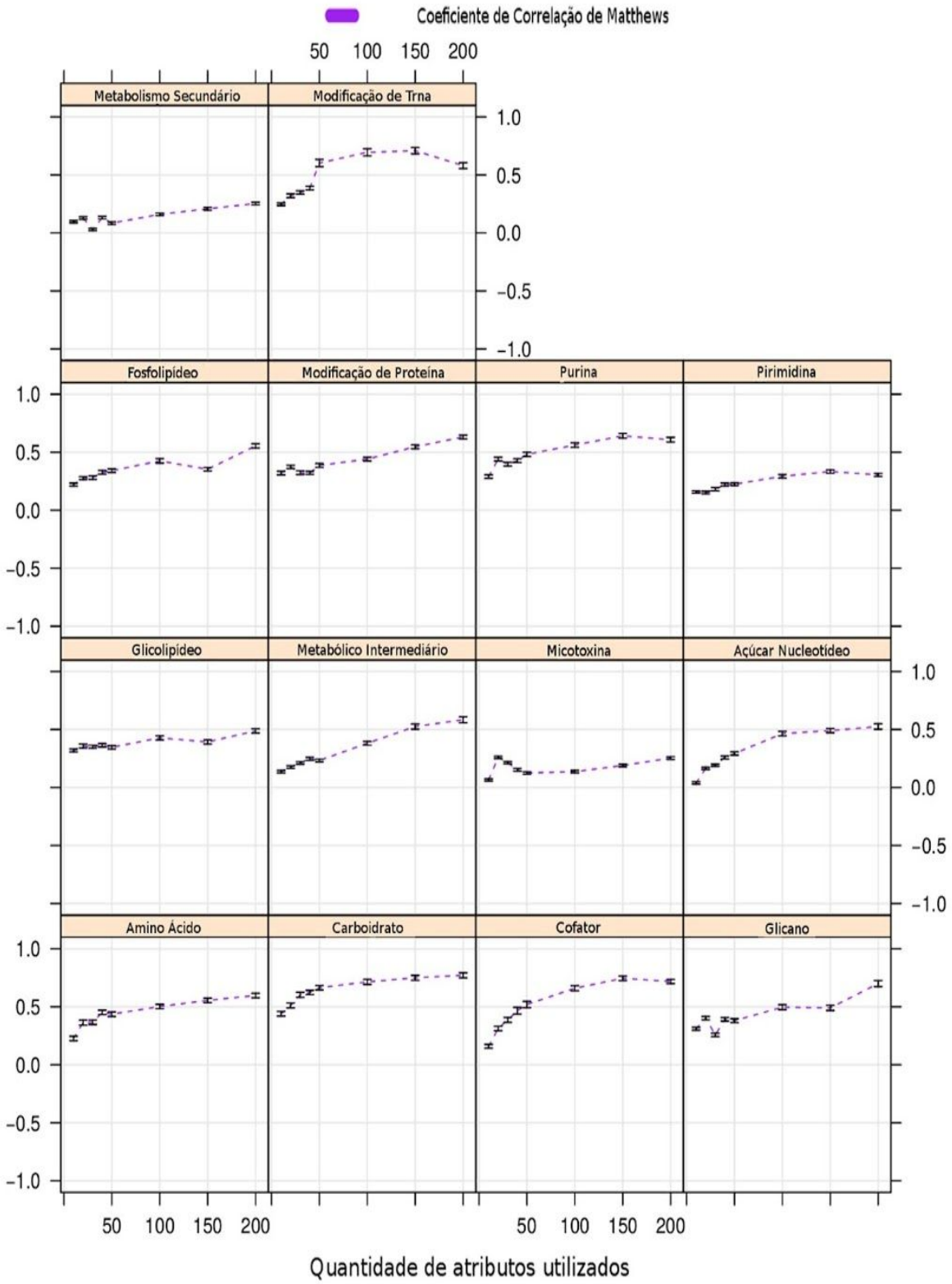
Após criar os modelos finais para cada tipo de rota metabólica, os conjuntos de dados de teste e validação tiveram seus rótulos positivo e negativo “mascarados”, ficando assim completamente desconhecidos para os algoritmos. Somado a isso, estes conjuntos de dados (teste e validação) foram subdivididos em 100 subconjuntos, para cada rota metabólica. Os subconjuntos foram gerados utilizando todas as instâncias positivas (sempre em menor quantidade) e selecionando 80% das instâncias negativas aleatoriamente (podendo algumas destas instâncias se repetirem ao longo dos subconjuntos formados), para cada tipo de rota metabólica. Além disso, o conjunto de validação foi dividido em dois subconjuntos sendo que cada um contenha apenas uma única espécie, formando assim três conjuntos de validação: 1- dados de duas espécies escolhidas; 2- dados somente da espécie A; 3- dados somente da espécie B. Em seguida, os modelos foram aplicados sobre todos subconjuntos acima

descritos (Figura 10A-H e Material Suplementar T16). As espécies selecionadas no dado de validação serão posteriormente descritas.

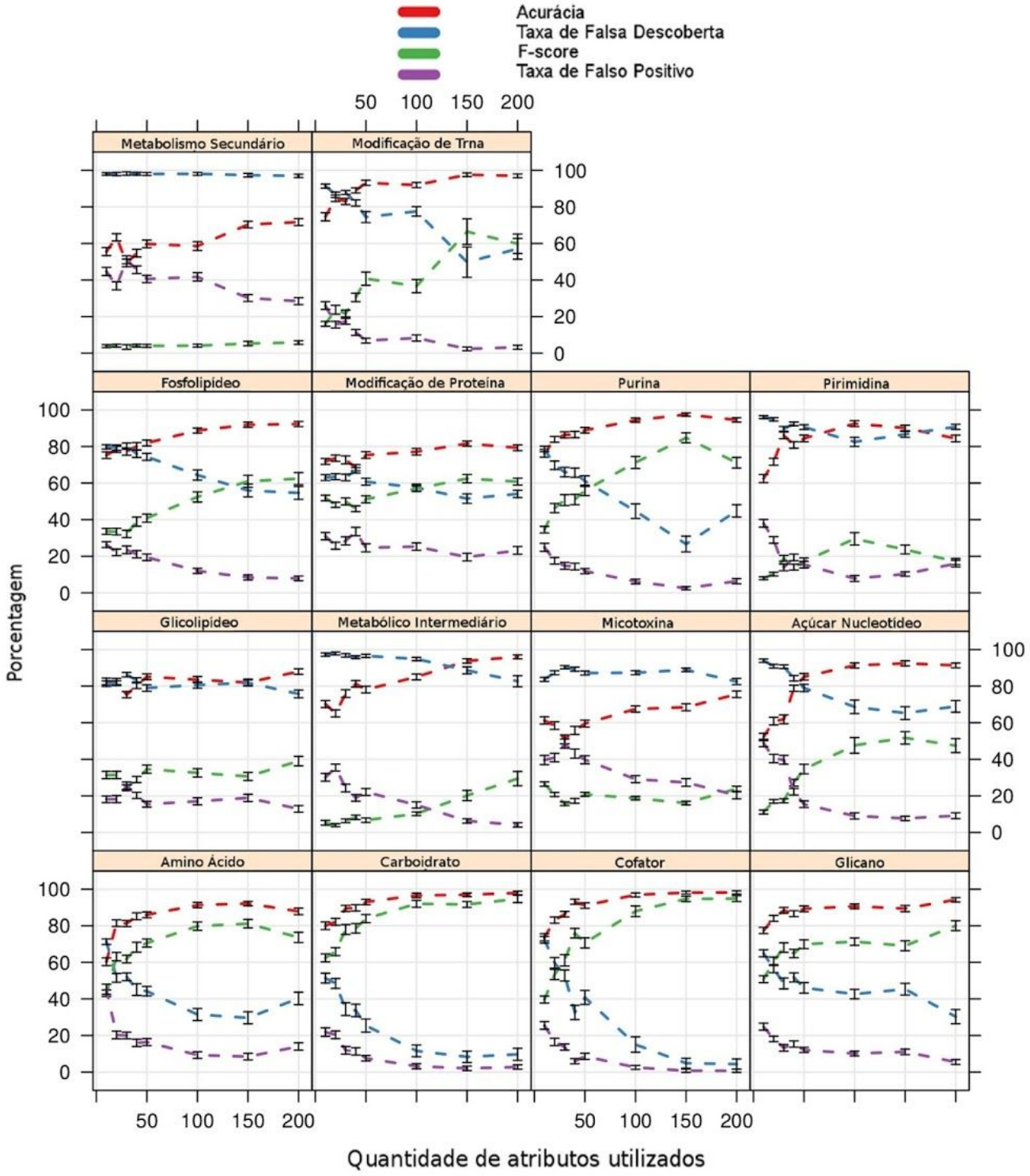
A


Acurácia
Taxa de Falsa Descoberta
F-score
Taxa de Falso Positivo
Metabolismo Secundário
Modificação de Trna
Fosfolipídeo
Modificação de Proteína
Purina
Pirimidina
Glicolipídeo
Metabólico Intermediário
Micotoxina
Açúcar Nucleotídeo
Amino Ácido
Carboidrato
Cofator
Glicano
Porcentagem
Quantidade de atributos utilizados

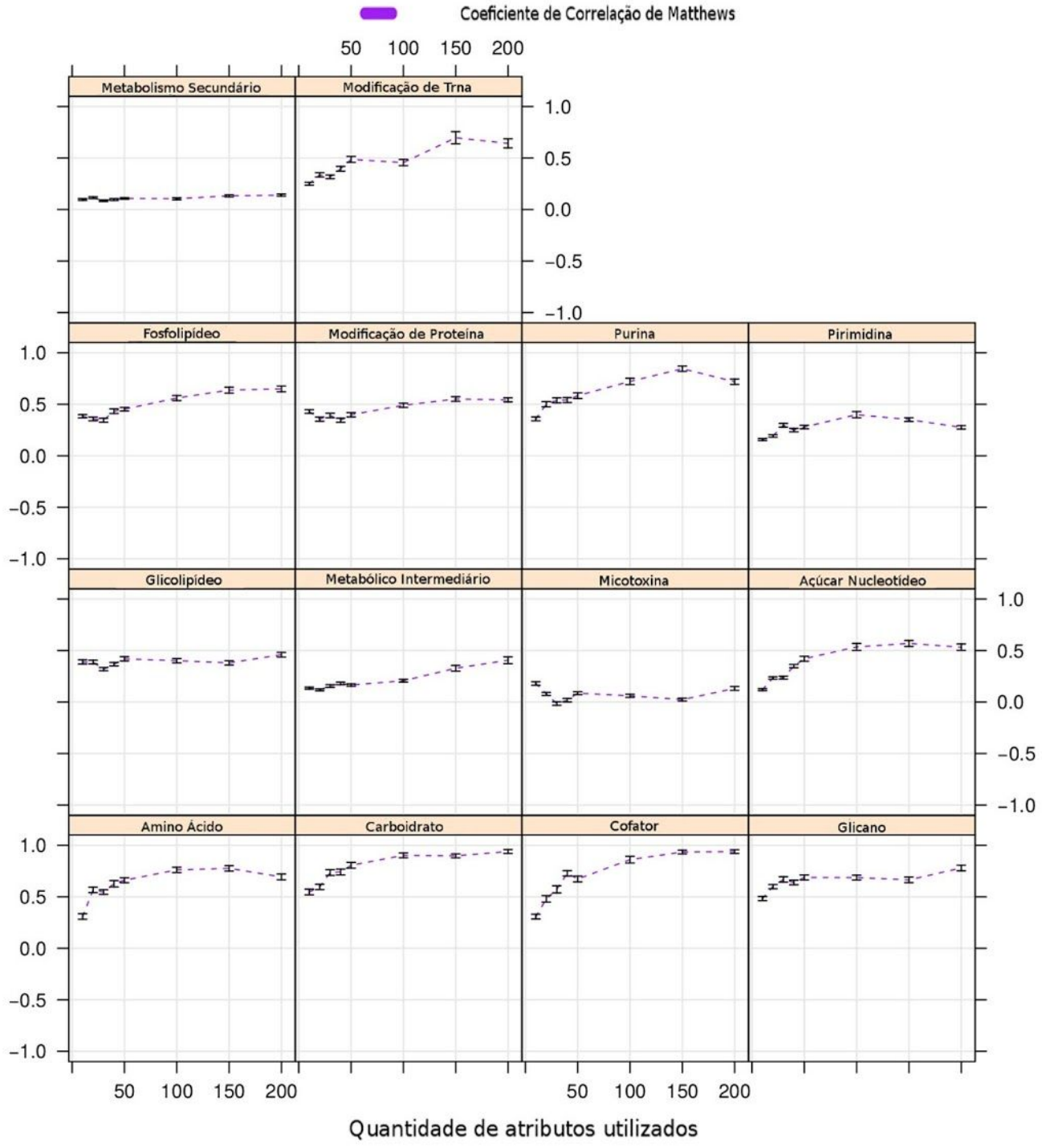
B


Coeficiente de Correlação de Matthews
Metabolismo Secundário
Modificação de Trna
Fosfolipídeo
Modificação de Proteína
Purina
Pirimidina
Glicolipídeo
Metabólico Intermediário
Micotoxina
Açúcar Nucleotídeo
Amino Ácido
Carboidrato
Cofator
Glicano
50 100 150 200
1.0 0.5 0.0 −0.5 −1.0
Quantidade de atributos utilizados

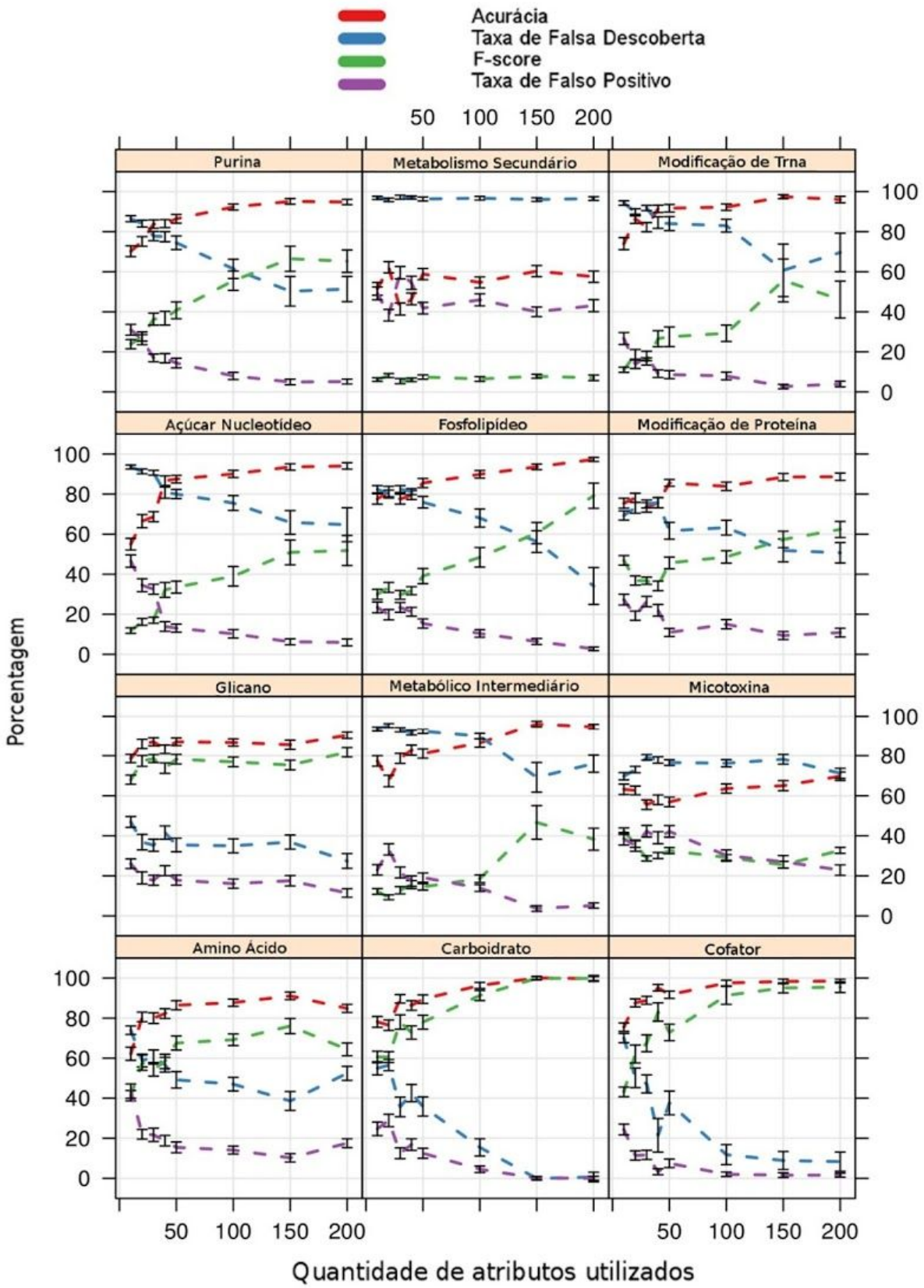
C


Acurácia
Taxa de Falsa Descoberta
F-score
Taxa de Falso Positivo
Metabolismo Secundário
Modificação de Trna
Fosfolipídeo
Modificação de Proteína
Purina
Pirimidina
Glicolipídeo
Metabólico Intermediário
Micotoxina
Açúcar Nucleotídeo
Amino Ácido
Carboidrato
Cofator
Glicano
Porcentagem
Quantidade de atributos utilizados

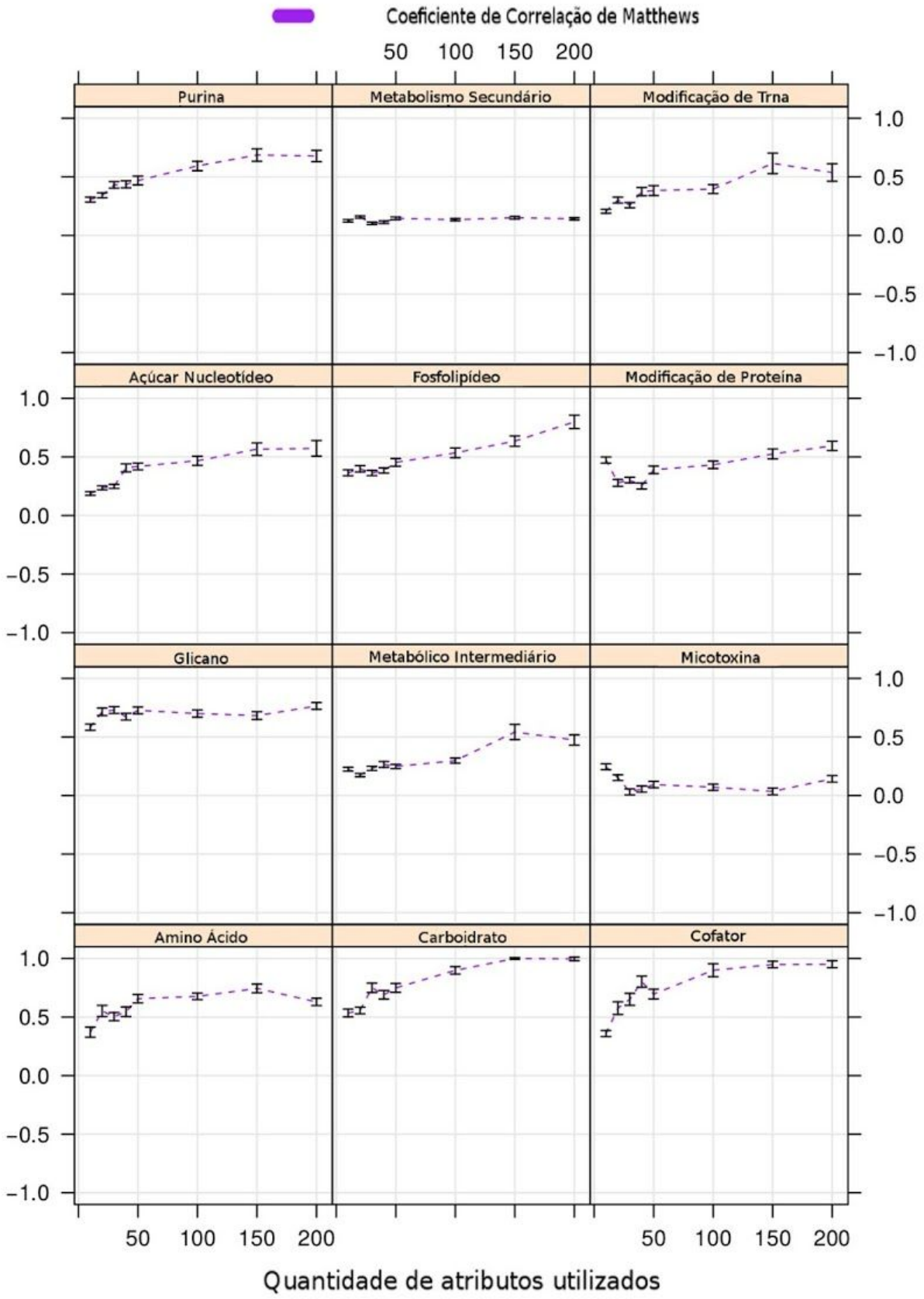
D


Coeficiente de Correlação de Matthews
Metabolismo Secundário
Modificação de Trna
Fosfolipídeo
Modificação de Proteína
Purina
Pirimidina
Glicolipídeo
Metabólico Intermediário
Micotoxina
Açúcar Nucleotídeo
Amino Ácido
Carboidrato
Cofator
Glicano
50 100 150 200
1.0 0.5 0.0 −0.5 −1.0
Quantidade de atributos utilizados

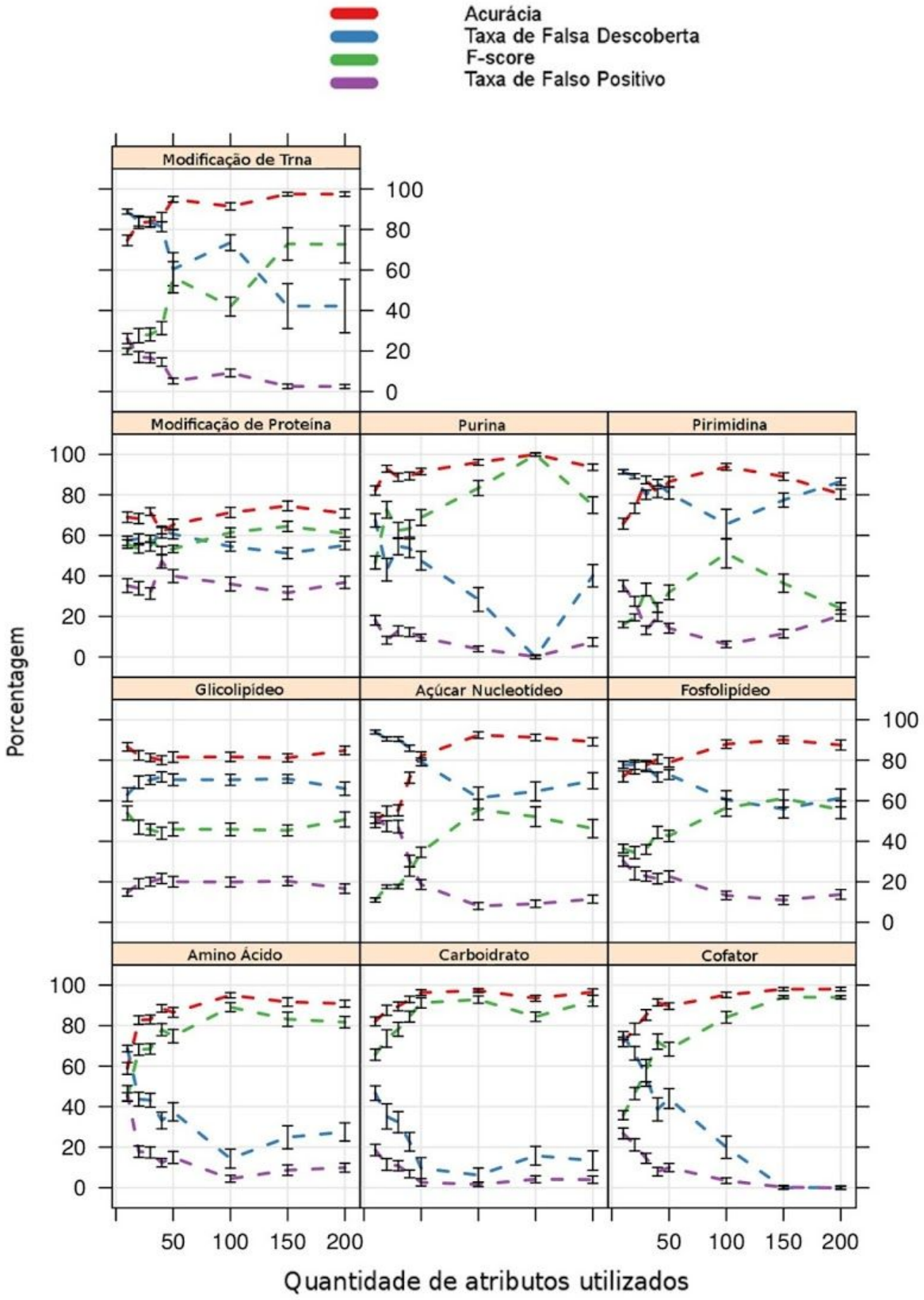
E


Acurácia
Taxa de Falsa Descoberta
F-score
Taxa de Falso Positivo
Purina
Metabolismo Secundário
Modificação de Trna
Açúcar Nucleotídeo
Fosfolipídeo
Modificação de Proteína
Glicano
Metabólico Intermediário
Micotoxina
Amino Ácido
Carboidrato
Cofator
Porcentagem
Quantidade de atributos utilizados

F


Coeficiente de Correlação de Matthews
Purina
Metabolismo Secundário
Modificação de Trna
Açúcar Nucleotídeo
Fosfolipídeo
Modificação de Proteína
Glicano
Metabólico Intermediário
Micotoxina
Amino Ácido
Carboidrato
Cofator
50 100 150 200
1.0 0.5 0.0 −0.5 −1.0
Quantidade de atributos utilizados

G


Acurácia
Taxa de Falsa Descoberta
F-score
Taxa de Falso Positivo
Modificação de Trna
Modificação de Proteína
Purina
Pirimidina
Glicolipídeo
Açúcar Nucleotídeo
Fosfolipídeo
Amino Ácido
Carboidrato
Cofator
Porcentagem
Quantidade de atributos utilizados
0
20
40
60
80
100
50
100
150
200

H

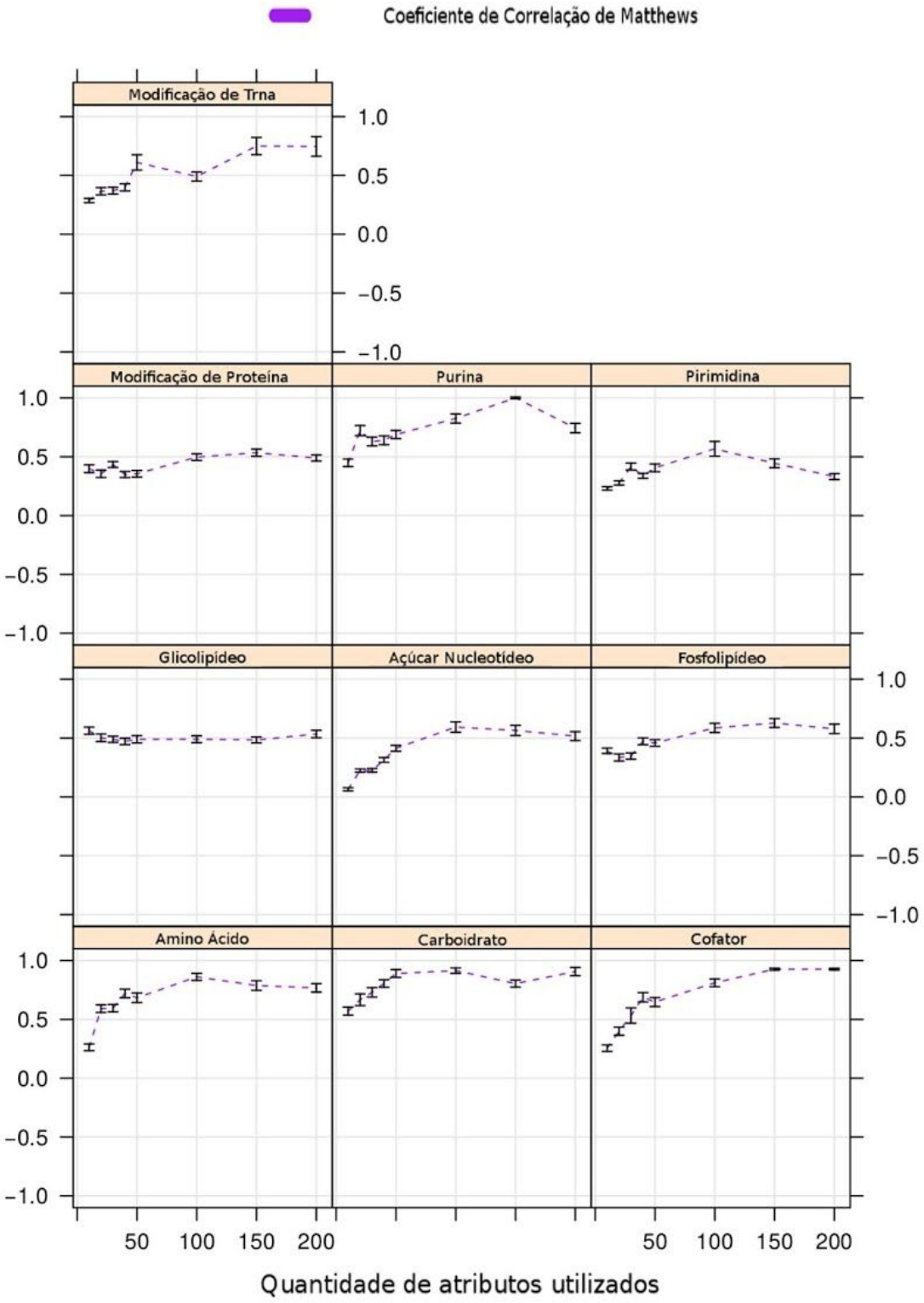

Coeficiente de Correlação de Matthews
Modificação de Trna
Modificação de Proteína
Purina
Pirimidina
Glicolipídeo
Açúcar Nucleotídeo
Fosfolipídeo
Amino Ácido
Carboidrato
Cofator
1.0
0.5
0.0
−0.5
−1.0
50
100
150
200
Quantidade de atributos utilizados

Figura 10. Aplicação dos modelos gerados para cada rota metabólica sobre os conjunto de dados. A e B - Conjunto de teste; C e D - Conjunto de validação com as duas espécies juntas; E e F - Conjunto de validação (espécie *Aspergillus clavatus* (strain ATCC 1007 CBS 513.65 DSM 816 NCTC)); G e H - Conjunto de validação (espécie *Debaryomyces hansenii* (strain ATCC 36239 CBS 767 JCM 1990 NBRC)).

Esse procedimento foi realizado para visualizar a real performance do modelo preditor, obtendo assim a média das performances juntamente com os devidos desvios padrões. Outro importante detalhe é que as instâncias positivas (para cada rota metabólica e para cada conjunto de dados) são as mesmas em todos os subconjuntos. Entretanto, cada subconjunto têm todas suas instâncias (positivas e negativas) aleatorizadas, tornando assim cada subconjunto diferente dos demais. As duas espécies, selecionadas aleatoriamente na etapa de formação do conjunto de validação, foram as espécies *Aspergillus clavatus* (strain ATCC 1007 CBS 513.65 DSM 816 NCTC) e *Debaryomyces hansenii* (strain ATCC 36239 CBS 767 JCM 1990 NBRC), sendo ausente as anotações da rota metabólica de glicolipídeo e pirimidina na primeira espécie citada; e glicano, metabólico intermediário, micotoxina e metabolismo secundário para a segunda espécie citada.

As métricas escolhidas para realizar a avaliação da performance dos resultados foram a acurácia, taxa de falsa descoberta, F-*score*, falso positivo, coeficiente de correlação de Matthews, sensitividade (recall) e especificidade. Com isso, dentro de cada conjunto de dados de teste e validação, foram selecionados os maiores valores de acurácia, F-*score*, coeficiente de correlação de Matthews, sensitividade (recall) e especificidade, ao mesmo tempo que foram selecionados os menores valores falso positivo e de taxa de falsa descoberta (Material Suplementar T16). Realizando esse procedimento, podemos verificar que a maior frequência de melhores performances se encontra com os modelos gerados com conjunto de treinamento contendo de 150 a 200 atributos (geralmente) para a maioria das rotas metabólicas (Tabela 8).

Tabela 8. Quantidade de atributos utilizados (maior frequência) nos modelos preditivos aplicados nos conjuntos de dados de teste e de validação. Ac – Acurácia; FD – Taxa de falsa descoberta; FP – Taxa de falso positivo; CCM – Coeficiente de correlação de Matthews.

Rota Metabólica	Teste					Validação				
	Ac	FD	F-score	FP	CCM	Ac	FD	F-score	FP	CCM
Amino Ácido	150	200	200	200	200	150	150 ou 200	150	100 ou 150	150
Carboidrato	150 ou 200	150 ou 200	200	200	200	200	150	200	150	200
Cofator	150	150	150	150	150	150 ou 200	150 ou 200	150 ou 200	150 ou 200	150 ou 200
Glicano	200	200	200	200	200	200	200	200	200	200
Glicolipídeo	200	200	200	200	200	200	200	200	200	200
Metabólico Intermediário	200	200	200	200	200	200	200	200	200	200
Micotoxina	200	200	200	200	200	200	10 ou 200	150 ou 200	200	10
Açúcar-Nucleotídeo	200	200	200	200	200	150	150	150	150	150
Fosfolipídeo	200	200	200	200	200	150 ou 200	150 ou 200	150 ou 200	150 ou 200	150 ou 200
Modificação de Proteína	200	200	200	150 ou 200	200	150	150	150	200	150 ou 200
Purina	150 ou 200	150	150	150 ou 200	150	150	150	150	150	150

Pirimidina	150	150	150	150	150	100	100	100	100	100
Metabolismo Secundário	200	200	200	200	200	200	10 até 100	150 ou 200	200	150 ou 200
Modificação de Trna	100 ou 150	100	100 ou 150	100	150	150	150	150	150	150

Com isso, os resultados médios obtidos do conjunto de validação foram de 94, 67 e 6% para acurácia, F-*score* e taxa de falso positivo, respectivamente (Tabela 9).

Tabela 9. Performance preditiva dos modelos aplicados sobre conjunto de dados de validação. Atrib - Quantidade de atributos utilizados para compor o modelo; Ac - Acurácia; FD - Taxa de falsa descoberta; FP - Taxa de falso positivo; CCM - Coeficiente de correlação de Matthews; Sens - Sensitividade (Recall); Esp - Especificidade.

	Atrib	Ac	FD	F-score	FP	CCM	Sens	Esp
amino ácido	150	0.92	0.30	0.81	0.08	0.78	0.96	0.92
carboidrato	200	0.98	0.08	0.95	0.03	0.94	1.00	0.97
cofator	150	0.98	0.05	0.95	0.01	0.93	0.94	0.99
glicano	200	0.94	0.30	0.80	0.06	0.78	0.94	0.95
metabólico intermediário	200	0.96	0.83	0.29	0.04	0.41	1.00	0.96
açúcar nucleotídeo	150	0.92	0.65	0.52	0.08	0.57	1.00	0.92
fosfolipídeo	150	0.92	0.56	0.61	0.08	0.64	1.00	0.92
modificação de proteína	150	0.82	0.52	0.62	0.20	0.55	0.91	0.77
purina	150	0.97	0.27	0.85	0.03	0.84	1.00	0.98
pirimidina	100	0.93	0.83	0.30	0.08	0.40	1.00	0.92
modificação de trna	150	0.98	0.50	0.66	0.02	0.70	1.00	0.98

Média	0.94	0.44	0.67	0.06	0.69	0.98	0.93

Fica evidenciado que ao realizar uma redução dimensional muito severa dos dados, tal como utilizar somente 10 a 50 atributos de maior relevância (aproximadamente 1 a 5% dos atributos), há uma considerável perda de performance preditiva. Em contrapartida, utilizando 150 e 200 atributos de maior relevância (aproximadamente 15 a 20% dos atributos), observamos as melhores performances de predição. Isso mostra a importância do processo de seleção de atributos para melhorar a performance de predição, conforme afirmado por Guyon e Elisseeff (2003) e Saeys et al. (2007).

Após esta etapa, baseado nos resultados obtidos do conjunto de validação, os melhores modelos de cada rota metabólica foram selecionados para compor a ferramenta mAppLe (*Metabolic Prediction of Pathway of Enzymes*), sendo:

- Aminoácido = utilizando 150 atributos;
- Carboidrato = utilizando 200 atributos;
- Cofator = utilizando 150 atributos;
- Glicano = utilizando 200 atributos;
- Glicolipídeo = utilizando 200 atributos;
- Metabólico Intermediário = utilizando 200 atributos;
- Micotoxina = utilizando 200 atributos;
- Açúcar-Nucleotídeo = utilizando 150 atributos;
- Fosfolipídeo = utilizando 150 atributos;
- Modificação de Proteína = utilizando 150 atributos;
- Purina = utilizando 150 atributos;
- Pirimidina = utilizando 100 atributos;
- Metabolismo Secundário = utilizando 200 atributos;
- Modificação de Trna = utilizando 150 atributos.

Das 16 rotas metabólicas inicialmente selecionadas para este trabalho, foram removidos das análises as rotas metabólicas de lipídeo e de enxofre, pois ambas

apresentaram problemas na geração dos modelos devido a alta exigência de *hardware*, tornando-se assim computacionalmente inviáveis.

5.7. Ferramenta mAppLe, sua aplicação e comparação com outros programas

Nesta etapa, podemos notar que instâncias positivas para uma determinada rota metabólica X podem ser confundidas como positivas em uma rota metabólica Y (Figura 11, instância Q0V118), indicando que as características (atributos) desta enzima X são muito próximas das características da enzima Y; ainda, podemos supor um comprometimento do modelo, o qual pode não estar sendo capaz de diferenciar as rotas metabólicas destas duas enzimas. De fato, a segunda explanação é mais plausível para justificar a retirada dos modelos referentes às rotas metabólicas de glicolipídeo, micotoxina e metabólito secundário (abordado na seção 4.9), onde estes modelos obtiveram resultados médios de 0,522 , 0,54 e 0,61 (respectivamente) de predição positiva para os seus controles.

Identificação	predição para amino ácido	predição para carboidrato	predição para cofator	predição para glicano	predição para metabólico intermediário	predição para açúcar nucleotídeo	predição para fosfolipídeo	predição para modificação de proteína	predição para purina	predição para pirimidina	predição para modificação de tma	Anotação Real
A5DTU8	0.852	0.442	0.311	0.187	0.061	0.124	0.364	0.523	0.362	0.531	0.615	amino ácido
K3VI22	0.300	0.746	0.142	0.221	0.144	0.451	0.309	0.505	0.438	0.309	0.420	carboidrato
A0A254UFM5	0.237	0.304	0.886	0.251	0.288	0.502	0.452	0.481	0.358	0.504	0.616	cofator
Q9HEZ1	0.214	0.350	0.160	0.594	0.318	0.392	0.342	0.458	0.400	0.250	0.211	glicano
G9P656	0.519	0.305	0.307	0.342	0.991	0.330	0.392	0.420	0.419	0.570	0.325	metabólico intermediário
A1CPD0	0.264	0.346	0.218	0.316	0.336	1.000	0.333	0.522	0.487	0.471	0.558	açúcar nucleotídeo
Q0V118	0.232	0.489	0.280	0.247	0.071	0.314	0.513	0.408	0.509	0.686	0.679	fosfolipídeo
I8I8C4	0.187	0.300	0.203	0.167	0.143	0.617	0.337	0.710	0.408	0.600	0.472	modificação de proteína
A0A060SWD4	0.120	0.397	0.235	0.143	0.270	0.559	0.304	0.390	0.744	0.402	0.353	purina
P05035	0.238	0.453	0.288	0.245	0.398	0.129	0.442	0.443	0.436	0.740	0.613	pirimidina
A0A0P9EXR1	0.315	0.456	0.256	0.147	0.034	0.116	0.312	0.390	0.454	0.505	0.976	modificação de tma
mAppLe-amino_ácido	0.866	0.252	0.204	0.263	0.061	0.395	0.400	0.520	0.440	0.323	0.447	
mAppLe-carboidrato	0.141	0.889	0.142	0.277	0.147	0.059	0.298	0.253	0.375	0.502	0.191	
mAppLe-cofator	0.358	0.216	0.934	0.189	0.118	0.003	0.247	0.476	0.344	0.437	0.255	
mAppLe-glicano	0.206	0.316	0.143	0.680	0.088	0.000	0.392	0.468	0.405	0.469	0.246	
mAppLe-metabólico_intermediário	0.500	0.330	0.342	0.276	0.954	0.427	0.451	0.406	0.419	0.410	0.569	
mAppLe-açúcar_nucleotídeo	0.185	0.352	0.186	0.271	0.103	1.000	0.387	0.493	0.381	0.469	0.448	
mAppLe-fosfolipídeo	0.238	0.361	0.255	0.224	0.067	0.003	0.747	0.344	0.346	0.275	0.254	
mAppLe-modificação_de_proteína	0.103	0.257	0.183	0.147	0.345	0.005	0.408	0.786	0.424	0.536	0.674	
mAppLe-purina	0.124	0.345	0.157	0.184	0.517	0.494	0.345	0.385	0.824	0.406	0.380	
mAppLe-pirimidina	0.357	0.534	0.302	0.335	0.393	0.010	0.257	0.338	0.449	0.906	0.172	
mAppLe-modificação_de_tma	0.132	0.259	0.222	0.188	0.118	0.522	0.308	0.484	0.428	0.364	0.966	

Figura 11. Exemplo de resultado fornecido pela ferramenta mAppLe (subconjunto 1 - Material Suplementar T18-19), mostrando as probabilidades de cada instância em pertencer a uma determinada rota metabólica, juntamente com as instâncias controle da ferramenta (destacado em amarelo).

Os exemplos utilizados para a aplicação da ferramenta mAppLe (como o que foi utilizado para gerar a figura 9) são subconjuntos do conjunto avaliador 2, que contém sequências de aminoácidos de enzimas das espécies que não foram selecionadas para serem utilizadas no trabalho (nas etapas de treinamento, teste e validação), conforme

descrito na seção 4.2. A ferramenta mAppLe realiza a devida preparação dos dados, aplica os modelos gerados no processo de aprendizagem supervisionada e fornece uma tabela de resultados das predições de uma instância para cada uma das 11 rotas metabólicas, não necessitando assim, realizar todo o procedimento de treino. Utilizando 10 subconjuntos do conjunto avaliador 2 (Material Suplementar T18), cada um contendo uma enzima (sequência de aminoácido) de cada uma das 11 rotas metabólicas, mAppLe obteve uma média de acerto de 76.4%, com desvio padrão de 13.6%.

Alguns programas para predição de enzimas (funções enzimáticas) foram publicados. Por exemplo, o programa PRIAM (CLAUDEL-RENARD et al., 2003) é capaz de predizer se uma proteína é uma enzima (especificidade e sensitividade média de 86 e 89%, respectivamente), assim como sua provável função enzimática (*EC number*), de genomas sequenciados, utilizando descritores oriundo do banco de dados ENZYME. Além disso, este programa reporta em qual rota metabólica a enzima predita atua, utilizando os dados e gráficos do banco de dados KEGG. Já o programa EFFICAZ (KUMAR and SKOLNICK, 2012) prediz enzimas e seu provável *EC number* sobre dados genômicos (acurácia e sensitividade média de 92 e 82%, respectivamente), baseados em seis descritores de enzimas, incluindo a detecção de funcionalidade de resíduos, domínios de proteínas (dos bancos de dados Pfam e Prosite) e uso de *Support Vector Machine* (aprendizado de máquina). ComPath (CHOI and KIM, 2008) é uma ferramenta *on line* (*Web Server*) baseado na integração de informações de diversos banco de dados (KEGG, Pfam, Prosite, SCOPEC e PDB), onde a sequência/*motif*/filogenia são comparadas para predizer as funções enzimáticas e suas rotas metabólicas (especificidade e sensitividade média de 18 e 86%, respectivamente). Outro programa *on line*, Pathway Analyst (PIREDDU et al., 2006) realiza predições de rotas bioquímicas utilizando *Support Vector Machine* , *Hidden Markov Models* e *Blast* (precisão e sensitividade média de 78.3 e 92.6%, respectivamente).

Todos os programas mencionados e similares, mostram altos valores de performance, mas sempre abordando duas métricas, nos quais tais informações podem confundir um leitor com pouca experiência na área. Por exemplo, dados que mostram altos valores de especificidade e sensitividade (como o trabalho de Claudel-Renard et al., (2003) e Choi e

Kim (2008)) podem ter também altos valores de taxa de falsa descoberta, baixo valor para coeficiente de correlação de Matthews e uma considerável taxa de falso positivo; dados que mostram altos valores de acurácia e sensitividade (como o trabalho de Kumar and Skolnick (2012)), podem ter também altos valores de taxa de falsa descoberta, baixo valor para coeficiente de correlação de Matthews; dados que mostram altos valores precisão e sensitividade (como o trabalho de Pireddu et al. (2006)) podem ter também uma considerável taxa de falsa descoberta. Apresentar mais de duas métricas permite melhor compreensão dos resultados, além de poder ressaltar as características referentes à base de dados utilizado. Conforme mostrado na tabela 9, as médias obtidas pelos modelos que compõem a ferramenta mAppLe foram de 94% de acurácia, 44% de taxa de falsa descoberta, 67% de F-score, 98% de sensitividade, 93% de especificidade e 0,69 para coeficiente de correlação de Matthews. A métrica coeficiente de correlação de Matthews é utilizada para mensurar a qualidade de classificações binárias, variando seus valores de -1 a +1, onde coeficientes mais próximos de +1 representam uma predição consistente, próximos de 0 representam predição randômica e próximas de -1 representam uma predição inconsistente (em desacordo). Percebe-se também da impossibilidade de comparar as performances (*benchmark*) da ferramenta mAppLe, uma vez que mAppLe realiza as predições das rotas metabólicas, enquanto os demais programas citados realizam a predição da função enzimática (*EC number*).

Os programas anteriormente citados se baseiam em alinhamentos de sequência/estrutura, busca por similaridades por *Blast* e *Hmmer*, descrições de rotas metabólicas utilizando *EC numbers* de banco de dados (como KEGG) e busca por *motifs* e domínios de proteínas. Todos esses parâmetros e abordagens não são comuns, aplicáveis, ou confiáveis a todo tipo de sequência. Sendo assim, se torna impossível determinar a rota metabólica de uma enzima sem que haja similaridade com uma enzima bem definida e conhecida nos bancos de dados ou que não contenha os dados de sua função enzimática (*EC number*). Além disso, predições de rotas bioquímicas ou de funções enzimáticas baseados somente na similaridade de sequências podem não ser tão úteis, pois similaridades podem não estar relacionadas à homologia, analisando por uma perspectiva evolutiva (PEARSON, 2013). Contudo, análises filogenéticas em larga escala são muito

complexas (REGIER et al., 2013) e considerando a era pós-genômica, fica evidente que a maioria das espécies tem diversas sequências específicas (LIEW et al., 2010). Com isso, a maior parte dos programas podem restringir um número de enzimas propensas a uma determinada rota metabólica.

De modo a evitar problemas com similaridade de sequências, como exposto acima, mAppLe é totalmente baseado em características da estrutura primária das enzimas e com uma abordagem de aprendizado de máquina, tornando-o aplicável à maioria das enzimas de fungos. Utilizado com sucesso em vários campos da ciência, o aprendizado de máquina vem se destacando cada vez mais, como mencionado na Reunião de Anotação Funcional sobre os métodos utilizados em predições de função de proteínas (JIANG et al., 2016).

Usuários do mAppLe podem utilizar sequências de enzimas com rota metabólica desconhecida (em arquivos do tipo fasta, tab ou csv) e aplicar os modelos nele existente. Durante o processo, mAppLe automaticamente calcula e determina os atributos, realiza a normalização dos dados, seleciona os atributos específicos para cada modelo e ao final, fornece uma tabela com probabilidades de cada rota metabólica para cada sequência (Figura 9).

6. CONCLUSÃO

As performances médias de predição (sobre o conjunto de validação) dos modelos que compõe a ferramenta mAppLe foram de 94% de acurácia, 44% de taxa de falsa descoberta, 67% de F-*score*, 98% de sensitividade, 93% de especificidade e 0,69 para coeficiente de correlação de Matthews. A taxa de acerto de classificação correta da ferramenta mAppLe foi de 76,4%. Com uma abordagem completamente diferenciada, esta ferramenta poderá superar os problemas encontrados por outros programas (por se basearem na similaridade de sequências). Entretanto, futuras melhorias ainda serão feitas no mAppLe, de modo a ampliar as rotas metabólicas a serem preditas, aumentar a performance de predição e inferir a função enzimática das amostras analisadas.

7. REFERÊNCIAS

ALBERTS, B., JOHNSON, A., LEWIS, J., MORGAN, D., RAFF, M., ROBERTS, K., WALTER, P. Cell chemistry and bioenergetics. In: Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P. **Molecular biology of the cell**. New York: Garland Science, Taylor and Francis group, 6th edition, 2015, p. 51-73.

ANGERMUELLER, C., PÄRNAMAA, T., PARTS, L., STEGLE, O. Deep learning for computational biology. **Molecular Systems Biology**, v. 12, p. 878, 2016.

ARAKAKI, A.K., HUANG, Y., SKOLNICK, J. EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. **BMC bioinformatics**, 10:107, 2009.

AZUAJE, F. Computational models for predicting drug responses in cancer research. **Briefings in Bioinformatics**, v. 18, 820-829, 2017.

BULJAN, M., BATEMAN, A. The evolution of protein domain families. **Biochemical Society Transactions**, v. 37, p. 751-755, 2009.

CLAUDEL-RENARD, C., CHEVALET, C., FARAUT, T., KAHN, D. Enzyme-specific profiles for genome annotation: PRIAM. **Nucleic Acids Research**, 31, p. 6633-6639, 2003.

DOBSON, P.D., DOIG, A.J. Predicting Enzyme Class From Protein Structure Without Alignments. **Journal of Molecular Biology**, v. 345, p. 187-199, 2005.

FABRIS, F., MAGALHÃES, J.P.F., ALEX, A. A review of supervised machine learning applied to ageing research. **Biogerontology**, v. 18, p. 171-188, 2017.

FILHO, M.A.C.P.P. Metagenômica: princípios e aplicações. In: Faleiro, F.G., Andrade, S.R.M., Reis Júnior, F.B. Biotecnologia: estado da arte e aplicações na agropecuária. Planaltina-DF: Embrapa Cerrados, 2011, p. 174-193.

FREILICH, S., SPRIGGS, R.V., GEORGE, R.A., AL-LAZIKANI, B., SWINDELLS, M., THORNTON, J.M. The complement of enzymatic sets in different species. **Journal of Molecular Biology**, v. 349, p. 745-763, 2005.

FRIEDBERG, I. Automated protein function prediction - The genomic challenge. **Briefings in Bioinformatics**, v. 7, p. 225-242, 2006.

GINSBURG, H. Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium. **Trends in Parasitology**, v. 25, p. 37-43, 2009.

GLASNER, M.E., GERLT, J.A., BABBITT, P.C. Evolution of enzyme superfamilies. **Current Opinion in Chemical Biology**, v. 10, p. 492-497, 2006.

GUTERL, J.K., GARBE, D., CARSTEN, J., STEFFLER, F., SOMMER, B., REISSE, S., PHILIPP, A., HAACK, M., RÜHMANN, B., KOLTERMANN, A., KETTLING, U., BRÜCK, T., SIEBER, V. Cell-free metabolic engineering: Production of chemicals by minimized reaction cascades. **ChemSusChem**, v. 5, p. 2165-2172, 2012.

GUYON, I., ELISSEEFF, A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, v.3, p.1157-1182, 2003.

HECKMANN, D., SCHLÜTER, U., WEBER, A.P.M. Machine Learning Techniques for PredictingCrop Photosynthetic Capacity from Leaf Reflectance Spectra. **Molecular Plant**, v. 10, p. 878-890, 2017.

JIANG, Y., ORON, T.R., CLARK, W.T., BANKAPUR, A.R., D'ANDREA, D., LEPORE, R., FUNK, CHRISTOPHER, S., KAHANDA, I., VERSPOOR, K.M., BEN-HUR, A., et al.. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. **Genome biology**, 17:184, 2016.

KANDOI, G., ACENCIO, M.L., LEMKE, N. Prediction of druggable proteins using machine learning and systems biology: A mini-review. **Frontiers in Physiology**, v. 6, p. 366, 2015.

KANEHISA, M., GOTO, S., KAWASHIMA, S., NAKAYA, A. The KEGG databases at GenomeNet. **Nucleic Acids Research**, v. 30, p. 42-46, 2002.

KARIMPOUR-FARD, A., EPPERSON, L.E., HUNTER, L.E. A survey of computational tools for downstream analysis of proteomic and other omic datasets. **Human Genomics**, v. 9 (1), p. 28, 2015.

KARP, P.D., RILEY, M., SAIER, M., PAULSEN, I.T., PALEY, S.M.; PELLEGRINI-TOOLE, A. The EcoCyc and MetaCyc databases. **Nucleic Acids Research**, v. 28, p. 56-59, 2000.

KORMAN, T.P., Opgenorth, P.H., Bowie, J.U. A synthetic biochemistry platform for cell free production of monoterpenes from glucose. **Nature Communications**, v. 8, p. 15526, 2017.

LARRAÑAGA, P., CALVO, B., SANTANA, R., BIELZA, C., GALDIANO, J., INZA, I., LOZANO, J.A., ARMAÑANZAS, R., SANTAFÉ, G., PÉREZ, A., ROBLES, V. Machine learning in bioinformatics. **Briefings in Bioinformatics**, v. 7, p. 86-112, 2006.

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

LODISH, H., BERK, A., MATSUDAIRA, P., KAISER, C.A., KRIEGER, M., SCOTT , M.P., ZIPURSKY, S.L., DARNELL, J. Protein structure and function. In: Lodish, H., Berk, A., Matsudaira, P., Kaiser, C.A., Krieger, M., Scott , M.P., Zipursky, S.L., Darnell, J., Molecular Cell Biology. W.H. Freeman and Company, 5th edition, 2004,p. 59-99.

MALHIS, N., WONG, E.T.C., NASSAR, R., GSPONER, J. Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule. **PlosOne**, v. 10, p. 1-15, 2015.

McDONALD, A.G., BOYCE, S. AND TIPTON, K.F. ExplorEnz: the primary source of the IUBMB enzyme list. **Nucleic Acids Research**, v. 37, p. 593-597, 2009.

NELSON, D.L.; COX, M.M. Enzimas. In: NELSON, D.L.; COX, M.M. Lehninger Principles of Biochemistry. New York: W. H. Freeman and Company, 5th edition, 2011. p. 183-234

NIELSEN, J., FUSSENEGGER, M., KEASLING, J., LEE, S.Y., LIAO, J.C., PRATHER, K., PALSSON, B. Engineering synergy in biotechnology. **Nature Chemical Biology**. v. 10, p. 319-322, 2014.

ORTH, J.D., THIELE, I., PALSSON, B.Ø. What is flux balance analysis? **Nature Biotechnology**. v. 28, p. 245-248, 2010.

PEREGRIN-ALVAREZ, J., MANUEL TSOKA, S., OUZOUNIS, C.A. The phylogenetic extent of metabolic enzymes and pathways. **Genome Research**, v. 13, p. 422-427, 2003.

PIREDDU, L., SZAFRON, D., LU, P., GREINER, R. The Path-A metabolic pathway prediction web server. **Nucleic Acids Research**, v.34, p.714-719, 2006.

PLANES, F.J.,BEASLEY, E. Path finding approaches and metabolic pathways. **Discrete Applied Mathematics**. v.157, p. 2244-2256, 2009.

POPTSOVA, M.S., GOGARTEN, J.P. Using comparative genome analysis to identify problems in annotated microbial genomes. **Microbiology**, v.156, p.1909-1917, 2010.

QUESTER, S., SCHOMBURG, D. EnzymeDetector: an integrated enzyme function prediction tool and database. **BMC Bioinformatics**, 12:376, 2011.

ROST, B. Enzyme function less conserved than anticipated. **Journal of Molecular Biology**, v. 318, p. 595-608, 2002.

SAEYS, Y., INZA, I., LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, v. 23, p. 2507-2517, 2007.

SCHOMBURG, I., CHANG, A. SCHOMBURG, D. BRENDA, enzyme data and metabolic information. Nucleic Acids Research, v. 30, p. 47-49, 2002.

SCHREIBER, F. Visual comparison of metabolic pathways. **Journal of Visual Languages and Computing**. v. 14, p.327-340, 2003.

STEPHANOPOULOS, G., KEASLING, J., GONZALEZ, R. Metabolic Engineering and Synthetic Biology in Strain Development. **ACS Synthetic Biology**, v. 1, p. 491-492, 2012.

TANIGUCHI, H., OKANO, K., HONDA, K. Modules for *in vitro* metabolic engineering: Pathway assembly for bio-based production of value-added chemicals. Synthetic and Systems **Biotechnology**, v. 2, p. 65-74, 2017.

THE UNIPROT CONSORTIUM. UniProt: the universal protein knowledgebase, **Nucleic Acids Research**, v. 45, p. 158-169, 2017.

TIAN, W., ARAKAKI, A.K., SKOLNICK, J. EFICAz: A comprehensive approach for accurate genome-scale enzyme function inference. **Nucleic Acids Research**. v. 32, p. 6226-6239, 2004.

TIPTON, K., BOYCE, S. History of the enzyme nomenclature system. **Bioinformatics**, v. 16, p. 34-40, 2000.

VARSHAVSKY, R., GOTTLIEB, A., LINIAL, M., HORN, D. Novel unsupervised feature filtering of biological data. **Bioinformatics**, v. 22, p. 507-513, 2006.

WU, G., YAN, Q., JONES, J.A., TANG, Y.J., FONG, S.S., KOFFAS, M.A.G. Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. **Trends in Biotechnology**, v. 34, p. 652-664, 2016.

YANG, J., YAN, R., ROY, A., XU, D., POISSON, J., ZHANG, Y. The I-TASSER Suite: protein structure and function prediction. **Nature Methods**, v. 12, p. 7-8, 2014.

ZHANG, L., TAN, J., HAN, D., ZHU, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. **Drug Discovery Today**, 2017, In press. https://doi.org/10.1016/j.drudis.2017.08.010

ZHANG, Y.H.P., SUN, J., ZHONG, J.J. Biofuel production by *in vitro* synthetic enzymatic pathway biotransformation. **Current Opinion in Biotechnology**, v. 21, p. 663-669, 2010.

8. PRODUÇÕES CIENTÍFICAS

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review of combinations of experimental and computational techniques to identify and understand genes involved in innate immunity and effector-triggered defence

Henrik U. Stotz [a,*], Rodrigo de Oliveira Almeida [b], Neil Davey [c], Volker Steuber [c], Guilherme T. Valente [b]

[a] School of Life and Medical Sciences, University of Hertfordshire, Hatfield AL10 9AB, UK
[b] Department of Bioprocess and Biotechnology, São Paulo State University (Unesp), School of Agriculture, Botucatu, Brazil
[c] Centre for Computer Science and Informatics Research, University of Hertfordshire, Hatfield AL10 9AB, UK

ARTICLE INFO

Article history:
Received 14 April 2017
Received in revised form 24 August 2017
Accepted 28 August 2017
Available online xxxx

Keywords:
Breeding
Graph theory
Receptor-like protein
Systems biology

ABSTRACT

The innate immune system includes a first layer of defence that recognises conserved pathogen-associated molecular patterns that are essential for microbial fitness. Resistance (*R*) gene-based recognition of pathogen effectors, which function in modulation or avoidance of host immunity, activates a second layer of plant defence. In this review, experimental and computational techniques are considered to improve understanding of the plant immune system. Biocomputation contributes to discovery of the molecular genetic basis of host resistance against pathogens. Sequenced genomes have been used to identify *R* genes in plants. Resistance gene enrichment sequencing based on conserved protein domains has increased the number of *R* genes with nucleotide-binding site and leucine-rich repeat domains. Network analysis will contribute to an improved understanding of the innate immune system and identify novel genes for partial disease resistance. Machine learning algorithms are expected to become important in defining aspects of the immune system that are less well characterised, including identification of *R* genes that lack conserved protein domains.

Chromosoma (2017) 126:73–81
DOI 10.1007/s00412-016-0613-6

REVIEW

B chromosomes: from cytogenetics to systems biology

Guilherme T. Valente[1] · Rafael T. Nakajima[2] · Bruno E. A. Fantinatti[2] · Diego F. Marques[2] · Rodrigo O. Almeida[1] · Rafael P. Simões[1] · Cesar Martins[2]

Received: 1 July 2016 / Revised: 10 August 2016 / Accepted: 15 August 2016 / Published online: 24 August 2016

Abstract Though hundreds to thousands of reports have described the distribution of B chromosomes among diverse eukaryote groups, a comprehensive theory of their biological role has not yet clearly emerged. B chromosomes are classically understood as a sea of repetitive DNA sequences that are poor in genes and are maintained by a parasitic-drive mechanism during cell division. Recent developments in high-throughput DNA/RNA analyses have increased the resolution of B chromosome biology beyond those of classical and mo-

The legacy of B chromosomes

B chromosomes (Bs) are enigmatic accessory elements to the regular chromosome set (A) and, since their discovery at the beginning of the twentieth century (Wilson 1907), Bs have ranked among the main topics of chromosome biology. The importance of B chromosomes is illustrated by the series of conferences on B chromosomes that have been organized during the last three decades (1st, 2nd and 3rd B chromosome

SCIENTIFIC REPORTS

OPEN

HIV Reverse Transcriptase and Protease Genes Variability Can Be a Biomarker Associated with HIV and Hepatitis B or C Coinfection

Received: 17 January 2018
Accepted: 10 May 2018
Published online: 29 May 2018

Natália Mirele Cantão[1], Lauana Fogaça de Almeida[2], Ivan Rodrigo Wolf[2], Rodrigo Oliveira Almeida[2], Andressa Alves de Almeida Cruz[1], Caroline Nunes[1], Alexandre Naime Barbosa[1], Guilherme Targino Valente[2], Maria Inês de Moura Campos Pardini[1] & Rejane Maria Tommasini Grotto[1,2]

Variability of the HIV reverse transcriptase (RT) and protease (PR) genes has been used as indicators of drug resistance and as a mean to evaluate phylogenetic relationships among circulating virus. However, these studies have been carried in HIV mono-infected populations. The goal of this study