UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

Renan dos Reis

Redes de meta-modelagem e suas aplicações no estudo de anotações de proteínas

São Carlos

2023

Renan dos Reis

Redes de meta-modelagem e suas aplicações no estudo de anotações de proteínas

Dissertação apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Mestre em Ciências.

Área de concentração: Física Biomolecular

Orientador: Prof. Dr. Luciano da Fontoura Costa

Versão corrigida

(Versão original disponível na Unidade que aloja o Programa)

São Carlos

2023

Reis, Renan dos
Redes de meta-modelagem e suas aplicações no estudo de anotações de proteínas / Renan dos Reis; orientador Luciano da Fontoura Costa - versão corrigida -- São Carlos, 2023.
110 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Física Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2023.

1. Meta-modelagem. 2. Ciência de redes. 3. Reconhecimento de padrões. 4. Anotação de proteínas. 5. Enzimas ativas em carboidratos. I. Costa, Luciano da Fontoura, orient. II. Título.

Este trabalho é dedicado aos amigos que me acompanharam, me divertiram e me escutaram, deixando esses últimos dois anos muito melhores.

AGRADECIMENTOS

Agradeço à minha família, pelo grande apoio ao longo de toda minha carreira acadêmica.

Agradeço ao meu irmão, pela sintonia, amizade e conversas que me fazem ver a vida de uma maneira mais simples.

Agradeço aos meus amigos, sobretudo o João Paulo, a Giulia, o Danilo e o Yuri, que me ajudaram a passar por todos os momentos difíceis e celebraram comigo todos os bons momentos ao longo desses anos.

Agradeço ao meu orientador Prof. Dr. Luciano da Fontoura Costa, por orientar a minha pesquisa em diversos temas interessantíssimos, com dedicação e atenção.

Agradeço ao Instituto de Física de São Carlos da Universidade de São Paulo, pela oportunidade de seguir uma carreira acadêmica multidisciplinar.

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

"Nada na biologia faz sentido exceto à luz da evolução"
Theodosius Dobzhansky

"Todos os modelos estão errados, mas alguns são úteis"
George Box

RESUMO

REIS, R. **Redes de meta-modelagem e suas aplicações no estudo de anotações de proteínas**. 2023. 110p. Dissertação (Mestrado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2023.

A crescente disponibilidade de dados tem motivado o desenvolvimento de novas abordagens de modelagem para sua análise e interpretação, incluindo métodos estatísticos, mineração de dados e aprendizado de máquina. Apesar de serem particularmente eficazes, os modelos baseados em dados tendem a ter interpretabilidade limitada, o que pode dificultar a compreensão de suas previsões. Para lidar com essas questões, este trabalho foca na extensão e aplicação de uma abordagem formal de meta-modelagem que possa fornecer subsídios para caracterizar, melhorar e integrar modelos baseados em dados. O procedimento proposto consiste na aplicação de ciência de redes na construção de uma rede de meta-modelagem que conecta conjuntos de dados a modelos científicos. Primeiro, a meta-modelagem envolve a delimitação de três domínios: um domínio de universo que contém todos os dados acessíveis para modelagem, um ambiente de dados com conjuntos de dados organizados, e uma estrutura de modelagem capaz de explicar esse ambiente de dados. Depois disso, a rede é construída com base em duas operações: a associação bijetiva entre conjuntos de dados e modelos (resumida no conceito de cartucho) e a conexão entre os elementos de cada conjunto de dados e cada modelo. Com essas propriedades, a rede permite avaliar quantitativamente a interação entre modelos na mesma estrutura de modelagem, além de facilitar a criação de novos modelos por meio da correspondência entre operações lógicas de modelos e operações entre conjuntos de dados. Esta abordagem foi aplicada a dois problemas de modelagem. No primeiro caso, o foco estava no reconhecimento de padrões em sequências binárias. Nele, descrevemos detalhadamente a interação entre seis modelos de padrões, além de derivar um modelo preciso para um conjunto de dados usando uma composição lógica de modelos pré-existentes, o que mostra o potencial dessa abordagem para estudar a detecção de padrões em sequências de símbolos. No segundo caso, o método foi aplicado para auxiliar a análise exploratória da anotação de domínios de proteínas em enzimas ativas em carboidratos, presente no banco de dados CAZy. O estudo desse meta-modelo revelou informações sobre a modularidade das classes funcionais e suas relações evolutivas e funcionais. Coletivamente, esses resultados indicam que a rede de meta-modelagem desenvolvida tem potencial para auxiliar na caracterização e aprimoramento da modelagem científica em múltiplas áreas, com aplicações promissoras para a análise de anotação de proteínas.

Palavras-chave: Meta-modelagem. Ciência de redes. Reconhecimento de padrões. Anotação de proteínas. Enzimas ativas em carboidratos.

ABSTRACT

REIS, R. **Meta-modeling networks and their applications in the study of protein annotations**. 2023. 110p. Dissertation (Master in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2023.

The growing availability of data has motivated the development of new modeling approaches for its analysis and interpretation, including statistical methods, data mining, and machine learning. Despite being particularly effective, data-driven models tend to have limited interpretability, which can make their predictions difficult to understand. To deal with these issues, this work focuses on the extension and application of a meta-modeling formal approach that can provide subsidies to characterize, improve and integrate data-based models. The proposed procedure consists of applying network science in the construction of a meta-modeling network that connects datasets to scientific models. First, the meta-modeling involves delimiting three domains: a universe domain that contains all data accessible for modeling, a data environment with organized datasets, and a modeling framework capable of explaining this data environment. After that, the network is built based on two operations: the bijective association between datasets and models (summarized in the concept of cartouche) and the connection between the elements of each dataset and each model. With these properties, the network enables to quantitatively evaluate the interaction between models in the same modeling structure, in addition to facilitating the creation of new models through the correspondence between logical operations between models and set operations between datasets. This approach was applied to two modeling problems. In the first case, the focus was on pattern recognition in binary sequences. In this problem, we describe in detail the interaction between six models of patterns, in addition to deriving an accurate model for a dataset using a logical composition of pre-existing models, which shows the potential of this approach to study pattern detection in sequences of symbols. In the second case, the method was applied to aid the exploratory analysis of a protein domain annotation in carbohydrate-active enzymes, available in the CAZy database. The study of this meta-model revealed information about the modularity of functional classes and their evolutionary and functional relationships. Collectively, these results indicate that the developed meta-modeling network has the potential to aid in the characterization and improvement of scientific modeling in multiple areas, with promising applications for protein annotation analysis.

Keywords: Meta-modeling. Network science. Pattern recognition. Protein annotation. Carbohydrate-active enzymes.

LISTA DE FIGURAS

Figura 1 – Os principais domínios da modelagem científica considerados neste trabalho. Para modelar um fenômeno de interesse, começa-se observando o fenômeno, coletando conjuntos de dados e criando modelos baseados nesses dados. O conjunto Ω corresponde ao universo de dados, que contém todos os dados acessíveis relacionados a um fenômeno. Dados de interesse são coletados experimentalmente do universo Ω e organizados em conjuntos de dados, que serão dispostos no ambiente de dados A. A partir desses conjuntos, modelos podem ser construídos e combinados em uma estrutura de modelagem, E, que é aplicada para explicar o fenômeno em questão. 24

Figura 2 – Grafo representando uma rede social entre membros de um clube universitário de karatê. 26

Figura 3 – Representação de domínios encontrada em uma proteína hipotética. . . 29

Figura 4 – Alguns tipos de grafos. 32

Figura 5 – Exemplo prático de meta-modelagem de talheres. 33

Figura 6 – Exemplo de meta-modelagem de quadriláteros. 37

Figura 7 – A incorporação de novos conjuntos de dados no ambiente de dados A. . 39

Figura 8 – A diversidade das conexões $\mathcal{D}$ de um modelo m indica o número de conjuntos de dados que podem se conectar a um modelo $\tilde{m}$ com uma distribuição uniforme de pesos p, mantendo a mesma entropia observada nas conexões de m. 45

Figura 9 – Quatro exemplos de sequências de 9 símbolos binários. 52

Figura 10 – Rede de meta-modelagem de padrões em sequências de símbolos binários. 58

Figura 11 – As 27 sequências de símbolos binários explicados pela composição lógica dos modelos $m_X(m_5) = \left(\bigwedge_{i\neq 6} \neg m_i\right) \vee (((m_1 \wedge m_2) \vee m_4) \wedge m_5)$ que não são explicadas pelo modelo m_5. 61

Figura 12 – Rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos. 63

Figura 13 – Distribuição dos 8680 elementos inespecíficos do conjunto de dados da classe GT nos subconjuntos de dados aos quais pertencem. 70

Figura 14 – Distribuição dos 5490 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e GH. 71

Figura 15 – Distribuição dos 2006 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e CE. 73

Figura 16 – Distribuição dos 1942 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e CBM. 74

Figura 17 – Distribuição dos 6 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e PL. 76
Figura 18 – Distribuição dos 3 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e AA. 77
Figura 19 – Distribuição dos 2907 elementos inespecíficos do conjunto de dados da classe PL nos subconjuntos de dados aos quais pertencem. 78
Figura 20 – Distribuição dos 2270 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes PL e CBM. 79
Figura 21 – Distribuição dos 543 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes PL e CE. 80
Figura 22 – Distribuição dos 126 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes PL e GH. 81
Figura 23 – Distribuição dos 10391 elementos inespecíficos do conjunto de dados da classe CE nos subconjuntos de dados aos quais pertencem. 82
Figura 24 – Distribuição dos 6173 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes CE e GH. 83
Figura 25 – Distribuição dos 2667 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes CE e CBM. 85
Figura 26 – Distribuição dos 25 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes CE e AA. 86
Figura 27 – Distribuição dos 174778 elementos inespecíficos do conjunto de dados da classe GH nos subconjuntos de dados aos quais pertencem. 87
Figura 28 – Distribuição dos 164004 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GH e CBM. 88
Figura 29 – Distribuição dos 30 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GH e AA. 90
Figura 30 – Distribuição dos 5576 elementos inespecíficos do conjunto de dados da classe AA nos subconjuntos de dados aos quais pertencem. 92
Figura 31 – Distribuição dos 5536 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes AA e CBM. 93
Figura 32 – Distribuição dos 176112 elementos inespecíficos do conjunto de dados da classe CBM nos subconjuntos de dados aos quais pertencem. 94

LISTA DE TABELAS

Tabela 1 – As principais correspondências entre operações de conjuntos e operações lógicas. 41
Tabela 2 – Exemplos de relações entre operações de conjuntos de dados no ambiente A e operações lógicas entre modelos na estrutura de modelagem E. . . . 41
Tabela 3 – Descrição dos modelos respectivos a cada conjunto de dados de sequências de símbolos binários. 51
Tabela 4 – Descrição dos modelos respectivos a cada conjunto de dados de domínios funcionais de enzimas ativas em carboidratos. 56
Tabela 5 – Métricas descritivas da rede de meta-modelagem de padrões em sequências de símbolos binários. 59
Tabela 6 – Métricas descritivas para cada par de conjunto de dados e modelo disponíveis na rede de meta-modelagem de padrões em sequências de símbolos binários. 59
Tabela 7 – Valores máximos de coincidência comparando cada par de conjunto de dados e modelos com combinações de conjuntos e modelos alternativos presentes na rede de meta-modelagem de padrões em sequências de símbolos binários. 60
Tabela 8 – Métricas descritivas da rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos. 63
Tabela 9 – Métricas descritivas para cada par de conjunto de dados e modelo disponíveis na rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos. 64
Tabela 10 – Valores máximos de coincidência comparando cada par de conjunto de dados e modelos com combinações de conjuntos e modelos alternativos presentes na rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos. 67

LISTA DE ABREVIATURAS E SIGLAS

HMM	Modelos Ocultos de Markov, do inglês *Hidden Markov Models*
CAZy	Banco de dados de enzimas ativas em carboidratos, do inglês *Carbohydrate Active Enzymes*
AA	Atividades Auxiliares
PL	Liases de Polissacarídeos, do inglês *Polysaccharide Lyases*
CE	Esterases de Carboidratos, do inglês *Carbohydrate Esterases*
CBM	Módulos de Ligação a Carboidrato, do inglês *Carbohydrate-Binding Module*
GT	Glicosiltransferases
GH	Glicosidases, do inglês *Glycoside Hydrolase*

LISTA DE SÍMBOLOS

Ω	Domínio universo
A	Domínio do ambiente de dados
E	Domínio da estrutura de modelagem
$\mathcal{G}$	Grafo
$\mathcal{V}$	Conjunto de vértices de um grafo
$\mathcal{A}$	Conjunto de arestas de um grafo
$\mathcal{M}$	Rede de meta-modelagem e meta-modelo
ω	Conjunto de dados
m	Modelo
$\in$	Pertence
$\cup$	União
$\cap$	Intersecção
$()^c$	Complemento
$\vee$	Disjunção, ou
$\wedge$	Conjunção, e
$\neg$	Negação, não
$\bigcup$	União de uma coleção de conjuntos
$\bigcap$	Intersecção de uma coleção de conjuntos
$\bigvee$	Disjunção de uma coleção de sentenças lógicas
$\bigwedge$	Conjunção de uma coleção de sentenças lógicas
N	Número de elementos no ambiente de dados
n	Número de conjuntos de dados
C	Cardinalidade de um conjunto de dados
CI	Número de elementos inespecíficos

PI	Percentual de elementos inespecíficos
$\bar{U}$	Multiplicidade média de elementos
H	Entropia
$\mathcal{D}$	Diversidade de conexões
$\mathcal{D}_{\mathcal{T}}$	Diversidade total de conexões
$\mathcal{J}$	Índice de Jaccard
$\mathcal{I}$	Índice de interioridade, coeficiente de Szymkiewicz–Simpson ou índice de sobreposição
$\mathcal{C}$	Índice de coincidência

SUMÁRIO

1 INTRODUÇÃO 23
1.1 Modelagem Baseada em Dados 23
1.2 Ciência de Redes 25
1.3 Anotação de Domínios Proteicos 27
1.4 Objetivos 30

2 CONCEITOS E MÉTODOS 31
2.1 Grafos Bipartidos 31
2.2 Redes de Meta-modelagem 32
2.2.1 Domínios, Mapas e Representações da Meta-modelagem 32
2.2.2 Relação entre Elementos e Modelos 35
2.2.3 Incorporação de Novos Conjuntos de Dados e Modelos 37
2.2.4 Álgebra Pareada entre Operações de Conjuntos e Operações Lógicas Aplicada à Meta-modelagem 40
2.2.5 Métodos para Quantificar a Especificidade de Estruturas de Modelagem . . 42
2.2.5.1 Cardinalidade, número e percentual de elementos inespecíficos dos conjunto de dados 43
2.2.5.2 Multiplicidade média de elementos 44
2.2.5.3 Diversidade de conexões 44
2.2.6 Aplicando a Álgebra Pareada para Construir e Avaliar Modelos 46
2.2.7 Análise de Redes de Meta-modelagem 48
2.3 Experimentos 50
2.3.1 Padrões em Sequências de Símbolos Binários 51
2.3.2 Domínios Funcionais de Enzimas Ativas em Carboidratos 53
2.3.2.1 O Banco de Dados CAZy 53
2.3.2.2 Construção e Análise da Rede de Meta-modelagem 54

3 RESULTADOS E DISCUSSÕES 57
3.1 Rede de Meta-modelagem de Padrões em Sequências de Símbolos Binários 57
3.2 Rede de Meta-modelagem dos Domínios Funcionais de Enzimas Ativas em Carboidratos 62
3.2.1 Análise de Conjuntos de Dados e Modelos na Rede de Meta-Modelagem dos Domínios Funcionais de Enzimas Ativas em Carboidratos 64
3.2.2 Análise de Subconjuntos de Dados e Submodelos na Rede de Meta-Modelagem dos Domínios Funcionais de Enzimas Ativas em Carboidratos 68

3.2.2.1 Interações envolvendo a classe GT . . . 69
3.2.2.2 Interação entre as classes GT e GH . . . 69
3.2.2.3 Interação entre as classes GT e CE . . . 72
3.2.2.4 Interação entre as classes GT e CBM . . . 73
3.2.2.5 Interação entre as classes GT e PL . . . 75
3.2.2.6 Interação entre as classes GT e AA . . . 76
3.2.2.7 Interações envolvendo a classe PL . . . 78
3.2.2.8 Interação entre as classes PL e CBM . . . 78
3.2.2.9 Interação entre as classes PL e CE . . . 78
3.2.2.10 Interação entre as classes PL e GH . . . 81
3.2.2.11 Interações envolvendo a classe CE . . . 82
3.2.2.12 Interação entre as classes CE e GH . . . 83
3.2.2.13 Interação entre as classes CE e CBM . . . 84
3.2.2.14 Interação entre as classes CE e AA . . . 84
3.2.2.15 Interações envolvendo a classe GH . . . 87
3.2.2.16 Interação entre as classes GH e CBM . . . 87
3.2.2.17 Interação entre as classes GH e AA . . . 89
3.2.2.18 Interações envolvendo a classe AA . . . 91
3.2.2.19 Interação entre as classes AA e CBM . . . 92
3.2.2.20 Interações envolvendo a classe CBM . . . 92
3.2.2.21 As possibilidades da análise de subconjuntos e submodelos em redes de meta-modelagem de proteomas anotados . . . 94

4 CONSIDERAÇÕES FINAIS . . . 97
4.1 Conclusões . . . 97
4.2 Lista das Principais Contribuições . . . 99
4.3 Futuros Desenvolvimentos . . . 100

REFERÊNCIAS . . . 103

1 INTRODUÇÃO

1.1 Modelagem Baseada em Dados

A modelagem de objetos e fenômenos presentes no mundo real constitui uma atividade central na ciência.(1–3) Através da modelagem, os cientistas podem estudar uma classe inteira de fenômenos utilizando uma construção simplificada e única, o modelo, auxiliando-os a identificar e explicar comportamentos e características essenciais dos objetos modelados.(1–3) A modelagem científica é amplamente utilizada em todas as áreas científicas, especialmente nas ciências naturais.(1,3–7)

Normalmente, abordagens de modelagem possuem três etapas principais. Ela começa com as observações empíricas de um fenômeno de interesse. Em segundo lugar, com base nessas observações, os dados experimentais são organizados. Em terceiro lugar, com base nos dados organizados, é concebida uma emulação simplificada e útil desses dados, chamada de *modelo*. A Figura 1 mostra um diagrama que representa essas três etapas da modelagem científica. O primeiro domínio do diagrama, representado como Ω, é o universo que contém todos os dados relacionados ao fenômeno estudado, acessíveis por meio de observações empíricas. Ao observar o fenômeno de interesse, torna-se possível coletar dados, que são progressivamente organizados em conjuntos de dados no segundo domínio, o ambiente de dados A. Com base nisso, modelos preliminares podem ser construídos e aprimorados, resultando em uma explicação do fenômeno de interesse. O conjunto de modelos desenvolvidos ao longo da abordagem de modelagem constitui o terceiro domínio, a estrutura de modelagem E.

Quando um modelo é abstraído a partir da coleção bruta de dados experimentais em vez de princípios científicos fundamentais, a abordagem de modelagem pode ser compreendida como sendo *baseada em dados* — *data-driven*, em inglês.(3,8,9) As abordagens de modelagem baseadas em dados têm a vantagem de extrair padrões e correlações entre observações sem conhecimento prévio dos princípios físicos subjacentes.(3,8,9) Isso permite o uso de técnicas de análise de dados para elucidar novas leis físicas dos objetos estudados e até mesmo fazer previsões confiáveis sem conhecimento adicional da natureza desses objetos.(3,8) Exemplos de técnicas de modelagem baseadas por dados incluem análise de regressão, reconhecimento de padrões, redes neurais e várias outras abordagens estatísticas, de mineração de dados e de aprendizado de máquina.(3,8–10)

O maior limitante para a modelagem científica baseada em dados é a disponibilidade de dados experimentais e a capacidade de processamento desses dados. Por isso, houve um aumento significativo de novas técnicas de modelagem e modelos baseados em dados a partir do início do século XXI, quando entramos na Era do *Big Data*.(11,12) Nas

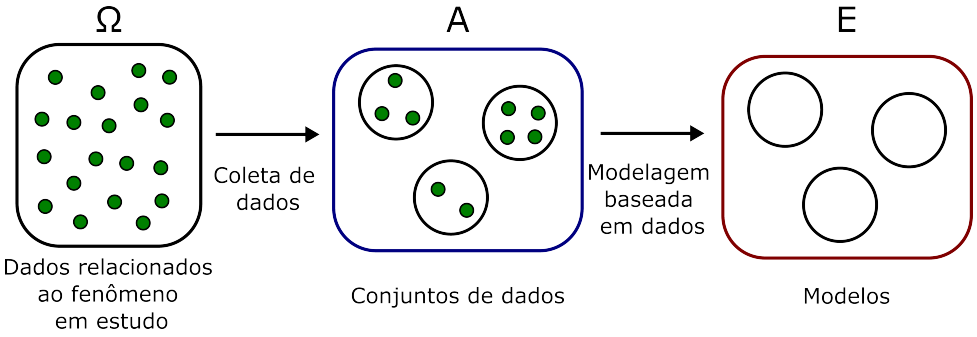


Figura 1 – Os principais domínios da modelagem científica considerados neste trabalho. Para modelar um fenômeno de interesse, começa-se observando o fenômeno, coletando conjuntos de dados e criando modelos baseados nesses dados. O conjunto Ω corresponde ao universo de dados, que contém todos os dados acessíveis relacionados a um fenômeno. Dados de interesse são coletados experimentalmente do universo Ω e organizados em conjuntos de dados, que serão dispostos no ambiente de dados A. A partir desses conjuntos, modelos podem ser construídos e combinados em uma estrutura de modelagem, E, que é aplicada para explicar o fenômeno em questão.

Fonte: Elaborada pelo autor.

últimas décadas, houve um crescimento acelerado na capacidade de capturar, armazenar, gerenciar e analisar dados, resultando em maior disponibilidade de dados em múltiplas áreas científicas.(12) Por exemplo, atualmente, conjuntos massivos de dados são utilizados em estudos baseados em dados em geologia (13), química (14), medicina (15), ecologia (16), biologia molecular (17) e várias outras ciências.(9, 12, 18)

Os principais métodos que foram desenvolvidos para enfrentar os desafios de lidar com grandes quantidades de dados são baseados em técnicas de mineração de dados e aprendizado de máquina. Essas abordagens são altamente eficazes, mas podem ter limitações de interpretabilidade, o que pode dificultar a compreensão de suas previsões.(3, 8, 9) Para superar essa limitação, uma nova tendência em ciência de dados está surgindo: o desenvolvimento de modelos que integram diversas abordagens baseadas em dados com princípios fundamentais e estruturas matemáticas - incluindo modelos estatísticos, equações matemáticas, processos estocásticos, lógica, leis físicas e descrições qualitativas de objetos.(3, 8, 9, 19, 20)

No entanto, os modelos integrados também têm suas limitações. Eles frequentemente requerem conhecimento prévio dos dados, requerem esforços de compatibilizar diferentes tipos de dados e modelos em uma única estrutura, podem ter generalização limitada, e podem ter custos computacionais mais elevados para desenvolver e usar.(8, 20) Além disso, eles ainda possuem um problema comum em modelagem de dados, que é a sensibilidade à qualidade dos dados.(20) Independentemente da abordagem escolhida, ao modelar vários conjuntos de dados associados a um mesmo fenômeno, a utilização da mesma técnica de

modelagem pode resultar em modelos de precisão distintos, dependendo da qualidade de cada conjunto de dados.(20) A avaliação dessas diferenças de qualidade é, por si só, um desafio complexo no processo de modelagem.

Para abordar essas questões, buscamos criar um procedimento que nos permita examinar como modelos podem ser combinados para explicar uma coleção de conjuntos de dados. Para alcançar esse objetivo, sugerimos desenvolver um *meta-modelo* baseado em princípios de integração de modelos e de ciência de redes. Esse meta-modelo será usado para retratar o relacionamento entre conjuntos de dados e modelos — por isso o uso do prefixo *meta* — nos permitindo avaliar a qualidade de modelos, ampliar a interpretabilidade deles e criar novos modelos com base nos pré-existentes. Esse trabalho desenvolverá ainda mais a metodologia proposta anteriormente por Costa.(1)

1.2 Ciência de Redes

A ciência de redes é uma área de estudo que busca desenvolver métodos para compreender as propriedades e comportamentos de redes complexas.(21, 22)

Para entender o conceito de redes complexas, antes precisamos recorrer ao conceito de grafos. Grafos são estruturas compostas por um conjunto de entidades e suas interações.(23) Matematicamente, um grafo é formado por um conjunto de vértices (também chamados de nós) e um conjunto de arestas (também chamadas de conexões), que conecta esses vértices em pares. Cada vértice representa uma entidade, enquanto as arestas representam algum tipo de interação entre essas entidades.(23) Já uma rede complexa é caracterizada por um grafo que apresenta topologia particularmente intrincada, cujas conexões entre nós não seguem um padrão regular, o que resulta em várias propriedades que não são facilmente deduzidas a partir de sua estrutura.(21, 22)

A Figura 2, a seguir, apresenta um exemplo de grafo que representa uma rede complexa.

Vários sistemas presentes no cotidiano e na ciência podem ser entendidos como exemplos de redes complexas.(22, 25) Um exemplo notável é a Internet, que é composta de computadores e servidores que interagem entre si via conexões físicas, que incluem cabos coaxiais, fibras ópticas e ondas de rádio.(26) Outro exemplo, um pouco mais sutil, são as redes tróficas, em que organismos interagem entre si via relações de predação.(27) Um terceiro exemplo, mais abstrato, são as redes de similaridade entre proteínas, em que proteínas se relacionam entre si por grau de semelhança.(28)

Apesar da natureza diversa dos sistemas acima mencionados, em todos eles, a forma como cada par de entidades interage entre si é o que revela as principais características e funções das redes. Um par de computadores conectado por fios podem trocar informação (26), um par de animais conectados trocam energia e biomassa em um ato de

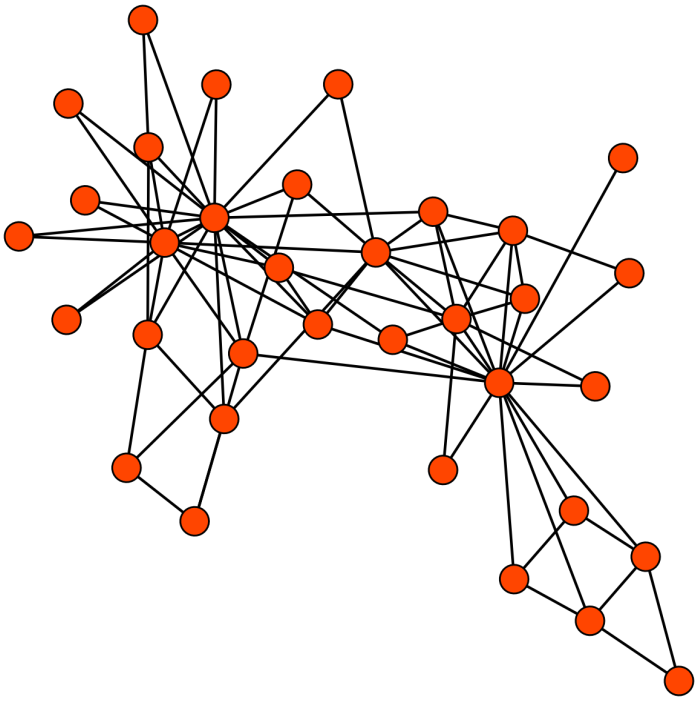


Figura 2 – Grafo representando uma rede social entre membros de um clube universitário de karatê. Cada vértice do diagrama (pontos vermelhos) representa um indivíduo, e cada aresta (retas pretas) indica uma relação de amizade entre pares de indivíduos. Os dados originais desta rede foram coletados e publicados por Zachary, em 1977.(24)

Fonte: Elaborada pelo autor.

predação (27), e um par de proteínas conectadas por uma relação de semelhança possui alguma relação funcional e evolutiva.(28) Essa propriedade em comum permite que esses sistemas, assim como inúmeros outros, sejam descritos, explicados, ou modelados, igualmente como redes complexas.(22, 25) Isso possibilita analisá-los usando uma abordagem unificada, a abordagem de *ciência de redes*.(21, 22)

A ciência de redes integra aspectos de múltiplas áreas do conhecimento que estudam sistemas complexos, com muitos objetos interagentes, incluindo matemática, sociologia, biologia e física.(21, 22) Nesse campo de pesquisa, diversas ferramentas são utilizadas para construir, analisar e visualizar os mais diversos tipos de redes a fim de revelar propriedades e padrões implícitos em suas estruturas.(21,22) Exemplos dessas ferramentas e alguns casos de aplicações incluem: métodos de representação e análise topológica baseadas em teoria de grafos, usados genericamente para estudar vários tipos de redes (21, 22); algoritmos de detecção de comunidades para identificar grupos de nós mais conectados (21, 22), usados em redes sociais para identificar grupos de amizades e influências (29) e em redes de similaridade de proteínas para identificar grupos com relação evolutiva (28); e estimadores de centralidade de nós (21,22), aplicados em redes de colaboração científica na identificação

de pesquisadores mais influentes (30) e em redes tróficas na avaliação da importância ecológica de espécies.(27)

Devido ao amplo aparato metodológico da ciência de redes para estudar interação entre objetos, decidimos tratar a meta-modelagem da relação entre modelos e conjuntos de dados também como um problema de redes complexas. Por isso, como proposta de meta-modelo, sugerimos desenvolver uma *rede de meta-modelagem*, a qual poderemos usar para a visualização e análise topológica das interações entre uma estrutura de modelagem e um ambiente de dados experimentais.

1.3 Anotação de Domínios Proteicos

Desde o início do século, houve avanços significativos nas tecnologias de captura, armazenamento e processamento de dados biológicos, resultando em uma enorme quantidade de conjuntos de dados disponíveis para análise.(31) Esse aumento de dados é notado, principalmente, em bancos de dados de sequências de ácidos nucleicos, bem como de sequências de aminoácidos de proteínas.(32–34) Baseado nessa abundância de dados, surgiram diversas explicações e modelos do funcionamento e estrutura de macromoléculas.(31, 35)

Devido à abundância de dados e modelos explicativos disponíveis, o estudo de sequências biológicas é um excelente candidato para explorar a relação entre conjuntos de dados e modelos utilizando a nossa proposta de meta-modelagem. Particularmente, optamos por analisar sequências de proteínas.

As proteínas formam uma classe muito importante de macromoléculas biológicas, desempenhando funções estruturais e catalíticas nos organismos.(36) Quimicamente, elas são polipeptídeos, polímeros de aminoácidos sintetizados pelos organismos seguindo uma ordem específica codificada em seus genes.(36) Uma proteína e todas as suas características — incluindo, mas não se limitando à sua estrutura, estabilidade, solubilidade e, principalmente, função — dependem da ordem de aminoácidos em sua estrutura polipeptídica.(36) Por isso, a sequência de aminoácidos das proteínas é uma informação tão importante.

Existem 20 aminoácidos padrões que formam as proteínas encontradas na maioria dos organismos, cada um sendo representado graficamente por uma letra: Alanina — A, Cisteína — C, Aspartato — D, Glutamato — E, Fenilalanina — F, Glicina — G, Histidina — H, Isoleucina — I, Lisina — K, Leucina — L, Metionina — M, Asparagina — N, Prolina — P, Glutamina — Q, Arginina — R, Serina — S, Treonina — T, Valina — V, Triptofano — W, e Tirosina — Y.(36) A partir da correspondência entre aminoácidos e letras, cada proteína pode ser identificada por uma sequências de caracteres, respectiva à sua sequência de aminoácidos. Isso permite o desenvolvimento de um rol de análises computacionais extremamente práticas e informativas sobre a função e estrutura de proteínas.(31, 35)

Dentro de uma sequência proteica existem blocos de aminoácidos que interagem mais entre si, formando unidades estruturais chamadas de *domínios*.(31, 36) Eles se organizam tridimensionalmente em estruturas bem definidas dentro de uma proteína, desempenhando um papel estrutural e funcional próprio.(31, 36) Além disso, um mesmo tipo de domínio proteico pode fazer parte de proteínas diversas, em diferentes combinações e diferentes ordens, podendo conter pequenas variações na sua constituição de aminoácidos.(31,36) Essa característica permite analisar as proteínas em termos dos domínios identificados em suas estruturas, facilitando correlacionar diferentes proteínas, inferir suas funções e compreender suas evoluções.(31)

Entre as funções que domínios proteicos podem desempenhar, a mais importante é a capacidade de atuar como catalisadores. Especificamente, as proteínas que possuem domínios catalíticos são chamadas de *enzimas*. Uma enzima tem a capacidade de aumentar a velocidade a qual uma reação bioquímica específica alcança o seu equilíbrio, possibilitando que ela ocorra nas taxas necessárias para a sobrevivência dos organismos.(36) Para isso, os domínios interagem com substratos e estabilizam os estados de transição da reação, facilitando a conversão deles em produtos.(36)

Na Figura 3, há um exemplo de proteína representada por uma sequência de aminoácidos. Nessa sequência, são identificados dois domínios, cada um com uma função distinta.

Para analisar a grande quantidade de sequências proteicas disponíveis, bioinformatas e biólogos computacionais utilizam diversas abordagens baseadas em dados.(31, 35) No geral, essas técnicas consistem em detectar padrões funcionais dentro de sequências de aminoácidos a fim de obter informações biológicas relevantes.(31, 35) Para isso, primeiro, padrões em sequências de proteínas conhecidas são relacionados a alguma função por meio de experimentos em bancada.(35) Feita essa relação, a detecção desse padrão nas sequências de outras proteínas pode ser utilizada para inferir suas funções.(35) Esses métodos, apesar de serem menos confiáveis, oferecem uma forma muito mais prática de se obter informações funcionais relevantes de proteínas do que a experimentação *in vivo* ou *in vitro*.

O processo de identificar os domínios presentes em uma proteína é chamado de *anotação de domínios proteicos*.(37) Os métodos mais comuns de se realizar a anotação de domínios são baseados em alinhamento de sequências e modelos ocultos de Markov (HMM, do inglês, *hidden markov models*).(31, 35, 37) Além desses métodos serem aplicados para inferir a função de proteínas, eles também são usados para criar perfis para detecção de domínios homólogos, isto é, relacionados evolutivamente.(35, 37) As proteínas que possuem domínios homólogos detectados por esses métodos são classificadas em famílias proteicas, formando grupos que possuem relação funcional e evolutiva.(35, 37)

Muitos bancos de dados concentram informações sobre a anotação de proteínas em

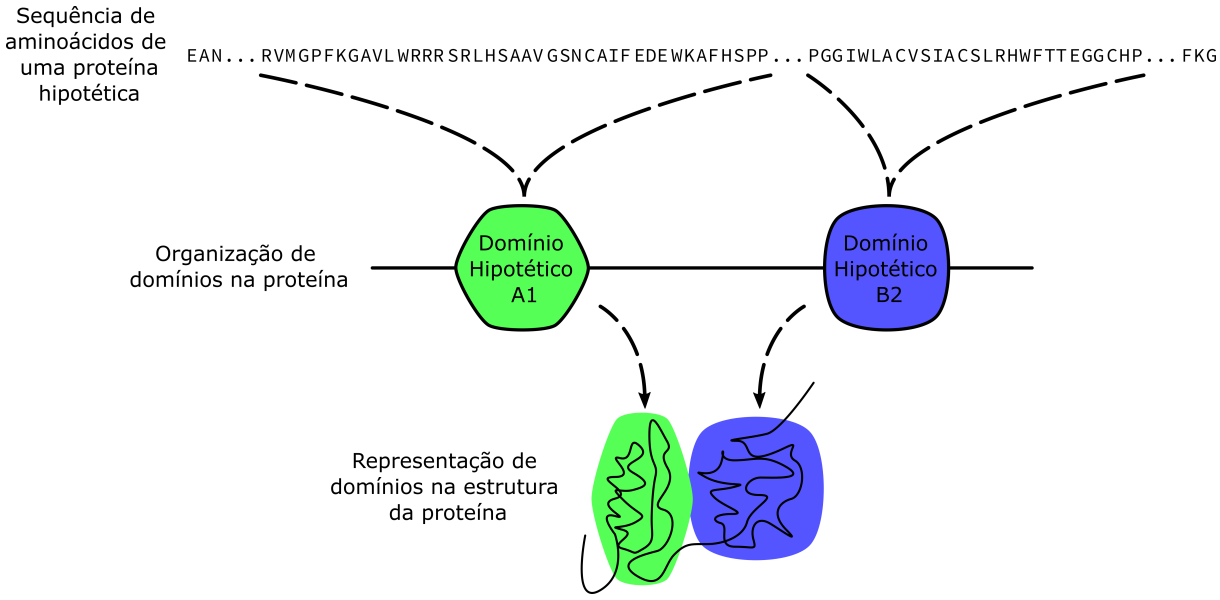


Figura 3 – Representação de domínios encontrada em uma proteína hipotética. Dentro da sequência de aminoácidos de uma proteína, indicada pela sequência de caracteres no topo da figura, há blocos de aminoácidos que interagem mais entre si. Esses blocos formam unidades estruturais chamadas de domínios. Dois trechos da sequência de aminoácidos formam dois domínios hipotéticos, chamados aqui de A1 e B2. Os domínios são representados no diagrama por formas geométricas organizadas sobre uma linha, correspondente à cadeia polipeptídica da proteína. Abaixo, há uma representação da estrutura da proteína indicando os domínios como unidades estruturais distintas.

Fonte: Elaborada pelo autor.

famílias baseadas em métodos de alinhamento de sequências e HMM. Alguns exemplos incluem: Interpro, banco de dados que concentra informações sobre famílias proteicas diversas (38); RedoxiBase/PeroxiBase, um banco de dados de proteínas reguladas por espécies reativas de oxigênio (39, 40); MEROPS, baseada em famílias de enzimas proteolíticas (41); ThYme, de enzimas que compõem os ciclos de síntese de ácidos graxos (42); e CAZy, o banco de dados de enzimas ativas em carboidratos.(43, 44)

Nesses bancos de dados, cada proteína é classificada em famílias a partir da combinação de domínios funcionais que elas apresentam. Como algumas proteínas podem apresentar múltiplos domínios funcionais, então é possível que uma mesma proteína seja classificada em múltiplas famílias. Isso cria uma relação complexa entre famílias proteicas, seus conjuntos de dados e seus modelos, e a análise dessa relação pode revelar informações sobre a evolução e função dessas famílias.

A análise da interação entre domínios em conjuntos de proteínas pode ser feita utilizando uma rede de coocorrência de domínios proteicos.(45, 46) Nessa rede, cada nó representa um domínio proteico, e as conexões entre os nós indicam a coocorrência desses

domínios na mesma sequência proteica.(45,46) Ao analisar a topologia e os padrões dentro da rede, é possível identificar famílias de domínios proteicos intimamente associadas, o que pode auxiliar na compreensão de associações funcionais e relações evolutivas.(45,46) No entanto, com esse tipo de rede, a forma como os modelos de domínios funcionais são construídos e associados não é diretamente abordada. Por isso, acreditamos que uma abordagem de meta-modelagem pode se complementar à análise de redes de coocorrência, proporcionando uma análise mais aprofundada da modelagem de domínios proteicos.

1.4 Objetivos

Este trabalho tem como objetivos principais a extensão e aplicação de uma abordagem de meta-modelagem que permita integrar informações entre modelos e conjuntos de dados. Para isso, consideraremos a abordagem de Costa (1) a ser complementada através da consideração de métodos disponíveis em ciência de redes, incluindo a visualização dos dados em grafos e o uso de medidas topológicas para extrair informações. Essa abordagem será utilizada para avaliar a qualidade de modelos, ampliar as suas interpretabilidades e criar novos modelos com base nos pré-existentes.

Esses objetivos serão aplicados mais especificamente na análise exploratória de anotação de domínios proteicos, uma vez que essa área lida com a modelagem de grandes volumes de dados. Ao alcançar esses objetivos, espera-se contribuir para a criação de uma ferramenta que possibilite a compreensão mais abrangente da função e evolução de famílias de proteínas.

2 CONCEITOS E MÉTODOS

2.1 Grafos Bipartidos

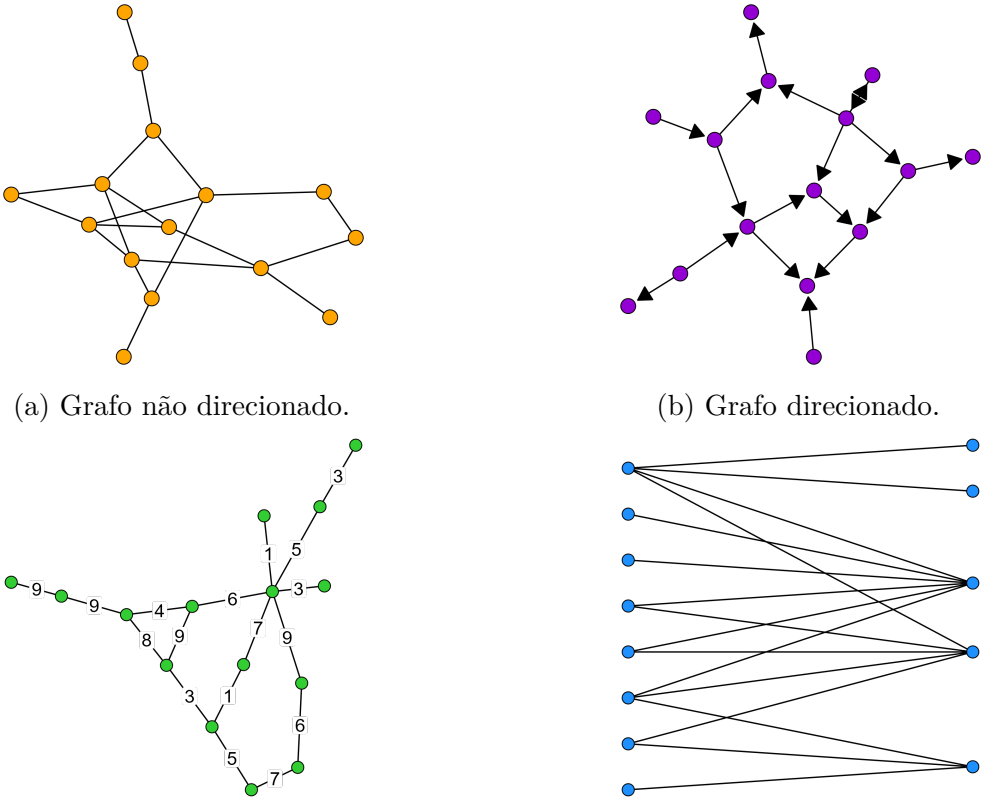
Um grafo $\mathcal{G}$ é definido por um par ordenado $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ composto por um conjunto de vértices (ou nós) $\mathcal{V}$ conectados aos pares por arestas (ou conexões) $\mathcal{A}$.(23) Eles podem ser não direcionados, quando as arestas conectam vértices de forma simétrica, ou direcionados, quando as arestas conectam dois vértices de maneira assimétrica.(23) Por exemplo, a rede de Internet pode ser representada como um grafo não direcionado, uma vez que a conexão entre um computador A e um computador B através de cabos implica que B também está conectado a A.(26) No entanto, uma rede trófica deve ser representada por um grafo direcionado, uma vez que as relações tróficas entre as espécies são assimétricas: quando uma espécie A é predadora de uma espécie B, normalmente a espécie B não é a predadora da espécie A.(27)

Além disso, grafos podem ser ponderados, quando cada aresta entre vértices possui um peso associado, ou não ponderados, quando não há pesos.(23) Redes de similaridade entre proteínas, por exemplo, são representadas por grafos ponderados, onde os vértices simbolizam proteínas, arestas indicam relações de similaridade e os pesos nas arestas indicam a relevância estatística dessa similaridade.(28)

Alguns exemplos de grafos de diversos tipos estão ilustrados na Figura 4.

Outro tipo importante é o de grafos bipartidos. Os grafos bipartidos são caracterizados por dividir seus vértices em dois subconjuntos distintos, $\mathcal{V}_A$ e $\mathcal{V}_B$, de modo que todas as arestas do grafo conectam um vértice em $\mathcal{V}_A$ a um vértice em $\mathcal{V}_B$, sem haver arestas que conectem dois vértices presentes em um mesmo subconjunto.(23) Um grafo bipartido é representado na Figura 4(d). Eles são particularmente úteis para representar redes que possuem dois tipos de entidades interagentes.(22, 47) Por exemplo, grafos bipartidos podem representar redes tróficas, em que predadores interagem com presas (47); redes de fármacos, em que fármacos interagem com alvos moleculares (47); redes co-autoria, em que autores são conectados a artigos (22); e muitas outras.

Grafos bipartidos serão usados na construção da rede de meta-modelagem, já que eles permitem representar a interação entre dois tipos distintos de estruturas: os conjuntos de dados e os modelos. O uso desses grafos permite a visualização e análise dessas interações de forma clara e sistemática.



(a) Grafo não direcionado.

(b) Grafo direcionado.

(c) Grafo ponderado.

(d) Grafo bipartido.

Figura 4 – Alguns tipos de grafos.

Fonte: Elaborada pelo autor.

2.2 Redes de Meta-modelagem

Nesta seção, descreveremos como construir uma rede dedicada a representar — ou modelar — o processo de modelagem de dados, chamada de *rede de meta-modelagem* $\mathcal{M}$. A rede de meta-modelagem é um meta-modelo que permite representar a integração entre um ambiente de dados e uma estrutura de modelagem. Ela é construída considerando três aspectos dos estrutura de modelagem baseadas em dados, a saber: (i) o relacionamento entre dados, conjuntos de dados e modelos; (ii) a possibilidade de atualizar estruturas de modelagem para incorporar novos dados; e (iii) a possibilidade de combinar modelos dentro de uma estrutura de modelagem. Cada um desses aspectos é discutido a seguir com base no trabalho desenvolvido por Costa.(1)

2.2.1 Domínios, Mapas e Representações da Meta-modelagem

Modelos científicos baseados em dados são diretamente relacionados a *conjuntos de dados* cujos *elementos* representam objetos do mundo real. Cada conjunto de dados contém elementos agrupados a partir de uma porção de características observáveis em comum que seja de interesse científico. Em um processo de modelagem, cada conjunto de

dados é um candidato a ser explicado por um modelo científico, que é construído a partir dos atributos dos elementos estudados.

Um exemplo prático para ilustrar essas ideias é a modelagem de um conjunto de dados que contém todas as informações relevantes sobre diferentes exemplos de talheres presentes em uma gaveta. Um cientista pode, inicialmente, coletar informações sobre diferentes colheres em um único conjunto de dados e, a partir dela, construir um modelo de colher que represente e explique essa coleção de informações. Ele pode repetir isso para outros tipos de talheres, formando uma coleção de conjuntos de dados e seus respectivos modelos. Essa atividade é representada na Figura 5(a).

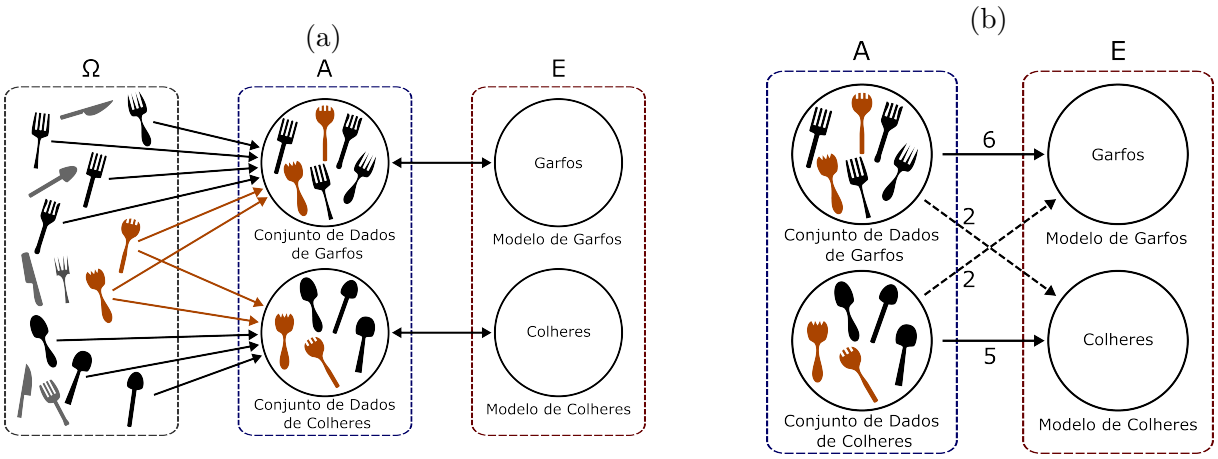


Figura 5 – Exemplo prático de meta-modelagem de talheres. (a) Representação de mapeamento entre domínios. (b) Representação de rede de meta-modelagem. O *domínio universo* Ω contém virtualmente todos os elementos de dados disponíveis no mundo real, como, por exemplo, todos os talheres disponíveis em uma gaveta. Os elementos de dados são retirados de Ω para formar conjuntos de dados, que estão disponíveis no *domínio do ambiente de dados* A. Cada conjunto de dados é modelado por um respectivo modelo disponível no *domínio da estrutura de modelagem* E. A representação de mapeamento entre domínios (a) representa a meta-modelagem com duas operações de mapeamento: primeiro, cada elemento de dados em Ω é mapeado em cada conjunto de dados em A que o contém; segundo, cada conjunto de dados em A é mapeado em seu respectivo modelo em E por meio de uma associação bijetiva (A, E). O mapeamento entre Ω e A não é injetivo, pois os dois elementos em laranja pertencem a dois conjuntos de dados em A. A representação de rede de meta-modelagem (b) substitui o mapeamento (A, E) por conexões ponderadas direcionadas de A para E, cujos pesos destacam o número de elementos em cada conjunto de dados que se encaixa em cada modelo disponível. Em (b), as conexões com linhas tracejadas mostram que dois elementos, os elementos em laranja, podem ser associados a um modelo alternativo, pois também fazem parte do conjunto de dados alternativo.

Fonte: Elaborada pelo autor.

No meta-modelo $\mathcal{M}$, o *domínio universo* é o conjunto Ω, e ele contém virtualmente todos os elementos de dados no mundo real que podem ser selecionados para formar conjuntos de dados de interesse científico. Na Figura 5(a), os elementos de Ω representam o número finito de talheres presente em uma gaveta. Dois conjuntos de dados são construídos a partir do domínio Ω: o *conjunto de dados de garfos*, selecionando 6 elementos, e o *conjunto de dados de colheres*, selecionando 5 elementos. Observe que os conjuntos de dados disponíveis serão, por definição, subconjuntos do universo Ω.

O conjunto de todos os conjuntos de dados coletados constitui o *domínio do ambiente de dados* A. Portanto, A representa o conjunto de subconjuntos de Ω disponíveis para a análise científica. Dessa forma, o maior A possível é o conjunto de partes de Ω, contendo 2^N elementos, onde N é o número de elementos de Ω.(1)

A estrutura de modelagem é construída progressivamente com base nos conjuntos de dados disponíveis em A, cada um apoiando a construção de um modelo diferente. No caso mostrado na Figura 5(a), para cada conjunto de dados disponível em A, um modelo do tipo rótulo foi construído: o conjunto de garfos é modelado por um rótulo "Garfo", enquanto o conjunto de colheres é modelado por um rótulo "Colher". Esses rótulos podem ser acompanhados por descrições diversas, mas vamos usá-los de uma maneira simplificada durante esse exemplo. O conjunto composto pelos modelos construídos é definido, então, como o *domínio da estrutura de modelagem* E.

Uma forma de representar as interações entre os três domínios envolvidos no meta-modelo, Ω, A e E, é considerando duas operações de *mapeamento* entre eles. Primeiramente, os elementos de dados em Ω são mapeados nos conjuntos de dados em A. Em seguida, os conjuntos de dados disponíveis em A são mapeados em seus respectivos modelos em E. A segunda operação de mapeamento, representada como (A, E), é a base da construção de uma estrutura de modelagem e representa o *conhecimento atual* dos objetos em estudo. Ela destaca os conjuntos de dados coletados do universo e os modelos que puderam ser construídos com base neles, capturando a ideia de que uma estrutura de modelagem depende dos conjuntos de dados disponíveis.

Sendo tão importante para a concepção do meta-modelo $\mathcal{M}$, impomos algumas propriedades matemáticas que o mapeamento entre os domínios A e E deve possuir. Primeiro, estabelecemos que o mapeamento (A, E) seja uma *associação bijetiva*. Matematicamente, uma associação bijetiva entre A e E nos permite identificar não apenas *modelos por meio de conjuntos de dados*, mas também *conjuntos de dados por meio de modelos*, sem perda de informação. Ao definirmos o mapeamento (A, E) como uma associação bijetiva, o meta-modelo $\mathcal{M}$ pode representar estruturas de modelagem com consistência e simplicidade, evitando a representação de modelos não verificados (modelos sem conjunto de dados), redundantes (dois modelos explicando o mesmo conjunto de dados) ou ambíguos (modelos com dois conjuntos de dados distintos com características diferentes).

Na Figura 5(a), o mapeamento de elementos de Ω em conjuntos de dados em A é representado por setas unidirecionais. A associação bijetiva entre conjuntos de dados em A e modelos em E é representada por setas bidirecionais. Essa representação é chamada de representação de *mapeamento entre domínios*.

2.2.2 Relação entre Elementos e Modelos

Na estrutura de meta-modelagem, cada conjunto de dados no ambiente de dados A está conectado ao seu respectivo modelo na estrutura de modelagem E. Como cada elemento retirado de Ω fará parte de um conjunto de dados em A, e cada conjunto de dados em A está conectado a um modelo em E, podemos conectar indiretamente cada elemento individual a um modelo em A.

Como mencionado, cada elemento coletado para fazer parte de um conjunto de dados modelável representa um objeto do mundo real. O que determina a união de diferentes elementos em um mesmo conjunto de dados são as semelhanças observáveis apresentadas por seus respectivos objetos no mundo real. Essas semelhanças serão traduzidas em dados quantitativos e qualitativos dos elementos de cada conjunto presente no ambiente de dados A, e serão levadas em conta durante a construção de modelos da estrutura de modelagem E.

No entanto, um mesmo objeto no mundo real pode compartilhar características com objetos cujos elementos respectivos pertencem a diferentes conjuntos de dados. Nesse caso, o elemento que representa esses objetos fará parte de mais de um conjunto de dados no ambiente A. Consequentemente, os atributos de um único elemento será considerado durante a construção de múltiplos modelos. Dessa forma, é possível haver uma relação não injetiva entre elementos de dados e modelos. Por exemplo, no exemplo prático de meta-modelagem de talheres da Figura 5(a), as *garfolheres* laranjas — talheres semelhantes com uma parte côncava semelhante a colheres e com dentes semelhantes a garfos — são mapeadas tanto nos conjuntos de dados de colheres quanto de garfos, resultando em uma sobreposição entre os seus modelos respectivos.

Essa relação não injetiva aumenta significativamente a complexidade da relação entre conjuntos de dados e modelos. Para visualizar essa relação de forma mais direta, podemos construir uma *rede de meta-modelagem* representada por um *grafo bipartido ponderado*, ilustrado na Figura 5(b). O grafo bipartido da rede de meta-modelagem conecta vértices do domínio A, que representam conjuntos de dados, a vértices do domínio E, que representam modelos, por meio de arestas ponderadas e direcionadas. O peso de uma aresta entre um conjunto de dados e um modelo é igual ao número de elementos do conjunto de dados que se encaixa no modelo. Essa representação é chamada de representação de rede de meta-modelagem.

A rede destaca como diferentes modelos podem explicar parcialmente o mesmo

conjunto de dados, revelando possíveis redundâncias na estrutura de modelagem. Essas redundâncias podem ser consideradas indesejáveis em algumas modelagens científicas, pois revelam uma falta de *especificidade* nos modelos construídos. No entanto, conforme implicado pelo mapeamento entre domínios, elementos de dados que satisfazem vários modelos não afetarão a *consistência* das estruturas de modelagem, pois a associação bijetiva entre um conjunto de dados e seu modelo respectivo ainda é mantida.

Para que as duas representações do meta-modelo $\mathcal{M}$ — de mapeamento entre domínios e de rede — e as suas características sejam consistentes entre si, podemos definir uma abordagem na qual um conjunto de dados, suas conexões ponderadas com a estrutura de modelagem e seu modelo respectivo são analisados em conjunto como parte do que chamamos aqui de *cartucho*.(1)

O *cartucho* corresponde à coleção de todos os elementos de dados envolvidos em um determinado conjunto ω_i, seu respectivo modelo m_i e a conexão ponderada $w_{i,i}$ entre ω_i e m_i.(1) Algumas condições são necessárias para determinar quando um *cartucho* está *completo*. A primeira delas é que, embora elementos possam estar indiretamente conectados a qualquer modelo, o *cartucho* só é considerado completo quando todos os elementos de ω_i se conectam a m_i, isto é, quando $w_{i,i}$ é igual ao número de elementos de ω_i. Esse critério garante a associação bijetiva (ω_i, m_i) que constitui a representação de mapeamento entre domínios. Essa abordagem concilia os mapeamentos não injetivos de elementos de dados para modelos e assegura a consistência da relação entre os domínios A e E.

Na Figura 6(a), há as representações de meta-modelagem de quadriláteros com apenas dois conjuntos de dados: o conjunto de dados de *quadrados* e o conjunto de dados de *retângulos*. Como os quadrados são um caso particular de retângulos, os dois conjuntos de dados compartilham elementos que são explicados por ambos os modelos, conforme ilustrado na Figura 6(b). Enquanto dois elementos de dados do conjunto de dados *retângulos* são mapeados para o modelo de *quadrados*, a totalidade do conjunto de dados *retângulos* é associada ao modelo de *retângulos*. Portanto, o *cartucho* entre o conjunto de dados de *retângulos* e o modelo de *retângulos* é completo, justificando a associação bijetiva na Figura 6(a).

Também há a possibilidade de um conjunto de dados ser totalmente mapeado em dois ou mais modelos. Conforme ilustrado na representação de rede de meta-modelagem na Figura 6(b), isso ocorre com o conjunto de dados de *quadrados*, que está completamente contido no conjunto de dados de *retângulos*. Consequentemente, o conjunto de dados de *quadrados* pode ser explicado pelos dois modelos disponíveis, embora com diferentes níveis de particularidade: o modelo de *quadrados*, um modelo mais específico, e o modelo de *retângulos*, um modelo mais geral.

Quando o domínio A inclui conjuntos de dados que são subconjuntos de outros, como nos conjuntos de dados de quadriláteros do exemplo anterior, a tradução da representação

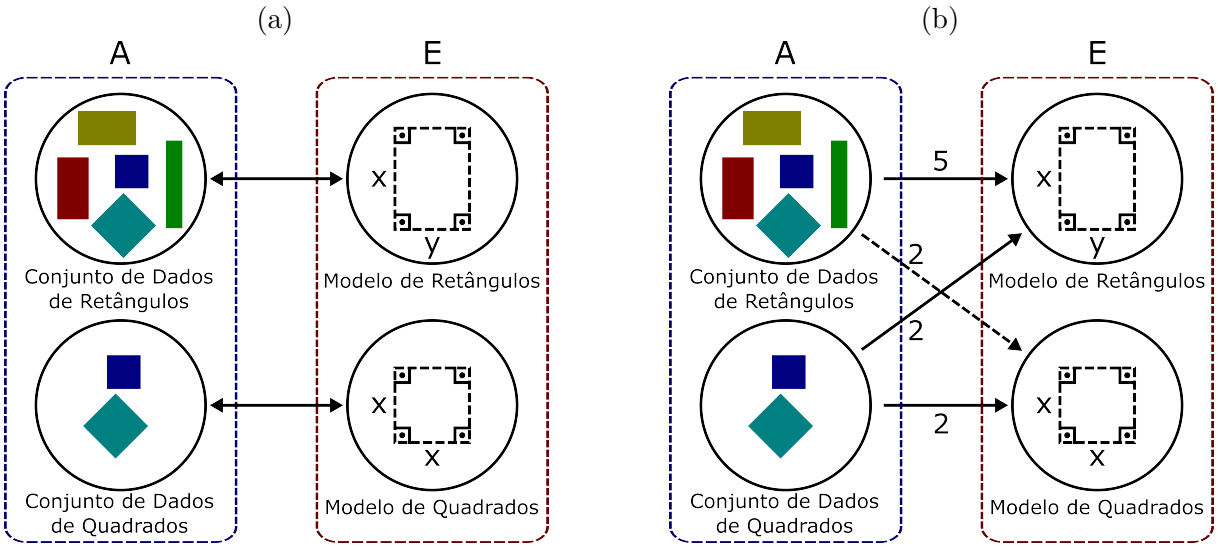


Figura 6 – Exemplo de meta-modelagem de quadriláteros. (a) Representação de mapeamento entre domínios. (b) Representação de rede de meta-modelagem. Embora todos os elementos do conjunto de dados de *quadrados* sejam explicados tanto pelo modelo de *retângulos* quanto pelo modelo de *quadrados*, o conjunto de dados de *quadrados* é mapeado apenas para o segundo. Isso segue a regra de que, para garantir o mapeamento bijetivo entre A e E, um modelo é associado de forma bijetiva ao maior conjunto de dados que ele pode explicar completamente. Portanto, sendo o menor dos dois conjuntos de dados que satisfazem o modelo de *retângulos*, o conjunto de dados de *quadrados* é mapeado apenas para o modelo de *quadrados*.

Fonte: Elaborada pelo autor.

de rede de meta-modelagem de volta para a representação de mapeamento entre domínios se torna ambígua. Isso ocorre porque o *cartucho* do conjunto de dados de *quadrados* pode ser satisfeito ao considerar o modelo de *retângulos* como seu modelo respectivo. A solução para esse problema é criar uma segunda condição necessária para considerar um *cartucho* completo: um modelo m_i é associado apenas ao maior conjunto de dados ω_i que ele pode explicar completamente.

Seguindo as duas condições apresentadas para definir um *cartucho* completo, podemos reconstruir a Figura 6(a) a partir da Figura 6(b) conectando o modelo de *retângulos* ao conjunto de dados de *retângulos* (5 elementos — o maior conjunto de dados que ele pode explicar completamente) e o modelo de *quadrados* ao conjunto de dados de *quadrados* (2 elementos — o único conjunto de dados que ele pode explicar completamente).

2.2.3 Incorporação de Novos Conjuntos de Dados e Modelos

Abordagens de modelagem baseadas em dados começam com observações preliminares que se desenvolvem por meio da coleta e análise de novos dados.(1,3) O surgimento

de um novo conjunto de dados que não pode ser suficientemente explicado pelos modelos já estabelecidos — ou seja, não pode ser classificado, processado e/ou previsto corretamente pelos modelos disponíveis — deve incentivar a construção de novos modelos, expandindo e consolidando a estrutura de modelagem para essa nova circunstância. Isso representa a incorporação de novos conhecimentos em uma estrutura de modelagem, uma característica fundamental da modelagem científica.(1,3)

No meta-modelo $\mathcal{M}$, o domínio A representa todos os conjuntos de dados inicialmente disponíveis, enquanto o domínio E engloba os modelos baseadas por esses dados.(1) Uma coleta de dados adicionais pode ser representada como um novo mapeamento entre os elementos do universo Ω em um novo conjunto de dados em A, que deve ser mapeado em um modelo recém-criado em E. Esse procedimento transmite a ideia de que as estruturas de modelagem podem se expandir, o que é adequado para retratar abordagens baseadas em dados na modelagem científica.

Existem duas maneiras principais de se incorporar novos conjuntos de dados ao ambiente de dados e novos modelos à estrutura de modelagem. A primeira maneira se aplica a novos conjuntos de dados que não são explicados por nenhum modelo disponível. Nesse caso, um novo modelo pode ser construído especificamente para esse conjunto de dados, garantindo assim a associação bijetiva entre os domínios A e E. Como a construção de modelos pode ser difícil, esse novo modelo inicial pode ser pouco informativo ou interpretável — por exemplo, ele pode ser gerado usando algoritmos tradicionais de aprendizado de máquina, com pouca ou nenhuma informação das características físicas dos objetos estudados, ou ele pode ser um mero rótulo provisório para o conjunto de dados, que deve ser desenvolvido em um modelo mais informativo posteriormente.

Vamos voltar ao exemplo prático de meta-modelagem de talheres. A Figura 7 ilustra diferentes formas de incorporar novos conjuntos de dados no domínio de ambiente de dados A. Se um novo conjunto de dados composto por *facas* for coletado de Ω, então a estrutura de modelagem pode ser consolidada incorporando um modelo de *rótulo para facas* em E (Figura 7(a)).

A segunda forma pela qual um conjunto de dados podem ser incorporados no ambiente de dados A do meta-modelo $\mathcal{M}$ se aplica a conjuntos de dados novos que são completamente explicados por um modelo já estabelecido em E. Por exemplo, suponha que um novo conjunto de dados ω_i seja totalmente explicado por um modelo m_k já associado a um conjunto de dados ω_k. Nesse caso, devemos unir os conjuntos de dados ω_i e ω_k em um único conjunto $\omega_\cup = \omega_i \cup \omega_k$, que será mapeado de forma bijetiva ao modelo m_k. Embora o procedimento anterior, de criar um novo modelo para o novo conjunto de dados, também possa ser usado, um novo modelo m_i para o conjunto de dados ω_i pode compartilhar várias características com o modelo m_k já existente. Portanto, esse segundo procedimento, de combinar ω_i e ω_k, pode ser preferível, pois evita redundâncias no ambiente de dados e na

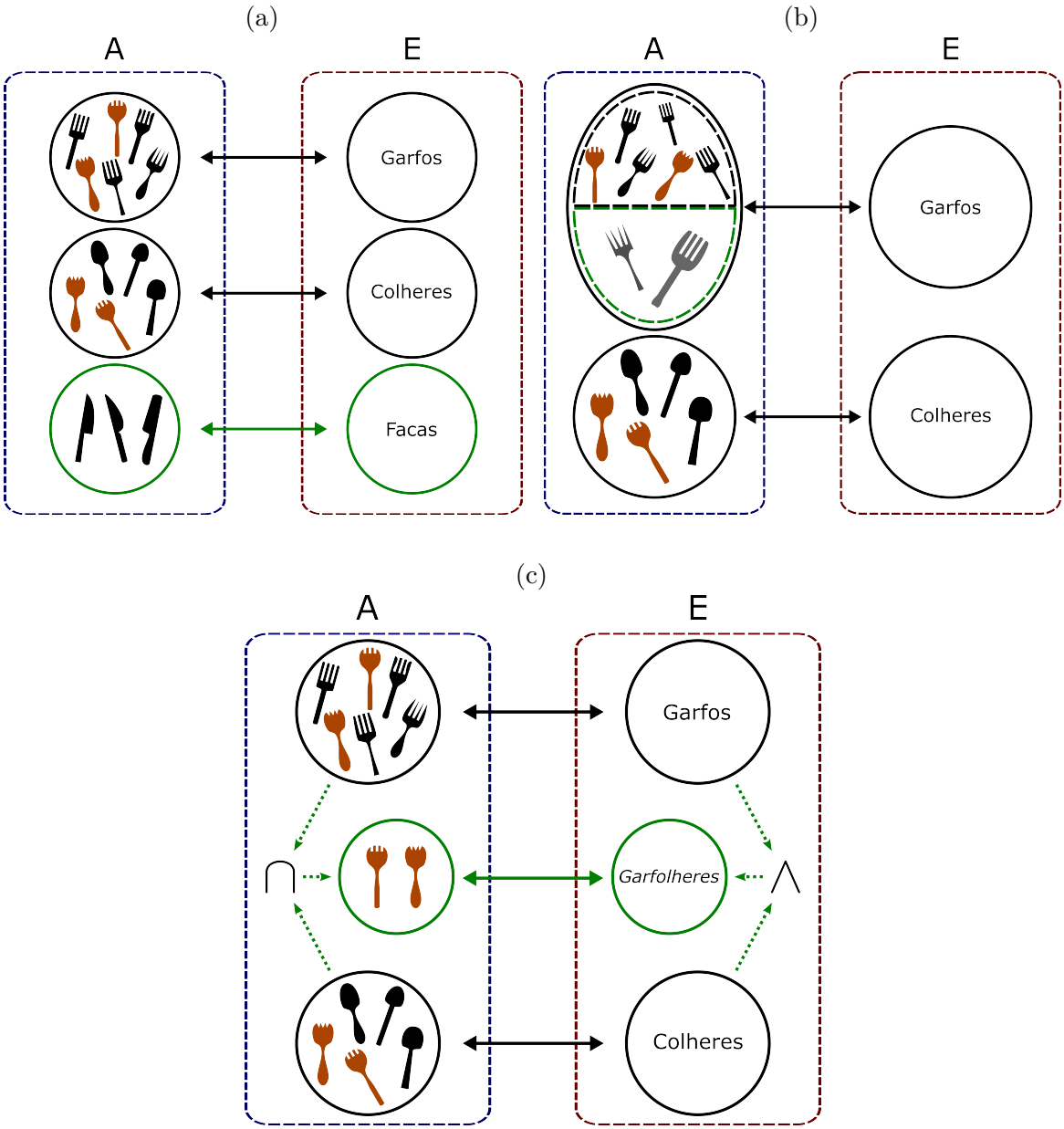


Figura 7 – A incorporação de novos conjuntos de dados no ambiente de dados A. (a) O novo conjunto de dados (verde) não é explicado por nenhum modelo disponível, o que leva à construção de um novo modelo, o modelo de "Facas". (b) O novo conjunto de dados (linhas tracejadas verdes) é totalmente explicado por um modelo disponível em E que já possui um conjunto de dados respectivo em A (linhas tracejadas pretas), levando à união dos dois conjuntos de dados em um só. (c) O novo conjunto de dados (verde) é uma combinação dos dois conjuntos de dados disponíveis em A, e pode ser explicado por uma composição lógica dos dois modelos já disponíveis em E.

Fonte: Elaborada pelo autor.

estrutura de modelagem.

Voltando ao exemplo dos talheres, se um novo conjunto de dados composto por "garfos da cor cinza" for coletado de Ω, qualquer uma das duas abordagens pode ser adotada dependendo da decisão de levar em consideração ou não a cor dos garfos. Se a cor for levada em consideração, o primeiro método deve ser utilizado, resultando em um novo modelo para "garfos da cor cinza". Caso contrário, o segundo procedimento deve ser adotado, com o conjunto de dados de "garfos da cor cinza" sendo unido ao conjunto de dados já existente de "garfos" (Figura 7(b)).

Para garantir a consistência de ambos os procedimentos propostos, todos os conjuntos de dados disponíveis em A devem ser atualizados após cada expansão do meta-modelo $\mathcal{M}$.(1) Cada elemento de dado em A deve ser continuamente verificado em relação a cada par de conjuntos de dados e modelos respectivos, e deve ser adicionado a todos os conjuntos de dados cujo modelo seja satisfeito. Essa abordagem garante consistência ao garantir que um conjunto de dados, em relação a um modelo, sempre contenha todos os elementos de dados que o satisfaçam.

Uma consequência desse procedimento é que, na representação de rede de meta-modelagem, o peso da conexão entre um conjunto de dados ω_i e um modelo m_k deve ser sempre igual ao peso da conexão entre ω_k e m_i. Isso ocorre porque todos os elementos de ω_i que satisfazem m_i e m_k também devem estar em ω_k, assim como os elementos de ω_k que satisfazem m_k e m_i também devem estar em ω_i.

2.2.4 Álgebra Pareada entre Operações de Conjuntos e Operações Lógicas Aplicada à Meta-modelagem

Conforme discutido anteriormente, alguns novos conjuntos de dados extraídos de Ω podem não ser completamente explicados por modelos pré-existentes, o que leva à construção de novos modelos. No entanto, há casos em que novos conjuntos de dados podem ser explicados pela composição de vários modelos disponíveis. Por exemplo, os elementos de um novo conjunto de dados podem satisfazer tanto os modelos m_1 quanto m_2, satisfazendo assim a composição lógica dos modelos $m_1 \vee m_2$. Isso motiva um procedimento formal para explicar novos conjuntos de dados exclusivamente por meio da composição lógica entre modelos pré-existentes.

Esse procedimento explora a associação bijetiva entre conjuntos de dados e modelos que é incorporada ao meta-modelo $\mathcal{M}$. Por exemplo, vamos considerar novamente o caso das *garfolheres*, os talheres híbridos entre garfos e colheres. Um conjunto de dados composto por *garfolheres* pode ser compreendido como a operação entre conjuntos *conjunto de colheres* $\cap$ *conjunto de garfos*. Devido à associação bijetiva entre os domínios A e E, esse novo conjunto de dados é imediatamente relacionado a um modelo correspondente à conjunção lógica *modelo de colheres* $\wedge$ *modelo de garfos*. Essa composição, ilustrada na Figura 7(c),

reflete uma álgebra pareada entre *operações de conjuntos de dados* e *operações lógicas entre modelos*.(1)

A álgebra pareada entre operações de conjuntos e operações lógicas é um produto da já conhecida conexão entre teoria dos conjuntos, lógica e álgebra booliana.(48, 49) Ela é imediatamente observada quando avaliamos, usando sentenças lógicas, o pertencimento de um elemento a um conjunto resultante de uma operação de conjuntos. Por exemplo, x pertence à intersecção dos conjuntos ω_a e ω_b — expressão $x \in (\omega_a \cap \omega_b)$ — se x pertence à ω_a e x pertence à ω_b — expressão lógica $(x \in \omega_a) \wedge (x \in \omega_b)$. A Tabela 1 indica três correspondências principais entre operações de conjuntos e operações lógicas.(48, 49)

Tabela 1 – As principais correspondências entre operações de conjuntos e operações lógicas.(48, 49)

Operação de Conjuntos	Operação Lógica	Correspondência
União $\cup$	Disjunção, "ou", $\vee$	$x \in (\omega_a \cup \omega_b)$ se é verdadeiro que $(x \in \omega_a) \vee (x \in \omega_b)$
Intersecção $\cap$	Conjunção, "e", $\wedge$	$x \in (\omega_a \cap \omega_b)$ se é verdadeiro que $(x \in \omega_a) \wedge (x \in \omega_b)$
Complemento $()^c$	Negação, "não", $\neg$	$x \in (\omega_a)^c$ se é verdadeiro que $\neg(x \in \omega_a)$

Fonte: Elaborada pelo autor.

A aplicação desse repertório teórico à rede de meta-modelagem $\mathcal{M}$ permite associar a montagem de novos conjuntos de dados via operações de conjuntos à construção de novos modelos via operações lógicas. Além do exemplo na Figura 7(b), a Tabela 2 ilustra alguns exemplos de como associar um novo modelo m_k para um conjunto de dados ω_k criado via operações entre conjuntos $\omega_{i \neq k}$.(1) Nesses casos, a operação de complemento, $()^c$, considera o universo de dados como a união de todos os elementos disponíveis no ambiente de dados A, incluindo elementos em ω_k.

Tabela 2 – Exemplos de relações entre operações de conjuntos de dados no ambiente A e operações lógicas entre modelos na estrutura de modelagem E.(1)

A	E
$\omega_k = \omega_i$	$m_k = m_i$
$\omega_k = \omega_i \cup \omega_j$	$m_k = m_i \vee m_j$
$\omega_k = \omega_i \cap \omega_j$	$m_k = m_i \wedge m_j$
$\omega_k = \omega_i^c$	$m_k = \neg m_i$
$\omega_k = (\omega_i \cup \omega_j) \cap \omega_l$	$m_k = (m_i \vee m_j) \wedge m_l$
$\omega_k = (\omega_i \cup \omega_j)^c \cup \omega_l$	$m_k = \neg(m_i \vee m_j) \vee m_l$
...	...

Fonte: Elaborada pelo autor.

Agora, podemos formalizar um procedimento para criar um modelo provisório para um novo conjunto de dados a partir da rede de meta-modelagem. Basta expressar o novo conjunto de dados como o resultado de operação entre conjuntos de dados já modelados no ambiente A, seguido pela tradução dessa expressão para uma composição lógica de modelos na estrutura E.(1) O resultado dessa composição lógica será um novo modelo adequado ao novo conjunto de dados.

Esse procedimento também pode ser usado para verificar como os modelos já disponíveis na estrutura de modelagem estão associados entre si. Por exemplo, dado um conjunto de dados ω_i e seu modelo m_i, é possível combinar modelos alternativos $m_{j \neq i}$ em um novo modelo $\tilde{m}_i$ e verificar em que medida seu poder de explicação se sobrepõe ao modelo m_i. Isso cria a oportunidade de compartilhar os conhecimentos e interpretações disponíveis para alguns modelos já consolidados em E com novos modelos que carecem de informações. Portanto, essa abordagem oferece uma maneira não só de construir novos modelos, como também de caracterizar modelos já disponíveis na estrutura de modelagem em termos dos outros.

A seguir, formalizamos os métodos envolvidos na caracterização da especificidade de uma estrutura de modelagem, bem como os métodos para avaliar o potencial de se usar uma rede de meta-modelagem $\mathcal{M}$ na construção de novos modelos.

2.2.5 Métodos para Quantificar a Especificidade de Estruturas de Modelagem

Em geral, elementos de dados extraídos de Ω podem fazer parte de mais de um conjunto de dados, cada um explicado por um respectivo modelo. Portanto, um modelo pode estar relacionado a elementos que também pertencem a conjuntos de dados de outros modelos, o que indica uma falta de *especificidade* entre os elementos de dados modelados e a estrutura de modelagem E. É importante ressaltar que, embora seja desejável que as estruturas de modelagem sejam compostas apenas por modelos específicos, nem sempre esse é o caso — como no exemplo da meta-modelagem de quadriláteros, da Figura 6, em que dois modelos tem a mesma capacidade de explicar um mesmo conjunto de dados, mas com níveis de particularidade diferentes.

Para quantificar a especificidade dos modelos em uma estrutura de modelagem, podemos considerar algumas métricas: a *cardinalidade, número e percentual de elementos inespecíficos*, a *multiplicidade média de elementos*, e a *diversidade de conexões* na rede de meta-modelagem. Cada uma dessas métricas se concentra em um aspecto diferente da especificidade da estrutura de modelagem e de seus modelos.

2.2.5.1 Cardinalidade, número e percentual de elementos inespecíficos dos conjunto de dados

Dentro de cada conjunto de dados em A existem dois tipos de elementos: os *elementos específicos*, que estão presentes apenas em um conjunto de dados e que só é explicado por um único modelo; e os *elementos inespecíficos*, que estão presentes em mais de um conjunto de dados e que, portanto, podem ser explicados por múltiplos modelos.

É importante lembrar que cada conjunto de dados em A é agrupado a partir de características ou fenômenos observados no mundo real, e um modelo para explicar essas características com base nos dados é construído e incorporado em E. Assim sendo, um conjunto de dados com vários elementos inespecíficos representa a união de elementos com características diversas que já são explicadas, parcial ou totalmente, por modelos alternativos presentes na rede. Em contrapartida, um modelo construído com base em um conjunto de dados com muitos elementos inespecíficos irá ser capaz de explicar, parcial ou totalmente, vários conjuntos alternativos, possivelmente sendo redundante em relação aos outros modelos em E.

Assim sendo, é interessante saber o tamanho de cada conjunto de dados, o número de elementos inespecíficos que eles possuem e qual é a proporção, ou o percentual, que eles representam.

A *cardinalidade* de um conjunto de dados é número de elementos que ele contém, e é denotada pela Equação:

$$C_i = |\omega_i| \tag{2.1}$$

Representamos a cardinalidade do conjunto ω_i simplesmente por C_i para facilitar a notação.

O *número de elementos inespecíficos* de um conjunto ω_i é a cardinalidade do conjunto de elementos inespecíficos presentes em ω_i. O conjunto de elementos inespecíficos é a intersecção entre o conjunto ω_i e a união de todos os conjuntos alternativos $\omega_{j \neq i}$ presentes no ambiente de dados A. Portanto, o número de elementos inespecíficos de um conjunto ω_i é dado pela equação:

$$CI_i = \left| \omega_i \cap \left(\bigcup_{j \neq i} \omega_j \right) \right| \tag{2.2}$$

Portanto, o *percentual de elementos inespecíficos* em um conjunto ω_i é estimado pela equação:

$$PI_i = 100 * CI_i / C_i \% \tag{2.3}$$

2.2.5.2 Multiplicidade média de elementos

A *multiplicidade* de um elemento é o número de vezes que o elemento se repete em diferentes conjuntos de dados. Como cada conjunto de dados possui um modelo respectivo específico, se o mesmo elemento de dados aparecer em vários conjuntos de dados, esse elemento se ajustará a diferentes modelos. Em outras palavras, a multiplicidade de um elemento indicará o número de modelos indiretamente ligados a ele. A *multiplicidade média de elementos* é a média de quantos modelos são usados para explicar cada elemento de dados disponível em A, sendo assim uma métrica de especificidade geral da estrutura de modelagem E.

Se um determinado meta-modelo $\mathcal{M}$ possui N elementos de dados distintos distribuídos em n conjuntos de dados, a multiplicidade média de elementos será igual à soma da cardinalidade C_i de cada conjunto de dados ω_i dividida por N. Isso é calculado pela a Equação 2.4:

$$\bar{U}(\mathcal{M}) = \frac{1}{N}\sum_{i=1}^{n} C_i \tag{2.4}$$

Essa métrica possui um limite inferior igual a 1, onde cada elemento pertence a apenas um conjunto de dados, e um limite superior igual a n, correspondendo ao caso extremo em que todos os elementos pertencem a todos os conjuntos em A.

2.2.5.3 Diversidade de conexões

A *diversidade de conexões* é uma medida de especificidade calculada sobre cada modelo individual em E. Essa métrica é baseada na forma como são distribuídos os pesos das conexões presentes entre os vértices da rede de meta-modelagem.

Para estimar a diversidade de conexões, tratamos os pesos das conexões recebidas por um modelo como uma distribuição discreta de valores e, em seguida, calculamos a exponencial da entropia dessa distribuição. A entropia $H(m_i)$ dos pesos das conexões recebidas por um modelo m_i é calculada usando a Equação 2.5, enquanto a *diversidade de conexões* $\mathcal{D}(m_i)$ é obtida a partir da Equação 2.6:

$$H(m_i) = -\sum_{k=1}^{n}\left(\frac{w_{k,i}}{\sum_{j=1}^{n} w_{j,i}} \ln \frac{w_{k,i}}{\sum_{j=1}^{n} w_{j,i}}\right) \tag{2.5}$$

$$\mathcal{D}(m_i) = \exp(H(m_i)) \tag{2.6}$$

onde n representa o número de conjuntos de dados no ambiente de dados A, $w_{k,i}$ é o peso da conexão entre o conjunto ω_k e o modelo m_i na rede de meta-modelagem, ln é o logaritmo natural. Nas equações, $0 \leq H(m_i)$ e $1 \leq \mathcal{D}(m_i)$.

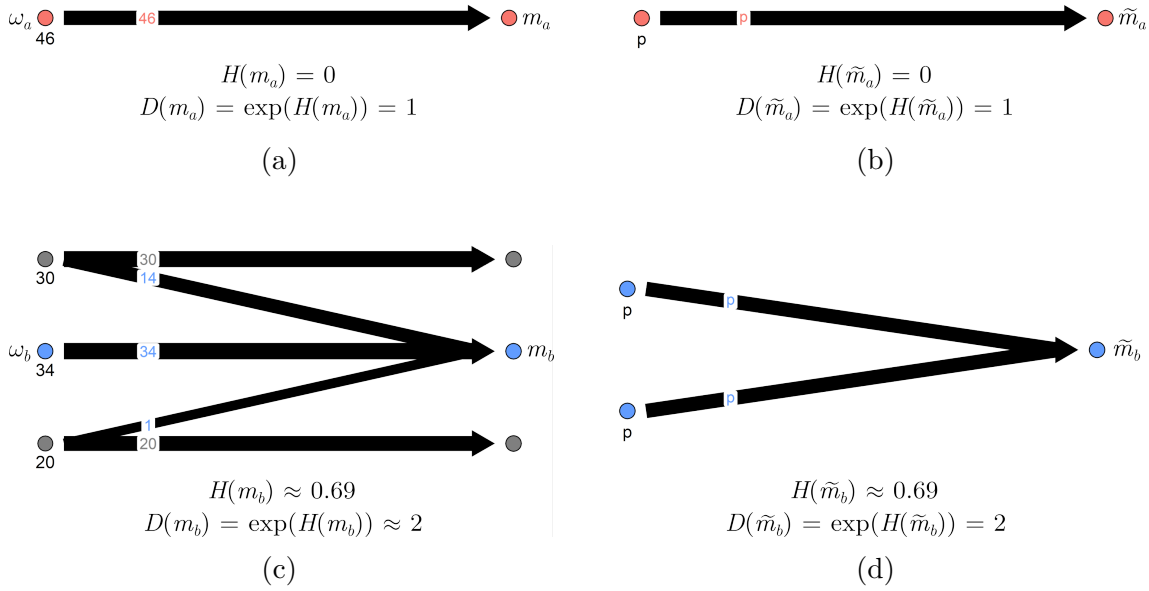


Figura 8 – A diversidade das conexões $\mathcal{D}$ de um modelo m indica o número de conjuntos de dados que podem se conectar a um modelo $\tilde{m}$ com uma distribuição uniforme de pesos p, mantendo a mesma entropia observada nas conexões de m. O modelo m_a na figura (a) só se conecta ao seu conjunto de dados respectivo, ω_a. Ele possui uma distribuição de pesos com entropia nula, $H(m_a) = 0$, tendo uma diversidade de conexões $D(m_a)$ igual a 1. Este modelo possui entropia equivalente à do modelo arbitrário $\tilde{m}_a$, que se conecta a apenas um conjunto de dados com peso p, representado na figura (b). O modelo m_b na figura (c) se conecta a três conjuntos de dados, incluindo seu respectivo conjunto de dados ω_b, com uma distribuição de pesos com entropia $H(m_b) \approx 0.69$. A diversidade das conexões de m_b é $D(m_b) \approx 2$, que é o número aproximado de conjuntos de dados conectados uniformemente ao modelo arbitrário $\tilde{m}_b$ com a mesma entropia de conexões, representado na figura (d).

Fonte: Elaborada pelo autor.

Como a entropia de uma distribuição é uma medida de incerteza, a entropia das conexões ponderadas recebidas por um modelo partindo de diferentes conjuntos de dados mostra o nível de incerteza desse modelo sobre os conjuntos de dados disponíveis. Este princípio pode ser exemplificado pelos modelos m_a e m_b das duas redes de meta-modelagem representados na Figura 8. O modelo m_a possui apenas uma conexão ponderada proveniente de apenas um conjunto de dados, a entropia de sua distribuição de pesos é igual a 0 e não há incerteza sobre sua conexão com os conjuntos de dados (Figura 8(a)). Alternativamente, o modelo m_b tem múltiplas conexões ponderadas, produzindo uma distribuição de entropia diferente de zero, que revela uma maior incerteza em relação aos conjuntos de dados (Figura 8(c)).

A exponencial de uma entropia — chamada de entropia exponencial — é igual a x quando aplicada a uma distribuição uniforme de x classes. Essa propriedade pode ser usada

para interpretar o valor da entropia exponencial de uma distribuição aleatória como o número de classes de uma distribuição uniforme equivalente. Por esses motivos, a entropia exponencial é frequentemente adotada como medida da diversidade de distribuições: quanto maior o valor da entropia exponencial, maior a incerteza e mais diversa é sua distribuição uniforme equivalente.(50–52)

Quando aplicada ao nosso caso, a entropia exponencial das conexões ponderadas fornecerá a *diversidade de conexões* de um modelo m_i. Utilizando a propriedade da entropia exponencial, a *diversidade de conexões* indica o número de categorias (conexões) de uma distribuição uniforme equivalente de pesos. Podemos aprofundar essa interpretação considerando que a *diversidade de conexões* de um modelo m_i é igual ao número de conjuntos de dados diferentes que esse modelo poderia explicar com o mesmo nível de incerteza, conforme ilustrado na Figura 8.

Na Figura 8(a), o modelo m_a com apenas uma conexão ponderada possui uma entropia exponencial igual a 1, ou seja, esse modelo tem a incerteza equivalente a se conectar a apenas um conjunto de dados — assim como o modelo arbitrário $\tilde{m}_a$ na Figura 8(b). Na Figura 8(c), o modelo com duas conexões ponderadas tem sua entropia exponencial aproximadamente igual a 2, ou seja, possui a incerteza equivalente a se conectar uniformemente a dois conjuntos de dados diferentes do mesmo tamanho — como o modelo arbitrário $\tilde{m}_b$ na Figura 8(d)

Podemos usar as métricas anteriores para caracterizar a estrutura de modelagem da rede de meta-modelagem de talheres da Figura 5. Todo o meta-modelo tem uma *multiplicidade média de elementos* de aproximadamente 1,22, mostrando que a maioria dos elementos é bastante específica para seus conjuntos de dados. O *conjunto de dados de garfos* possui 6 elementos, 2 deles sendo inespecíficos, 33,33% de seus dados. Já o *conjunto de dados de colheres* possui 5 elementos, também com 2 elementos inespecíficos, que agora representam 40% dos dados. A *diversidade de conexões* dos *modelos de garfos* e de *colheres* é 1,75 e 1,82, respectivamente. Ambos os conjuntos de dados são bastante inespecíficos, pois possuem uma diversidade quase tão alta quanto estarem igualmente conectados a dois conjuntos de dados diferentes. No entanto, podemos notar que o *modelo de garfos* é um pouco mais específico que o de *colheres*, pois possui um proporção menor de elementos inespecíficos.

2.2.6 Aplicando a Álgebra Pareada para Construir e Avaliar Modelos

Como visto anteriormente, podemos utilizar a álgebra pareada entre operações de conjuntos e operações lógicas para construir modelos m_k respectivos a novos conjuntos de dados ω_k. Para isso, basta expressar ω_k em termos de uma operação entre conjuntos de dados $\omega_{i\neq k}$ que já são modelados por modelos $m_{i\neq k}$ e, então, traduzir essa expressão para uma operação lógica entre modelos $m_{i\neq k}$.

Esse procedimento é ótimo quando um novo conjunto de dados ω_k é *idêntico* ao resultado de uma operação de conjunto entre outros conjuntos de dados $\omega_{i \neq k}$. Nesses casos, o modelo m_k construído via operações lógicas necessariamente explicará todo ω_k. Isso se aplica aos vários exemplos apresentados na Tabela 2. No entanto, a situação mais comum é que ω_k é apenas parcialmente coberta pela combinação de outros conjuntos de dados, sendo possível obter apenas um conjunto de dados aproximado $\omega_{k(aprox)}$ e um modelo aproximado $m_{k(aprox)}$ através de operações lógicas.

Para quantificar o quanto um modelo aproximado $m_{k(aprox)}$ está próximo do modelo exato desejado, m_k, respectivo ao conjunto de dados ω_k, adotamos o *índice de coincidência.*(53, 54) O *índice de coincidência* compara dois conjuntos de dados com uma combinação do índice de Jaccard e do índice de interioridade, incorporando informações sobre a intersecção relativa e a interioridade relativa entre os conjuntos.(53) No nosso caso, compararemos os modelos m_k e o modelo $m_{k(aprox)}$ através do índice de coincidência entre seus respectivos datasets exatos, o conjunto ω_k e o conjunto $\omega_{k(aprox)}$.

O índice Jaccard (53–57) é usado para quantificar a similaridade entre dois conjuntos de dados, ω_a e ω_b, com base no número de elementos que eles têm em comum frente ao número de elementos disponíveis em ambos:

$$\mathcal{J}(\omega_a, \omega_b) = \frac{|\omega_a \cap \omega_b|}{|\omega_a \cup \omega_b|} \tag{2.7}$$

O índice de interioridade (53, 54), também chamado de coeficiente de Szymkiewicz–Simpson ou índice de sobreposição (56, 58, 59), quantifica o quanto um conjunto de dados é interior ao outro:

$$\mathcal{I}(\omega_a, \omega_b) = \frac{|\omega_a \cap \omega_b|}{min\{|\omega_a|, |\omega_b|\}} \tag{2.8}$$

O índice de coincidência (53, 54) entre dois conjuntos de dados, ω_a e ω_b, pode ser calculado utilizando a Equação 2.9:

$$\mathcal{C}(\omega_a, \omega_b) = \mathcal{J}(\omega_a, \omega_b)\, \mathcal{I}(\omega_a, \omega_b) \tag{2.9}$$

onde $0 \leq \mathcal{C}(\omega_a, \omega_b) \leq 1$.

Este índice de coincidência permite comparar diferentes modelos aproximados $m_{k(aprox)}$ para um novo conjunto de dados ω_k, redefinindo o problema de construção do modelos como um problema de otimização. Nesse problema, busca-se encontrar a combinação $\omega_{k(aprox)}$ de conjunto de dados em A com a maior coincidência ao conjunto ω_k, fornecendo o modelo parcial $m_{k(aprox)}$ com o maior poder de explicação possível entre as composições lógicas de modelos presentes em E.(1)

Além disso, o índice de coincidência pode ser usado para quantificar quanto de um conjunto de dados ω_k é explicado pela composição lógica de modelos alternativos

$m_{j\neq k}$ disponíveis na estrutura de modelagem E. Usando a álgebra pareada entre lógica e operações de conjuntos, essa quantificação é realizada comparando combinações de conjuntos de outros conjuntos de dados $\omega_{j\neq k}$ com o ω_k. Se um conjunto de dados ω_k e uma dada combinação de conjuntos $\omega_{j\neq k}$ tiverem valor de coincidência igual a 1, então o modelo m_k pode ser visto como redundante, afinal seu conjunto de dados é suficientemente explicado com a ajuda da composição entre modelos alternativos $m_{j\neq k}$. Além disso, se os modelos alternativos $m_{j\neq k}$ tiverem descrições e interpretações sólidas, é possível extrair informações de sua composição lógica e compartilhá-las com o modelo m_k, que pode ter um comportamento de difícil compreensão.

Esta análise complementa a caracterização de especificidade da estrutura de modelagem E focando em acessar o nível de redundância de seus modelos. Além disso, fornece uma maneira de interpretar melhor um modelo em termos dos outros. Este procedimento também pode indicar modelos que podem ser decompostos e simplificados na estrutura de modelagem, aumentando sua simplicidade.

2.2.7 Análise de Redes de Meta-modelagem

Para analisar e caracterizar uma rede de meta-modelagem $\mathcal{M}$ e todos os modelos em sua estrutura, utilizamos as seguintes métricas:

a) *Cardinalidade, número e percentual de elementos inespecíficos*, calculados usando as Equações 2.1, 2.2, 2.3. Essas métricas caracterizam cada conjunto de dados em A e auxiliam a quantificar a especificidade dos seus modelos respectivos em E.

b) *Multiplicidade média de elementos*, $\bar{U}(\mathcal{M})$, calculado usando a Equação 2.4. Essa métrica caracteriza os domínios A e E como um todo, fornecendo o número médio de repetições de um elemento entre os conjuntos de dados disponíveis em A. Quanto maior o valor de $\bar{U}(\mathcal{M})$, mais repetitivo são os elementos nos conjuntos de dados, indicando uma sobreposição do poder de explicação dos modelos disponíveis.

c) *Diversidade de conexões*, $\mathcal{D}(m_i)$, calculada usando a Equação 2.6. Essa métrica caracteriza cada modelo m_i presente na estrutura de modelagem E. Ela indica o número de conjuntos de dados distintos que m_i poderia explicar com o mesmo nível de incerteza contido em suas conexões com o ambiente de dados A.

d) *Diversidade total de conexões*, $\mathcal{D}_\mathcal{T}(\mathcal{M})$, estimada usando a Equação:

$$\mathcal{D}_\mathcal{T}(\mathcal{M}) = \sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{n} \frac{w_{i,i}}{\sum_{j=1}^{n} w_{j,i}} \log \frac{w_{i,i}}{\sum_{j=1}^{n} w_{j,i}} \right) \tag{2.10}$$

em que $n \leq \mathcal{D}_\mathcal{T}(\mathcal{M})$. A *diversidade total de conexões* é a soma da diversidade de cada modelo. Portanto, ela indica o número total de conjuntos de dados

disjuntos que poderiam ser totalmente conectados aos modelos disponíveis na estrutura de modelagem E com o mesmo nível de incerteza contido nas conexões presentes no meta-modelo. Quando $\mathcal{D}_{\mathcal{T}}(\mathcal{M}) = n$, os modelos de E são altamente específicos, e cada um deles possui a incerteza equivalente a ter apenas um conjunto de dados conectado. Quando $\mathcal{D}_{\mathcal{T}}(M) > n$, os modelos de E são inespecíficos, e pelo menos um deles possui a incerteza equivalente a ter mais de um conjunto de dados conectado.

Além disso, cada modelo m_k foi comparado a modelos m_X, construídos através de composições lógicas de modelos alternativos $m_{j\neq k}$ presentes na estrutura de modelagem. Essa comparação foi realizada a fim de avaliar a redundância de cada modelo m_k frente a composições lógicas m_X na explicação de seu conjunto de dados $omega_k$. Essa redundância é quantificada calculando o *índice de coincidência* (Equação 2.9) entre o conjunto ω_k e o conjunto ω_X, respectivo a m_X, como discutido na Subseção 2.2.6.

Cada modelo m_X é construído seguindo a equação lógica:

$$m_X(m_k) = \left(\bigwedge_{i\neq k} \neg m_i \right) \vee \left(m_O(m_k) \right) \tag{2.11}$$

Usando a rede de meta-modelagem, podemos traduzir as operações lógicas entre modelos na Equação 2.11 para definir operações entre seus respectivos conjuntos de dados (consulte as Tabelas 1 e 2), obtendo:

$$\omega_X(\omega_k) = \left(\bigcap_{i\neq k} \omega_i^c \right) \cup \left(\omega_O(\omega_k) \right) \tag{2.12}$$

O primeiro termo representa os elementos específicos do conjunto de dados ω_k, ou seja, que não são explicados por nenhum outro modelo além de m_k. Esses casos são comuns em estruturas de modelagem formadas por modelos mutuamente exclusivos, o que significa que podemos construir um modelo simplesmente combinando o negativo dos outros. O segundo termo, ω_O, é qualquer combinação de um número O de conjuntos de dados alternativos usando qualquer uma das três operações básicas entre conjuntos: união — $\cup$; intersecção — $\cap$; e complemento — $()^c$.

Obtivemos todos os conjuntos de dados ω_O possíveis usando $O = 1$ a $O = 4$, seguindo a construção iterativa:

$$\omega_{O=1}(\omega_k) = \left(\omega_{j\neq k}^{\square} \right) \tag{2.13}$$

$$\omega_{O=2}(\omega_k) = \omega_{O=1}(\omega_k) \blacksquare \left(\omega_{j\neq k}^{\square} \right) \tag{2.14}$$

$$\omega_{O=3}(\omega_k) = \omega_{O=2}(\omega_k) \blacksquare \left(\omega_{j\neq k}^{\square} \right) \tag{2.15}$$

$$\omega_{O=4}(\omega_k) = \omega_{O=3}(\omega_k) \blacksquare \left(\omega_{j\neq k}^{\square} \right) \tag{2.16}$$

onde $()^{\square}$ pode ser a operação $()^{c}$, e $\blacksquare$ é $\cup$ (união) ou $\cap$ (intersecção). Observe que equações mais complexas, como aquelas com operações entre parênteses, não são incluídas — por exemplo, a equação $((\omega_1 \cup \omega_2) \cap (\omega_3 \cap \omega_4))$ não é avaliada, mas a equação $(((\omega_1 \cup \omega_2) \cap \omega_3) \cap \omega_4)$ é.

O *índice de coincidência máximo* entre ω_k e cada ω_X possível representa a coincidência máxima entre o modelo m_k e qualquer composição lógica de modelos $m_{j \neq k}$. Essa quantia é estimada pela Equação 2.17:

$$\mathcal{C}_{max}(m_k) = \max_{\omega_X(\omega_k)} \mathcal{C}(\omega_k, \omega_X(\omega_k)) \tag{2.17}$$

em que $\max_{\omega_X(\omega_k)}$ denota o valor máximo obtido para qualquer $\omega_X(\omega_k)$ avaliado, e $\mathcal{C}$ denota a função do índice de coincidência, mostrado na Equação 2.9.

Se $\mathcal{C}_{max}(m_k)$ estiver próximo de 1, então m_k é suficientemente explicado pela combinação de modelos alternativos disponíveis, o que indica que esse modelo é redundante dentro de sua estrutura de modelagem. Se esse número for baixo, então m_k é um modelo único que não pode ser adquirido com as informações embutidas nos outros modelos.

2.3 Experimentos

Construímos e analisamos redes de meta-modelagem $\mathcal{M}$ de duas estruturas de modelagem: (i) uma estrutura para reconhecimento de padrões em sequências de símbolos binários; e (ii) uma estrutura que classifica domínios funcionais de enzimas ativas em carboidratos.

Em cada exemplo, começamos com uma coleção inicial de seis conjuntos de dados que já possuem modelos explicativos preliminares. Esses conjuntos são dispostos no domínio do ambiente de dados A e seus modelos integram a estrutura de modelagem E. Em seguida, construímos a rede de meta-modelagem ao fixar conexões ponderadas entre os conjuntos em A e os modelos em E com base no número de elementos presente em cada conjunto que se encaixam nas definições de cada modelo. Após a construção da rede, caracterizamos a estrutura de modelagem utilizando as métricas apresentadas na subseção 2.2.7.

Nos dois casos, os conjunto de dados e os seus modelos respectivos foram escolhidos por apresentarem características bem definidas. Isso permitiu validar os resultados obtidos pela análise das redes de meta-modelagem. No entanto, vale ressaltar que esse não é o caso da maioria das abordagens de modelagem, que podem ter modelos incompletos e pouco informativos.

O primeiro caso estudado é um exemplo de teste a fim de explorar as principais análises permitidas pela rede de meta-modelagem. Nesse exemplo, analisamos sequências de símbolos binários divididas em conjuntos de dados a partir dos padrões que elas apresentam. Essas sequências são analogias simplificadas de proteínas, que nada mais são que sequências

de aminoácidos, ou símbolos, classificadas a partir de domínios funcionais, ou padrões, contidos em suas estruturas. Já no segundo caso, aplicamos a rede de meta-modelagem para analisar a anotação e classificação de domínios funcionais em enzimas ativas em carboidratos.

2.3.1 Padrões em Sequências de Símbolos Binários

O primeiro exemplo aborda a modelagem de padrões em seis conjuntos de dados de sequências de 9 símbolos binários, representados por quadrados pretos ou brancos.

Inicialmente, geramos todas as 512 sequências possíveis usando apenas quadrados pretos ou brancos, e depois as dividimos em seis conjuntos de dados de acordo com seis tipos de padrões que poderiam estar presentes nelas. Dessa forma, cada conjunto é explicado por um modelo que define o padrão contido em seus elementos.

Os conjuntos de dados analisados são representados por ω_i, enquanto seus modelos respectivos são representados por m_i, com i variando de 1 a 6. Os padrões presentes em cada conjunto de dados e explicados por cada modelo são descritos na Tabela 3.

Tabela 3 – Descrição dos modelos respectivos a cada conjunto de dados de sequências de símbolos binários.

Conjunto de Dados (ω_i)	Modelo (m_i)
ω_1	m_1: Sequências palindrômicas, isto é, a sequência permanece a mesma quando ordenada de trás para frente.
ω_2	m_2: Sequências cujos símbolos nas posições 2, 3 e 4 são iguais aos símbolos 6, 7 e 8.
ω_3	m_3: Sequências cujos símbolos nas posições 1, 2 e 6 são iguais aos símbolos 9, 4 e 8.
ω_4	m_4: Sequências que possuem quadrados pretos em pelo menos uma das posições 1, 3, 7 ou 9.
ω_5	m_5: Sequências que possuem quadrados pretos em pelo menos uma das posições 2, 4, 6 ou 8.
ω_6	m_6: Sequências em que mais da metade dos símbolos são quadrados pretos.

Fonte: Elaborada pelo autor.

A Figura 9 apresenta quatro instâncias de sequências classificadas usando a estrutura de modelagem anterior. Cada uma das sequências representadas pertence a mais de um conjunto de dados, pois satisfazem mais de um padrão apresentado na Tabela 3. Por exemplo, a sequência rotulada como 495 na Figura 9 pertence a quatro conjuntos de dados

diferentes: ω_2, pois os símbolos da posição 2, 3 e 4 são iguais aos símbolos 6, 7 e 8; ω_4, pois possui um quadrado preto nas posições 1, 3 e 7; ω_5, pois possui um quadrado preto nas posições 2, 4, 6 e 8; e ω_6, pois possui mais da metade de seus símbolos são quadrados pretos.

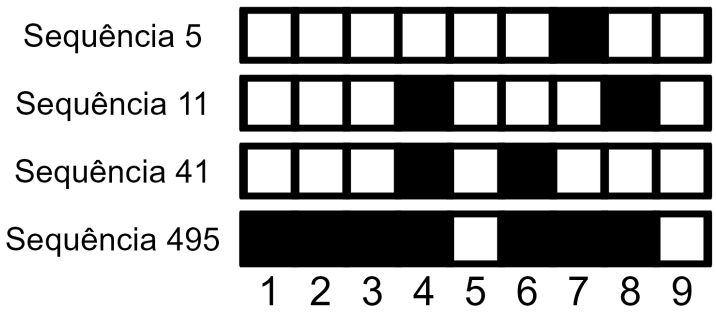


Figura 9 – Quatro exemplos de sequências de 9 símbolos binários. Os padrões são rotulados por números na ordem em que são criados usando um algoritmo de permutação. Os padrões mostrados foram incluídos em diferentes conjuntos de dados de acordo com os padrões apresentados na Tabela 3. A sequência 5 está incluída no conjunto de dados ω_3, pois possui símbolos nas posições 1, 2 e 6 iguais aos símbolos das posições 9, 4 e 8, e no conjunto ω_4, pois possui um quadrado preto na posição 7. A sequência 11 está contida no conjunto ω_2, pois os símbolos 2, 3 e 4 são iguais aos 6, 7 e 8, e no conjunto ω_5, pois possui quadrados pretos nas posições 2 e 8. A sequência 41 está contida no conjunto ω_1 e ω_5, pois sua sequência é palindrômica e há quadrados pretos nas posições 4 e 6. A sequência 495 pertence aos conjuntos ω_2, ω_4, ω_5 e ω_6, possui a trinca de símbolos 2, 3, 4 igual à trinca 6, 7, 8, e a trinca 1, 2, 6 igual à 9, 4, 8, além de possui quadrados pretos em mais da metade da sequência, incluindo as posições 2, 4, 6 e 8.

Fonte: Elaborada pelo autor.

Os conjuntos de dados e modelos analisados são baseados em padrões de simetria e repetição de símbolos nas sequências, mas poderiam ser mais complexos. Por exemplo, modelos válidos poderiam ser definidos pela detecção de alguma série de poucos símbolos encontrada embutida no início das sequências analisadas. Eles também poderiam ser modelos híbridos, que integram padrões de símbolos e informações adicionais de cada elemento. De qualquer forma, a rede de meta-modelagem $\mathcal{M}$ depende apenas da relação entre os modelos e seus conjuntos de dados, e não de como esses modelos foram obtidos.

A coleção desses seis conjuntos de dados de sequências forma o ambiente de dados A e a coleção de modelos de padrões compõe a estrutura de modelagem E, que foram usadas para construir a rede de meta-modelagem $\mathcal{M}$. Após a construção da rede, exploramos os métodos de quantificação de especificidade dos modelos e discutimos a formação de uma

novo modelo a partir da composição lógica de modelos disponíveis na estrutura E, como é apresentado na subseção 2.2.7.

2.3.2 Domínios Funcionais de Enzimas Ativas em Carboidratos

No segundo exemplo, apresentada nessa subseção, exploramos o potencial de aplicação de redes de meta-modelagem $\mathcal{M}$ para extrair informações funcionais e evolutivas de anotações de domínios proteicos. Para esse fim, escolhemos analisar a anotação de domínios funcionais de enzimas ativas em carboidratos, disponível no banco de dados CAZy.(43, 44)

A anotação e classificação de domínios presente no banco de dados CAZy é considerada de alto nível, pois acopla diferentes métodos automáticos de anotação de proteínas com curadoria manual de especialistas.(43, 60) Além disso, há um grande volume de estudos sobre função, estrutura e evolução de domínios de enzimas ativas carboidratos disponíveis na literatura científica.(43, 60) Isso permite tratar esse banco de dados como um protótipo robusto para validar novos métodos de análise exploratória, afinal, cada hipótese concebida a partir de um novo método de análise pode ser validada (ou descartada) com base em estudos publicados.

Portanto, a análise da anotação de domínios funcionais do banco CAZy será tratada como um experimento de validação da rede de meta-modelagem $\mathcal{M}$ como um novo método de análise para anotações de proteínas.

2.3.2.1 O Banco de Dados CAZy

O banco de dados CAZy, disponível *online* desde 1998, reúne informações genômicas e funcionais de enzimas ativas em carboidratos, isto é, enzimas que degradam, modificam ou criam ligações glicosídicas.

Esse banco classifica quase 3 milhões de enzimas em classes de acordo com os tipos de domínios funcionais anotados em suas estruturas.(43, 44) Atualmente, são reconhecidas seis classes funcionais, listadas a seguir:

a) Atividades Auxiliares (AA);

b) Liases de Polissacarídeos (PL, do inglês *Polysaccharide Lyases*);

c) Esterases de Carboidratos (CE, do inglês *Carbohydrate Esterases*);

d) Módulos de Ligação a Carboidratos (CBM, do inglês *Carbohydrate-Binding Module*);

e) Glicosiltransferases (GT);

f) Glicosidases (GH, do inglês *Glycoside Hydrolase*).

Dentro do banco CAZy, cada uma dessas classes forma um conjunto de dados que reúne enzimas cujos domínios funcionais possuem um grau mínimo de semelhança. Por exemplo, todas as enzimas classificadas no grupo GH possuem, pelo menos, um domínio com atividade hidrolítica. No entanto, é importante ressaltar que isso permite que uma mesma enzima seja classificada em diferentes classes, uma vez que ela pode conter domínios proteicos com atividades funcionais características de classes distintas.

Mais especificamente, cada classe funcional é subdivididas em famílias, que agrupam enzimas com semelhanças ainda maiores em termos de sequência de aminoácidos, estrutura, função e mecanismo catalítico.(43, 61) Famílias ainda podem ser subdivididas em subfamílias, a depender das funções catalíticas e das relações filogenéticas entre seus membros.(43, 61–63) Na prática, a classificação em famílias e subfamílias é a mais importante no banco CAZy, formando subconjuntos de dados com características mais homogêneas dentro dos conjuntos respectivos às classes funcionais.

A partir de cada conjunto e subconjunto de dados, o banco CAZy constrói modelos que servirão de base para integrar informações e explicar comportamentos dos domínios funcionais. Para construir esses modelos, são usados métodos de classificação de sequências de aminoácidos, incluindo alinhamento e HMM, com auxílio de dados estruturais e bioquímicos das enzimas analisadas.(43,60,64) Uma vez construídos, esses modelos podem ser usados para detectar domínios em novas enzimas obtidas a partir do sequenciamento de novos genomas.(43,60) Essas enzimas serão, então, integradas nos conjuntos e subconjuntos de dados cabíveis e irão ser usadas para atualizar os modelos.(43, 64, 65)

2.3.2.2 Construção e Análise da Rede de Meta-modelagem

Para construir a rede de meta-modelagem $\mathcal{M}$ da anotação de domínios funcionais de enzimas ativas em carboidratos, obtivemos os dados da última versão do banco CAZy (versão atualizada em 23 de maio de 2023 (44)) a partir da função *Download CAZy* disponível no site.(44) Com ela, obtivemos a anotação e classificação de domínios presentes em 2'831'365 enzimas.

Para construir a rede de meta-modelagem $\mathcal{M}$, escolhemos analisar as seis classes de domínios funcionais, de forma que cada classe ocupa um conjunto diferente no ambiente de dados A. Por sua vez, seus modelos respectivos são incorporados na estrutura de modelagem E. Nessa rede, as famílias e subfamílias correspondem a subconjuntos dos conjuntos presentes em A. A elas podemos associar submodelos, que são instâncias mais particulares dos modelos em E.

Portanto, as enzimas, tratadas como elementos de dados, foram divididas em seis conjuntos de dados conforme a classe de seus domínios funcionais. É importante ressaltar que cada enzima pode conter mais de um domínio funcional, podendo ser inserida em múltiplos conjuntos de dados. Nós também mantivemos a informação sobre as famílias

e subfamílias as quais essas enzimas são classificadas, a fim de obter informações mais detalhadas das relações observadas entre as classes funcionais.

Os seis pares de conjuntos de dados e modelos respectivos a cada classe são descritos na Tabela 4. Os seis conjuntos presentes no ambiente de dados A são representados pelos símbolos ω_{XX}, em que XX pode ser AA, PL, CE, CBM, GT ou GH — a sigla das seis classes de domínios funcionais em enzimas ativas em carboidratos. Os seus modelos respectivos na estrutura de modelagem E são representados pelos símbolos m_{XX}.

Após a construção da rede, exploramos os métodos de quantificação de especificidade e interação dos modelos disponíveis na estrutura E, como é apresentado na subseção 2.2.7. Cada um dos resultados foram debatidos com base nos atributos biológicos das classes funcionais e das enzimas estudadas. Além disso, analisamos como os subconjuntos de dados respectivos a famílias e subfamílias proteicas, que ficaram implícitos no ambiente A, se relacionam com modelos e submodelos alternativos presentes na estrutura de modelagem E. Com isso, levantamos e discutimos hipóteses sobre a função e evolução de diversos grupos de enzimas com base em estudos presentes na literatura científica.

Tabela 4 – Descrição dos modelos respectivos a cada conjunto de dados de domínios funcionais de enzimas ativas em carboidratos.

Conjunto de Dados (ω_{XX})	Modelo (m_{XX})
ω_{AA}	m_{AA}: Domínios da classe de Atividades Auxiliares (AA). Agrupa domínios catalíticos com o potencial de auxiliar domínios de outras classes a acessar substratos polissacarídicos presentes na estrutura da parede celular vegetal.(65) O conjunto de dados dessa classe de domínios é subdividido em 26 subconjuntos de dados, correspondentes a 17 famílias de proteínas e suas subfamílias, mais 1 subconjunto de dados de enzimas sem subclassificação.
ω_{PL}	m_{PL}: Domínios da classe de Liases de Polissacarídeos (PL, do inglês *Polysaccharide Lyases*). Agrupa domínios catalíticos que utilizam um mecanismo de β-eliminação lítica para quebrar cadeias de polissacarídeos que contêm ácido urônico.(66) O conjunto de dados dessa classe é subdividido em 107 subconjuntos de dados, correspondentes 41 famílias de proteínas e suas subfamílias, mais 1 subconjunto de dados de enzimas sem subclassificação.
ω_{CE}	m_{CE}: Domínios da classe de Esterases de Carboidratos (CE, do inglês *Carbohydrate Esterases*). Agrupa domínios catalíticos do tipo esterase, que catalisa a liberação de grupos acil ou alquil ligados por ligação éster a carboidratos.(67) O conjunto de dados dessa classe é subdividido em 19 subconjuntos de dados, correspondentes a 19 famílias de proteínas, mais 1 subconjunto de dados de enzimas sem subclassificação.
ω_{CBM}	m_{CBM}: Domínios da classe de Módulos de Ligação a Carboidratos (CBM, do inglês *Carbohydrate-Binding Module*). Agrupa domínios funcionais que não possuem atividade catalítica, mas que se liga a substratos polissacarídicos e os direciona a domínios catalíticos de outras classes de enzimas.(68) O conjunto de dados dessa classe é subdividido em 95 subconjuntos de dados, correspondentes a 95 famílias de proteínas, mais 1 subconjunto de dados de enzimas sem subclassificação.
ω_{GT}	m_{GT}: Domínios da classe de Glicosiltransferases (GT). Agrupa domínios catalíticos responsáveis por catalisar a transferência de grupos glicosil de moléculas doadoras ativadas para moléculas receptoras nucleofílicas.(69) O conjunto de dados dessa classe é subdividido em 113 subconjuntos de dados, correspondentes a 113 famílias de proteínas, mais 1 subconjunto de dados de enzimas sem subclassificação.
ω_{GH}	m_{GH}: Domínios da classe de Glicosidases (GH, do inglês *Glycoside Hydrolase*). Agrupa domínios catalíticos que hidrolisam ligações glicosídicas.(70) O conjunto de dados dessa classe é subdividido em 412 subconjuntos de dados, correspondentes a 175 famílias de proteínas e suas subfamílias, mais 1 subconjunto de dados de enzimas sem subclassificação.

Fonte: Elaborada pelo autor.

3 RESULTADOS E DISCUSSÕES

Neste capítulo, apresentamos e analisamos as redes de meta-modelagem de dois casos: a modelagem de padrões em sequências de símbolos binários e a modelagem de domínios enzimáticos ativos em carboidratos. Os modelos abordados nos dois casos são baseados na comparação e classificação de sequências de símbolos, mas com graus distintos de complexidade.

No primeiro caso, dados artificiais foram criados para representar uma versão simplificada de sequências de aminoácidos. Desta forma, pudemos verificar o potencial de redes de meta-modelagem para fornecer informações sobre sistemas de modelos de padrões de sequências. Construímos a rede de meta-modelagem de 6 pares de conjuntos de dados e seus respectivos modelos, que incorporam, no total, 512 elementos distintos. Cada elemento corresponde a uma sequência de 9 símbolos binários, representados graficamente por quadrados brancos ou pretos. Esses elementos foram classificados e modelados de acordo com padrões presentes em suas sequências. Os elementos são simplificações análogas a sequências de aminoácidos, os dois símbolos que podem fazer parte das sequências são análogos aos 22 aminoácidos que podem compôr uma proteína, e os padrões detectados e usados para classificar esses elementos são análogos à sequências de aminoácidos que formam um domínio enzimático.

No segundo caso, montamos uma rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos, cuja anotação está disponível no banco de dados CAZy.(43, 44) Nesse sistema de modelos, mais de 2,8 milhões de enzimas são classificadas em seis conjuntos de dados a partir da anotação de domínios catalíticos e de ligação a carboidratos detectados nas suas sequências de aminoácidos. A rede de meta-modelagem permite uma análise exploratória detalhada desses conjuntos de dados e seus respectivos modelos, possibilitando predizer o funcionamento de enzimas, detectar subgrupos de enzimas semelhantes e sugerir tendências evolutivas implícitas nos modelos de domínios proteicos.

3.1 Rede de Meta-modelagem de Padrões em Sequências de Símbolos Binários

Nesta seção, analisamos a rede de meta-modelagem $\mathcal{M}$ referente à detecção de padrões em sequências de símbolos binários. O meta-modelo abrange a relação entre seis conjuntos de dados de sequências de 9 quadrados pretos ou brancos, e seis modelos que explicam os padrões contidos nessas sequências. A descrição detalhada de cada conjunto de dados e de seus modelos respectivos está presenta na Tabela 3.

A Figura 10 representa a rede de meta-modelagem construída para o ambiente de

dados A e a estrutura de modelagem E de padrões em sequências de símbolos binários. As métricas de especificidade calculadas sobre toda a estrutura de modelagem — ou seja, *multiplicidade média de elementos* e *diversidade total de conexões* —, bem como outras métricas descritivas do meta-modelo, são apresentadas na Tabela 5.

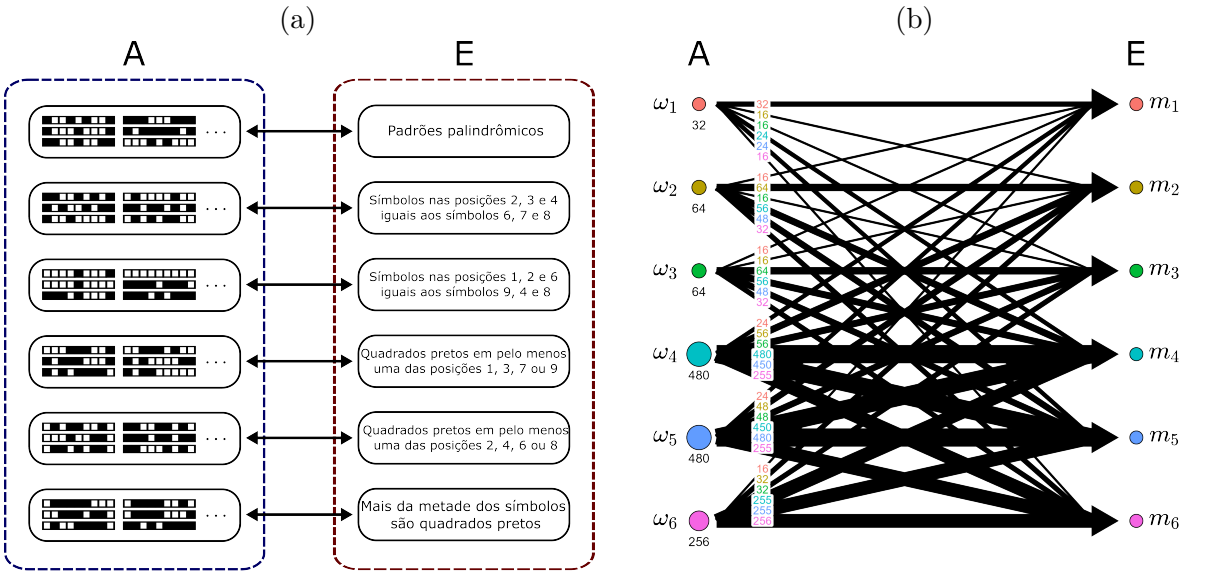


Figura 10 – Rede de meta-modelagem de padrões em sequências de símbolos binários. (a) Representação do mapa entre domínios, mostrando alguns exemplos de sequências presentes em cada conjunto de dados. (b) Representação de rede de meta-modelagem. Os conjuntos de dados ω_i no ambiente de dados A são grupos de sequências de 9 símbolos binários, representados por quadrados pretos e brancos, que podem ser explicadas coletivamente por um modelo m_i respectivo, disponível na estrutura de modelagem E. Cada par de conjunto de dados e modelo compreende diferentes padrões de símbolos binários de acordo com as regras presentes na Tabela 3. Os círculos do lado esquerdo da representação em rede representam conjuntos de dados, e os números abaixo deles representam as suas respectivas cardinalidades. Os círculos do lado direito da representação estão associados aos modelos correspondentes a cada conjunto de dados disponível. A relação entre cada conjunto de dados e cada modelo alternativo é representada por conexões ponderadas e orientadas, cujos pesos representam o número de elementos no conjunto de dados de saída que podem ser explicados pelo modelo de destino.

Fonte: Elaborada pelo autor.

A rede de meta-modelagem na Figura 10 revela uma estrutura de modelagem altamente inespecífica. Apenas 24 dos 512 elementos distribuídos no ambiente de dados são específicos, ou seja, pertencem a apenas um conjunto de dados. Em média, cada elemento pertence a mais de dois conjuntos de dados diferentes, ou seja, podem ser explicados por dois modelos distintos. Como indicado na Tabela 5, pode-se observar quatro vezes mais diversidade do que seria necessário para um mapeamento específico entre conjuntos de

Tabela 5 – Métricas descritivas da rede de meta-modelagem de padrões em sequências de símbolos binários.

Número de elementos em A	512
Número de elementos inespecíficos em A	488 (95,31% dos dados)
Multiplicidade média de elementos ($\bar{U}$)	2,69
Diversidade total de conexões ($\mathcal{D}_{\mathcal{T}}$)	28,41

Fonte: Elaborada pelo autor.

dados e modelos, o que enfatiza o alto nível de inter-relação entre os diferentes modelos.

As métricas descritivas para cada par de conjunto de dados e modelo são apresentadas na Tabela 6. Os seis modelos possuem alta diversidade de conexões, com mais de três vezes a diversidade esperada para um mapeamento específico entre conjuntos de dados e modelos. Os modelos 1, 2, 3 e 6 são totalmente compostos por elementos que também pertencem a outros conjuntos de dados, sendo os modelos menos específicos da estrutura de modelagem.

Tabela 6 – Métricas descritivas para cada par de conjunto de dados e modelo disponíveis na rede de meta-modelagem de padrões em sequências de símbolos binários.

Conjunto de Dados (ω_k) e Modelo (m_k)	Cardinalidade	Elementos Inespecíficos (% do Total)	Diversidade de Conexões ($\mathcal{D}$)
ω_1 / m_1	32	32 (100,00 %)	5,778
ω_2 / m_2	64	64 (100,00 %)	5,293
ω_3 / m_3	64	64 (100,00 %)	5,293
ω_4 / m_4	480	472 (98,33 %)	4,027
ω_5 / m_5	480	464 (96,67 %)	3,937
ω_6 / m_6	256	256 (100,00 %)	4,086

Fonte: Elaborada pelo autor.

Curiosamente, os modelos m_2 e m_3 possuem os mesmos valores para cada uma das métricas de caracterização consideradas. Isso pode ser explicado usando o conhecimento sobre a estrutura desses modelos. Nos dois casos, as sequências possuem uma repetição de três símbolos em dois trios de posições diferentes em suas sequências: enquanto o modelo m_2 explica as sequências cujos símbolos presentes nas posições 2, 3 e 4 são iguais aos das posições 6, 7 e 8, o modelo m_3 explica sequências cujos símbolos presentes nas posições 1, 2 e 6 se repetem nas posições 9, 4 e 8. Portanto, para cada elemento presente no conjunto de dados respectivo ao modelo m_2, há uma sequência permutada deste elemento presente no conjunto de dados do modelo m_3, e vice-versa. Isso cria uma simetria entre esses dois conjuntos de dados que é refletida em suas métricas descritivas.

Realizamos a análise de redundância na estrutura de modelagem buscando o valor máximo de coincidência entre cada modelo disponível e cada composição de modelos

alternativos, conforme descrito nas Equações 2.11, 2.12, e 2.17. O valor máximo do índice de coincidência alcançado para cada modelo, a respectiva composição lógica de modelos alternativos usados para obter esse valor de coincidência e a respectiva operação entre conjuntos de dados são apresentados na Tabela 7.

Tabela 7 – Valores máximos de coincidência comparando cada par de conjunto de dados e modelos com combinações de conjuntos e modelos alternativos presentes na rede de meta-modelagem de padrões em sequências de símbolos binários.

Conjunto de Dados (ω_k) e Modelo (m_k)	Coincidência Máxima ($\mathcal{C}_{max}$)	Operação entre Conjuntos de Dados Alternativos (ω_X)
ω_1 / m_1	0,500	$\left(\bigcap_{i\neq 1} \omega_i^c\right) \cup (\omega_2 \cap \omega_3)$
ω_2 / m_2	0,250	$\left(\bigcap_{i\neq 2} \omega_i^c\right) \cup (\omega_1 \cap \omega_3)$
ω_3 / m_3	0,250	$\left(\bigcap_{i\neq 3} \omega_i^c\right) \cup (\omega_1 \cap \omega_2)$
ω_4 / m_4	0,939	$\left(\bigcap_{i\neq 4} \omega_i^c\right) \cup (((\omega_6 \cup \omega_1^c) \cup \omega_2^c) \cup \omega_5^c)$
ω_5 / m_5	0,947	$\left(\bigcap_{i\neq 5} \omega_i^c\right) \cup (((\omega_6 \cup \omega_1^c) \cup \omega_2^c) \cup \omega_4^c)$
ω_6 / m_6	0,561	$\left(\bigcap_{i\neq 6} \omega_i^c\right) \cup (((\omega_1 \cap \omega_2) \cup \omega_4) \cap \omega_5)$
Conjunto de Dados (ω_k) e Modelo (m_k)	**Coincidência Máxima ($\mathcal{C}_{max}$)**	**Composição Lógica de Modelos Alternativos (m_X)**
ω_1 / m_1	0,500	$\left(\bigwedge_{i\neq 1} \neg m_i\right) \vee (m_2 \wedge m_3)$
ω_2 / m_2	0,250	$\left(\bigwedge_{i\neq 2} \neg m_i\right) \vee (m_1 \wedge m_3)$
ω_3 / m_3	0,250	$\left(\bigwedge_{i\neq 3} \neg m_i\right) \vee (m_1 \wedge m_2)$
ω_4 / m_4	0,939	$\left(\bigwedge_{i\neq 4} \neg m_i\right) \vee (((m_6 \vee \neg m_1) \vee \neg m_2) \vee \neg m_5)$
ω_5 / m_5	0,947	$\left(\bigwedge_{i\neq 5} \neg m_i\right) \vee (((m_6 \vee \neg m_1) \vee \neg m_2) \vee \neg m_4)$
ω_6 / m_6	0,561	$\left(\bigwedge_{i\neq 6} \neg m_i\right) \vee (((m_1 \wedge m_2) \vee m_4) \wedge m_5)$

Fonte: Elaborada pelo autor.

O valor máximo de coincidência da rede de meta-modelagem é $0,947$. Esse valor de coincidência é obtido ao comparar o conjunto de dados respectivo ao modelo m_5 e um modelo $m_X(m_5)$ criado por meio da composição lógica de quatro modelos alternativos (m_1, m_2, m_4 e m_6) unida à negação de todos os modelos alternativos ($\bigwedge_{i\neq 5} \neg m_i$). Portanto, o modelo m_5 pode ser visto como redundante, pois os demais modelos podem ser utilizados para explicar grande parte de seu conjunto de dados. Dado que o modelo m_5 é o mais redundante entre os seis modelos, centraremos a nossa atenção, doravante, na sua respectiva discussão.

Usando as definições anteriores para cada modelo listado no início da seção, $m_X(m_5)$ é o modelo que explica: (i) sequências que não são palindrômicas, em que mais da metade dos símbolos são quadrados brancos, que possuem quadrados brancos nas posições 1, 3, 7 e 9, e que não repetem o mesmo trio de símbolos nas posição 2, 3, 4 e 6, 7, 8 ou 1, 2, 6 e 9, 4, 8; ou (ii) sequências em que mais da metade dos símbolos são quadrados pretos; ou (iii) sequências que não são palindrômicas; ou (iv) sequências cujos símbolos das posições 2, 3, 4 não são iguais aos das posições 6, 7, 8; ou (v) sequências que possuem quadrados brancos nas posições 1, 3, 7, 9. Este amplo modelo explica 507 sequências, incluindo todas as 480 sequências explicadas pelo modelo m_5 mais 27 outras, representados na Figura 11.

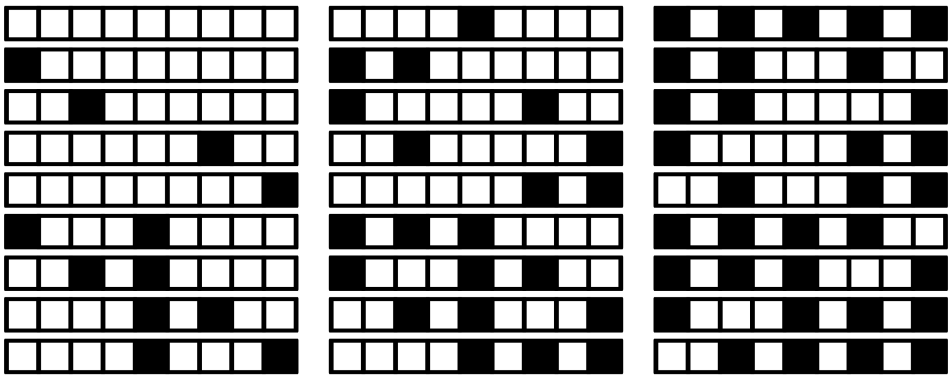


Figura 11 – As 27 sequências de símbolos binários explicados pela composição lógica dos modelos $m_X(m_5) = \left(\bigwedge_{i\neq 6} \neg m_i\right) \vee (((m_1 \wedge m_2) \vee m_4) \wedge m_5)$ que não são explicadas pelo modelo m_5. Elas incluem sequências que *não* possuem quadrados pretos nas posições 2, 4, 6 e 8 e satisfazem pelo menos uma das seguintes afirmações: (i) sequências que não são palindrômicas, em que mais da metade dos símbolos são quadrados brancos, que possuem quadrados brancos nas posições 1, 3, 7 e 9, e que não repetem o mesmo trio de símbolos nas posição 2, 3, 4 e 6, 7, 8 ou 1, 2, 6 e 9, 4, 8; (ii) sequências em que mais da metade dos símbolos são quadrados pretos; (iii) sequências que não são palindrômicas; (iv) sequências cujos símbolos das posições 2,3,4 não são iguais aos das posições 6,7,8; (v) sequências que possuem quadrados brancos nas posições 1, 3, 7, 9.

Fonte: Elaborada pelo autor.

As 27 sequências extras explicadas por $m_X(m_5)$ podem formar um novo conjunto de dados ω_{Extra}, com um respectivo modelo m_{Extra} definido por meio de inspeção direta de seus elementos. O conjunto de dados ω_{Extra} possui sequências com quadrados pretos restritos às posições 1, 3, 5, 7 e 9, com a maior distância entre qualquer símbolos pretos igual ou menor a 6 posições.

Podemos associar esta definição do modelo m_{Extra} às informações disponíveis do modelo $m_X(m_5)$ a fim de criar uma composição lógica exata de modelos equivalente a m_5.

Isso é obtido combinando $m_X(m_5)$ e a negação de m_{Extra}, conforme expresso abaixo:

$$m_5 = \left(\left(\bigwedge_{i \neq 5} \neg m_i \right) \vee (((m_6 \vee \neg m_1) \vee \neg m_2) \vee \neg m_4) \right) \wedge \neg m_{Extra} \tag{3.1}$$

Essa composição lógica e sua interpretação é muito mais complexa do que a explicação direta do modelo m_5 (modelo que descreve sequências que possuem quadrados pretos em pelo menos uma das posições 2, 4, 6 ou 8). Porém, caso essa explicação direta não esteja disponível, a composição lógica exata para representar m_5 pode ser útil como um modelo geral e interpretável das 480 sequências no conjunto de dados ω_5.

Esses resultados ilustram como a rede de meta-modelagem pode ser efetivamente usada para extrair informações de conjuntos de dados de sequências de símbolos e de seus modelos de detecção de padrões. Além disso, a rede possibilita a construção de novos modelos por meio de composições lógicas dentro de uma estrutura de modelagem definida. Isso motiva o uso de meta-modelagem para auxiliar modelos complexos de análise e detecção de padrões em sequências. Essa abordagem é especialmente relevante em problemas como a anotação de domínios proteicos em sequências de aminoácidos.

3.2 Rede de Meta-modelagem dos Domínios Funcionais de Enzimas Ativas em Carboidratos

Nesta seção, analisamos a rede de meta-modelagem $\mathcal{M}$ referente à anotação e classificação de domínios funcionais de enzimas ativas em carboidratos, disponível no banco de dados CAZy.(43, 44) O meta-modelo abrange a relação entre conjuntos de dados de seis classes de domínios funcionais, e seis modelos que explicam diversas características dessas classes. A descrição detalhada de cada conjunto de dados e de seus modelos respectivos está presenta na Tabela 4.

A Figura 12 representa a rede de meta-modelagem construída para o ambiente de dados A e a estrutura de modelagem E dos domínios funcionais de enzimas ativas em carboidratos. As métricas de especificidade calculadas sobre toda a estrutura de modelagem — ou seja, *multiplicidade média de elementos* e *diversidade total de conexões* —, bem como outras métricas descritivas do meta-modelo, são apresentadas na Tabela 5.

A rede de meta-modelagem na Figura 12 revela uma estrutura de modelagem altamente específica. Apenas 6,66% dos elementos distribuídos no ambiente de dados são inespecíficos, ou seja, pertencem a mais de um conjunto de dados. No entanto, como indicado na Tabela 8, a rede apresenta uma diversidade total de conexões superior a 9, isto é, há uma distribuição de conexões equivalente a um mapeamento entre 9 pares de conjuntos de dados e modelos. Isso revela um nível considerável de entrelaçamento entre as classes de domínios funcionais observadas nas enzimas ativas em carboidratos.

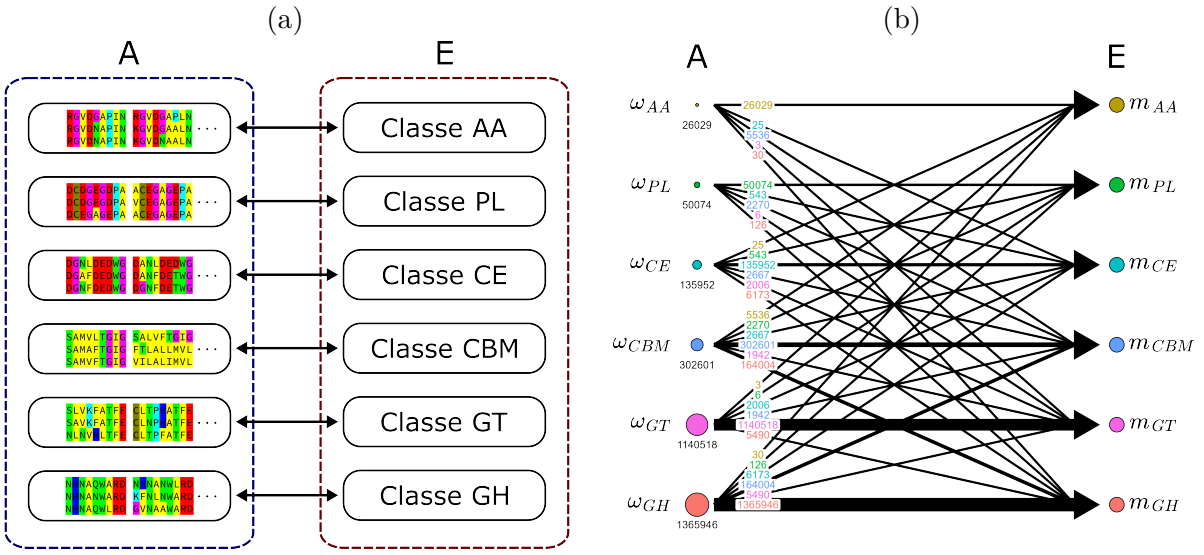


Figura 12 – Rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos. (a) Representação do mapa entre domínios, mostrando alguns trechos de sequências proteicas presentes em cada conjunto de dados. (b) Representação de rede de meta-modelagem. Os conjuntos de dados ω_{XX} no ambiente de dados A são conjuntos de enzimas agrupadas a partir das classes de seus domínios funcionais. Cada conjunto de dados é explicado por um modelo m_{XX} respectivo, disponível na estrutura de modelagem E. Cada par de conjunto de dados e modelo compreende a uma das classes AA, PL, CE, CBM, GT e GH, descritas na Tabela 4. Os círculos do lado esquerdo da representação em rede representam conjuntos de dados, e os números abaixo deles representam as suas respectivas cardinalidades. Os círculos do lado direito da representação estão associados aos modelos correspondentes a cada conjunto de dados disponível. A relação entre cada conjunto de dados e cada modelo alternativo é representada por conexões ponderadas e orientadas, cujos pesos representam o número de elementos no conjunto de dados de saída que podem ser explicados pelo modelo de destino.

Fonte: Elaborada pelo autor.

Tabela 8 – Métricas descritivas da rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos.

Número de elementos em A	2’831’365
Número de elementos inespecíficos em A	188’689, (6,66% dos dados)
Multiplicidade média de elementos ($\bar{U}$)	1,07
Diversidade total de conexões ($\mathcal{D}_{\mathcal{T}}$)	9,06

Fonte: Elaborada pelo autor.

Na Subseção 3.2.1, a seguir, as características dos seis modelos na rede foram integradas e explicadas utilizando atributos biológicos de cada classe funcional em estudo. Isso permitiu revelar aspectos implícitos desses modelos e de suas interações.

Na Subseção 3.2.2, realizamos uma análise mais detalhada, em que exploramos cada interação entre um conjunto de dados e um modelo alternativo em termos dos subconjuntos — correspondentes a famílias proteicas — envolvidos. Essa análise proporcionou uma maior compreensão das enzimas que possuem domínios de múltiplas classes funcionais, permitindo-nos formular hipóteses sobre suas origens evolutivas e seus mecanismos de funcionamento.

3.2.1 Análise de Conjuntos de Dados e Modelos na Rede de Meta-Modelagem dos Domínios Funcionais de Enzimas Ativas em Carboidratos

Algumas métricas descritivas para cada par de conjunto de dados e modelo disponíveis na rede de meta-modelagem da Figura 12 são apresentadas na Tabela 9.

Tabela 9 – Métricas descritivas para cada par de conjunto de dados e modelo disponíveis na rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos. Os pares de conjuntos de dados e modelos possuem diversidades de conexões e proporção de elementos inespecíficos variáveis coerentes com as funções biológicas dos domínios funcionais a quais eles se referem.

Conjunto de Dados (ω_k) e Modelo (m_k)	Cardinalidade	Elementos Inespecíficos (% do Total)	Diversidade de Conexões ($\mathcal{D}$)
ω_{AA} / m_{AA}	26’029	5’576 (21,42 %)	1,614
ω_{PL} / m_{PL}	50’074	2’907 (5,81 %)	1,286
ω_{CE} / m_{CE}	135’952	10’391 (7,64 %)	1,434
ω_{CBM} / m_{CBM}	302’601	176’112 (58,20 %)	2,193
ω_{GT} / m_{GT}	1’140’518	8’680 (0,76 %)	1,057
ω_{GH} / m_{GH}	1’365’946	174’778 (12,80 %)	1,475

Fonte: Elaborada pelo autor.

Os modelos m_{GT} e m_{PL}, com 1,057 e 1,286 de diversidade de conexão, respectivamente, são os mais específicos da estrutura de modelagem analisada. Eles possuem diversidades de conexão próximas de 1, ou seja, os elementos de seus conjuntos de dados respectivos são distribuídos ordenadamente para pouco mais de um modelo.

O modelo m_{GT} é o mais específico, com menor diversidade e proporção de elementos inespecíficos. Isso é especialmente interessante porque o conjunto de dados ω_{GT} é o segundo maior do ambiente de dados, o que significa que a baixa interconexão entre o par ω_{GT} / m_{GT} e os pares de conjuntos de dados e modelos alternativos é uma característica intrínseca ao modelo usados para definir essa classe de domínios funcionais. Do ponto de vista biológico, isso significa que as enzimas glicosiltransferases são menos modulares e, geralmente, não possuem domínios alternativos com funções catalíticas em carboidrato. Hipoteticamente, o que explica a baixa modularidade de enzimas com domínios GT é a incompatibilidade

funcional com os domínios de outras classes. Enquanto a classe GT agrupa domínios que catalisam a síntese de cadeias polissacarídicas (69,71), enzimas da classe GH (72), PL (66), CE (67) e AA (65) atuam na degradação dessas cadeias, catalisando reações de quebra de ligações covalentes. Com funções tão antagônicas, é plausível que seja mais rara a evolução de enzimas que combinem, de forma não deletéria, domínios GT e domínios de outras classes.

O modelo m_{PL} é o segundo menor em diversidade de conexões e tem um conjunto de dados respectivo bem pequeno, com pouco mais de 50 mil elementos. Ele engloba os domínios da classe PL utilizam um mecanismo de β-eliminação lítica para quebrar cadeias de polissacarídeos que contêm ácido urônico, caso de pectinas, presente em grande quantidade na parede celular de plantas.(66) A baixa cardinalidade do conjunto de dados ω_{PL} indica uma raridade desse tipo de domínio entre enzimas ativas em carboidratos. Essa peculiaridade é conhecida desde 2010, quando um estudo que analisou mais de 1300 genomas mostrou uma baixa detecção relativa de domínios PLs frente a domínios de outras classes.(63) Conjecturou-se que isso é consequência da baixa proporção dos polissacarídeos contendo ácidos urônicos entre polissacarídeos naturais.(63) Indo além, podemos supor que, atuando sobre uma variedade menor e mais rara de substratos, enzimas que possuem domínios PL evoluíram em uma quantidade menor de organismos e atuam em vias metabólicas mais restritas, dificultando a associação com outros tipos de domínios funcionais. Essa hipótese explica a pequena diversidade de conexões do modelo m_{PL}.

Os modelos m_{CE}, m_{GH} e m_{AA}, com 1,434, 1,475 e 1,614 de diversidade de conexão, respectivamente, são modelos com especificidade mediana. Eles possuem diversidades de conexão significativamente maior que 1, mas seus elementos não são explicados tão genericamente por mais de um modelo. Nos três casos, isso acontece pois há uma interconexão substancial entre esses modelos e um modelo alternativo em especial. No caso do modelo m_{CE}, seu conjunto de dados ω_{CE} compartilha uma fração significativa de elementos com o conjunto de dados ω_{GH}. Já os modelos m_{GH} e m_{AA} explicam muitos elementos que se encaixam no conjunto de dados respectivo ao modelo m_{CBM}.

O modelo m_{CE} é respectivo à classe CE, que agrupa domínios catalíticos do tipo esterase. Na maior parte das enzimas em que há domínios da classe CE e domínios de outras classes, os domínios funcionam de forma sequencial para processar um substrato. Além disso, a subclassificação dos domínios dessas enzimas se concentra em poucas famílias. A maior parte das enzimas classificadas como CE e GH possuem um par de domínios catalíticos das famílias CE4 e GH153, que podem trabalhar sequencialmente no processamento de poliacetilglicosaminas.(64,73) As enzimas com domínios da classe CE e CBM provavelmente utilizam os módulos CBM para otimizar a atividade de esterases em quitina e xilano.(67,74) Enzimas com domínios CE e GT podem usar os dois módulos para a síntese de poliacetilglicosaminas e para o processamento de quitina.(74–76) A maior

parte das enzimas classificadas em CE e PL possuem um par de domínios das famílias CE8 e PL1, que trabalham no processamento de pectina.(77) E a pequena sobreposição entre as classes CE e AA se deve a enzimas que apresentam domínios AA5 e CE3, possivelmente com o segundo domínio fazendo papéis estruturais e não catalíticos.(78)

O modelo m_{GH} é respectivo à classe GH. Essa classe agrupa domínios que catalisam a hidrólise de ligações glicosídicas. Assim como no caso da classe CE, as enzimas que possuem domínios GH e domínios de outras classes utilizam esses domínios de forma sequencial para obter produtos de maneira otimizada ou produtos que seriam mais difíceis de produzir com duas enzimas diferentes. As classes GH e CBM tem uma relação extensiva, envolvendo 370 famílias e subfamílias de ambas as classes, já que muitos domínios da classe GH tem sua atividade catalítica otimizada na presença de módulos CBM.(68) A maior parte das enzimas com domínios GH e GT possuem um par de domínios das famílias GT2 e GH17 ou GT84 e GH94, possivelmente usando esses domínios para sintetizar β-glucanos cíclicos.(79–84) Enzimas com domínios GH e PL podem combiná-los para facilitar a degradação de carboidratos, primeiro pela ação do domínio PL, que cria uma terminação insaturada que se torna disponível para a hidrólise pelo domínio GH.(85–87) Já as poucas enzimas que associam domínios das classes GH e AA provavelmente o fazem para facilitar a degradação de quitina.(88, 89)

Os domínios da classe AA, respectiva ao modelo m_{AA}, catalisam reações de oxirredução e atuam, geralmente, na degradação da parede celular vegetal.(65) A alta diversidade desse modelo se deve à divisão do conjunto de dados em duas tendências principais: enzimas que possuem domínios apenas da classe AA, enzimas que possuem domínios da classe AA e CBM. Em sua maior parte, as enzimas com domínios da classe AA e CBM possuem domínios classificados nas famílias AA10, CBM5 e CBM73. Enquanto a família AA10 atua na degradação de quitina (88, 90), os módulos CBM5 e CBM73 se ligam à quitina, aumentando a atividade catalítica dos domínios AA10.(91) Mais uma vez, a diversidade do modelo se deve à interação sinergética de domínios de diferentes classes estarem presentes em uma mesma enzima.

O modelo m_{CBM}, com 2,193 de diversidade de conexão, é o modelo mais inespecífico. Ele possui diversidades de conexão superior a 2, ou seja, a distribuição dos elementos de seus conjuntos de dados são próximas a de um conjunto de dados explicado por dois modelos alternativos. Os domínios funcionais da classe CBM tem a função de auxiliar domínios das outras classes a facilitarem o direcionamento do substrato ao seus sítios ativos.(68) Isso justifica a alta inespecificidade do modelo m_{CBM}, já que diferentes pressões seletivas induziram a evolução de enzimas que agrupem domínios catalíticos das classes GT, GH, PL, CE e AA e módulos de ligação aos seus substratos, da classe CBM.

Realizamos a análise de redundância na estrutura de modelagem buscando o valor máximo de coincidência entre cada modelo disponível e cada composição de modelos

alternativos, conforme descrito nas Equações 2.11, 2.12, e 2.17. O valor máximo do índice de coincidência alcançado para cada modelo, a respectiva composição lógica de modelos alternativos usados para obter esse valor de coincidência e a respectiva operação entre conjuntos de dados são apresentados na Tabela 10.

Tabela 10 – Valores máximos de coincidência comparando cada par de conjunto de dados e modelos com combinações de conjuntos e modelos alternativos presentes na rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos.

Conjunto de Dados (ω_k) e Modelo (m_k)	Coincidência Máxima ($\mathcal{C}_{max}$)	Operação entre Conjuntos de Dados Alternativos (ω_X)
ω_{AA} / m_{AA}	0,786	$\left(\bigcap_{XX \neq AA} \omega^c_{XX}\right)$
ω_{PL} / m_{PL}	0,942	$\left(\bigcap_{XX \neq PL} \omega^c_{XX}\right)$
ω_{CE} / m_{CE}	0,924	$\left(\bigcap_{XX \neq CE} \omega^c_{XX}\right)$
ω_{CBM} / m_{CBM}	0,418	$\left(\bigcap_{XX \neq CBM} \omega^c_{XX}\right) \cup (\omega_{GH} \cap \omega_{AA})$
ω_{GT} / m_{GT}	0,992	$\left(\bigcap_{XX \neq GT} \omega^c_{XX}\right)$
ω_{GH} / m_{GH}	0,898	$\left(\bigcap_{XX \neq GH} \omega^c_{XX}\right) \cup ((\omega_{CBM} \cap \omega^c_{AA}) \cap \omega^c_{PL})$
Conjunto de Dados (ω_k) e Modelo (m_k)	Coincidência Máxima ($\mathcal{C}_{max}$)	Composição Lógica de Modelos Alternativos (m_X)
ω_{AA} / m_{AA}	0,786	$\left(\bigwedge_{XX \neq AA} \neg m_{XX}\right)$
ω_{PL} / m_{PL}	0,942	$\left(\bigwedge_{XX \neq PL} \neg m_{XX}\right)$
ω_{CE} / m_{CE}	0,924	$\left(\bigwedge_{XX \neq CE} \neg m_{XX}\right)$
ω_{CBM} / m_{CBM}	0,418	$\left(\bigwedge_{XX \neq CBM} \neg m_{XX}\right) \vee (m_{GH} \vee m_{AA})$
ω_{GT} / m_{GT}	0,992	$\left(\bigwedge_{XX \neq GT} \neg m_{XX}\right)$
ω_{GH} / m_{GH}	0,898	$\left(\bigwedge_{XX \neq GH} \neg m_{XX}\right) \vee ((m_{CBM} \vee \neg m_{AA}) \vee \neg m_{PL})$

Fonte: Elaborada pelo autor.

O valor máximo de coincidência da rede de meta-modelagem é $0,947$. Esse valor de coincidência é obtido ao comparar o conjunto de dados respectivo ao modelo m_{GT} e um modelo $m_X(m_{GT})$ criado apenas pela negação de todos os modelos alternativos $(\bigwedge_{XX \neq GT} \neg m_{XX})$. Isso significa que o conjunto de dados ω_{GT} é praticamente igual ao conjunto de enzimas que possuem apenas domínios GT. Além disso, aponta que não há composição significativa de conjuntos de dados compostos por enzimas de outras classes que se assemelha em maior grau ao conjunto de dados desse modelo. Isso é uma consequência

da alta especificidade do modelo m_{GT} e da maioria dos modelos analisados na rede de meta-modelagem. Esse é o mesmo caso dos modelos m_{PL}, m_{CE} e m_{AA}.

Diferente da análise da rede de meta-modelagem de padrões em sequências de símbolos binários, o modelo com o maior índice de coincidência não pode ser visto como redundante. Apesar do modelo m_{GT} explicar todos os elementos que não se encaixam em nenhum modelo alternativo, a sua natureza não é essa. Diferente de um modelo que explica genericamente elementos que não se encaixam a modelo algum, o modelo m_{GT} explica um grupo bem definido de domínios de enzimas ativas em carboidratos, que compartilham diversas características entre si. Essa distinção é necessária, pois uma nova enzima ativa em carboidratos que não possui domínios funcionais das classes AA, PL, CE, CBM ou GH não necessariamente será incluída na classe GT.

Além dos casos acima, o modelo m_{GH} é altamente similar a uma composição lógica envolvendo os modelos das classes CBM, AA e PL unida à negação de todos os modelos alternativos ($\bigwedge_{XX \neq GH} \neg m_{GH}$). Nesse caso, a composição de modelos alcançada inclui enzimas que possuem exclusivamente domínios GH e enzimas que possuem domínios CBM, mas não possuem domínios AA e PL. Isso é uma consequência da relação distante entre as classes AA, PL e GH e a relação próxima entre as classes CBM e GH, em que mais de 50% das enzimas que possuem módulos CBM também possuem um domínio GH.

Esses resultados ilustram como a análise da rede de meta-modelagem pode ser utilizada para encontrar tendências evolutivas e relações funcionais a partir da anotação de domínios proteicos. Também é possível obter avaliações quantitativas e qualitativas da interação entre modelos de domínios funcionais, o que pode auxiliar na construção de uma classificação mais robusta de proteínas.

3.2.2 Análise de Subconjuntos de Dados e Submodelos na Rede de Meta-Modelagem dos Domínios Funcionais de Enzimas Ativas em Carboidratos

Cada conjunto de dados disponível no ambiente de dados A — ω_{AA}, ω_{PL}, ω_{CE}, ω_{CBM}, ω_{GT} e ω_{GH} — agrupa enzimas que possuem, pelo menos, um domínio da mesma classe funcional. Cada um desses conjuntos de dados é explicado por um modelo respectivo presente na estrutura de modelagem E — m_{AA}, m_{PL}, m_{CE}, m_{CBM}, m_{GT}, m_{GH}.

Mesmo dentro de uma mesma classe funcional, há uma variabilidade significativa de funções e de sequências de aminoácidos. Por isso, as classes são subdivididas em famílias com maior semelhança de sequências e, possivelmente, com mesma origem evolutiva.(43) Mais além, algumas famílias também são subdivididas em subfamílias, dependendo da quantidade e heterogeneidade dos domínios presentes nelas.(43)

Dentro da rede de meta-modelagem, as famílias e subfamílias representam subconjuntos dentro dos conjuntos de dados respectivos a cada classe. A partir de um subconjunto de dados, podemos atribuir a cada família e subfamília um respectivo submodelo que

explica as suas características. Os submodelos são versões mais particulares dos modelos que descrevem toda a classe funcional a qual as famílias fazem parte.

Nesta subseção, iremos aprofundar a análise das interações entre conjuntos de dados e modelos na rede de meta-modelagem dos domínios funcionais, que é representada na Figura 12. Nosso foco será a análise dos subconjuntos e submodelos envolvidos nessas interações. Vamos analisar as famílias e subfamílias de cada enzima que possui domínios funcionais de classes distintas, levantando hipóteses sobre a origem evolutiva e o comportamento funcional desses domínios.

Há 28 interações entre conjuntos de dados e modelos alternativos, com 14 delas sendo únicas. Isso ocorre pois a relação entre um conjunto de dados XX e um modelo alternativo YY é igual à relação entre um conjunto de dados YY e um modelo alternativo XX. As 14 relações distintas, em termos das classes as quais elas pertencem, ocorrem entre: GT e PL, GT e CE, GT e GH, GT e AA, GT e CBM, PL e CE, PL e GH, PL e CBM, CE e GH, CE e AA, CE e CBM, GH e AA, GH e CBM, AA e CBM.

Apresentaremos as análises das interações entre as classes em ordem de diversidade de conexões de seus modelos na rede de meta-modelagem. Primeiro, apresentaremos todas as interações que envolvem a classe GT; depois as interações restantes que envolvem a classe PL; depois as interações restantes que envolvem a classe CE; e seguindo da mesma forma para as classes GH, AA e CBM.

3.2.2.1 Interações envolvendo a classe GT

A classe GT tem a menor diversidade no meta-modelo, igual a 1,057. Ela compartilha 8680 elementos, apenas 0,76% de seu conjunto de dados, com outras classes funcionais. A Figura 13 apresenta a distribuição desses elementos nos subconjuntos de dados da classe GT a quais eles pertencem.

3.2.2.2 Interação entre as classes GT e GH

Enzimas da classe GT compartilham características funcionais com as enzimas da classe GH: as duas classes possuem enzimas que catalisam transferências de grupos glicosil.(69) As enzimas da classe GT catalisam a transferência de grupos glicosil de açúcares doadores que contém nucleosídeos fosfatados para grupos aceptores nucleofílicos, incluindo carboidratos, lipídeos e peptídeos.(69,71) Já enzimas com domínios GH atuam na hidrólise de grupos glicosil, liberando-os de seu grupo doador.(70,72) Uma terceira classe de enzimas também atua transferindo grupos glicosil, chamada de fosforilases. As fosforilases catalisam o corte da ligação glicosídica por meio da substituição por fosfato.(71) No entanto, fosforilases não formam um grupo distinto de classe funcional na classificação CAZy, sendo classificadas em famílias das classes GT e GH.(71)

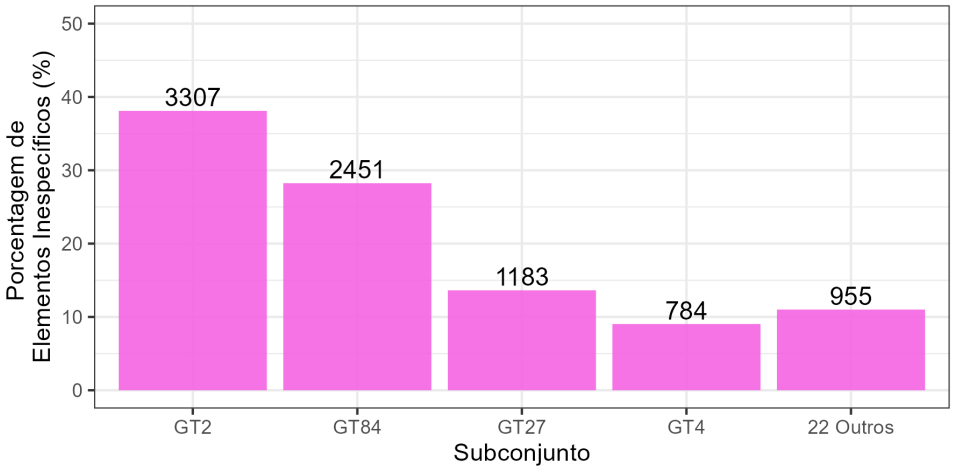


Figura 13 – Distribuição dos 8680 elementos inespecíficos do conjunto de dados da classe GT nos subconjuntos de dados aos quais pertencem.

Fonte: Elaborada pelo autor.

Na rede de meta-modelagem, os conjuntos de dados e modelos de GT e GH compartilham 5490 enzimas. A Figura 14 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

A família da classe GT que mais compartilha enzimas com a classe GH é a família GT84. Essa família tem 2584 proteínas e compartilha 2451 delas, todas com a família GH94. Já a família GH94 possui 5482 proteínas e compartilha 2451 delas, todas com a família GT84. Esse nível de relação entre GT84 e GH94 revela uma profunda relação evolutiva e funcional dos domínios dessas famílias. Podemos levantar a hipótese de que as enzimas com domínios GT84 necessitam de um domínio auxiliar de GH94 para realizar sua função biológica de maneira efetiva.

Todos os membros da família GT84 atuam como sintases de β-1,2-glucanos cíclicos.(43) Essas sintases são presentes em múltiplos clados de bactérias gram-negativas.(43, 84) Elas possuem um domínio de glicosiltransferase que catalisa o alongamento de cadeias de β-1,2-glucanos (o domínio GT84) e um domínio de fosforilase que ajusta o tamanho dessas cadeias (o domínio GH94), além de um domínio extra que catalisa a ciclização do polissacarídeo.(79, 80, 84) Um estudo recente, ainda não publicado, de enzimas com domínios GT84 e GH94 ressalta que a falta parcial ou total do domínio GH94 tem efeito negativo na atividade da sintase.(92) Isso reforça nossa hipótese levantada a partir da análise da rede de meta-modelagem.

As classes GT e GH também compartilham outras 2463 enzimas por meio da família GT2, a família da classe GT com maior quantidade de proteínas com domínios GH. No entanto, enzimas representam menos de 1% da família GT2.

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
GH94 e GT84	2451	44.71%	94.85%	-
GH17 e GT2	1511	19.08%	0.45%	-
CE4 e GH18 e GT2	752	1.72%	2%	0.22%
60 Outras Relações	776	-	-	-

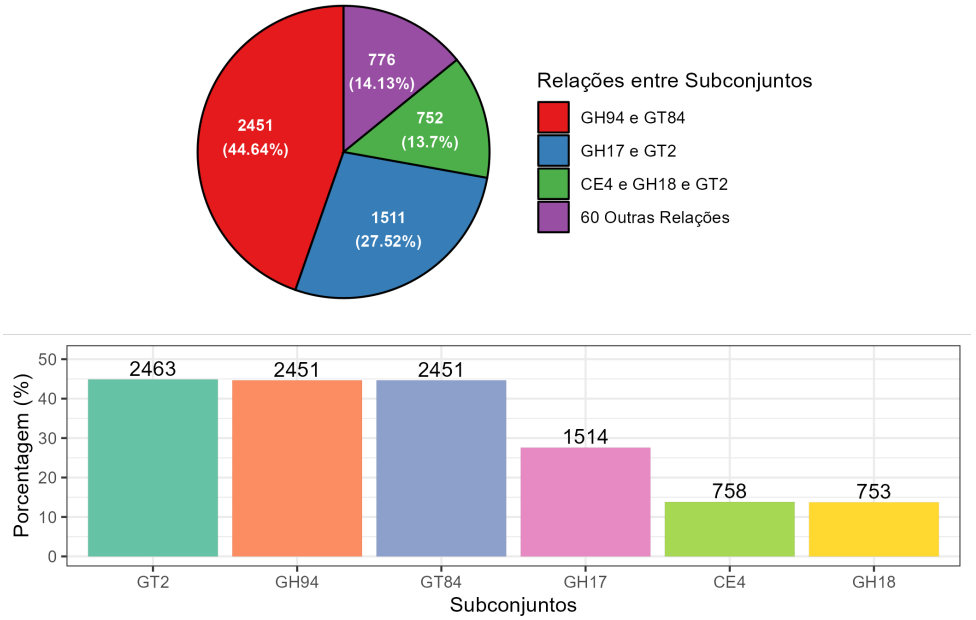


Figura 14 – Distribuição dos 5490 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e GH. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios GTXX e GHZZ é contabilizada na relação "GTXX e GHZZ", enquanto uma enzima com domínios GTXX, GTYY e GHZZ é contada apenas na relação "GTXX, GTYY e GHZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 5490. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contabilizada em mais de uma barra. Uma enzima com domínios GTXX, GTYY e GHZZ seria contada três vezes, nas barras "GTXX", "GTYY" e "GHZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

Essa relação numericamente grande mas proporcionalmente pequena entre a família GT2 e a classe GH pode ser consequência dessa família ser um dos grupos mais ancestrais da classe GT.(93) Isso permitiria o aparecimento e conservação de múltiplos domínios de classes alternativas de domínios funcionais ao longo da sua história evolutiva.

A segunda relação mais numerosa entre subconjuntos de dados das classes GT e GH ocorre entre as famílias GT2 e GH17, com 1511 proteínas. Entre as proteínas que

compartilham domínios GT2 e GH17 já caracterizadas na literatura, se encontram as proteínas de proteobactérias homólogas a proteína NdvB de *Bradyrhizobium japonicum*, que atuam como sintases de β-1,3-glucanos cíclicos.(81–83)

A relação entre domínios GT2 e GH17 pode ser análoga à relação entre domínios GT84 e GH94. Afinal, a própria família GT84 é relacionada estruturalmente, funcionalmente e sequencialmente com a família GT2.(79) Possivelmente, estes dois grupos de proteínas são exemplos de glicosiltransferases que foram selecionadas de forma a sintetizar β-glucanos cíclicos com auxílio de domínios da classe GH. No caso das sintases com domínios GT84 e GH94, as glicosiltransferases evoluíram com domínios fosforilíticos que atuam em ligações β-1,2. No caso de sintases com domínios GT2 e GH17, as glicosiltransferases evoluíram com domínios hidrolíticos que atuam nas ligações β-1,3.

3.2.2.3 Interação entre as classes GT e CE

Os conjuntos de dados das classes GT e CE compartilham 2006 enzimas. A Figura 15 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

As famílias GT2, GT4 e CE4 desempenham os papeis principais na relação entre as classes GT e CE, apesar dessas enzimas representarem uma pequena parcela do total de enzimas classificadas em cada uma dessas famílias. Isso pode ocorrer pois essas famílias representam grupos heterogêneos de domínios catalíticos, agrupando enzimas com múltiplas funções presentes em uma vasta diversidade de organismos.

As famílias GT2 e GT4 representam os grupos mais ancestrais da classe GT, a partir das quais praticamente todas as enzimas dessa classe evoluíram.(93, 94) Isso pode ter permitido o surgimento, ao longo de suas histórias evolutivas, de múltiplos domínios funcionais que auxiliem as funções primárias dos domínios dessas famílias. Já a família CE4 é a maior família da classe CE, contendo enzimas pertencentes a bactérias, arqueobactérias, eucariotos e até vírus — sendo a única família de CE a ter representantes de origem viral.(43, 74) A diversidade de organismos que possuem enzimas dessa família indica uma ancestralidade que permitiria o surgimento e seleção de domínios CE4 ao lado de domínios GT2 e GT4.

Particularmente, domínios GT2 e CE4 podem atuar sobre quitina, um dos carboidratos mais comuns da natureza, presentes na parede celular de fungos e diatomáceas, exoesqueleto de artrópodes e endoesqueleto de moluscos.(74, 76) Além disso, eles podem atuar em poliacetilglicosaminas, um carboidrato fundamental da matriz extracelular de fungos e bactérias.(75) Essas relações funcionais podem ser a chave para entender o surgimento de enzimas modulares com domínios GT2 e CE4.

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
CE4 e GH18 e GT2	752	1.72%	2%	0.22%
CE4 e GT2	694	1.59%	0.2%	-
CE4 e GT4	467	1.07%	0.18%	-
9 Outras Relações	93	-	-	-

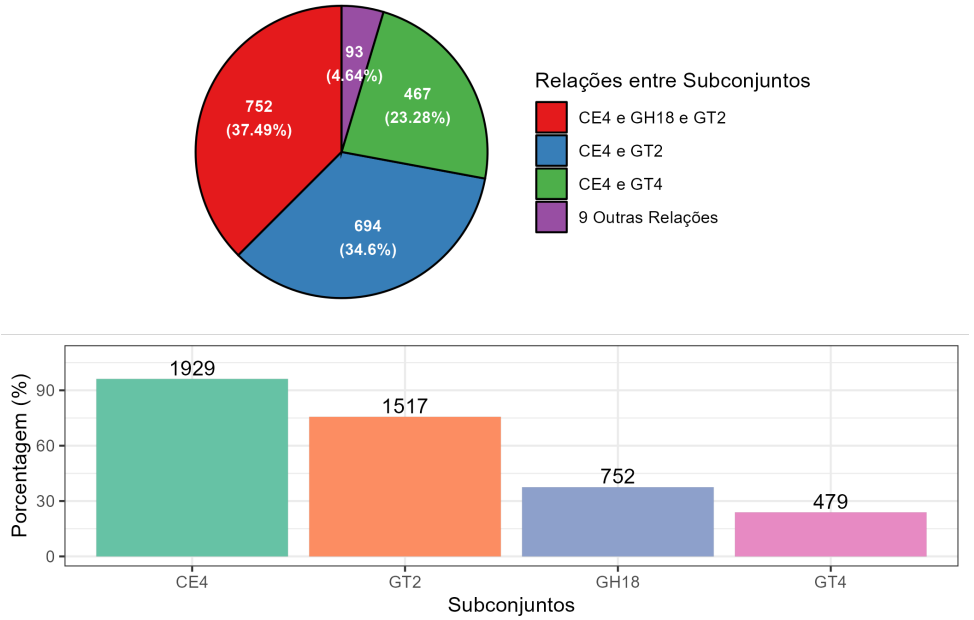


Figura 15 – Distribuição dos 2006 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e CE. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios GTXX e CEZZ é contada na relação "GTXX e CEZZ", enquanto uma enzima com domínios GTXX, GTYY e CEZZ é contada apenas na relação "GTXX, GTYY e CEZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 2006. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios GTXX, GTYY e CEZZ seria contada três vezes, nas barras "GTXX", "GTYY" e "CEZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

3.2.2.4 Interação entre as classes GT e CBM

Apenas 1942 proteínas possuem domínios GT e CBM. A Figura 16 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas conforme os subconjuntos de dados dos domínios que elas possuem

Existem dois pares de famílias predominantes na maioria das enzimas com domínios

das classes GT e CBM: enzimas com domínios das famílias GT27 e CBM13, e das famílias GT54 e CBM94.

Os domínios da família GT27 atuam como transferases de N-acetilgalactosaminil, enquanto o domínio CBM13 interage com resíduos acetilgalactosamina, que fazem parte do substrato do domínio GT27.(95) Há 1180 enzimas compartilhadas entre as famílias GT27 e CBM13, o que corresponde à maior parte da família GT27 (73.20%). Há, portanto,

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
CBM13 e GT27	1180	7.44%	73.2%	-
CBM94 e GT54	282	97.92%	84.18%	-
27 Outras Relações	480	-	-	-

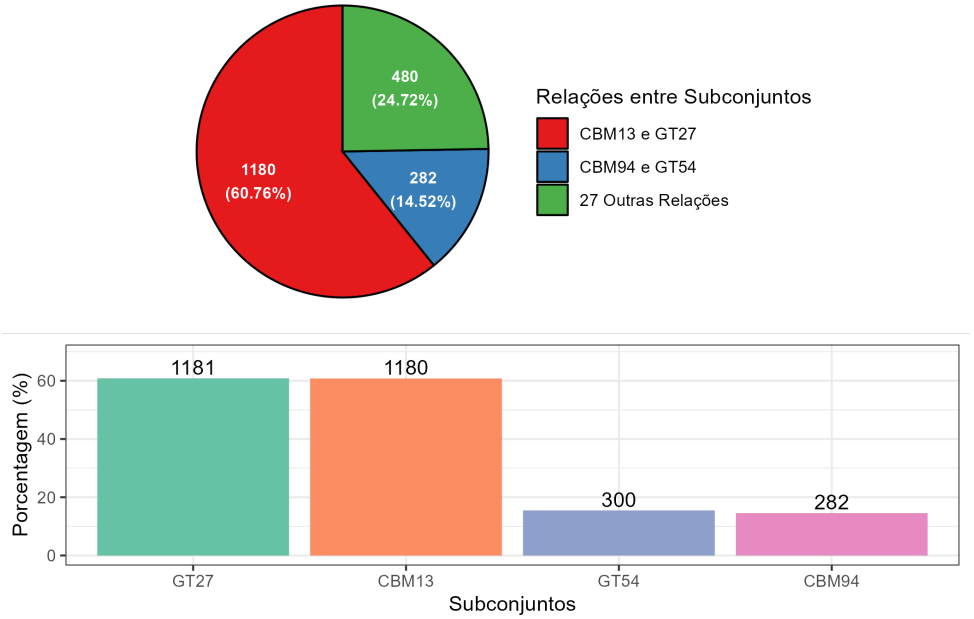


Figura 16 – Distribuição dos 1942 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e CBM. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios GTXX e CBMZZ é contada na relação "GTXX e CBMZZ", enquanto uma enzima com domínios GTXX, GTYY e CBMZZ é contada apenas na relação "GTXX, GTYY e CBMZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 1942. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios GTXX, GTYY e CBMZZ seria contada três vezes, nas barras "GTXX", "GTYY" e "CBMZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

uma provável vantagem funcional das enzimas GT27 em possuírem um domínio CBM13, resultando na alta proporção de enzimas GT27 com tal módulo acessório. Estudos adicionais podem fornecer informações biotecnológicas sobre como melhorar a atividade das enzimas GT27 que não possuem o módulo CBM13.

De maneira análoga ao caso anterior, os domínios da família GT54 atuam como transferases de N-acetilglicosaminil e o domínio CBM94 se liga a resíduos acetilglicosamina.(95) Existem 282 enzimas catalogadas com domínios GT54 e CBM94. Esse é um número absoluto pequeno de proteínas, mas corresponde à maior parte das enzimas detectadas com qualquer um desses domínios. Isso é um forte indício de que essas duas famílias evoluíram juntas em consequência de uma provável dependência funcional da atividade catalítica de GT54 e dos módulos de ligação CBM94.

Os domínios das famílias GT27 e GT54 possuem origens evolutivas diferentes (93, 96), mas tem similaridades funcionais: ambas são transferases de aminoaçúcares acetilados (acetilgalactosaminil e acetilglicosamina) que se beneficiam da presença de módulos CBM.(95) Isso indica um padrão evolutivo que se repetiu para otimizar a catálise de um mesmo tipo de reação orgânica em substratos sutilmente diferentes.

3.2.2.5 Interação entre as classes GT e PL

Apenas seis enzimas possuem, ao mesmo tempo, domínios GT e PL. A Figura 17 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

Dentre essas seis enzimas, quatro são proteínas hipotéticas, isto é, preditas a partir do sequenciamento de genomas sem a presença de evidências experimentais de que de fato são expressas nos organismos; uma proteína não anotada (número de acesso no GenBank (97): CAE5966419.1), isto é, que há evidência de sua existência, mas não de sua função; e uma proteína anotada computacionalmente (número de acesso no GenBank (97): QBY01811.1), que possui função predita computacionalmente. A enzima anotada é uma alginato liase de uma bactéria do gênero *Boseongicolae*, que possui um domínio atribuído à família PL38 e um domínio não atribuído a nenhuma família GT (família GT0).

O aparecimento de apenas algumas enzimas na interação entre as classes GT e PL indica um evento raro na evolução de enzimas ativas em carboidratos. Portanto, a caracterização dessas enzimas pode ser de grande interesse científico e, possivelmente, tecnológico.

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
GT2 e PL38	3	0%	0.18%	-
GT0 e PL38	1	0%	0.06%	-
GT1 e PL1_1	1	0%	0.03%	-
GT4 e PL12_3	1	0%	0.21%	-

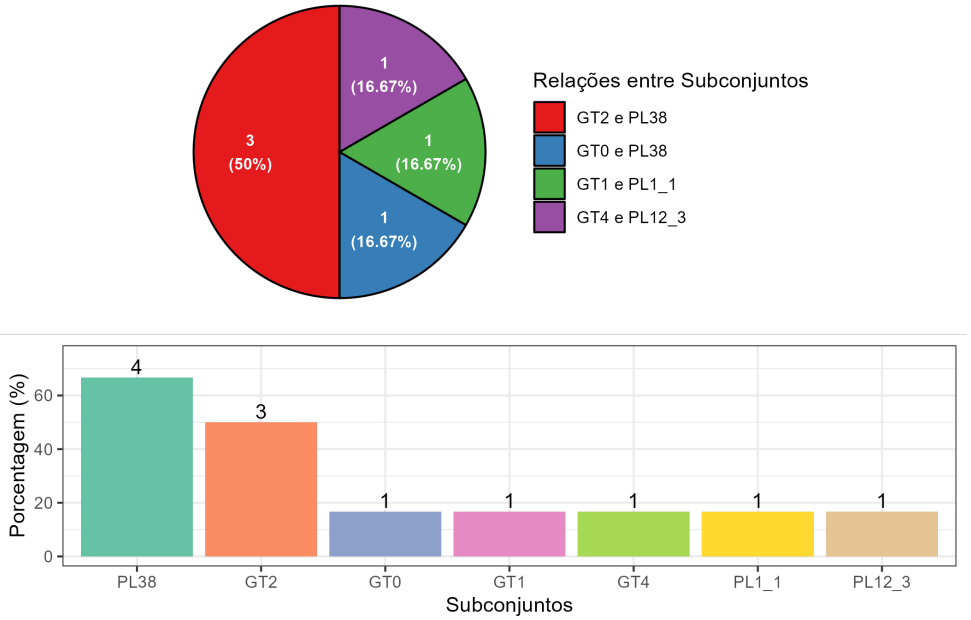


Figura 17 – Distribuição dos 6 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e PL. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios GTXX e PLZZ é contada na relação "GTXX e PLZZ", enquanto uma enzima com domínios GTXX, GTYY e PLZZ é contada apenas na relação "GTXX, GTYY e PLZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 6. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios GTXX, GTYY e PLZZ seria contada três vezes, nas barras "GTXX", "GTYY" e "PLZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

3.2.2.6 Interação entre as classes GT e AA

Existem apenas três proteínas catalogadas com domínios GT e AA, todas elas hipotéticas (números de acesso no GenBank (97): ARX86487.1, QMW38021.1 e QMW25940.1). Assim como no caso anterior, isso revela que a mistura de domínios das classes GT com AA são uma raridade na evolução de enzimas ativas em carboidratos e a caracterização dessas enzimas pode ser de interesse científico e tecnológico. A Figura 18 apresenta uma tabela,

um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
AA3_2 e GT4	2	0.08%	0%	-
AA5_0 e GT2	1	1.01%	0%	-

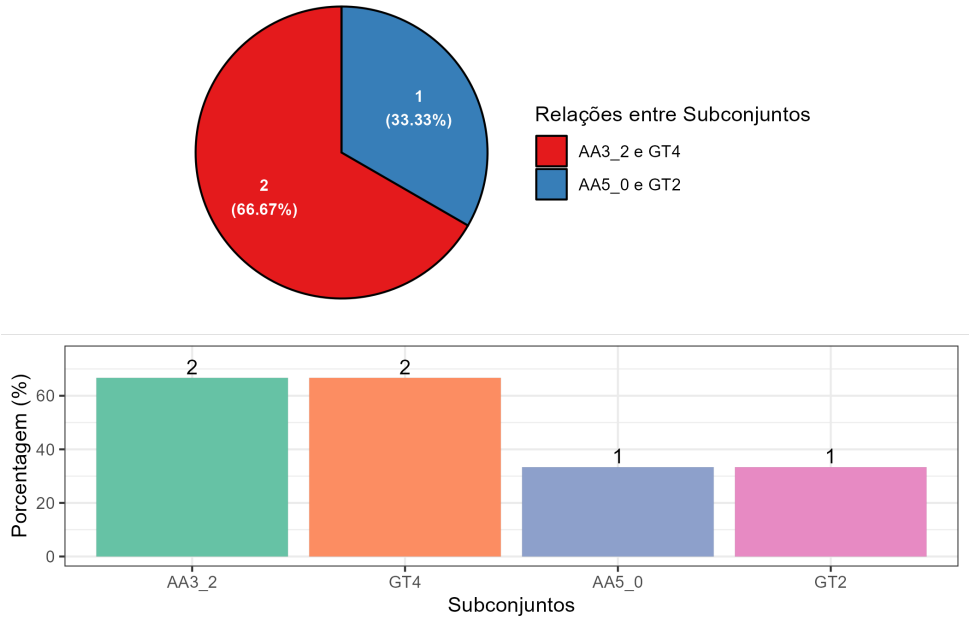


Figura 18 – Distribuição dos 6 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GT e AA. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios GTXX e AAZZ é contada na relação "GTXX e AAZZ", enquanto uma enzima com domínios GTXX, GTYY e AAZZ é contada apenas na relação "GTXX, GTYY e AAZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 6. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios GTXX, GTYY e AAZZ seria contada três vezes, nas barras "GTXX", "GTYY" e "AAZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

3.2.2.7 Interações envolvendo a classe PL

Depois da classe GT, PL é a classe com menor diversidade do meta-modelo, igual a 1,286. Ela compartilha 2907 elementos, 5,81% de seu conjunto de dados, com outras classes funcionais. A Figura 19 apresenta a distribuição desses elementos nos subconjuntos de dados da classe PL a quais eles pertencem.

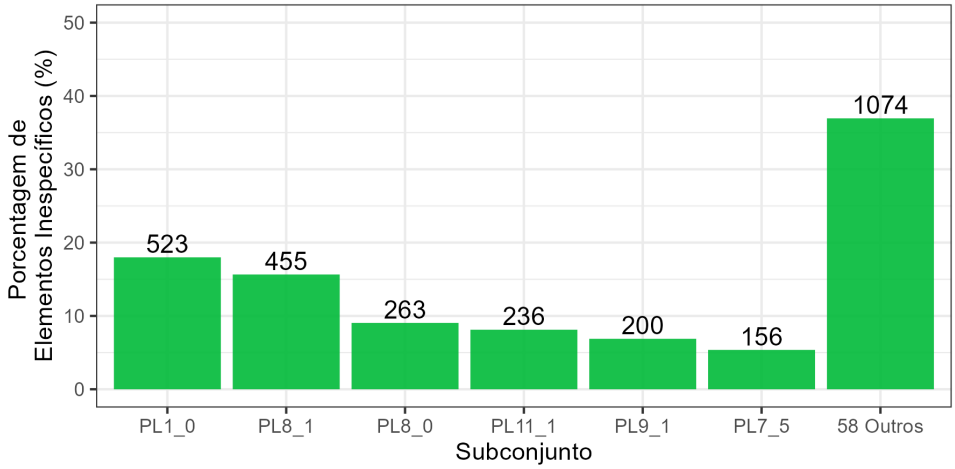


Figura 19 – Distribuição dos 2907 elementos inespecíficos do conjunto de dados da classe PL nos subconjuntos de dados aos quais pertencem.

Fonte: Elaborada pelo autor.

3.2.2.8 Interação entre as classes PL e CBM

A classe PL compartilha 2270 enzimas com a classe CBM. A Figura 20 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

Domínios da subfamília PL8_1 tem como uma possível atividade catalítica a atividade de liase de hialuronatos (63), enquanto a família CBM70 tem função de ligação ao hialuronato.(98) Essa relação é interessante porque os elementos compartilhados correspondem a 86,01% da classe CBM70, o que sugere uma estreita relação evolutiva entre as origens dos módulos CBM70 e a função de liase de hialuronatos.

3.2.2.9 Interação entre as classes PL e CE

A classe PL compartilha 543 enzimas com a classe CE. A Figura 21 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

A relação entre essas duas classes se deve principalmente aos 418 elementos da família CE8 que também possuem um domínio PL, das quais 85 enzimas possuem um

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
CBM70 e PL8_1	455	86.01%	32.2%	-
225 Outras Relações	1815	-	-	-

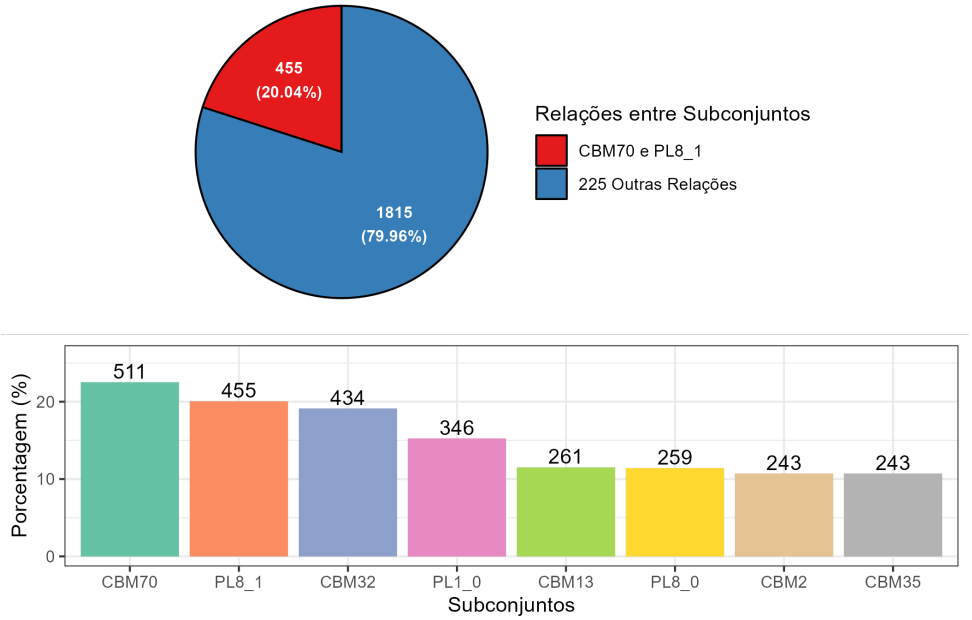


Figura 20 – Distribuição dos 2270 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes PL e CBM. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios PLXX e CBMZZ é contada na relação "PLXX e CBMZZ", enquanto uma enzima com domínios PLXX, PLYY e CBMZZ é contada apenas na relação "PLXX, PLYY e CBMZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 2270. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios PLXX, PLYY e CBMZZ seria contada três vezes, nas barras "PLXX", "PLYY" e "CBMZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

domínio da subfamília PL1_2 e 173 enzimas possuem domínios PL1 sem subclassificação (PL1_0).

A família CE8 é uma família de metilesterases de pectina (67), enquanto a família PL1 é composta de liases de pectato.(99) A família PL1 é dividida em 13 subfamílias, conforme uma classificação baseada na filogenia de seus membros.(99) Há indícios que domínios CE8 e PL1 podem operar de forma conjunta para processar pectina, primeiro

com o domínio CE8 convertendo pectina em pectato e, depois, com o domínio PL1 atuando sobre o pectato produzido.(77) Dado que uma grande quantidade de enzimas contendo esses dois domínios tem um domínio PL1 sem subclassificação, uma análise filogenética dessas enzimas poderia revelar detalhes ainda não descritos sobre a evolução da família PL1.

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
CE8 e PL1_0	173	1.55%	8.36%	-
CE8 e PL1_2	85	0.76%	4.34%	-
CE12 e PL11_1	66	1.35%	3.41%	-
24 Outras Relações	219	-	-	-

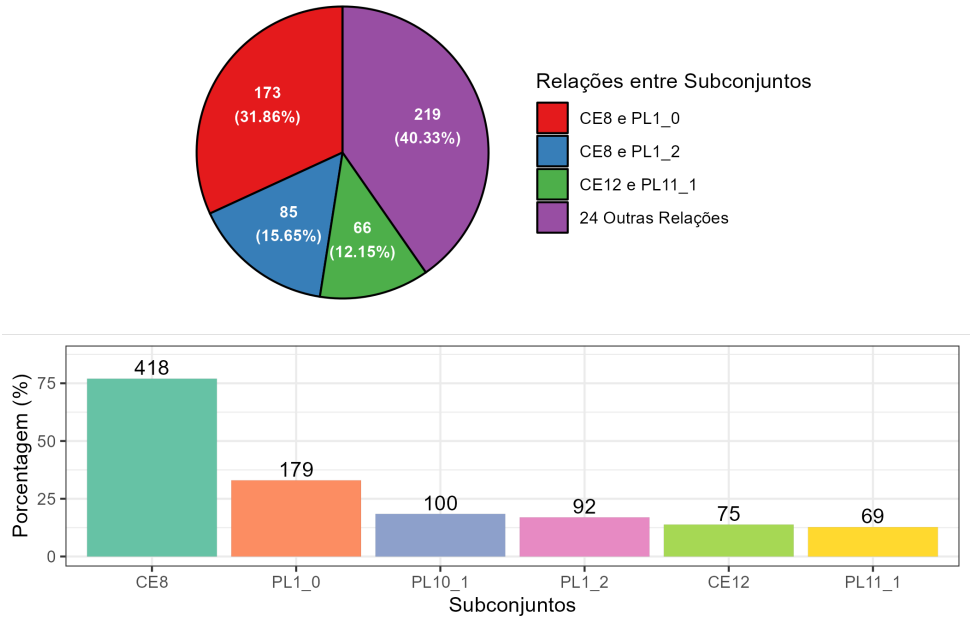


Figura 21 – Distribuição dos 543 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes PL e CE. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios PLXX e CEZZ é contada na relação "PLXX e CEZZ", enquanto uma enzima com domínios PLXX, PLYY e CEZZ é contada apenas na relação "PLXX, PLYY e CEZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 543. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios PLXX, PLYY e CEZZ seria contada três vezes, nas barras "PLXX", "PLYY" e "CEZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

3.2.2.10 Interação entre as classes PL e GH

As classes PL e GH compartilham apenas 126 elementos. A Figura 22 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
GH105 e PL33_2	41	0.45%	12.81%	-
GH88 e PL38	18	0.34%	1.05%	-
34 Outras Relações	67	-	-	-

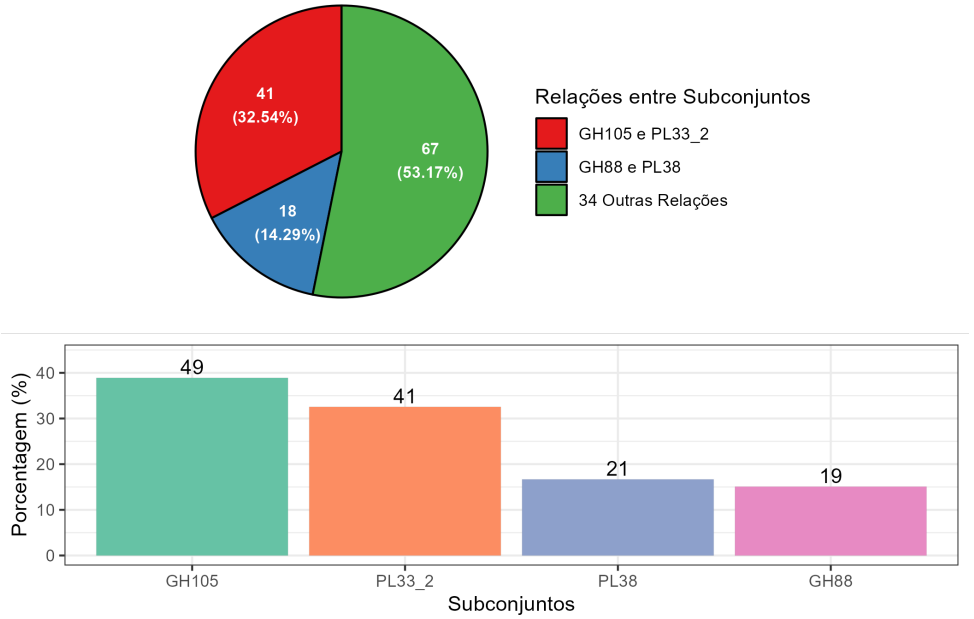


Figura 22 – Distribuição dos 126 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes PL e GH. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios PLXX e GHZZ é contada na relação "PLXX e GHZZ", enquanto uma enzima com domínios PLXX, PLYY e GHZZ é contada apenas na relação "PLXX, PLYY e GHZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 126. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios PLXX, PLYY e GHZZ seria contada três vezes, nas barras "PLXX", "PLYY" e "GHZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

Quase um terço dos elementos compartilhados entre essas classes possuem domínios da família GH105 e da subfamília PL33_2, e mais um sétimo deles possuem domínios das

famílias GH88 e PL38. A pequena quantidade de elementos inespecíficos entre essas duas classes, além de grande parte desses elementos estarem relegados a apenas quatro famílias de enzimas, indica uma novidade evolutiva ainda pouco estudada na literatura científica.

As enzimas que possuem os pares de domínios GH105 e PL33_2 ou GH88 e PL38 são originadas de bactérias da ordem bacteroidales, bactérias gram-negativas presentes na microbiota intestinal humana. A família PL33 inclui liases de gelano, sulfato de condroitina e hialuronanos, enquanto a família PL38 agrupa liasas de glucuronano.(86,87) A atividade catalítica das duas famílias de liases resulta na produção de terminações insaturadas em polissacarídeos.(86,87) As famílias GH105 e GH88 incluem glicosidases que atuam sobre sacarídeos insaturados, inclusive os produzidos após a ação de liases.(85,86) Isso sugere que, na ordem bacteroidales, podem ter evoluído enzimas modulares que atum sequencialmente para degradar um substrato: primeiro pela atuação do domínio liase e, depois, do domínio glicosidase sobre a terminação insaturada produzida pela liase.

3.2.2.11 Interações envolvendo a classe CE

A classe CE é a quarta em diversidade no meta-modelo, igual a 1,434. Ela compartilha 10'391 elementos, 7,64% de seu conjunto de dados, com outras classes. A Figura 23 apresenta a distribuição desses elementos nos subconjuntos de dados da classe CE a quais eles pertencem. A família CE4 está muito envolvida nas relações entre CE e os outros conjuntos de dados do meta-modelo. Essa família é a maior e mais diversa família de CEs, presente em uma grande variedade de espécies (43,74), o que pode explicar a existência de tantas enzimas com domínios CE4 e domínios de outras classes enzimáticas.

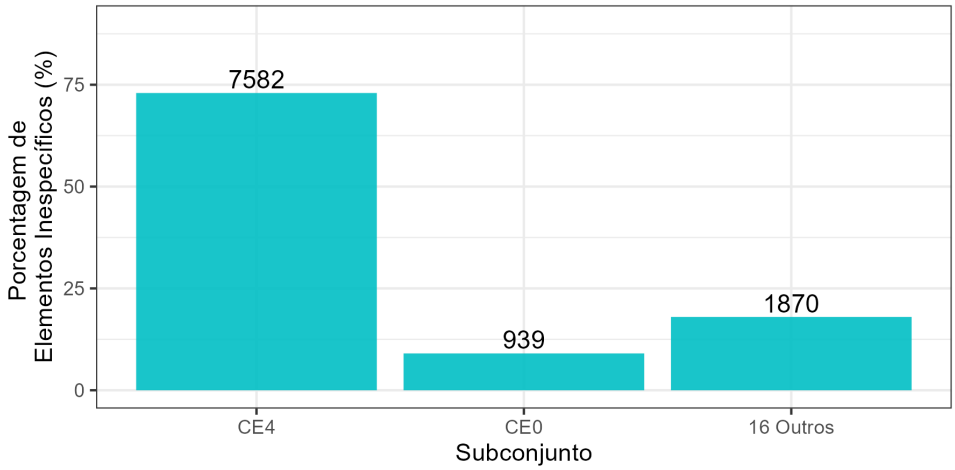


Figura 23 – Distribuição dos 10391 elementos inespecíficos do conjunto de dados da classe CE nos subconjuntos de dados aos quais pertencem.

Fonte: Elaborada pelo autor.

3.2.2.12 Interação entre as classes CE e GH

As classes CE e GH compartilham 6173 elementos. A Figura 24 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

A relação entre essas classes é interessante pois ela envolve, sobretudo, apenas duas famílias: GH153 e CE4. A família GH153 compartilha 4790 enzimas com a família CE4,

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
CE4 e GH153	4790	10.97%	98.86%	-
CE4 e GH18 e GT2	752	1.72%	2%	0.22%
129 Outras Relações	631	-	-	-

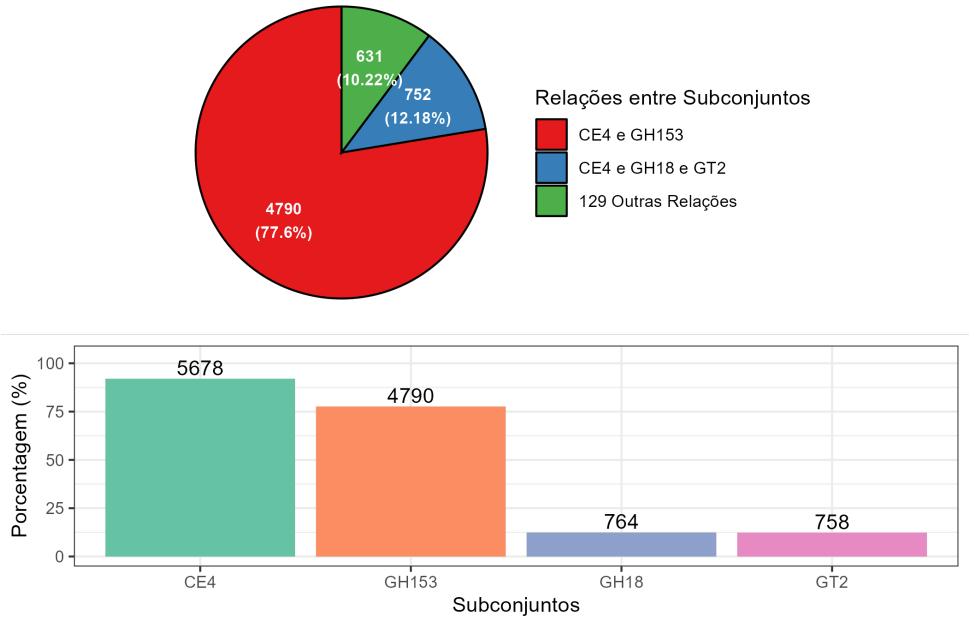


Figura 24 – Distribuição dos 6173 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes CE e GH. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios CEXX e GHZZ é contada na relação "CEXX e GHZZ", enquanto uma enzima com domínios CEXX, CEYY e GHZZ é contada apenas na relação "CEXX, CEYY e GHZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 6173. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios CEXX, CEYY e GHZZ seria contada três vezes, nas barras "CEXX", "CEYY" e "GHZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

correspondente a 98,86% de seus 4845 elementos. Isso indica que o submodelo de enzimas com domínios GH153 é quase completamente dependente da presença de um domínio CE4. De fato, a família GH153 foi definida a partir de duas enzimas que possuíam um domínio de atividade hidrolítica em poliacetilglicosaminas — os primeiros domínios identificados da família GH153 — e um domínio desacetilase da família CE4.(64, 73) Apesar disso, ainda existem 55 proteínas com domínios GH153 que não possuem um domínio CE4, o que sugere a ocorrência de uma divergência evolutiva ainda não descrita na literatura em relação à maior parte das enzimas da família GH153.

3.2.2.13 Interação entre as classes CE e CBM

A classe CE é a terceira classe com maior quantidade de proteínas com módulos CBM, contando com 2667 enzimas que os possuem. A Figura 25 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

A única relação numerosa entre subconjuntos de dados ocorre entre as famílias CBM12 e CE4, contendo 283 enzimas compartilhadas. Todos essas enzimas são originadas de gammaproteobactérias, indicando uma origem evolutiva comum entre elas. As famílias CBM12 e CE4 possuem membros que atuam sobre xilano, indicando a presença de módulos CBM12 para auxiliar o domínio catalítico CE4 a atuar nesse substrato.(68, 74) Uma caracterização mais minuciosa dessas enzimas pode revelar o porque da associação de módulos CE4 com módulos de ligação a xilano surgiu e foi fixada nessa clado de bactérias, diferente do que se observa na maioria dos organismos que possuem esterases de xilano.

Ademais, vários subconjuntos de dados apresentam, individualmente, mais de 10% dos elementos inespecíficos compartilhados pelas classes CE e CBM. São eles os subconjuntos de dados CBM2, CBM12, CBM13, CE1, CE3 e CE4, assim como o subconjunto de dados de enzimas CE sem subclassificação, CE0. Membros das famílias CBM2 tem afinidade por quitina e xilano, CBM12 por quitina e CBM13 por xilano.(43,67) Já a família CE1 atua em xilano e feruloil, CE3 em xilano, e CE4 em quitina e xilano.(43, 67, 74) Isso indica que as associações entre os domínios de CE e CBM se concentram, sobretudo, na otimização da atividade catalítica de esterases sobre quitina e xilano.

3.2.2.14 Interação entre as classes CE e AA

As classes CE e AA compartilham apenas 25 enzimas. A Figura 26 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem. Entre elas, três são compartilhadas por famílias distintas de AA e CE e pertencem a espécies distintas de fungos (gêneros *Flammulina*, *Ceratobasidium* e *Colletotrichum*), enquanto 22 enzimas são compartilhadas entre as famílias AA5 e CE3.

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
CBM12 e CE4	283	8.11%	0.65%	-
180 Outras Relações	2384	-	-	-

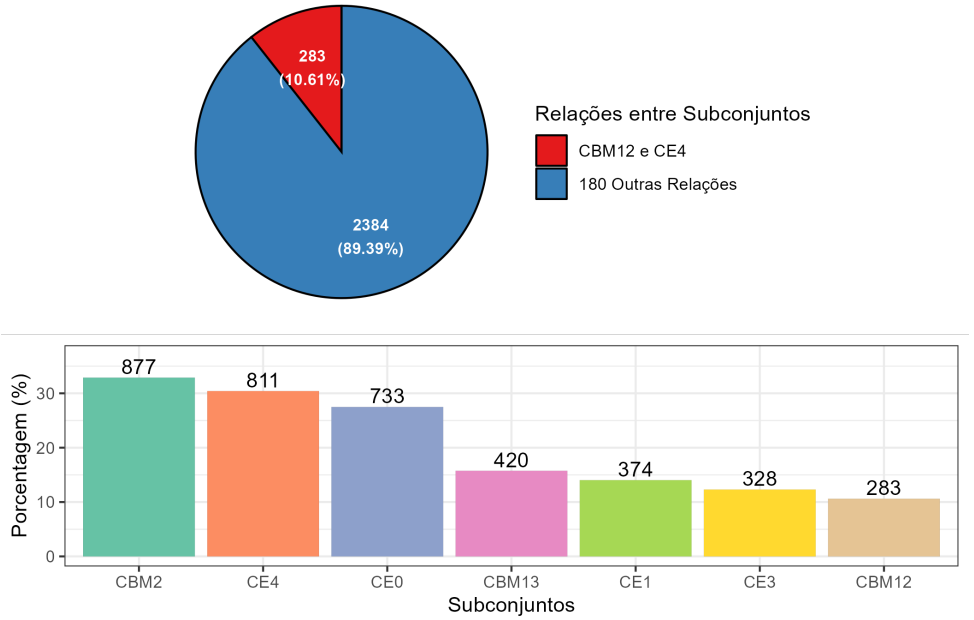


Figura 25 – Distribuição dos 2667 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes CE e CBM. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios CEXX e CBMZZ é contada na relação "CEXX e CBMZZ", enquanto uma enzima com domínios CEXX, CEYY e CBMZZ é contada apenas na relação "CEXX, CEYY e CBMZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 2667. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios CEXX, CEYY e CBMZZ seria contada três vezes, nas barras "CEXX", "CEYY" e "CBMZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

As 22 enzimas que possuem domínios AA5 e CE3 pertencem a uma mesma bactéria, *Burkholderia pseudomallei*, um patógeno oportunista gram-negativo.(78) Isso ressalta a raridade da associação desses dois domínios ao longo da evolução. A família AA5 é composta por oxidases dependentes de cobre e um radical, e catalisa a oxidação de alcoóis e aldeídos.(100) Já a família CE3 contém esterases de xilano.(67) Os substratos e produtos resultantes da atividade catalítica desses dois domínios são aparentemente

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
AA5_0 e CE3	22	22.22%	3.5%	-
3 Outras Relações	3	-	-	-

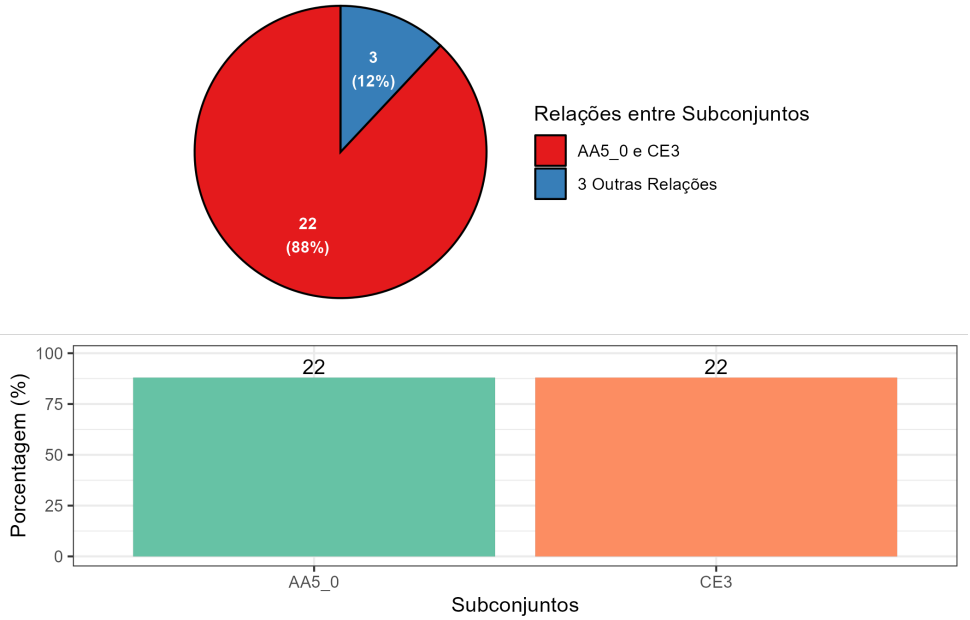


Figura 26 – Distribuição dos 25 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes CE e AA. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios CEXX e AAZZ é contada na relação "CEXX e AAZZ", enquanto uma enzima com domínios CEXX, CEYY e AAZZ é contada apenas na relação "CEXX, CEYY e AAZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 25. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios CEXX, CEYY e AAZZ seria contada três vezes, nas barras "CEXX", "CEYY" e "AAZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

incompatíveis para a uma atuação catalítica conjunta ou sequencial. Frente a essa questão, uma dessas enzimas foi caracterizada recentemente por Mazurkewich e colaboradores (78), e foi revelado que o domínio CE3 dela é cataliticamente inativo e, possivelmente, só possui funções estruturais.

A partir desse resultado podemos criar a hipótese de que o motivo para haver uma pequena quantidade de enzimas que contêm domínios AA5 e CE3 é a incompatibilidade

catalítica desses domínios, e o surgimento de enzimas desse tipo está condicionado à ocorrência de um evento evolutivo raro em que o domínio CE3 acoplado ao domínio AA5 possui mutações que o inativam — o que deve ter ocorrido na evolução da bactéria *Burkholderia pseudomallei*. Essa hipótese pode ser estendida para explicar o porque de haver apenas mais três enzimas com domínios AA e CE diversos em três espécies de fungos de gêneros filogeneticamente distantes.

3.2.2.15 Interações envolvendo a classe GH

A classe GH é a terceira com maior diversidade de conexões no meta-modelo, igual a 1,475, um pouco mais diversa que a classe CE. Ela compartilha 174778 elementos, 12,80% de seu conjunto de dados, com outras classes, a maior parte apenas com a classe CBM. A Figura 27 apresenta a distribuição desses elementos nos subconjuntos de dados da classe GH a quais eles pertencem. Como discutido anteriormente, a classe GH tem uma relação substancial com as classes GT e CE, principalmente devido a famílias de domínios enzimáticos muito inespecíficas, como GT2 e CE4. Em contraste, a classe GH compartilha poucos elementos com as classes AA e PL, apenas 30 e 126 enzimas, respectivamente.

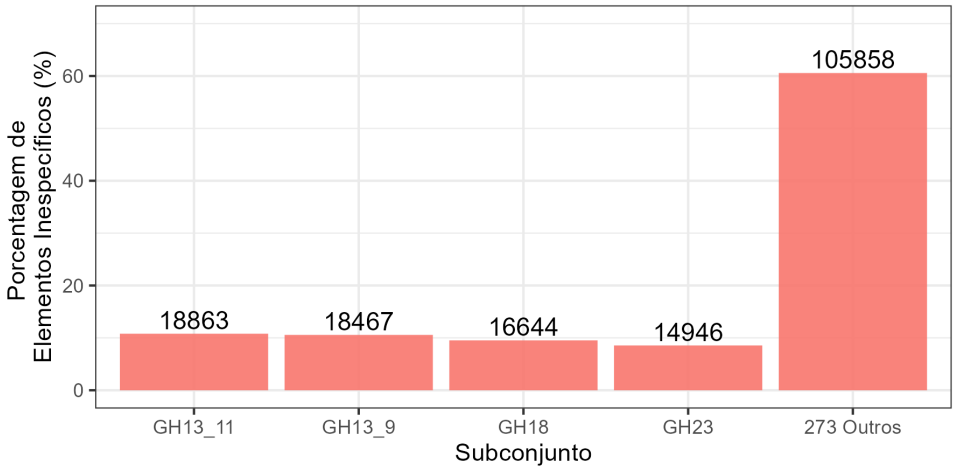


Figura 27 – Distribuição dos 174778 elementos inespecíficos do conjunto de dados da classe GH nos subconjuntos de dados aos quais pertencem.

Fonte: Elaborada pelo autor.

3.2.2.16 Interação entre as classes GH e CBM

O relacionamento entre as classes GH e CBM envolve 370 famílias e subfamílias diferentes e é, em números absolutos, o maior relacionamento entre quaisquer tipos de domínios enzimáticos analisados na rede de meta-modelagem. Proporcionalmente, GH é o segundo grupo com maior quantidade de enzimas contendo módulos CBM (12,01% de suas enzimas), perdendo apenas para a classe AA (21,27%). A Figura 28 apresenta

uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

A principal função de domínios da classe GH é degradar carboidratos, e sua efetividade é fortemente afetada pela sua capacidade de se acoplar e fixar nas cadeias polissacarídicas.(68) Nesse sentido, enzimas com domínios da classe GH se beneficiam de ter módulos de CBM, que se ligam ao carboidrato, disponíveis em sua estrutura

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
CBM48 e GH13_11	18814	32.82%	97.93%	-
CBM48 e GH13_9	18443	32.18%	98.87%	-
CBM50 e GH23	14899	13.51%	12.54%	-
1666 Outras Relações	111848	-	-	-

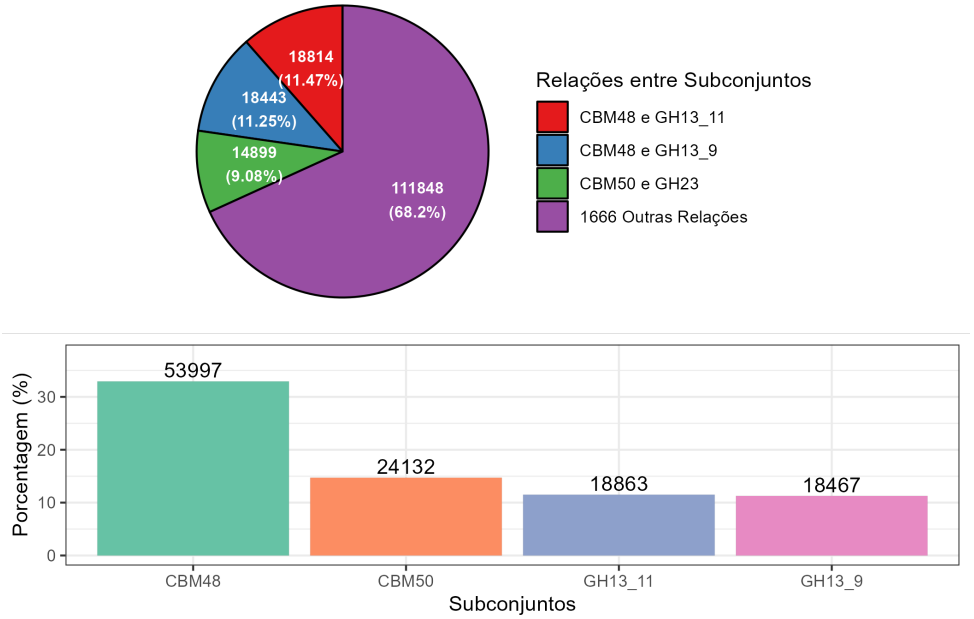


Figura 28 – Distribuição dos 164004 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GH e CBM. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios GHXX e CBMZZ é contada na relação "GHXX e CBMZZ", enquanto uma enzima com domínios GHXX, GHYY e CBMZZ é contada apenas na relação "GHXX, GHYY e CBMZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 164004. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios GHXX, GHYY e CBMZZ seria contada três vezes, nas barras "GHXX", "GHYY" e "CBMZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

proteica.(68)

Dois subconjuntos de dados da classe GH se destacam por agruparem a maior quantidade de enzimas com módulos CBM: GH13_11 e GH13_9. As duas subfamílias têm quase todos os seus elementos (mais de 18 mil enzimas cada) compartilhados exclusivamente com a família CBM48.

A família GH13 é a maior família da classe GH, e é constituída por enzimas que atuam em substratos contendo ligações α-glicosídicas, como o amido.(43, 62) Essa é uma das famílias mais diversas de enzimas ativas em carboidratos, agrupando domínios que catalisam mais de trinta reações químicas distintas.(43, 62) Particularmente, a subfamília GH13_11 é composta por isoamilases, enquanto a subfamília GH13_9 é formada por enzimas ramificadoras de glucanos.(62) A família CBM48 é composta por módulos que se ligam a diversos substratos relacionados e derivados de amido e glicogênio.(101) O fato de praticamente todas as enzimas das subfamílias GH13_11 e GH13_9 apresentarem domínios da família CBM48 sugere uma estreita relação evolutiva entre elas.

Confirmando essa hipótese, análises filogenéticas posicionam as subfamílias 9 e 11 em um clado monofilético, indicando uma origem evolutiva única para as subfamílias GH13_8 a GH13_14 e a subfamília GH13_41.(62, 102) Mais da metade de cada uma dessas subfamílias possui um módulo CBM48, mostrando que esse módulo CBM48 pode ter evoluído junto a este clado de subfamílias de GH13: 58,96% dos elementos da subfamília 8 possuem um domínio CBM48; 99,10% da 9; 93,84% da 10; 98,21% da 11; 84,52% da 12; 96,18% da 13; 84,14% da 14; 85,86% da 41.

A relação evolutiva e funcional entre subfamílias de GH13 e a família CBM48, a qual estamos sugerindo a partir da análise do meta-modelo, já é bem descrita na literatura científica.(101, 103) Isso revela o potencial da meta-análise dos conjuntos de dados e modelos CAZy de revelar relações evolutivas e funcionais válidas entre domínios enzimáticos, ajudando a elucidar características biológicas dos diferentes modelos usados para classificação de proteínas.

3.2.2.17 Interação entre as classes GH e AA

Os conjuntos de dados das classes GH e AA compartilham apenas 30 elementos. A Figura 29 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

A maior parte das enzimas que possuem domínios catalíticos dessas duas classes enzimáticas podem ser separadas em dois grupos principais. O primeiro grupo comporta 12 enzimas que possuem domínios das famílias GH18 e AA15, 7 com e 5 sem a presença de um módulo CBM14. O segundo grupo possui 6 enzimas com domínios GH18, AA10 e CBM12. As 12 enzimas restantes seguem tendências evolutivas mais peculiares. Entre

essas últimas, cabe ressaltar duas enzimas com domínios GH18, AA10 e CBM5, ambas da espécie *Jonesia denitrificans*. As 12 enzimas do primeiro grupo (domínios GH18 e AA15) pertencem a seis espécies diversas de eucariotos, enquanto as 6 enzimas do segundo grupo (domínios GH18 e AA10) pertencem a três espécies de gammaproteobactérias.

Em comum, esses dois grupos possuem domínios GH18. Essa família é composta de quitinases, que catalisam a hidrólise de quitina.(88) As quitinases normalmente

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
AA15 e CBM14 e GH18	7	1.64%	0.19%	0.02%
AA10 e CBM12 e GH18	6	0.06%	0.17%	0.02%
AA15 e GH18	5	1.17%	0.01%	-
8 Outras Relações	12	-	-	-

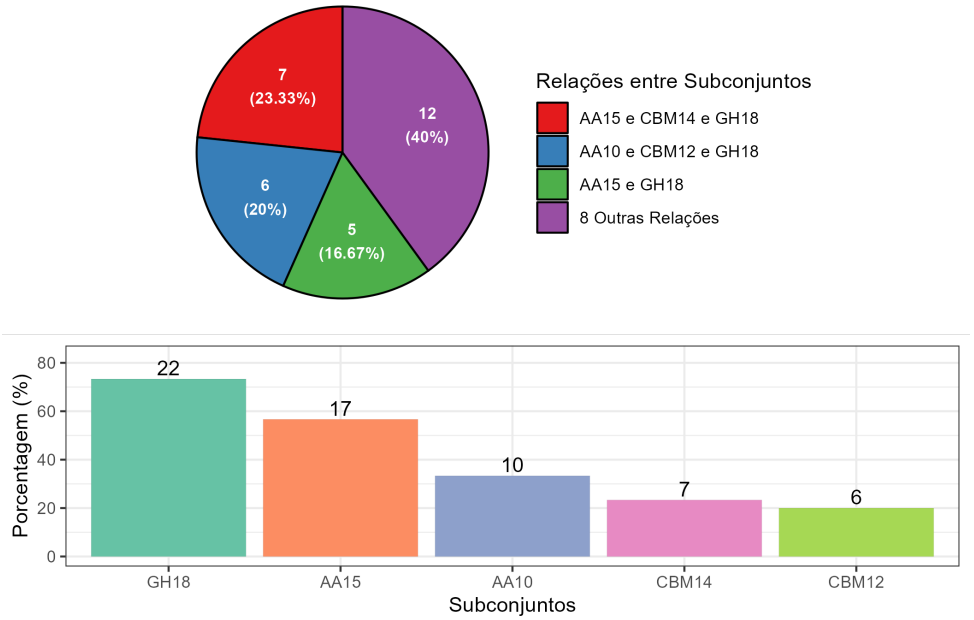


Figura 29 – Distribuição dos 30 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes GH e AA. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios GHXX e AAZZ é contada na relação "GHXX e AAZZ", enquanto uma enzima com domínios GHXX, GHYY e AAZZ é contada apenas na relação "GHXX, GHYY e AAZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 30. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios GHXX, GHYY e AAZZ seria contada três vezes, nas barras "GHXX", "GHYY" e "AAZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

só conseguem hidrolisar as terminações das cadeias de quitina, e, por isso, é comum encontrar espécies que associam a quitinases outras enzimas capazes de quebrarem ligações glicosídicas nas regiões cristalinas da quitina, produzindo novas terminações livres para a ação dos domínios GH18.(88) As principais enzimas capazes de executar essa função são mono-oxigenases líticas de polissacarídeos (LPMOs, do inglês *Lytic polysaccharide monooxygenases*), cujos domínios catalíticos são classificados nas famílias AA9 a AA17 da classe AA.(88)

O que difere os dois grupos principais de enzimas que compartilham domínios GH e AA é a classificação filogenética de seus domínios CBM e AA. Os módulos CBM14 e CBM12 são ligadores de quitina e comumente se associam a domínios GH18, facilitando a sua atuação.(104) No entanto, módulos CBM14 são normalmente encontrados em eucariotos, sobretudo insetos e mamíferos, enquanto módulos CBM12 são mais encontrados em bactérias.(104) As famílias AA15 e AA10 são famílias de LPMOs filogeneticamente próximas, que se distinguem das outras LPMOs por conterem uma fenilalanina conservada em seus sítios ativos.(88,105) Ambas as famílias são oxidantes de quitina e podem atuar criando terminações na região cristalina deste polissacarídeo.(88) A maior diferença entre elas é que, enquanto a família AA15 é presente principalmente em eucariotos, a família AA10 é encontrada, sobretudo, em bactérias.(88,105)

O paralelo funcional entre os dois pares de famílias CBM e AA, a função dos domínios AA10 e AA15 em facilitarem a atuação do domínio GH18, e a ocorrência dessas enzimas em espécies filogeneticamente diversas sugere que a associação dos domínios GH18, AA15 e CBM14 ou GH18, AA10 e CBM12 são soluções evolutivas convergentes e distintas para um mesmo problema bioquímico: a degradação eficiente de quitina.

Essa hipótese é reforçada pelos resultados obtidos por Mekasha e colaboradores (89) no estudo de enzimas da espécie *Jonesia denitrificans*. Foi mostrado que, nas enzimas com domínios GH18, AA10 e CBM5, há uma interação sinergética dos domínios hidrolíticos e da LPMO na degradação de quitina.

3.2.2.18 Interações envolvendo a classe AA

A classe de enzimas AA é a segunda com maior diversidade de conexões no metamodelo, igual a 1,614. Ela compartilha 5576 elementos, 21,42% do seu conjunto de dados, com outras classes, a maior parte apenas com a classe CBM. A Figura 30 apresenta a distribuição desses elementos nos subconjuntos de dados da classe AA a quais eles pertencem.

A classe AA agrupa domínios que atuam na degradação da parede celular vegetal a partir da catálise de reações de oxirredução.(65) Essa categoria foi adicionada à classificação CAZy apenas em 2013, a primeira após a criação do CAZy em 1998, e ainda há uma lacuna de dados experimentais sobre essa classe.(43,64,65,105) Essa falta de conhecimento pode

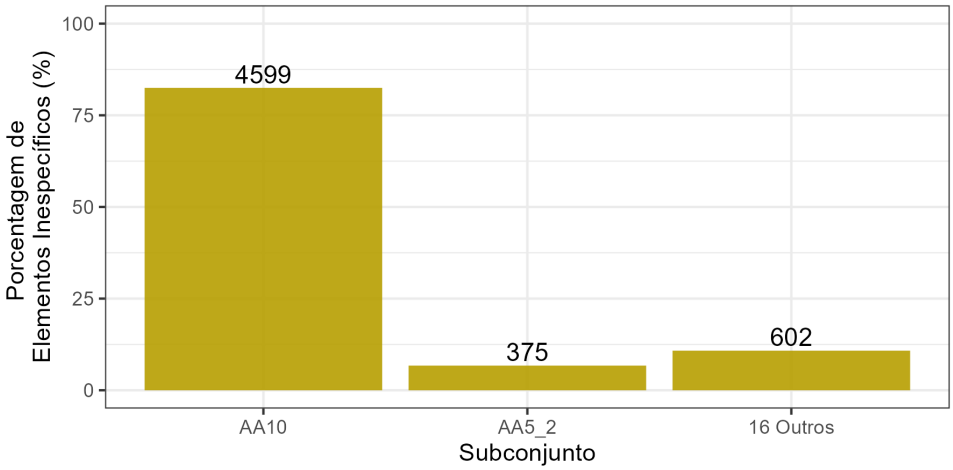


Figura 30 – Distribuição dos 5576 elementos inespecíficos do conjunto de dados da classe AA nos subconjuntos de dados aos quais pertencem.

Fonte: Elaborada pelo autor.

estar escondendo relações ainda mais amplas entre o AA e outras classes.

3.2.2.19 Interação entre as classes AA e CBM

A classe AA compartilha 5536 elementos com a classe CBM. A Figura 31 apresenta uma tabela, um gráfico de setores e um gráfico de barras que distribui essas enzimas nos subconjuntos de dados aos quais seus domínios pertencem.

Mais de 75% das enzimas com domínios AA e CBM possuem um domínio da família AA10, a maioria com módulos CBM5 ou CBM73. A família AA10 comporta LPMOs típicas de bactérias que atuam na degradação de quitina.(88, 90) As famílias CBM5 e CBM73 são módulos ligadores de quitina e derivados, com diferentes graus de afinidade dependendo do estado desse polissacarídeo.(91) As enzimas com módulos AA10 e CBM73 correspondem a quase um quinto do subconjunto de dados dessas LPMOs, além de ocupar dois quintos do subconjunto de dados de CBM73. Já as enzimas com domínios AA10 e CBM5 ocupam mais de 15% de cada um dos subconjuntos de dados correspondentes. Dados experimentais apontam que, na ausência desses módulos CBM, os domínios AA10 possuem atividade catalítica reduzida.(91) Isso sugere uma pressão seletiva positiva para manter módulos CBMs fusionados a essas LPMOs, o que explica a forte relação observada entre essas famílias enzimáticas.

3.2.2.20 Interações envolvendo a classe CBM

No meta-modelo analisado, a classe de CBMs possui a maior diversidade de conexões, igual a 2,193. Essa classe compartilha 176112 elementos, cerca de 58,20% de seu conjunto de dados, com outras classes. A Figura 32 apresenta a distribuição desses elementos nos

Principais Relações Entre Subconjuntos	Número de Elementos Compartilhados	% do Subdataset 1	% do Subdataset 2	% do Subdataset 3
AA10 e CBM73	1910	19.05%	40.68%	-
AA10 e CBM5	1719	17.15%	14.21%	-
50 Outras Relações	1907	-	-	-

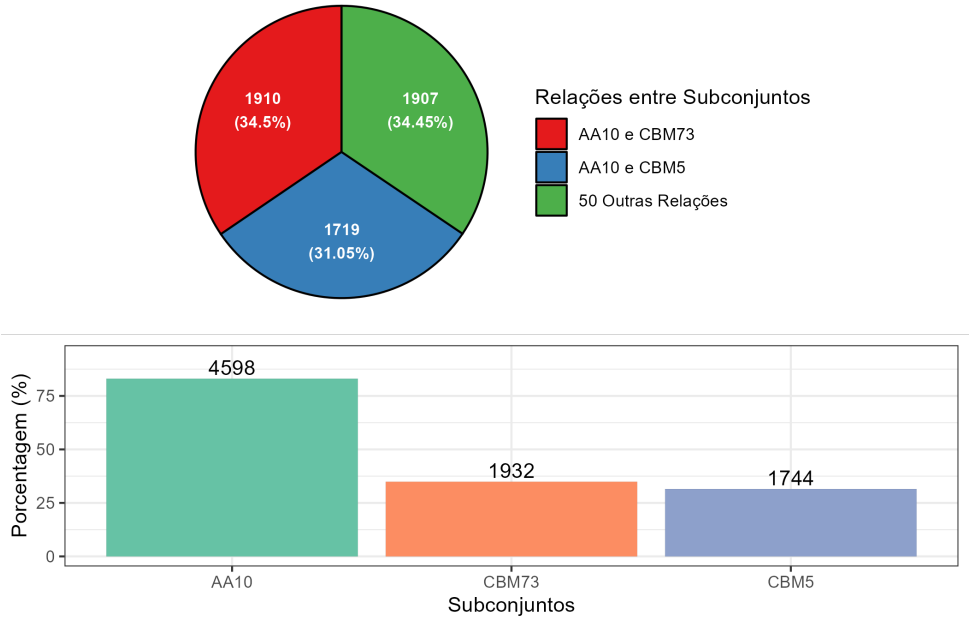


Figura 31 – Distribuição dos 5536 elementos inespecíficos compartilhados pelos subconjuntos de dados das classes AA e CBM. A tabela e o gráfico de setores dividem as enzimas em grupos mutualmente exclusivos de acordo com o perfil de famílias presentes em cada enzima. Por exemplo, uma enzima com domínios AAXX e CBMZZ é contada na relação "AAXX e CBMZZ", enquanto uma enzima com domínios AAXX, AAYY e CBMZZ é contada apenas na relação "AAXX, AAYY e CBMZZ". Dessa forma, os números de enzimas de cada relação presente nesses infográficos somam 5536. O gráfico de barras indica o número de enzimas com os domínios classificados em cada subconjunto de dados. Nesse caso, uma mesma enzima pode estar contada em mais de uma barra. Uma enzima com domínios AAXX, AAYY e CBMZZ seria contada três vezes, nas barras "AAXX", "AAYY" e "CBMZZ". São classificadas em "Outras Relações" todas as relações que possuem um número menor que 10% do total de elementos inespecíficos.

Fonte: Elaborada pelo autor.

subconjuntos de dados da classe CBM a quais eles pertencem.

A classe de módulos CBM é composta por domínios que não possuem atividade catalítica própria, e tem como principal função se ligar a substratos polissacarídicos e direcioná-los a domínios catalíticos de outras classes.(68, 106) Isso explica a alta frequência de enzimas que possuem módulos CBM associados a domínios de outras classes. Entre as famílias de CBM, vale destacar a CBM48 e a CBM50, que possuem a maior quantidade de

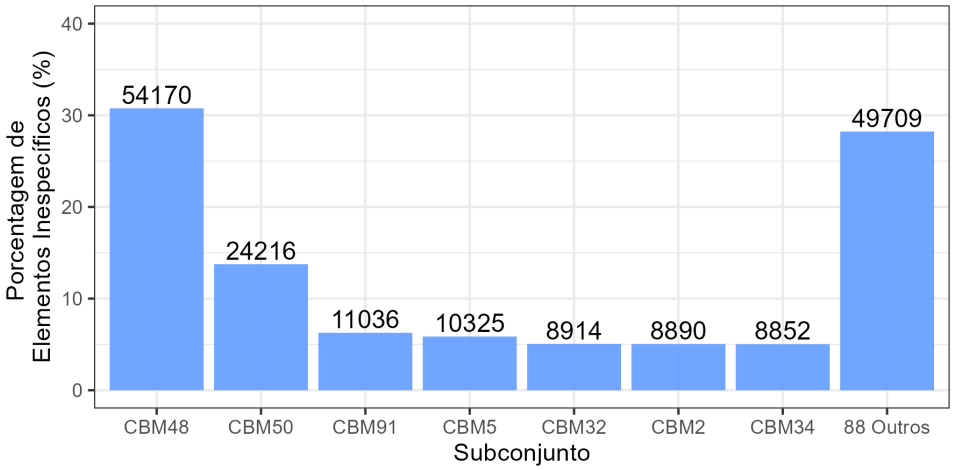


Figura 32 – Distribuição dos 176112 elementos inespecíficos do conjunto de dados da classe CBM nos subconjuntos de dados aos quais pertencem.

Fonte: Elaborada pelo autor.

enzimas catalogadas (57318 e 110257, respectivamente) e a maior frequência de associações com domínios de outras classes.

A família CBM50 se liga a peptidoglicanos e quitina, o primeiro é o principal componente da parede celular de bactérias e o segundo é presente na parede celular de fungos. A sua afinidade por carboidratos associados à parede celular a torna essencial para auxiliar funções metabólicas importantes, como a divisão celular, sinalização celular e atividade antimicrobiana e antifúngica (107, 108), justificando a abundância desta família.

A família CBM48 se liga a substratos relacionados a amido e glicogênio, os dois polissacarídeos com função de armazenamento energético mais comuns da natureza.(109) Os seus membros compartilham uma história evolutiva muito próxima da evolução de domínios GH envolvidos na degradação de amido.(101, 103, 109) Esses dois fatos explicam a abundância de elementos nessa família e a sua forte relação com domínios catalíticos de outras classes.

3.2.2.21 As possibilidades da análise de subconjuntos e submodelos em redes de metamodelagem de proteomas anotados

As análises abordadas nessa subseção tem caráter exploratório e permitiram inferir vários padrões funcionais e evolutivos de domínios de enzimas ativas em carboidratos. Todas as interações entre modelos de domínios funcionais observadas na rede de metamodelagem foram sistematicamente examinados em termos de subconjuntos de dados e submodelos, revelando detalhes de famílias enzimáticas que de outra forma estariam encobertas em meio à grande quantidade de informação presente no banco de dados.

Os dados de anotação do banco de dados CAZy (43,44) foram escolhidos justamente

para permitir que as hipóteses levantadas pudessem ser validadas na literatura científica. Praticamente todas as inferências retiradas da análise da rede de meta-modelagem foram confirmadas por estudos publicados na literatura, exceto algumas hipóteses que não puderam ser nem validadas, nem descartadas.

Assim sendo, a abordagem de meta-modelagem com análise de subconjuntos de dados e submodelos é uma ferramenta promissora para extrair informações de proteomas anotados computacionalmente, auxiliando ferramentas tradicionais de anotação de domínios proteicos para produzir hipóteses mais completas das enzimas estudadas.

4 CONSIDERAÇÕES FINAIS

4.1 Conclusões

O foco desse trabalho foi a extensão e a aplicação de uma abordagem de meta-modelagem para analisar modelos baseados em dados, a rede de meta-modelagem $\mathcal{M}$. A abordagem sugerida oferece um meio de quantificar a especificidade de modelos e facilita a criação de novos modelos por meio de composições lógicas. Devido aos seus princípios de integração de modelos, a rede de meta-modelagem $\mathcal{M}$ é projetada para auxiliar na avaliação, interpretação e melhoria de modelos científicos, especialmente quando aplicada em problemas de reconhecimento de padrões.

Para construir um meta-modelo, dividimos a abordagem de modelagem em três domínios principais e formalizamos operações de mapeamento entre eles: os domínios Ω, A e E. O primeiro domínio é o universo Ω, que abriga todos os dados disponíveis sobre um fenômeno que podem ser coletados experimentalmente. Cada elemento em Ω é mapeado para um conjunto de dados, que são organizados no domínio de ambiente de dados A. Cada conjunto de dados é mapeado em um modelo disponível no domínio da estrutura de modelagem E. Esses mapeamentos podem ser apreciados de forma eficaz a partir da representação de mapas entre domínios da abordagem de meta-modelagem $\mathcal{M}$.

Uma operação de mapeamento também pode ser descrita entre cada elemento nos conjunto de dados em A e cada modelo disponível em E, com base na capacidade desses modelos de explicarem cada elemento individual. Essa operação é a base para a formação de uma rede de meta-modelagem. Isso possibilita a utilização de métodos de ciência de rede para visualizar a interação ente conjuntos de dados e modelos e para avaliar quantitativamente essa interação. Particularmente, a rede foi representada por um grafo ponderado bipartido e, a partir dele, foram consideradas métricas de especificidade e redundância de modelos baseadas em análise de conjuntos e análises topológicas, incluindo *multiplicidade média dos elementos* e *diversidade total das conexões*.

Além disso, com a construção da rede de meta-modelagem $\mathcal{M}$, foi possível elaborar uma álgebra pareada de operações de conjuntos entre conjuntos de dados e operações lógicas entre modelos. Essa relação possibilitou a criação de novos modelos com base no ambiente de dados existente e na estrutura de modelagem estabelecida. Os modelos criados podem ser usados para auxiliar na interpretação de modelos com pouca informação, substituir modelos redundantes ou pouco específicos dentro da estrutura de modelagem, ou modelar temporariamente novos dados, fornecendo estratégias para melhorar abordagens de modelagem.

A abordagem de meta-modelagem foi aplicada em dois problemas de modelagem

distintos. No primeiro caso, o foco foi o reconhecimento de padrões em sequências binárias, consistindo em um exemplo de natureza mais didática e de avaliação preliminar dos métodos. No segundo caso, o método foi aplicado em um cenário do mundo real, especificamente na anotação de domínios funcionais em proteínas, com resultados promissores.

No primeiro caso, construímos a redes de meta-modelagem de padrões em sequências contendo 9 símbolos binários. Nessa rede, os vértices representam conjuntos de dados em A contendo sequências de 9 símbolos binários e modelos respectivos em E que explicam os padrões encontrados nessas sequências. As arestas da rede indicam o número de sequências em cada conjunto que se encaixam nas definições de padrões indicados por cada modelo. Essa rede mostrou uma grande inespecificidade dos seis modelos avaliados. Essa inespecificidade foi utilizada para substituir um modelo usado para explicar padrões com "quadrados pretos em pelo menos uma das posições 2, 4, 6 ou 8" por uma composição lógica dos outros cinco modelos disponíveis, além de um modelo adicional. A partir desse exemplo prático, observou-se o potencial da rede de meta-modelagem de aplicação na extração de informações de conjuntos de dados de sequências de símbolos e de modelos de detecção de padrões. Esses resultados são especialmente relevante para auxiliar modelos complexos de análise e detecção de padrões em sequências utilizados em bioinformática, como a anotação de domínios proteicos em sequências de aminoácidos.

No segundo caso, o método foi aplicado para construir uma rede de meta-modelagem de anotação de domínios proteicos em enzimas ativas em carboidratos, utilizando as informações contidas no banco de dados CAZy. Na rede, os vértices representam modelos e conjuntos de enzimas respectivos a seis classes funcionais (GT, PL, CE, GH, AA e CBM), e as arestas indicam enzimas que possuem domínios pertencentes a diferentes classes.

Os seis modelos identificados pela metodologia formam uma estrutura de modelagem com alta especificidade e baixa interação entre modelos. Isso revela que, no geral, domínios das seis classes funcionais se associam pouco em uma mesma proteína. A partir da análise de composição lógica de modelos, observou-se que a maioria dos modelos podem ser construídos satisfatoriamente por meio da negação lógica dos demais, uma consequência dos modelos serem praticamente mutualmente exclusivos, com apenas um modelo podendo ser comparado a uma combinação positiva da informação contida em outras modelos, o modelo GH.

Verificamos a baixa interação entre os conjuntos de dados das classes GT e PL e os modelos das outras classes, o que pode ser associada com a incompatibilidade funcional e de substratos entre os domínios GT e PL e os domínios de outras classes enzimáticas. Já os conjuntos de dados respectivos às classes CE, GH e AA se associam a módulos de outras classes para executar, de forma cooperativa, uma função no processamento de substratos. Isso é ressaltado pelo volume substancial de conexões entre esses conjuntos de dados e modelos alternativos, o que elevou a diversidade de conexões dos conjuntos dessas classes

a um nível próximo de 1,5. Enquanto isso, o conjunto de dados da classe CBM possui alta associação com modelos alternativos, com alta diversidade, o que está de acordo com a alta modularidade das enzimas dessa classe, que auxiliam a atividade de outros domínios de enzimas ativas em carboidratos.

Além disso, a rede foi usada como um instrumento de análise exploratória da anotação de famílias proteicas, revelando tendências evolutivas e relações funcionais entre as famílias que não seriam facilmente observados devido à grande quantidade de proteínas no banco de dados. A utilização dos dados de anotação do banco de dados CAZy possibilitou que muitas hipóteses levantadas a partir da rede pudessem ser validadas por meio de estudos científicos publicados. No entanto, foram levantadas algumas hipóteses que ainda devem ser exploradas.

Por exemplo, possíveis evoluções convergentes foram observadas entre diferentes domínios enzimáticos. Enzimas com domínios GT84/GH94 e GT2/GH17 podem ter evoluído de forma semelhante para a produção de beta-glucanos cíclicos, enquanto enzimas com domínios GT27/CBM13 e GT54/CBM94 evoluíram para desenvolver a atividade de transferase de aminoaçúcares. Além disso, existe uma possível proximidade evolutiva entre as famílias CBM70 e PL8_1, que atuam em hialuronatos, e sugere-se a existência de uma subfamília de PL1 ainda não descrita que se associa a domínios CE8 para o processamento de pectina. Também, é possível haver semelhanças nos mecanismos catalíticos de enzimas com domínios GH105/PL33_2 e GH88/PL38. A co-ocorrência de domínios AA5 e CE3 em uma mesma enzima pode estar relacionada à inatividade catalítica do módulo CE3. Por fim, há também uma possível associação sinergética entre domínios GH18 e domínios AA15 na degradação de quitina, e uma pressão seletiva para a associação de domínios da família AA10 com módulos CBM5 ou CBM73.

Assim sendo, acreditamos ter desenvolvido uma ferramenta promissora para a extração de informações a partir de anotação de proteínas.

4.2 Lista das Principais Contribuições

Lista das principais contribuições desta dissertação:

a) Extensão e aplicação de uma abordagem de meta-modelagem para analisar modelos baseados em dados, a rede de meta-modelagem $\mathcal{M}$. Em particular, foram propostos a construção de uma rede bipartida e, a partir dela, a quantificação de especificidade de modelos utilizando o índice de diversidade de conexões e o índice de coincidência entre modelos;

b) Aplicação da abordagem de meta-modelagem em um problema de reconhecimento de padrões em sequências binárias, o que permitiu a ilustração do potencial da metodologia. Foi possível descrever modelos de padrões a partir

de métricas de especificidade, permitindo, também, construir um novo modelo altamente específico a partir da combinação lógica de modelos inespecíficos, o que facilitou a interpretação dos padrões presentes nos banco de dados de sequências binárias. Este estudo também serviu como uma validação preliminar da abordagem;

c) Construção da rede de meta-modelagem para analisar uma anotação de domínios proteicos de enzimas ativas em carboidratos. A partir da rede, foi possível descrever modelos de classes de domínios funcionais a partir de métricas de especificidade e da comparação entre modelos e combinações lógicas de modelos alternativos. Isso permitiu quantificar o grau de associação entre domínios de diferentes classes, facilitando a identificação de diversas tendências evolutivas e relações funcionais entre as famílias de proteínas;

d) Mais especificamente, o estudo de famílias proteicas a partir da rede de meta-modelagem dos domínios funcionais de enzimas ativas em carboidratos permitiu o levantamento de algumas hipóteses ainda não completamente exploradas na literatura científica:

   – Possível evolução convergente de enzimas com domínios GT84 e GH94 e enzimas com domínios GT2 e GH17 para a produção de *beta*-glucanos cíclicos;
   – Possível evolução convergente de enzimas com domínios GT27 e CBM13 e enzimas com domínios GT54 e CBM94 para realização de atividade de transferase de aminoaçúcares;
   – Possível proximidade evolutiva entre as famílias CBM70 e PL8_1, que atuam sobre hialuronatos;
   – Possível existência de uma subfamília de PL1 ainda não descrita que se associa a domínios CE8 para o processamento de pectina;
   – Possível analogia entre os mecanismos catalíticos de enzimas com domínios GH105 e PL33_2 e enzimas com domínios GH88 e PL38 na degradação de substratos polissacarídicos;
   – Possivelmente, a co-ocorrência de domínios AA5 e CE3 em uma mesma enzima depende da inatividade catalítica do módulo CE3;
   – Possível associação sinergética de domínios GH18 e domínios AA15 na degradação de quitina;
   – Possível pressão seletiva para associação entre domínios da família AA10 e módulos CBM5 ou CBM73 para a degradação de quitina.

4.3 Futuros Desenvolvimentos

Espera-se que a metodologia descrita para a construção de rede de meta-modelagem de modelos baseadas em dados possa ser utilizada para auxiliar a caracterização e me-

lhoria da modelagem em múltiplas áreas científicas. Particularmente, a rede poderia ser amplamente utilizada para avaliar a especificidade de algoritmos de reconhecimento de padrões e *clustering*. Além disso, trabalhos futuros sobre a abordagem do meta-modelo $\mathcal{M}^*$ podem explorar a detecção de *outliers* em conjuntos de dados.

Por ora, duas principais possibilidades de trabalhos futuros podem ser diretamente delineados a partir dos conceitos e resultados apresentados neste trabalho. A primeira possibilidade é aplicar a rede de meta-modelagem em outros problemas dentro do campo das ciências biológicas, como análise de padrões em sequências de nucleotídeos e classificação de estruturas de proteínas, tipos de ligantes e classes de fármacos. A segunda possibilidade é aprofundar os métodos de análise das redes de meta-modelagem na anotação de domínios proteicos, incluindo a incorporação de medidas topológicas mais informativas para avaliar a relação entre famílias proteicas, a integração a rede de meta-modelagem com redes de coocorrência de domínios proteicos, e a comparação de redes de anotação de proteínas de diferentes espécies e classes de domínios funcionais.

REFERÊNCIAS

1 COSTA, L. da F. *An ample approach to data and modeling*. 2021. Disponível em: https://arxiv.org/abs/2110.01776. Acesso em: 29 June 2023.

2 FRIGG, R.; HARTMANN, S. Models in science. *In*: ZALTA, E. N. (ed.). *The Stanford Encyclopedia of Philosophy*. Stanford: The Metaphysics Research Lab, 2020.

3 MONTÁNS, F. J. *et al.* Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*, Elsevier, v. 347, n. 11, p. 845–855, Nov. 2019.

4 PEIERLS, R. Model-making in physics. *Contemporary Physics*, Informa UK Limited, v. 21, n. 1, p. 3–17, Jan. 1980.

5 HINCHLIFFE, A. (ed.). *Chemical modelling*. Cambridge: Royal Society of Chemistry, 2008.

6 ALLMAN, E. S.; RHODES, J. A. *Mathematical models in biology*. Cambridge: Cambridge University Press, 2012.

7 FRANTZ, F. K. A taxonomy of model abstraction techniques. *In*: CONFERENCE ON WINTER SIMULATION, 27., 1995, Arlington. *Proceedings [...]*. New York: IEEE Computer Society, 1995. p. 1413–1420. ISBN 0780330188.

8 BRADLEY, W. *et al.* Perspectives on the integration between first-principles and data-driven modeling. *Computers & Chemical Engineering*, Elsevier, v. 166, n. 107898, p. 107898, Oct. 2022.

9 BRUNTON, S. L.; KUTZ, J. N. *Data-driven science and engineering*: machine learning, dynamical systems, and control. 2nd ed. Cambridge: Cambridge University Press, 2022.

10 AL-JARRAH, O. Y. *et al.* Efficient machine learning for big data: a review. *Big Data Research*, Elsevier, v. 2, n. 3, p. 87–93, Sept. 2015.

11 CAO, L. Data science: a comprehensive overview. *ACM Computing Surveys*, Association for Computing Machinery, v. 50, n. 3, p. 1–42, May 2018.

12 SAGIROGLU, S.; SINANC, D. Big data: a review. *In*: INTERNATIONAL CONFERENCE ON COLLABORATION TECHNOLOGIES AND SYSTEMS, 2013, San Diego. *Proceedings [...]*. Danvers: IEEE Computer Society, 2013. p. 42–47. ISBN 978-1-4673-6404-1.

13 CHEN, L. *et al.* Review of the application of big data and artificial intelligence in geology. *Journal of Physics: conference series*, IOP Publishing, v. 1684, n. 1, p. 012007, Nov. 2020.

14 CHIANG, L.; LU, B.; CASTILLO, I. Big data analytics in chemical engineering. *Annual Review of Chemical and Biomolecular Engineering*, Annual Reviews, v. 8, n. 1, p. 63–85, June 2017.

15 BELLE, A. *et al.* Big data analytics in healthcare. *BioMed Research International*, Hindawi Limited, v. 2015, p. 370194, July 2015.

16 KELLING, S. *et al.* Data-intensive science: a new paradigm for biodiversity studies. *Bioscience*, Oxford University Press, v. 59, n. 7, p. 613–620, July 2009.

17 PERAKAKIS, N. *et al.* Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism*, Elsevier, v. 87, p. A1–A9, Oct. 2018.

18 WAMBA, S. F. *et al.* How big data can make big impact: findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, Elsevier, v. 165, p. 234–246, July 2015.

19 SOLOMATINE, D.; SEE, L. M.; ABRAHART, R. J. Data-driven modelling: concepts, approaches and experiences. *In*: ABRAHART, R. J.; SEE, L. M.; SOLOMATINE, D. P. (ed.). *Practical hydroinformatics*. Berlin: Springer, 2008. p. 17–30.

20 BELETE, G. F.; VOINOV, A.; LANIAK, G. F. An overview of the model integration process: from pre-integration assessment to testing. *Environmental Modelling & Software*, Elsevier, v. 87, p. 49–63, Jan. 2017.

21 BARABASI, A. *Network science*. Cambridge: Cambridge University Press, 2016.

22 NEWMAN, M. *Networks*. Oxford: Oxford University Press, 2018.

23 HARRIS, J. M.; HIRST, J. L.; MOSSINGHOFF, M. J. *Combinatorics and graph theory*. 2nd ed. New York: Springer, 2008.

24 ZACHARY, W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, University of Chicago Press, v. 33, n. 4, p. 452–473, 1977.

25 COSTA, L. da F. *et al.* Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, Taylor & Francis, v. 60, n. 3, p. 329–412, June 2011.

26 FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. *Computer Communication Review*, Association for Computing Machinery (ACM), v. 29, n. 4, p. 251–262, Oct. 1999.

27 JORDÁN, F.; LIU, W.-C.; DAVIS, A. J. Topological keystone species: measures of positional importance in food webs. *Oikos*, Wiley, v. 112, n. 3, p. 535–546, Mar. 2006.

28 ATKINSON, H. J. *et al.* Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLOS One*, Public Library of Science, v. 4, n. 2, p. e4345, Feb. 2009.

29 JACKSON, M. O. An overview of social networks and economic applications. *In*: BENHABIB, J.; BISIN, A.; JACKSON, M. O. (ed.). *Handbook of social economics*. Amsterdam: North-Holland, 2011. v. 1, p. 511–585.

30 NEWMAN, M. E. J. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Physical Review E*, American Physical Society, v. 64, n. 1, p. 016132, June 2001.

31 BAXEVANIS, A. D.; BADER, G. D.; WISHART, D. S. (ed.). *Bioinformatics*. 4th ed. Nashville: John Wiley & Sons, 2020.

32 CUMMINS, C. *et al.* The European nucleotide archive in 2021. *Nucleic Acids Research*, Oxford University Press, v. 50, n. D1, p. D106–D110, Nov. 2021.

33 SAYERS, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research*, Oxford University Press, v. 50, n. D1, p. D20–D26, Jan. 2022.

34 CONSORTIUM, T. U. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, Oxford University Press, v. 51, n. D1, p. D523–D531, Nov. 2022.

35 DURBIN, R. *et al. Biological sequence analysis*. Cambridge: Cambridge University Press, 1998.

36 VOET, D.; VOET, J. G. *Biochemistry*. 4th ed. Chichester: John Wiley & Sons, 2010.

37 EDDY, S. R. Profile hidden Markov models. *Bioinformatics*, Oxford University Press, v. 14, n. 9, p. 755–763, Jan. 1998.

38 PAYSAN-LAFOSSE, T. *et al.* Interpro in 2022. *Nucleic Acids Research*, Oxford University Press, v. 51, n. D1, p. D418–D427, Nov. 2022.

39 SAVELLI, B. *et al.* Redoxibase: a database for ros homeostasis regulated proteins. *Redox Biology*, Elsevier, v. 26, p. 101247, Sept. 2019.

40 FAWAL, N. *et al.* Peroxibase: a database for large-scale evolutionary analysis of peroxidases. *Nucleic Acids Research*, Oxford University Press, v. 41, n. D1, p. D441–D444, Nov. 2012.

41 RAWLINGS, N. D.; BARRETT, A. J.; BATEMAN, A. Merops: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, Oxford University Press, v. 40, n. D1, p. D343–D350, Nov. 2011.

42 CANTU, D. C. *et al.* Thyme: a database for thioester-active enzymes. *Nucleic Acids Research*, Oxford University Press, v. 39, n. Suppl. 1, p. D342–D346, Nov. 2010.

43 DRULA, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research*, Oxford University Press, v. 50, n. D1, p. D571–D577, Jan. 2022.

44 CARBOHYDRATE Active Enzymes Database. 2023. Disponível em: http://www.cazy.org/. Acesso em: 03 Jul. 2023.

45 WUCHTY, S.; ALMAAS, E. Evolutionary cores of domain co-occurrence networks. *BMC Evolutionary Biology*, Springer Nature, v. 5, n. 1, p. 24, Mar. 2005.

46 WANG, Z. *et al.* A protein domain co-occurrence network approach for predicting protein function and inferring species phylogeny. *PLOS One*, Public Library of Science, v. 6, n. 3, p. e17906, Mar. 2011.

47 PAVLOPOULOS, G. A. *et al.* Bipartite graphs in systems biology and medicine: a survey of methods and applications. *Gigascience*, Oxford University Press, v. 7, n. 4, p. giy014, Apr. 2018.

48 LEVITZ, K.; LEVITZ, H. *Logic and boolean algebra*. New York: Barron's Educational Series, 1979.

49 GREGG, J. *Ones and Zeros*: understanding boolean algebra, digital circuits, and the logic of sets. New York: Wiley-IEEE Press, 1998.

50 JOST, L. Entropy and diversity. *Oikos*, Wiley, v. 113, n. 2, p. 363–375, May 2006.

51 LEINSTER, T.; COBBOLD, C. A. Measuring diversity: the importance of species similarity. *Ecology*, Wiley, v. 93, n. 3, p. 477–489, Mar. 2012.

52 VIANA, M. P.; BATISTA, J. L. B.; COSTA, L. da F. Effective number of accessed nodes in complex networks. *Physical Review E*, American Physical Society, v. 85, p. 036105, Mar. 2012.

53 COSTA, L. da F. On similarity. *Physica A*, Elsevier, v. 599, p. 127456, Apr. 2022.

54 COSTA, L. da F. Multiset neurons. *Physica A*, Elsevier, v. 609, p. 128318, Jan. 2023.

55 JACCARD, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, Société Vaudoise des Sciences Naturelles, v. 37, p. 547–579, 1901.

56 RAMOS-GUAJARDO, A. B.; GONZÁLEZ-RODRÍGUEZ, G.; COLUBI, A. Testing the degree of overlap for the expected value of random intervals. *International Journal of Approximate Reasoning*, Elsevier, v. 119, p. 1–19, Apr. 2020.

57 BRUSCO, M.; CRADIT, J. D.; STEINLEY, D. A comparison of 71 binary similarity coefficients: the effect of base rates. *PLOS One*, Public Library of Science, v. 16, n. 4, p. 1–19, Apr. 2021.

58 SZYMKIEWICZ, D. Une contribution statistique à la géographie floristique. *Acta Societatis Botanicorum Poloniae*, Polish Botanical Society, v. 11, n. 3, p. 249–265, 1934.

59 VIJAYMEENA, M. K.; KAVITHA, K. A survey on similarity measures in text mining. *Machine Learning and Applications: an international journal*, Academy and Industry Research Collaboration Center (AIRCC), v. 3, n. 1, p. 19–28, Mar. 2016.

60 ZHANG, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, Oxford University Press, v. 46, n. W1, p. W95–W101, July 2018.

61 CANTAREL, B. L. *et al.* The Carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research*, Oxford University Press, v. 37, p. D233–8, Jan. 2009.

62 STAM, M. R. *et al.* Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Engineering, Design and Selection*, Oxford University Press, v. 19, n. 12, p. 555–562, Dec. 2006.

63 LOMBARD, V. *et al.* A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochemical Journal*, Portland Press Ltd., v. 432, n. 3, p. 437–444, Dec. 2010.

64 GARRON, M.; HENRISSAT, B. The continuing expansion of CAZymes and their families. *Current Opinion in Chemical Biology*, Elsevier, v. 53, p. 82–87, Dec. 2019.

65 LEVASSEUR, A. *et al.* Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels*, Springer Nature, v. 6, n. 1, p. 41, Mar. 2013.

66 GARRON, M.; CYGLER, M. Structural and mechanistic classification of uronic acid-containing polysaccharide lyases. *Glycobiology*, Oxford University Press, v. 20, n. 12, p. 1547–1573, Dec. 2010.

67 ARMENDÁRIZ-RUIZ, M. *et al.* Carbohydrate esterases: an overview. *In*: SANDOVAL, G. (ed.). *Lipases and phospholipases*: methods and protocols. New York: Springer, 2018. p. 39–68. ISBN 978-1-4939-8672-9.

68 GUILLÉN, D.; SÁNCHEZ, S.; RODRÍGUEZ-SANOJA, R. Carbohydrate-binding domains: multiplicity of biological roles. *Applied Microbiology and Biotechnology*, Springer Science and Business Media LLC, v. 85, n. 5, p. 1241–1249, Feb. 2010.

69 LAIRSON, L. L. *et al.* Glycosyltransferases: structures, functions, and mechanisms. *Annual Review of Biochemistry*, Annual Reviews, v. 77, n. 1, p. 521–555, Apr. 2008.

70 DAVIES, G.; HENRISSAT, B. Structures and mechanisms of glycosyl hydrolases. *Structure*, Elsevier, v. 3, n. 9, p. 853–859, Sept. 1995.

71 NAKAI, H. *et al.* Recent development of phosphorylases possessing large potential for oligosaccharide synthesis. *Current Opinion in Chemical Biology*, Elsevier, v. 17, n. 2, p. 301–309, Apr. 2013.

72 GLOSTER, T. M. *et al.* Divergence of catalytic mechanism within a glycosidase family provides insight into evolution of carbohydrate metabolism by human gut flora. *Chemistry & Biology*, v. 15, n. 10, p. 1058–1067, Oct. 2008.

73 LITTLE, D. J. *et al.* PgaB orthologues contain a glycoside hydrolase domain that cleaves deacetylated poly-β(1,6)-N-acetylglucosamine and can disrupt bacterial biofilms. *PLOS Pathogens*, Public Library of Science, v. 14, n. 4, p. e1006998, Apr. 2018.

74 NAKAMURA, A. M.; NASCIMENTO, A. S.; POLIKARPOV, I. Structural diversity of carbohydrate esterases. *Biotechnology Research and Innovation*, Editora Cubo, v. 1, n. 1, p. 35–51, Jan. 2017.

75 DISSEL, D. van *et al.* Production of poly-β-1,6-N-acetylglucosamine by MatAB is required for hyphal aggregation and hydrophilic surface adhesion by streptomyces. *Microbial Cell*, Shared Science Publishers OG, v. 5, n. 6, p. 269–279, June 2018.

76 ARAGUNDE, H.; BIARNÉS, X.; PLANAS, A. Substrate recognition and specificity of chitin deacetylases and related family 4 carbohydrate esterases. *International Journal of Molecular Sciences*, MDPI AG, v. 19, n. 2, p. 412, Jan. 2018.

77 HEHEMANN, J. *et al.* Aquatic adaptation of a laterally acquired pectin degradation pathway in marine gammaproteobacteria. *Environmental Microbiology*, John Wiley & Sons, v. 19, n. 6, p. 2320–2333, June 2017.

78 MAZURKEWICH, S.; SEVESO, A.; LARSBRINK, J. A unique AA5 alcohol oxidase fused with a catalytically inactive CE3 domain from the bacterium burkholderia pseudomallei. *FEBS Letters*, John Wiley & Sons, v. 597, n. 13, p. 1779–1791, May 2023.

79 CIOCCHINI, A. E. *et al.* Identification of active site residues of the inverting glycosyltransferase cgs required for the synthesis of cyclic beta-1,2-glucan, a brucella abortus virulence factor. *Glycobiology*, Oxford University Press, v. 16, n. 7, p. 679–691, July 2006.

80 GUIDOLIN, L. S. *et al.* Functional mapping of brucella abortus cyclic β-1,2-glucan synthase: Identification of the protein domain required for cyclization. *Journal of Bacteriology*, American Society for Microbiology, v. 191, n. 4, p. 1230–1238, Feb. 2009.

81 STANISICH, V. A.; STONE, B. A. Enzymology and molecular genetics of biosynthetic enzymes for (1,3)-β-glucans: prokaryotes. *In*: BACIC, A.; FINCHER, G. B.; STONE, B. A. (ed.). *Chemistry, biochemistry, and biology of 1-3 beta glucans and related polysaccharides*. San Diego: Elsevier, 2009. p. 201–232. ISBN 978-0-12-373971-1.

82 HREGGVIDSSON, G. O. *et al.* Exploring novel non-leloir β-glucosyltransferases from proteobacteria for modifying linear (β1->3)-linked gluco-oligosaccharide chains. *Glycobiology*, Oxford University Press, v. 21, n. 3, p. 304–328, Mar. 2011.

83 DOBRUCHOWSKA, J. M. *et al.* Modification of linear (β1→3)-linked gluco-oligosaccharides with a novel recombinant β-glucosyltransferase (trans-β-glucosidase) enzyme frombradyrhizobium diazoefficiens. *Glycobiology*, Oxford University Press, v. 26, n. Nov., p. 1157–1170, Nov. 2016.

84 NAKAJIMA, M. β-1,2-Glucans and associated enzymes. *Biologia*, Springer Science and Business Media LLC, v. 78, n. 7, p. 1741–1757, Oct. 2022.

85 GERMANE, K. L. *et al.* Structural analysis of clostridium acetobutylicum ATCC 824 glycoside hydrolase from CAZy family GH105. *Acta Crystallographica Section F*, International Union of Crystallography (IUCr), v. 71, n. 8, p. 1100–1108, Aug. 2015.

86 HELBERT, W. *et al.* Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 116, n. 13, p. 6063–6068, Mar. 2019.

87 KIKUCHI, M. *et al.* A bacterial endo-β-1,4-glucuronan lyase, CUL-I from brevundimonas sp. SH203, belonging to a novel polysaccharide lyase family. *Protein Expression and Purification*, Elsevier, v. 166, n. 105502, p. 105502, Feb. 2020.

88 COURTADE, G.; AACHMANN, F. L. Chitin-active lytic polysaccharide monooxygenases. *In*: YANG, Q.; FUKAMIZO, T. (ed.). *Targeting chitin-containing organisms*. Singapore: Springer, 2019. p. 115–129. ISBN 978-981-13-7318-3.

89 MEKASHA, S. *et al.* A trimodular bacterial enzyme combining hydrolytic activity with oxidative glycosidic bond cleavage efficiently degrades chitin. *Journal of Biological Chemistry*, American Society for Biochemistry and Molecular Biology, v. 295, n. 27, p. 9134–9146, July 2020.

90 CORRÊA, T. L. R.; SANTOS, L. V. dos; PEREIRA, G. A. G. AA9 and AA10: from enigmatic to essential enzymes. *Applied Microbiology and Biotechnology*, Springer Science and Business Media LLC, v. 100, n. 1, p. 9–16, Jan. 2016.

91 FORSBERG, Z.; COURTADE, G. On the impact of carbohydrate-binding modules (CBMs) in lytic polysaccharide monooxygenases (LPMOs). *Essays in Biochemistry*, Portland Press Ltd., v. 67, n. 3, p. 561–574, Apr. 2023.

92 SEDZICKI, J. *et al.* Structure-function analysis of the cyclic β-1,2-glucan synthase. 2023. Disponível em: https://doi.org/10.1101/2023.05.05.539553. Acesso em: 24 June 2023.

93 COUTINHO, P. M. *et al.* An evolving hierarchical family classification for glycosyltransferases. *Journal of Molecular Biology*, Elsevier, v. 328, n. 2, p. 307–317, Apr. 2003.

94 MARTINEZ-FLEITES, C. *et al.* Insights into the synthesis of lipopolysaccharide and antibiotics through the structures of two retaining glycosyltransferases from family GT4. *Chemistry & Biology*, Elsevier, v. 13, n. 11, p. 1143–1152, Nov. 2006.

95 PRABHAKAR, P. K. *et al.* Structural and biochemical insight into a modular β-1,4-galactan synthase in plants. *Nature Plants*, Springer Science and Business Media LLC, v. 9, n. 3, p. 486–500, Mar. 2023.

96 BROCKHAUSEN, I. Crossroads between bacterial and mammalian glycosyltransferases. *Frontiers in Immunology*, Frontiers Media SA, v. 5, p. 492, Oct. 2014.

97 SAYERS, E. W. *et al.* Genbank. *Nucleic Acids Research*, Oxford University Press, v. 48, n. D1, p. D84–D86, Jan. 2020.

98 SUITS, M. D. L. *et al.* Conformational analysis of the streptococcus pneumoniae hyaluronate lyase and characterization of its hyaluronan-specific carbohydrate-binding module. *Journal of Biological Chemistry*, Elsevier, v. 289, n. 39, p. 27264–27277, Sept. 2014.

99 ZHENG, L. *et al.* Pectinolytic lyases: a comprehensive review of sources, category, property, structure, and catalytic mechanism of pectate lyases and pectin lyases. *Bioresources and Bioprocessing*, Springer Science and Business Media LLC, v. 8, n. 1, p. 79, Dec. 2021.

100 KOSCHORRECK, K.; ALPDAGTAS, S.; URLACHER, V. B. Copper-radical oxidases: a diverse group of biocatalysts with distinct properties and a broad range of biotechnological applications. *Engineering Microbiology*, Elsevier, v. 2, n. 3, p. 100037, Sept. 2022.

101 JANEČEK, Š. *et al.* Starch-binding domains as CBM families-history, occurrence, structure, function and evolution. *Biotechnology Advances*, Elsevier, v. 37, n. 8, p. 107451, Dec. 2019.

102 MØLLER, M. S.; HENRIKSEN, A.; SVENSSON, B. Structure and function of α-glucan debranching enzymes. *Cellular and Molecular Life Sciences*, Springer Science and Business Media LLC, v. 73, n. 14, p. 2619–2641, July 2016.

103 MACHOVIČ, M.; JANEČEK, Š. Domain evolution in the GH13 pullulanase subfamily with focus on the carbohydrate-binding module family 48. *Biologia*, Springer Science and Business Media LLC, v. 63, n. 6, p. 1057–1068, Dec. 2008.

104 CHEN, W.; JIANG, X.; YANG, Q. Glycoside hydrolase family 18 chitinases: the known and the unknown. *Biotechnology Advances*, Elsevier, v. 43, n. 107553, p. 107553, Nov. 2020.

105 VANDHANA, T. M. *et al.* On the expansion of biological functions of lytic polysaccharide monooxygenases. *New Phytologist*, Wiley, v. 233, n. 6, p. 2380–2396, Mar. 2022.

106 ARMENTA, S. *et al.* Advances in molecular engineering of carbohydrate-binding modules. *Proteins*, Wiley, v. 85, n. 9, p. 1602–1617, Sept. 2017.

107 VISWESWARAN, G. R. R. *et al.* AcmD, a homolog of the major autolysin AcmA of lactococcus lactis, binds to the cell wall and contributes to cell separation and autolysis. *PLOS One*, Public Library of Science, v. 8, n. 8, p. e72167, Aug. 2013.

108 AKCAPINAR, G. B. *et al.* Molecular diversity of LysM carbohydrate-binding motifs in fungi. *Current Genetics*, Springer Science and Business Media LLC, v. 61, n. 2, p. 103–113, May 2015.

109 JANEČEK, Š.; SVENSSON, B.; MACGREGOR, E. A. Structural and evolutionary aspects of two families of non-catalytic domains present in starch and glycogen binding proteins from microbes, plants and animals. *Enzyme and Microbial Technology*, Elsevier, v. 49, n. 5, p. 429–440, Oct. 2011.