MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Programa de Pós-Graduação em Biologia Computacional e Sistemas

ESTUDO *IN SILICO* DA DIVERSIDADE ENTRE ENZIMAS RESPONSÁVEIS PELA RESISTÊNCIA BACTERIANA A DIFERENTES CLASSES DE ANTIMICROBIANOS COM FOCO PRINCIPAL EM BETA-LACTAMASES

MELISE CHAVES SILVEIRA

Rio de Janeiro
Maio de 2018

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

Melise Chaves Silveira

Estudo *in silico* da diversidade entre enzimas responsáveis pela resistência bacteriana a diferentes classes de antimicrobianos com foco principal em beta-lactamases

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Computacional e Sistemas

Orientador: Prof. Dr. Antônio Basílio de Miranda

RIO DE JANEIRO
Maio de 2018

Chaves Silveira, Melise .

Estudo *in silico* da diversidade entre enzimas responsáveis pela resistência a diferentes classes de antimicrobianos em bactérias / Melise Chaves Silveira. - Rio de janeiro, 2018.

172 f.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2018.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/ICICT com os dados fornecidos pelo(a) autor(a).

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: MELISE CHAVES SILVEIRA

ESTUDO IN SILICO DA DIVERSIDADE ENTRE ENZIMAS RESPONSÁVEIS PELA RESISTÊNCIA BACTERIANA A DIFERENTES CLASSES DE ANTIMICROBIANOS COM FOCO PRINCIPAL EM BETA-LACTAMASES

ORIENTADOR: Prof. Dr. Antônio Basílio de Miranda

Aprovada em: 04/05/2018

EXAMINADORES:

Prof. Dra. Ana Paula Carvalho-Assef - Presidente (Instituto Oswaldo Cruz)
Prof. Dra. Marisa Nicolás (Laboratório Nacional de Computação Científica)
Prof. Dr. DiogoTschoeke (Universidade Federal do Rio de Janeiro)
Prof. Dr. Philip Suffys -suplente (Instituto Oswaldo Cruz)
Prof. Dra. Ana Botelho -suplente (Universidade Federal do Rio de Janeiro)

Rio de Janeiro, 4 de maio de 2018

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

AGRADECIMENTOS

Em primeiro lugar agradeço à Deus por sempre guiar meus caminhos e ter me concedido as pessoas e oportunidades necessárias para que eu concluísse mais essa etapa.

Agradeço aos meus pais, por respeitarem e sustentarem meus sonhos e minhas escolhas, pela certeza de sempre contar com o amor e carinho deles.

À minha irmã, por ser minha melhor amiga e por me inspirar tanto.

Ao Rafa por trazer mais luz e amor nessa caminhada, além do companheirismo.

Aos meus amigos, de perto e de longe. Saber que posso contar com a torcida, o carinho e o ouvido de vocês é fundamental.

Ao meu orientador, Dr. Antônio Basílio, por ter me aceitado como aluna, pelos desafios e crescimento proporcionados, e por toda compreensão e apoio, principalmente na reta final.

Aos meus amigos do Laboratório de Biologia Computacional e Sistemas, Rangeline, Letícia e André, por terem sido boas companhias nesses quatro anos, pelas conversas, ajuda e trocas de experiência. Aprendi muito com vocês!

À todos os membros do Laboratório de Biologia Computacional e Sistemas, principalmente ao Dr. Rodrigo Jardim e ao Dr. Fábio Mota pela ajuda e conselhos.

À Pós-graduação em Biologia Computacional e Sistemas do Instituto Oswaldo Cruz, por buscar sempre o aprimoramento do programa e pelo suporte aos alunos.

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

À todos meu sincero muito obrigada!

“Acredite em si mesmo e chegará um dia em que os outros não terão outra escolha senão acreditar em você”

Cynthia Kersey

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

ESTUDO *IN SILICO* DA DIVERSIDADE ENTRE ENZIMAS RESPONSÁVEIS PELA RESISTÊNCIA BACTERIANA A DIFERENTES CLASSES DE ANTIMICROBIANOS COM FOCO PRINCIPAL EM BETA-LACTAMASES

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Melise Chaves Silveira

Bactérias podem se tornar resistentes à ação dos antibióticos por diferentes mecanismos, mas a produção de enzimas inativadoras é considerada o mecanismo mais eficiente e capaz de se disseminar com maior facilidade. Dentre essas enzimas, as beta-lactamases se destacam pelo seu impacto clínico, diversidade e distribuição entre espécies e ambientes distintos. As beta-lactamases foram utilizadas como modelo no presente estudo, para a concepção de uma metodologia baseada em similaridade de sequências capaz de identificá-las e classificá-las em níveis hierárquicos. Os critérios classificatórios foram então aplicados a outras atividades enzimáticas relacionadas com a resistência aos antibióticos. Inicialmente, um conjunto curado de beta-lactamases foi utilizado para determinar os *thresholds* de agrupamento, que em seguida foram validados utilizando um conjunto expressivo e não-redundante de sequências. Perfis de Modelos Ocultos de Markov (Hidden *Markov Models*, HMM) foram construídos para cada classe de beta-lactamase. Estes foram testados e aprimorados a partir de informações de bancos de dados como CATH e UniProt/Swiss-Prot. São descritos aqui cinco níveis classificatórios hierárquicos baseados na estrutura das beta-lactamases. Os perfis HMM apresentados são capazes de identificar as classes (terceiro nível) com alta especificidade. A classe SCD, composta de beta-lactamases bifuncionais, é proposta nesta tese. Todas as subclasses de beta-lactamases foram encontradas em *Proteobacteria*. A subclasse SD1 e a classe ME foram os grupos de beta-lactamases mais distribuídos entre os diferentes filos analisados. Através da análise dos demais grupos enzimáticos inativadores de antibióticos, concluímos que as beta-lactamases são as mais diversas. Esse trabalho permitiu analisar a diversidade de diferentes atividades enzimáticas relacionadas com a resistência sobre uma perspectiva comum, além de mostrar que a metodologia para uma anotação aprimorada de beta-lactamases conforme apresentada aqui pode contribuir na elaboração de estudos sobre evolução, dispersão e prevalência dessa atividade enzimática crítica.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

IN SILICO STUDY OF DIVERSITY BETWEEN ENZYMES RESPONSIBLE FOR BACTERIAL RESISTANCE TO DIFFERENT CLASSES OF ANTIMICROBIALS WITH MAIN FOCUS ON BETA-LACTMASES

ABSTRACT

PHD THESIS IN COMPUTATIONAL BIOLOGY AND SYSTEMS

Melise Chaves Silveira

Bacteria may become resistant to the action of antibiotics by different mechanisms, but the production of inactivating enzymes is considered the most efficient mechanism, with great dispersal. Among these enzymes, beta-lactamases stand out for their clinical impact, diversity and distribution between species and different environments. Beta-lactamases were used as a model in the present study, to design a methodology based on sequence similarity capable of identifying and classifying them in hierarchical levels. Classification criteria were then applied to other enzymatic activities related to antibiotic resistance. Initially, a cured set of beta-lactamases was used to determine the clustering thresholds, which were then validated using an expressive and non-redundant set of sequences. Profiles Hidden Markov Models (HMM) were constructed for each class of beta-lactamase. These have been tested and calibrated from database informations such as CATH and UniProt/Swiss-Prot. Five hierarchical classificatory levels based on the structure of beta-lactamases are described here. The presented profiles HMM can identify the classes (third level) with high specificity. The class SCD, composed of bifunctional beta-lactamases, is proposed in this thesis. All subclasses of beta-lactamases were found in *Proteobacteria*. The subclass SD1 and class ME were the groups of beta-lactamases most distributed among the different phyla analyzed. By analyzing the other enzymatic inactivating groups of antibiotics, we conclude that beta-lactamases are the more diverse group. This work allowed the analysis of the diversity of different enzymatic activities related to resistance from a common perspective, besides demonstrating that the methodology for an improved annotation of beta-lactamases presented here can contribute to the elaboration of studies on the evolution, dispersion and prevalence of this critical enzymatic activity.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AAC Aminoglicosídeo acetiltransferase

AAD Aminoglicosídeo adeniltransferase

ADP Adenosina Difosfato

AIM Adelaide Imipenemase

AmpC Ampicilinase

ANT Aminoglicosídeo nucleotidiltransferase

APH Aminoglicosídeo fosfotransferase

ARR *Rifampicin ADP-ribosylating transferase* (ADP transferase de rifampicina)

ATP Adenosina Trifosfato

BL Beta-lactamase

BlaR1 *Beta-lactamase regulatory protein* (Proteína reguladora de Beta-lactamase)

BLAST *Basic Local Alignment Search Tool* (Ferramenta de pesquisa de alinhamento local básico)

BNRB Base de dados Não-Redundante de Beta-lactamases

C Carbono

CARD Comprehensive Antibiotic Resistance Database

CAT Cloranfenicol acetiltransferases

CATH *Protein Structure Classification Database* (Banco de dados de classificação de estruturas de proteínas)

CBMAR *Comprehensive Beta-lactamase Molecular Annotation Resource* (Recurso Abrangente de Anotação Beta-lactamase)

CDD *Conserved Domain Database* (Banco de dados de domínios conservados)

CM Comprimento mínimo de cobertura

DNA Ácido Desoxirribonucléico

DP Densidade de Pontuação BLAST

EC *Enzyme Classification* (Classificação de enzima)

Ere Esterase de aminoglicosídeos

EsC Especificidade de Classe

EsF Especificidade de Função

Fos *Fosfomycin resistance protein* (Proteína de Resistência à fosfomicina)

GA Gathering threshold

GO Gene Ontology (Ontologia de Gene)

HGT *Horizontal Gene Transfer* (Transferência horizontal de genes)

HMM *Hidden Markov Models* (Modelos ocultos de Markov)

Inc Grupo de Incompatibilidade

LacED *Lactamase Engineering Database* (Banco de dados aplicado às lactamases)

Lnu *Lincosamide nucleotidyltransferase* (Nucleotidiltransferase de linsosamidas)

LRA *Beta-lactam resistance from Alaskan soil* (Resistências à beta-lactâmicos no solo do Alasca)

MBL Metalo-Beta-Lactamases

MBLED *Metallo-Beta-Lactamase Engineering Database* (Banco de dados aplicado às metalo-beta-lactamases)

MBP *Maltose-binding Protein* (proteína ligadora de Maltose)

MecR1 *Methicillin resistance protein* (Proteína de resistência à meticilina)

MLS Macrolídeos-Lincosamidas-Streptograminas

Mph *Macrolide phosphotransferase* (Fosfotransferase de macrolídeos)

MRCA *Most Recent Common Ancestors* (Ancestral comum mais recente)

N Nitrogênio

NCBI *National Center for Biotechnology Information* (Centro nacional de informação biotecnológica)

NDM Nova Deli metalo-beta-lactamase

NISE *Non-homologous Isofunctional Enzymes* (Enzimas isofuncionais não-homólogas)

O Oxigênio

OXA Oxacilinase

Pab87 Proteína de *Pyrococcus abyssi*

PBP *Penicillin-binding Protein* (Proteína ligadora de penicilinas)

PDB *Protein Data Bank* (Banco de dados de proteínas)

Pfam *Protein Families Database* (Banco de dados de famílias de proteínas)

Rep Proteínas iniciadoras de replicação

RNA Ácido Ribonucléico

RNAr Ácido Ribonucléico Ribossômico

SBL Serina-Beta-Lactamase

SBM *Serratia marcescens* beta-lactamase

SeF Sensibilade de Função

SPM São Paulo metalo-beta-lactamase

SSN *Sequence Similarity Network* (Rede de similaridade de sequências)

Tet *Tetracycline resistance protein* (Proteína de resistência à tetraciclina)

TrEMBL *Translated EMBL Nucleotide Sequence Data Library* (Biblioteca Traduzida de Dados de Sequência de Nucleotídeos do EMBL)

UDP Uracila Difosfato

UniProt *Universal Protein Resource* (Fonte universal de proteínas)

Zn Zinco

1 INTRODUÇÃO

1.1 Evolução e resistência aos antibióticos

Antibióticos são umas das classes de drogas mais prescritas pelo mundo. Em 2010, mais de 70 bilhões de doses clínicas de antibióticos foram administradas em todo mundo, um aumento de 36% quando comparado ao ano 2000. Brasil, Rússia, Índia, China e África do Sul juntos são responsáveis por 76% desse aumento. O consumo dessas drogas é um dos principais fatores responsáveis pelo desenvolvimento da resistência, que por sua vez leva ao uso de medicamentos mais caros e aumenta a morbidade e mortalidade dos pacientes (Van Boeckel et al., 2014).

As bactérias podem se tornar resistentes à ação dos antibióticos por diferentes mecanismos (Figura 1.1). Entre eles estão a modificação química ou mutacional dos alvos celulares, a diminuição da permeabilidade da membrana externa, a ação de sistemas de efluxo que expulsam a droga do interior da célula, e a inativação enzimática da droga através da sua desintegração ou modificação química. Os mecanismos enzimáticos são considerados os mais eficientes e capazes de se disseminar com maior facilidade entre as bactérias (Savjani et al., 2009).

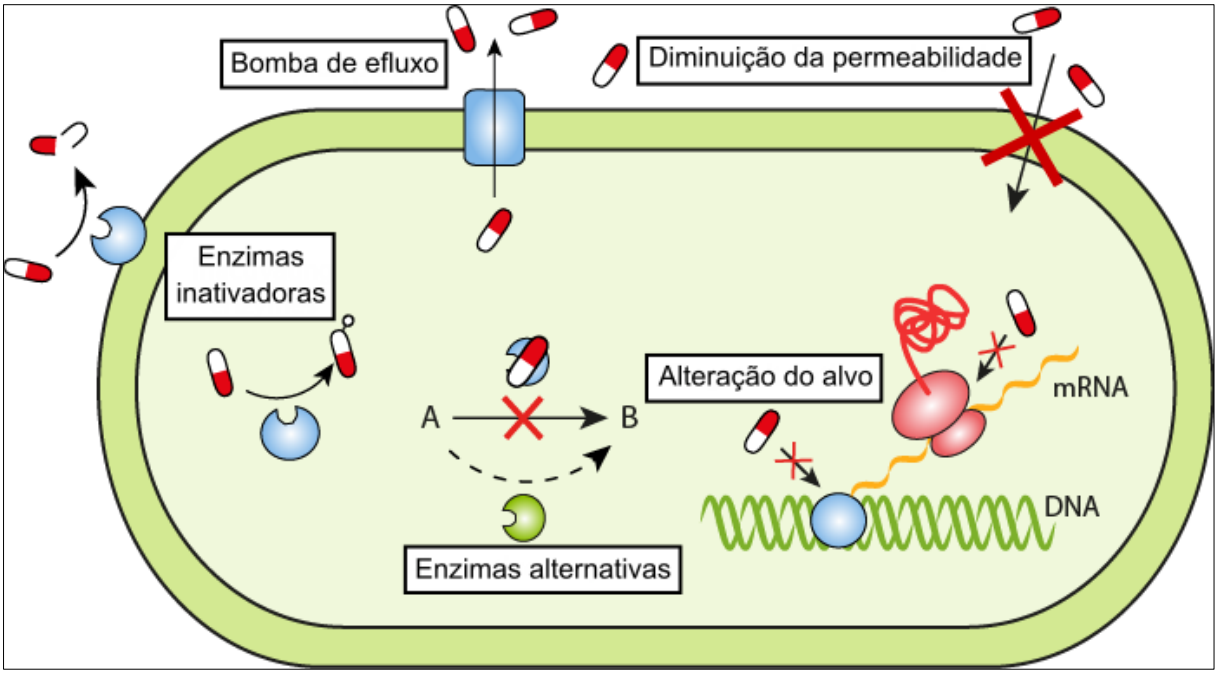


Figura 1.1: Principais mecanismos de resistência aos antibióticos em bactérias.

Fonte: Tradução de Erik Gullberg.

Enzimas são proteínas especializadas que catalisam reações nos sistemas biológicos (Nelson & Cox, 2005). Existe uma enorme flexibilidade de vias metabólicas essenciais na natureza, que incluem numerosas reações bioquímicas catalisadas por enzimas muito divergentes ou formas enzimáticas ainda não caracterizadas (Galperin and Koonin, 2000).

Dois eventos principais ocorrem no processo de evolução enzimática: a divergência e a convergência evolutiva. As proteínas em geral são organizadas em famílias com base na similaridade entre suas sequências, podendo ser combinadas em superfamílias quando apresentam motivos, atividades catalíticas e características estruturais parecidas. A divergência de sequências homólogas leva a diversificação funcional dentro da mesma superfamília de proteínas, enquanto a convergência funcional ocorre quando membros de diferentes superfamílias são recrutados para catalisar a mesma reação biológica (Galperin and Koonin, 2012).

Non-homologous Isofunctional Enzymes (NISE) é o termo que melhor descreve enzimas resultantes de convergência evolutiva que não possuem similaridade detectável entre suas estruturas primárias, mas catalisam a mesma reação biológica e por isso são identificadas com o mesmo *Enzyme*

Classification (EC) number (Omelchenko et al., 2010). Em vários casos, as NISE também não tem similaridade estrutural, sendo esse o indicador mais robusto de rotas evolucionárias diferentes para o cumprimento de uma mesma conversão metabólica (Omelchenko et al., 2010).

É possível notar que as NISE estão mais presentes entre as hidrolases, provavelmente porque o substrato dessas enzimas é uma pequena molécula universal (água) e a reação tipicamente não requer nenhuma coenzima. O dobramento do tipo *TIM-barrel* é significativamente superexpressado entre as NISE, o que pode estar vinculado a sua notável versatilidade bioquímica decorrente de sua simetria flexível (Omelchenko et al., 2010).

NISE são mais comuns em procariotos, principalmente em bactérias de vida livre (Omelchenko et al., 2010). Em relação ao tamanho dos genomas, microrganismos que possuem genomas pequenos tipicamente codificam uma única forma de qualquer enzima, enquanto aqueles com grandes tamanhos de genoma frequentemente carreiam NISE para certas etapas do metabolismo (Koonin et al., 1998).

Um bom exemplo de NISE são as beta-lactamases (BLs), importante causa da resistência aos antibióticos em bactérias (Gherardini et al., 2007). Essas enzimas clivam ligações similares através de diferentes mecanismos que podem atuar por meio de uma tríade catalítica, como as serino-proteases, ou por um mecanismo metal-dependente (Fabiane et al., 1998). Esses dois mecanismos realizam a quebra do anel beta-lactâmico de maneira tão eficiente que essas estratégias apareceram independentemente no curso da evolução (Gherardini et al., 2007).

A rápida evolução da resistência aos antibióticos em bactérias e sua distribuição nos diferentes ambientes é um conhecido problema de saúde mundial. O abandono, a utilização em doses subinibitórias e o uso excessivo de antibióticos geram uma incessante pressão seletiva sobre os genes relacionados com a resistência (Davies and Davies, 2010).

Bactérias possuem uma rápida capacidade de se adaptar ao estresse ambiental, e a evolução da resistência aos antibióticos é a evidência mais convincente disso. Populações bacterianas que evoluíram de forma independente e foram expostas à diferentes antibióticos mostraram variações

em suas sequências genômicas cujas consequências funcionais foram idênticas ou relacionadas. Essas observações reforçam a ocorrência do processo de convergência evolutiva como consequência da adaptação aos antibióticos utilizados contra as bactérias (Laehnemann et al., 2014).

Nesse mesmo raciocínio, pesquisadores estudaram a adaptação genética de bactérias à novos ambientes. Amostras de *Pseudomonas aeruginosa* foram isoladas de pacientes com fibrose cística durante aproximadamente quatro anos. Foram encontradas evidências de evolução molecular convergente entre alguns genes, dos quais grande parte possui função relacionada à resistência e susceptibilidade aos antibióticos (Marvig et al., 2015).

A diversidade entre os genes de resistência aos antibióticos é uma consequência direta das evidências de divergência e convergência evolutiva entre esses determinantes genéticos (Figura 1.2). Esses fenômenos, por sua vez, são ocasionados principalmente pela pressão seletiva relativamente recente gerada pelo uso de antibióticos, ainda que outros fatores também selecionem mecanismos de resistência, como a microbiota do habitat natural. Com isso faz-se necessária uma caracterização ampla, que inclua as relações entre esses genes, sua variedade e distribuição.

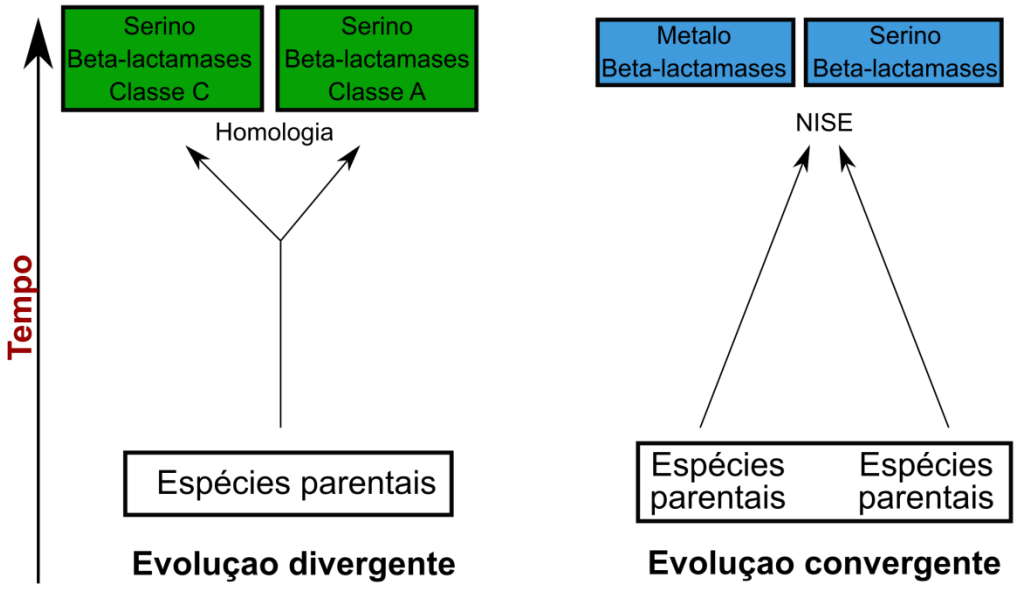


Figura 1.2: Convergência e divergência evolutivas e suas relações com os mecanismos enzimáticos resistência aos antibióticos.

Acredita-se que serino e metalo-beta-lactamases tenham evoluído a partir de ancestrais distintos, enquanto as enzimas do tipo serino-beta-lactamases classe A e classe C possuem um ancestral comum. NISE: *Non-homologous Isofunctional Enzymes* (Enzimas Isofuncionais Não-Homólogas).

1.2 Enzimas responsáveis pela resistência aos antibióticos

As enzimas que conferem resistência aos antibióticos são um grupo característico de proteínas adaptadas que utilizam um amplo esquema de estratégias enzimáticas (Tabela 1), podendo ser divididas em quatro grupos:

i) Oxirredutases: Catalisam reações de oxidação-redução;

ii) Transferases: Catalisam a transferência de grupos entre duas moléculas;

iii) Hidrolases: Catalisam a reação de hidrólise de várias ligações covalentes;

iv) Liases: Catalisam a clivagem de ligações C-C, C-O, C-N, entre outras, através de hidrólise ou oxidação.

Tabela 1.1 - Estratégias enzimáticas de resistência aos antibióticos e seus respectivos exemplos

Destruição do antibiótico		
Estratégia enzimática	*Classe do antibiótico alvo*	*Nome da enzima*
Hidrolase	Beta-lactamico	Beta-lactamase
Transferase	Fosfomicina	Tiol transferase
Oxirredutase	Tetraciclina	TetX
Liase	Estreptogramina tipo B	Vgb
Modificação do antibiótico		
Estratégia enzimática	*Classe do antibiótico alvo*	*Nome da enzima*
Transferase	Aminoglicosídeo	Acetiltransferase

A maior parte dos mecanismos enzimáticos caracterizados de resistência aos antibióticos é classificada nos grupos das hidrolases e das transferases, mas existe um número crescente de estratégias enzimáticas alternativas que estão sendo exploradas por bactérias a fim de eliminarem a ameaça dos antibióticos, como liases de estreptograminas e oxirredutases de tetraciclinas (Jacoby et al., 2009).

Várias enzimas possuem a capacidade de se ligar e quebrar ligações sensíveis à hidrólise presentes, por exemplo, em moléculas de beta-lactâmicos e macrolídeos. Como a água é o único substrato necessário para as hidrolases, elas podem ser excretadas e interceptar o antibiótico antes que ele entre em contato com a célula bacteriana (Wright, 2005).

O grupo das transferases é o maior e mais diverso entre as famílias de enzimas de resistência a antibióticos, catalisando modificações covalentes nas suas moléculas. Essas estratégias incluem tanto O- e N- acetilação, O- fosforilação, O- nucleotidilação, O- ribosilação e O- glicosilação. Essa tática de inativação de antibióticos requer a presença de co-substratos para a atividade enzimática, como por exemplo, acetil-CoA, ATP ou UDP-glicose. Consequentemente, a atividade enzimática está localizada no citosol bacteriano, e a reação de inativação é estável nesse ambiente. Sendo assim, essas reações que inativam os antibióticos são consideradas irreversíveis na ausência de outra enzima que as neutralize (Jacoby et al., 2009).

Os mecanismos enzimáticos de resistência aos antibióticos geram uma resposta altamente precisa. Por exemplo, a inativação de um antibiótico por uma enzima tem impacto amplo e significativo na antibioticoterapia, uma vez que elas diminuem efetivamente a concentração da droga no local, promovendo não só o crescimento do microrganismo detentor da enzima (resistente), como também das bactérias sensíveis adjacentes (Jacoby et al., 2009).

A seguir, examinaremos os principais grupos de enzimas responsáveis pela inativação das classes antibióticos que são ilustrados na Figura 1.3, como beta-lactamases; enzimas modificadoras de aminoglicosídeos, rifampicina, quinolonas e cloranfenicol; além de enzimas desintegradoras de fosfomicinas e tetraciclinas.

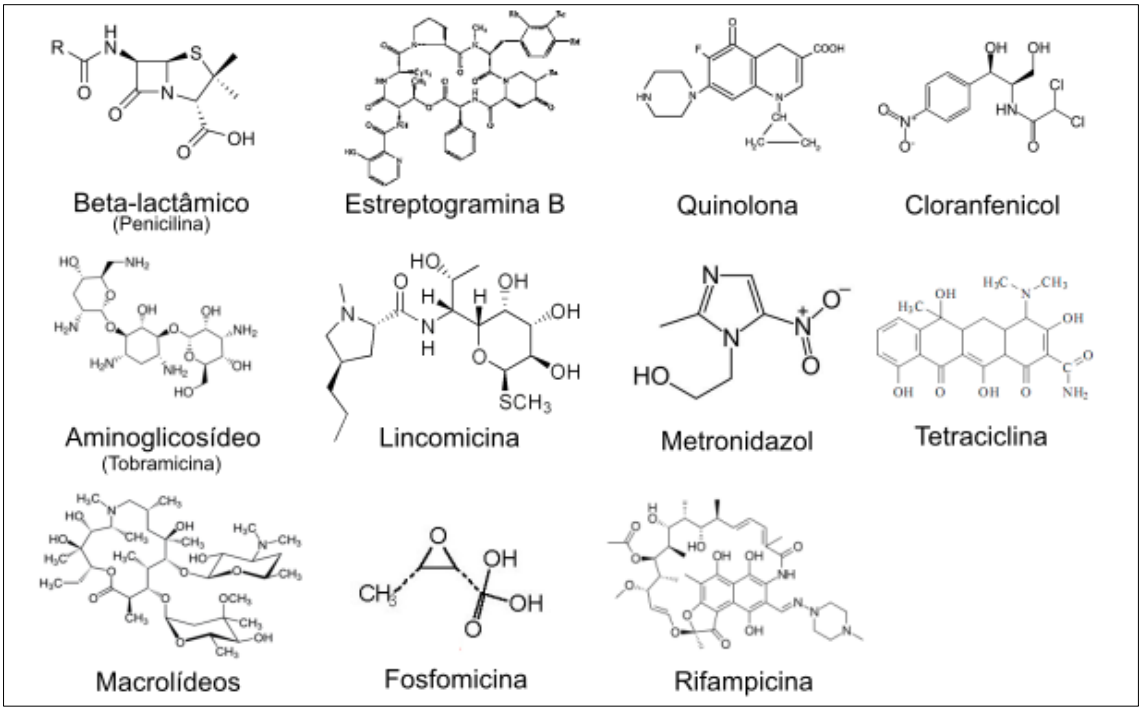


Figura 1.3: Estrutura molecular dos principais grupos de antibióticos abordados no trabalho.

Fonte: A estrutura molecular de casa antibiótico foi obtida da Wikipédia e em seguida agrupadas.

1.2.1 Beta-lactamases

Beta-lactâmicos são uma classe de antibióticos amplamente utilizados, que incluem penicilinas, cefalosporinas e carbapenemas, cujos espectros de ação aumentam nessa ordem (Bush, 2013). Todas agem através da modificação covalente das PBPs (*Penicillin-binding Protein*). As PBPs são enzimas associadas à membrana da célula bacteriana, importantes na montagem e manutenção da camada de peptideoglicano. Uma vez que a PBP é modificada por um beta-lactâmico, sua atividade enzimática é bloqueada, inibindo assim o metabolismo da parede celular, resultado na destruição da sua integridade e morte celular (Figura 1.4) (Jacoby et al., 2009).

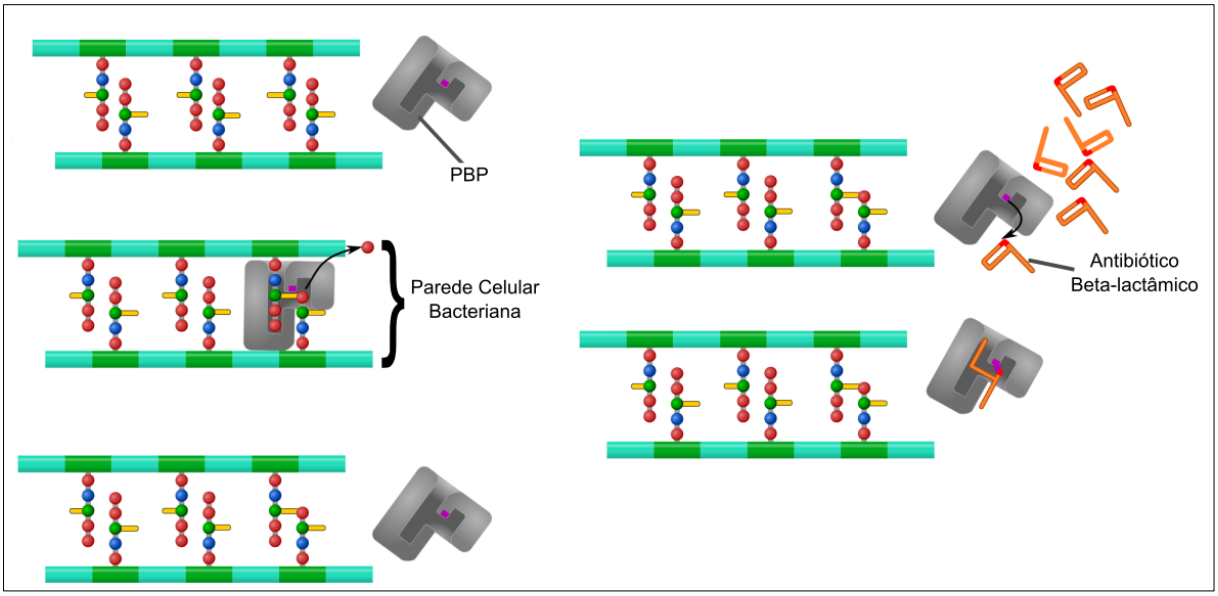


Figura 1.4: Inibição da transpeptidação do peptidoglicano da parede celular bacteriana por modificação covalente das *Penicillin-binding Proteins* (PBP) pelos antibióticos beta-lactâmicos.
Fonte: Adaptado de https://commons.wikimedia.org/wiki/File:Penicillin_inhibition.svg.

A principal causa de resistência aos beta-lactâmicos em bactérias Gram-negativas é a produção de beta-lactamases (BLs), que também estão distribuídas entre bactérias Gram-positivas (Eliopoulos and Bush, 2001) (Figura 1.5). O primeiro relato de uma BL ocorreu em 1940, por Abraham & Chain (Abraham and Chain, 1940). Atualmente é possível encontrar BLs em qualquer tipo de ambiente, incluindo o solo, a água e a microbiota de humanos e animais (Allen et al., 2009; Fróes et al., 2016; Gibson et al., 2014). Esse grupo de enzimas é numeroso, diverso e muito distribuído, principalmente porque são codificadas por genes presentes tanto no cromossomo como em elementos genéticos móveis (plasmídios e transposons, por exemplo). Muitas variantes de BLs já foram identificadas, chegando a mais de 1000 enzimas diferentes (Davies and Davies, 2010).

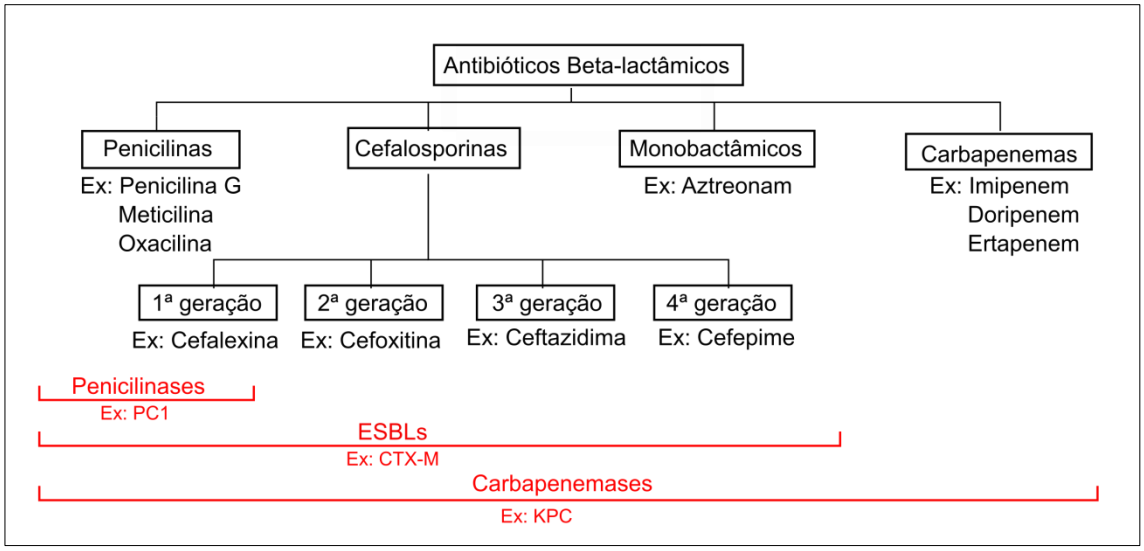


Figura 1.5: Grupos de antibióticos beta-lactâmicos e os respectivos tipos de beta-lamases resposáveis pela sua inativação.
Preto: antibióticos beta-lactâmicos; Vermelho: Beta-lactamases. ESBLs: Beta-lactamases de Espectro Ampliado.

Seu impacto clínico é particularmente crítico, uma vez que acarreta a dependência do uso de beta-lactâmicos mais potentes e caros para o tratamento de infecções simples (Bush, 2013). Importantes patógenos como *P. aeruginosa*, *Acinetobacter baumannii*, *Staphylococcus aureus* e as *Enterobacteriaceae* apresentam taxas significativas de resistência aos beta-lactâmicos no ambiente hospitalar em todo mundo (Davies and Davies, 2010). A produção de BLs com amplo espectro de ação por cepas de *Escherichia coli* e *Klebsiella* spp. isoladas na América Latina é um problema conhecido. Além disso, no Brasil, as taxas de bacilos Gram-negativos resistentes aos carbapenemas vêm aumentando muito (Gales et al., 2012). Por exemplo, cepas de *A. baumannii*, um patógeno oportunista associado a pacientes críticos, apresentaram taxa de resistência à carbapenemas igual a 76,8% entre pacientes internados em Unidades de Terapia Intensiva (UTI) em Goiás, Brasil (Castilho et al., 2017).

As BLs são designadas pelo comitê de nomenclatura da IUBMB (*International Union of Biochemistry and Molecular Biology*) como “um grupo de enzimas que hidrolisam beta-lactâmicos com especificidade variada”. Essas enzimas se dividem em dois grupos que não compartilham similaridade de estrutura ou sequência: as serina-beta-lactamases (SBLs) e metalo-beta-lactamases (MBLs).

As SBLs utilizam uma serina no sítio ativo para hidrolisar o anel beta-lactâmico, podendo ter preferência por antibióticos do tipo penicilinas, cefalosporinas (de espectro restrito ou ampliado) ou carbapenêmicos, dependendo da classe e da enzima em questão. As MBLs são capazes de hidrolisar todos os beta-lactâmicos, exceto antibióticos do tipo aztreonam. As MBLs necessitam de cátions divalentes como cofatores enzimáticos, sendo inibidas pela ação de agentes quelantes e por componentes derivados de tióis, como o ácido tiolático ou ácido 2-mercaptopropiônico (Bush et al., 1995).

Existem diferentes correntes sobre a origem evolutiva das BLs. A primeira delas acredita que essas enzimas foram primeiramente produzidas por bactérias ambientais a fim de protegê-las contra actinomicetos e bactérias produtores de beta-lactâmicos no solo (Massova et al., 1999). No entanto, outra corrente acredita que a função de proteção contra beta-lactâmicos não seria crítica para as bactérias, uma vez que esses compostos não se difundiriam o suficiente para serem combatidos por outras bactérias não produtoras. Essa segunda linha defende que as BLs teriam um papel principal, mas pouco entendido, na fisiologia bacteriana, possivelmente regulando o crescimento celular (Medeiros, 1998). Especula-se também que os beta-lactâmicos seriam um meio dos organismos conseguirem energia, e as BLs vieram depois como um sistema de defesa. Quando a biosfera começou a ficar rica em componentes orgânicos, não era mais necessário destruir organismos para conseguir alimento, então a produção de beta-lactâmicos foi cessada e os genes silenciados (Hall and Barlow, 2004).

1.2.2 Enzimas modificadoras de aminoglicosídeos

Os aminoglicosídeos são uma classe diversificada de antibióticos, com mais de 20 membros. Alguns são naturalmente encontrados na natureza e outros semissintéticos, como é o caso da amicacina. Sua atividade bactericida é exercida através da ligação ao RNAr e a subunidade 30S dos ribossomos, interferindo na síntese de proteínas (Figura 1.6) (Cox et al., 2015).

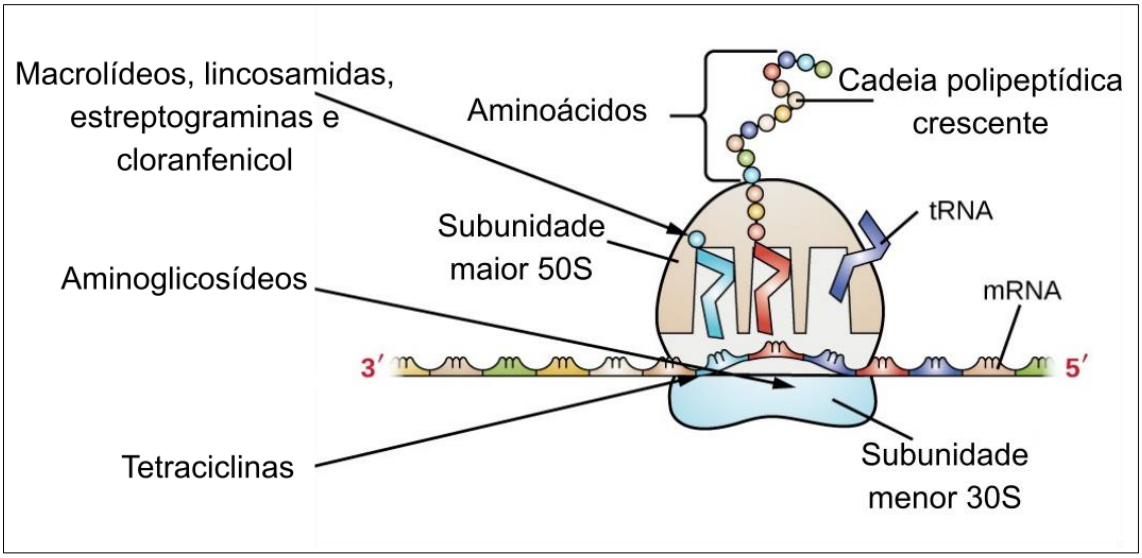


Figura 1.6: Sítio de ação das principais classes de antibióticos que inibem a síntese proteica em bactérias.

Fonte: Adaptado de https://courses.lumenlearning.com/microbiology/chapter/mechanisms-of-antibacterial-drugs/.

A modificação enzimática é o mecanismo mais prevalente de resistência aos aminoglicosídeos em bactérias causadoras de infecções (Ramirez and Tolmasky, 2010). Esse processo torna o antibiótico incapaz de se ligar de modo eficiente nos ribossomos e diminui sua interferência na síntese proteica (Cox et al., 2015). As enzimas modificadoras de aminoglicosídeos são transferases que catalisam a modificação covalente do antibiótico ou do seu alvo. Em isolados clínicos são mais comuns as N-acetilases, O-adenilases e O-fosforilases, que são muitas vezes adquiridas através de genes presentes em plasmídeos e/ou associados à transposons (Wright and Thompson, 1999). O esquema de classificação e nomenclatura dessas enzimas será abordado no item 1.3.2.

1.2.3 Enzimas inativadoras de Macrolídeos-Lincosamidas-Estreptograminas

Os antibióticos das classes macrolídeos, lincosamidas e estreptograminas (MLS), apesar de quimicamente distintos, geralmente são considerados juntos, pois compartilham uma sobreposição do sítio de ligação à subunidade 50S do ribossomo bacteriano, interferindo na síntese proteica (Figura 1.6) (Roberts, 2008).

A resistência a esse grupo de antibióticos inclui genes codificadores de metilases do RNAr e enzimas inativadoras dos antibióticos. As enzimas

inativadoras normalmente causam resistência a umas das três classes (M, L ou S), pertencendo a diferentes grupos como hidrolases, liases e transferases (Roberts, 2008). Elas são identificadas em cepas resistentes à MLS, e em microrganismos patogênicos o impacto dessas enzimas é desigual em termos de incidência e implicações clínicas (Jacoby et al., 2009).

A eritromicina e outros macrolídeos surgiram como uma tentativa de conter a resistência à meticilina (um beta-lactâmico) em Gram-positivos. Cepas resistentes por diferentes mecanismos já são amplamente disseminas (Davies and Davies, 2010). Existem dois tipos de esterases (um tipo de hidrolase) de eritromicina, codificadas pelos genes ortólogos *ere*A e *ere*B. Essas enzimas são capazes de linearizar o anel macrolídeo e consequentemente inativar sua ligação ao alvo. Embora não seja um mecanismo de resistência comum, causam níveis de resistência bastante elevados. Essas enzimas são mais relacionadas com Gram-negativos, podendo algumas vezes serem encontradas em Gram-positivos como *S. aureus* (Jacoby et al., 2009; Wright, 2005).

As enzimas Mph são fosforilases de macrolídeos, identificadas em vários patógenos humanos. Elas são codificadas por três genes diferentes: *mph*(A), *mph*(B) e *mph*(C). Quando expressos, todos causam resistência a eritromicina e telitromicina, além de azitromicina por *mph*(A), e espiramicina por *mph*(B) e *mph*(C) (Chesneau et al., 2007).

As lincosamidas, como a clindamicina e a lincomicina, são produzidas por várias espécies do gênero *Streptomyces* e utilizadas no tratamento de infecções por bactérias Gram-positivas. A resistência específica é realizada por enzimas inativadoras desses antibióticos. Fosforilação e nucleotidilação são detectadas em *Streptomyces*. As lincosamidas nucleotidiltransferases são codificadas pelo gene *lnu* (também chamados de *lin*), que incluem *lnu*(A) até *lnu*(G), além da lin_{AN2}. Já foram observadas assinaturas específicas[1] para subfamílias de proteínas Lnu, como para os homólogos mais próximos de Lnu(A) e Lnu(B). Análises filogenéticas mostram que Lnu(C) e Lnu(D) pertencem ao mesmo clado, assim como Lnu(F) e Lnu(G). Os clados Lnu(A), Lnu(E) e Lnu(C)/(D) são mais próximos. Já o clado de Lnu(B) está perto de Lnu(F)/(G) (Stogios et al., 2015).

[1] Regiões de conservação entre várias sequências

Os antibióticos do tipo estreptograminas podem ser divididos em duas classes distintas, tipo A e tipo B. As estreptograminas do tipo A são antibióticos que se ligam ao centro peptídeo-transferase do ribossomo, e a resistência a essa classe ocorre através da ação de acetiltransferases (Jacoby et al., 2009; Wright, 2005). O tipo B age ligando-se a subunidade 50S do RNAr, e a resistência ocorre principalmente pela enzima Vgb (tipos A e B), da classe liase, cuja ação resulta na abertura do anel peptídico. A enzima Vgb cliva peptídeos cíclicos, linearizando-os, e a estrutura resultante não possui mais afinidade ao ribossomo bacteriano (Mukhtar et al., 2001).

A modificação enzimática das estreptograminas pode ser realizada por genes carreados em plasmídeos, como os genes *saa* (acetiltransferase de estreptograminas do tipo A) e *sbh* (hidrolases de estreptograminas do tipo B) (Jacoby et al., 2009; Wright, 2005). Além disso, a acetiltransferase Vat(A-F) é capaz de modificar estreptograminas do tipo A, sendo algumas delas aparentemente relacionadas à acetiltransferases tipo B de cloranfencol (Roberts, 2008; Schwarz et al., 2004)

1.2.4 Enzimas desintegradoras das fosfomicinas

As fosfomicinas são antibióticos que possuem como alvo uma enzima que participa da biossíntese da parede celular, MurA. A modificação covalente de um resíduo Cys chave dessa enzima pela fosfomicina a inativa eficientemente. Por sua vez, para inativar a fosfomicina a bactéria utiliza-se de reações de abertura do anel epóxi através de modificações enzimáticas (Jacoby et al., 2009). Uma delas é catalisada pela enzima FosX, descrita como causadora de resistência na bactéria patogênica *Listeria monocytogenes*, e cujos homólogos já foram encontrados também em bactérias ambientais. Essa reação é um processo hidrolítico metal-dependente que gera o vicinal-diol (Fillgrove et al., 2007). Outro processo ocorre via uma reação dependente de tiol que causa abertura do anel, catalisada por enzimas que utilizam tióis abundantes no meio intracelular, como glutationa (Glutationa transferase FosA) e cisteína (Metalotiol transferase FosB). FosA é encontrada em bactérias Gram-negativas, enquanto FosB está em Gram-positivas. A estrutura cristalizada de FosA assemelha-se a de FosX. Todas essas estratégias resultam na destruição eficiente do centro reativo do

antibiótico, bloqueando assim sua ação em MurA (Jacoby et al., 2009; Wright, 2005).

1.2.5 Enzima modificadora de quinolonas

Quinolonas são agentes bactericidas de amplo espectro contra bactérias Gram-positivas e Gram-negativas. Seus alvos são as enzimas DNA girase e DNA topoisomerase IV, essenciais para a bactéria nos processos de replicação e reparo. Os mecanismos de resistência mais conhecidos a essa classe de drogas são mutações nos alvos enzimáticos e a extrusão do antibiótico por bombas de efluxo (Davies and Davies, 2010). O único mecanismo enzimático de resistência às quinolonas é a produção de AAC(6′)-Ib-cr, com apenas dois aminoácidos diferentes de uma acetiltransferase de aminoglicosídeos (AAC(6′)-Ib), capaz de causar resistência a ciprofloxacina e norfloxacina. Esse mecanismo é transmissível, e apesar de não causar altos níveis de resistência, favorece a seleção das mutações de resistência (Robicsek et al., 2006).

1.2.6 Redução do metronidazol

Metronidazol é uma droga introduzida em 1960, como uma opção terapêutica para o tratamento de várias bactérias anaeróbicas e microaerófilas, além de parasitas. Entre os anaeróbios, alguns Gram-positivos são inerentemente resistentes a essa droga, e virtualmente todos os Gram-negativos são conhecidos por serem sensíveis (Jacoby et al., 2009).

A 5'-Nitroimidazol é uma pró-droga que quando metabolizada interage com o DNA, o RNA e proteínas intracelulares, causando quebras nas fitas do DNA, inibindo sua reparação e interrompendo a transcrição, o que pode levar a morte celular. A resistência ao nitroimidazol é associada ao gene *nim*, com dez tipos descritos (A, B, C, D, E, F, G, H, I e J), todos capazes de reduzir a droga, transformando-a em aminoimidazol. Essas enzimas estão distribuídas em bactérias Gram-negativas e positivas além das arqueias, o que sugere uma origem bastante antiga. Entretanto, somente a presença do gene *nim* não determina a resistência ao nitroimidazol, devendo ser considerados outros mecanismos como um aumento na transcrição de genes de efluxo da droga e alterações no sistema de reparo do DNA (Husain et al., 2013). Além disso, outros

autores defendem a possibilidade de múltiplas oxirredutases estarem envolvidas na redução do nitroimidazol, não existindo, portanto, nenhum gene específico para a resistência ao mesmo (Diniz et al., 2004).

1.2.7 Enzimas modificadoras de rifampicina

Entre os antibióticos da classe rifamicina, a rifampicina, introduzida em 1968, é o membro mais largamente utilizado, tendo se tornado um componente integral do tratamento multi-antibiótico padrão-ouro para infecções causadas por *Mycobacterium tuberculosis*. Essa classe de drogas tem como alvo a subunidade *Beta* da RNA polimerase (Jacoby et al., 2009).

Apesar da resistência à rifampicina ocorrer normalmente por mecanismos mutacionais na RNA polimerase, existe também a inativação por modificação enzimática. As enzimas ARR realizam a modificação da rifampicina por um mecanismo de ADP-ribosil transferase, e estão amplamente distribuídas entre bactérias patogênicas e não patogênicas. Esse tipo de mecanismo de resistência só foi documentado para a rifampicina até o momento. As enzimas dessa família utilizam NAD+ como doador da porção ADP-ribosil, sendo conhecidas quatro proteínas (ARR-2, ARR-ms, ARR-sc e ARR-cb) (Marvaud and Lambert, 2017). A inativação de rifampicina por fosforilação e glicosilação já foi observada em algumas espécies bacterianas, no entanto os genes codificadores das enzimas responsáveis por essas modificações ainda não foram identificados (Jacoby et al., 2009).

1.2.8 Enzimas modificadoras de cloranfenicol

O cloranfenicol foi isolado pela primeira vez em 1947, produzido pela espécie *Streptomyces venezuelae*. Esse antibiótico possui amplo espectro de ação e é capaz de atravessar as membranas biológicas atingindo bactérias intracelulares, incluindo a barreira hematoencefálica. Ele atua inibindo a síntese proteica ao se ligar à subunidade 50S do ribossomo bacteriano (Figura 1.6), e pode interagir com ribossomos mitocondriais de células eucarióticas devido a semelhança estrutural com os ribossomos bacterianos. Algumas vezes, podem ser observadas reações adversas a esse antibiótico, variando entre erupções cutâneas até choque anafilático. Como já existem outras classes de antibióticos

menos tóxicas e com espectro de ação semelhante ao do cloranfenicol, seu uso em humanos atualmente está limitado, sendo utilizado como uma alternativa para o tratamento de meningites (Schwarz et al., 2004).

A resistência via inativação enzimática é na maior parte das vezes devido à presença de cloranfenicol acetiltransferases (CATs). Existem dois tipos de enzimas CATs que possuem sequências de aminoácidos e estruturas tridimensionais não relacionadas. Essas enzimas são trímeros composto por três monômeros idênticos que compartilham a estratégia molecular de O-acetilação (Wright, 2005).

1.2.9 Enzimas desintegradoras das tetraciclinas

As tetraciclinas têm sido utilizadas extensivamente por mais de meio século. Essa classe de antibióticos bloqueia a tradução em bactérias por se ligar a subunidade menor do ribossomo (Figura 1.6). Existe um mecanismo enzimático oxigênio-dependente capaz de inativar essa classe de antibióticos, apesar de não ser o meio principal de resistência às tetraciclinas. A oxirredução é uma estratégia molecular ainda pouco explorada pelas bactérias no processo de resistência. A enzima TetX facilita a mono-hidroxilação do antibiótico, quebrando eficientemente o sítio essencial de ligação ao metal na molécula (Jacoby et al., 2009). Esse gene foi encontrado em dois transposons intimamente relacionados com transposons de espécies do gênero *Bacteroides*, que também carreiam um gene de resistência à eritromicina (Speer et al., 1992). Através da análise de sequências foi identificada uma sequência similar a *tet*(X) nomeada como *tet*(37), que também é uma oxirredutase dependente de NADPH e foi isolada do microbioma oral de humanos (Diaz-Torres et al., 2006).

O significado clínico da resistência causada por TetX ainda não é claro. Essa enzima não confere resistência às cepas de *Bacteroides* onde foi originalmente encontrada, além de requerer altos níveis de aeração para funcionar como um fator de resistência em *E. coli*, o que pode significar níveis de resistência insignificantes no ambiente microaerofílico encontrado na maioria dos sítios humanos (Speer et al., 1992).

1.3 Sistemas de classificação

Por definição geral, uma classificação é o arranjo sistemático de entidades em categorias de acordo com diferentes características (http://www.biology-online.org/dictionary/Classification). Para Fleiss e Zubin (Fleiss and Zubin, 1969), chegar a uma descrição útil da amostra é válido para descobrir grupos ignorados que podem ser importantes.

A evolução esta inevitavelmente ligada à classificação. Podemos classificar as moléculas por filogenia ou pela sua função. Se mudanças funcionais importantes foram únicas na história, se surgiram uma única vez em resposta a uma pressão seletiva específica, então a classificação funcional estaria em concordância com a classificação evolutiva. No entanto, se a mudança funcional surgiu várias vezes, talvez em resposta a pressões seletivas similares, então membros do mesmo grupo funcional podem na verdade estar distantemente relacionados do ponto de vista evolutivo (Hall and Barlow, 2004).

Durante o processo de classificação de proteínas, existem dois pontos a serem considerados. O primeiro ponto diz respeito ao modo como a classificação deve ser feita, escolhendo entre uma classificação curada ou automatizada. A curadoria gera grupos de alta qualidade, mas o resultado pode não ser reproduzível e escalável para grandes volumes de novos dados. Já a automatização, apesar de poder gerar atribuições imprecisas, é completamente reproduzível e pode ser escalável. Idealmente, a automatização deve ser o objetivo final de todos os classificadores, com base na experiência de curadoria manual em menor escala. O segundo ponto do processo classificatório está relacionado à escolha do critério: (i) estrutural ou (ii) funcional. O primeiro (i) deve basear-se exclusivamente nos diferentes níveis da estrutura proteica, enquanto o segundo (ii) deve ser independente de atributos estruturais e permitir que tipos moleculares estruturalmente não relacionados pertençam à mesma classe funcional (Ouzounis et al., 2003).

Um dos mais antigos sistemas de classificação, tratando-se de uma classificação funcional hierárquica de enzimas, é o chamado *Enzyme Commission* (EC). Como descrito na seção 1.1, o número de EC consiste de quatro dígitos que descrevem a especificidade da reação enzimática pela definição do substrato/produto e do cofator de uma reação específica

(Omelchenko et al., 2010). Entretanto, as relações evolutivas, o postulado fundamental da biologia, não são consideradas pela classificação de EC. Portanto, a classificação baseada nos ECs permite que moléculas com origens evolutivas distintas, refletidas em sequências de aminoácidos e estruturas tridimensionais diferentes, sejam classificadas com o mesmo número, pois esta classificação leva em consideração apenas a reação enzimática catalisada.

1.3.1 Classificação das beta-lactamases

As BLs pertencem a duas superfamílias díspares, com características estruturais e funcionais diferentes, as SBLs e MBLs. No banco de dados CATH (*Protein Structure Classification Database*) (Sillitoe et al., 2015), as SBLs pertencem à superfamília DD-peptidase/beta-lactamase (CATH 3.40.710.10). Em relação à função, todos os membros dessa superfamília são hidrolases moderadamente diversas, tais como as próprias SBLs, as PBPs, glutaminases e transferases (Lee et al., 2016). As MBLs pertencem a uma superfamília mais heterogênea do CATH (3.60.15.10), incluindo um número grande de enzimas bastante distintas, como as próprias MBLs, hidroxiacilglutationa hidrolases e ribonucleases (Frère et al., 2005).

A diversidade e o impacto clínico das BLs levaram a várias tentativas de classificá-las, a partir de critérios funcionais ou estruturais. O primeiro sistema de classificação baseado em estruturas, elaborado por Ambler em 1980, é o mais utilizado (Ambler, 1980). Ele divide as BLs entre as classes A, B, C e D, de acordo com suas sequências de aminoácidos, que naquela época eram pouco disponíveis. Ambler originalmente especificou duas classes: a classe A, que apresenta o sítio ativo de SBLs, e a classe B, que emprega um ou dois íons Zn^{+2} em seu mecanismo catalítico (MBLs). Duas outras classes de SBL foram descritas mais tarde, as classes C e D (Jaurin and Grundström, 1981; Ouellette et al., 1987). A classe B foi dividida ainda nas subclasses B1, B2 e B3 com base nos seus motivos primários de ligação ao zinco (Galleni et al., 2001; Rasmussen and Bush, 1997).

Atualmente, cada classe inclui diferentes famílias de enzimas e suas variantes. As famílias costumam ser nomeadas por abreviações que demonstram suas diferentes origens, se referindo, por exemplo, ao substrato

beta-lactâmico preferencial da primeira variante descrita de cada família ou a nomes de lugares e pessoas relacionadas ao primeiro relato de determinada enzima (Brandt et al., 2017). Por exemplo, NDM se refere a uma família de MBLs cuja primeira variante (NDM-1) foi isolada de um paciente cuja infecção foi adquirida em Nova Deli, Índia (Yong et al., 2009).

A classe A é a mais abundante dentro das BLs, à qual pertence primeira BL relatada (Hall and Barlow, 2004). As enzimas da classe C, muitas vezes chamadas de AmpC, são amplamente encontradas no cromossomo da família *Enterobacteriaceae*, e a partir dos anos 1990 surgiram relatos da sua presença em plasmídios (Jacoby, 2009). Já a classe D é diversificada, presente tanto em plasmídios quanto em cromossomos de bactérias, sendo conhecidas também como oxacilinases (OXA) devido ao substrato preferencial da primeira variante descrita dessa classe (Hall and Barlow, 2004). As enzimas da classe B inativam eficientemente os beta-lactâmicos, além de serem frequentemente encontradas em plasmídios que facilitam sua disseminação (Widmann et al., 2012).

As estruturas terciárias de todas as SBLs são suficientemente similares entre si e por isso podem ser consideradas homólogas (Hall and Barlow, 2005). A separação entre as classes A, C e D é justificada pela diferença entre suas estruturas primárias e seus mecanismos de ação (Frère et al., 2005). Para as MBLs, que pertencem à classe B de Ambler, a mesma estrutura de quatro camadas alfa-beta/beta-alfa é compartilhada (Rasmussen and Bush, 1997). No entanto, existem discordâncias quanto à sua divisão em subclasses. Estudos indicam que as três subclasses de MBL não devem ser tratadas como grupos igualmente separados. As subclasses B1 e B2 apresentam similaridade detectável entre suas sequências, mas não com a subclasse B3 (Hall et al., 2004), e evidências estruturais sugerem fortemente que o ancestral comum mais recente (*Most Recent Common Ancestors*, MRCA) seja diferente para as subclasses B1/B2 e B3 (Alderson et al., 2014).

Novas possibilidades de subdivisões das classes de BLs vêm sendo sugeridas. Estudos baseados em filogenia e redes de similaridade de sequências (*Sequence Similarity Network*, SSN) se valeram da grande variedade de sequências disponíveis atualmente e mostraram que as classes A e D podem ser divididas em dois grupos distintos (A1 e A2, D1 e D2) (Brandt et al., 2017;

Philippon et al., 2016). A classe A seria dividida na subclasse A1, que inclui as famílias mais disseminadas pelo mundo como TEM e CTX-M; e na subclasse A2, encontrada principalmente no grupo *Cytophagales-Flavobacteriales-Bacteroidales* (CFB), mostrando ao menos quatro regiões de inserção em relação à A1 (Philippon et al., 2016). As enzimas da classe D2 são em sua maioria variantes intrínsecas que usualmente estão localizadas nos cromossomos bacterianos, formando um grupo bastante distinto em relação às demais oxacilinases da classe D (Brandt et al., 2017).

Devido a esses fatores citados acima, a classificação molecular de Ambler, que é a mais utilizada para BLs, não representa sua atual diversidade e não aponta uma definição precisa para várias classes e subclasses. Em 2005, Hall e Barlow sugeriram modificações dessa classificação (Hall and Barlow, 2005). De acordo com esse esquema, as BLs seriam classificadas em quatro níveis hierárquicos, da seguinte forma: no primeiro nível estão todas as BLs; no segundo, elas são divididas em dois grupos principais, estruturalmente não relacionados (S e M, se referindo as SBL e MBL); no terceiro nível encontram-se as classes de S e M, cujas enzimas não apresentam similaridade significativa entre suas sequências (classes SA, SC, SD, MB e ME); finalmente, no quarto nível estão os grupos de proteínas com similaridade de sequência (MB1 e MB2). Nesse esquema, as antigas subclasses B1 e B2 foram unidas e renomeadas como classe MB, enquanto a subclasse B3 foi renomeada como classe ME (Figura 1.7).

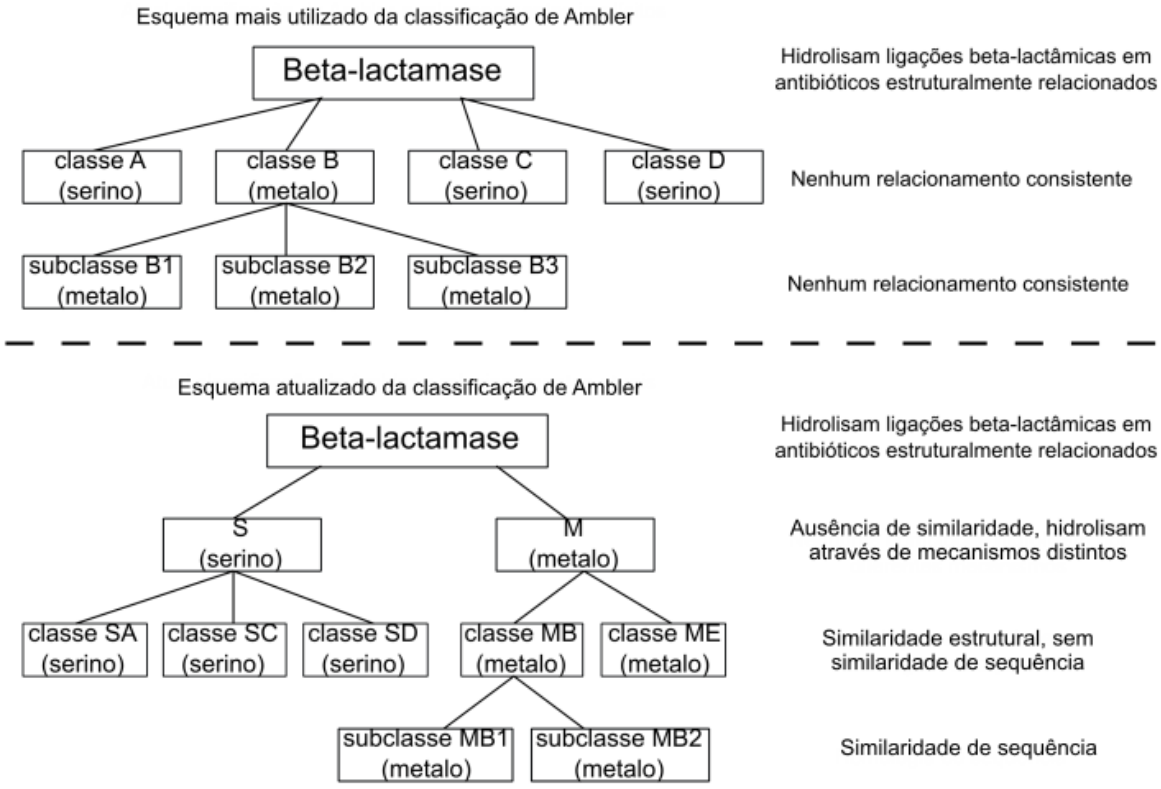


Figura 1.7: Diagrama esquemático dos relacionamentos implícitos no esquema de classificação estrutural de beta-lactamases, antes e após as modificações sugeridas por Hall e Barlow.
Fonte: Adaptado de Hall & Barlow, 2005.

Existe também a classificação funcional de Bush-Jacoby-Medeiros para as BLs. O objetivo dessa classificação é a descrição do papel clínico baseando-se nas interações com inibidores e perfil do subtrato, definindo subgrupos de acordo com a taxa de hidrólise de diferentes beta-lactâmicos e a resposta ao ácido clavulânico (Bush et al., 1995). Ela foi atualizada em 2013, adicionando propriedades microbiológicas, moleculares e bioquímicas das BLs (Bush, 2013). No entanto, uma classificação funcional não fornece um ponto de vista evolucionário preditivo, dificultando sua aplicação para identificação e classificação eficiente de proteínas.

1.3.2 Classificação das enzimas modificadoras de aminoglicosídeos

Existem dois sistemas de nomenclatura em uso para as enzimas modificadoras de aminoglicosídeos. O primeiro utiliza um identificador de três letras para o tipo de modificação enzimática (AAC, ANT e APH; acetiltransferase,

adeniltransferase e fosfotransferase, respectivamente) seguido pelo sítio de modificação em parênteses. Além disso, um número romano define o perfil de resistência único que é conferido à bactéria hospedeira, e por fim uma letra minúscula especifica proteínas individuais. Por exemplo, a enzima AAC(6')-Ia é uma acetiltransferase, que age no sítio 6' da molécula de aminoglicosídeo, gerando um perfil de resistência idêntico a outras enzimas do tipo I (Shaw et al., 1993).

Na segunda nomenclatura os genes são nomeados como *aac*, *aad* e *aph* (acetiltransferase, adeniltransferase e fosfotransferase, respectivamente), seguidos de uma letra maiúscula para definir o sítio de modificação. Um número também é adicionado no final para servir como identificador único para cada gene. Dessa forma, o gene aacA, aacB e aacC se referem a 6'-N-acetiltransferase, 2'-N-acetiltransferase e 3'-N-acetiltransferase de aminoglicosídeos, respectivamente (Novick et al., 1976). Um consenso sobre qual dessas nomenclaturas deveria ser adotada evitaria confusões e facilitaria o acompanhamento dos avanços no campo.

Essas nomenclaturas se baseiam na função das enzimas modificadoras de aminoglicosídeos, mas não esclarecem a relação estrutural entre os grupos. Shaw e colaboradores realizaram uma comparação utilizando as sequências dessas proteínas. Dentro de cada família (AAC, ANT e APH) as diferentes classes não apresentaram alta similaridade entre sí, só dentro da própria classe algumas vezes (Shaw et al., 1993)

As fosfotransferases ou quinases (APH) são divididas em sete famílias que compartilham muito pouca similaridada global entre suas sequências (Wright and Thompson, 1999). Existem quatro famílias de acetiiltransferases (AAC), que pertencem a uma mesma superfamília (Ramirez and Tolmasky, 2010). São conhecidas cinco famílias de adeniltransferases (ANT) e apenas três enzimas com a estrutura tridimensional resolvida, enquanto as AAC e APH são melhor caracterizadas (Cox et al., 2015).

Já foram descritos diferentes grupos dentro da classe AAC(6') de aminoglicosídeos. A relação filogenética entre os membros desses grupos indica a ausência de homologia. Situação semelhante à das SBLs, onde as três classes (SA, SC e SD) não exibem homologia entre suas sequências, porém clara

similaridade estrutural indicando um ancestral comum (Salipante and Hall, 2003). Dados estruturais mostraram que as famílias de AAC(6') apresentam sítios-ativos com arquiteturas muito diferentes, e por isso podem ser resultado de evolução funcional convergente, ainda que elas conservem o mesmo *fold* [2] (Stogios et al., 2017).

1.3.3 Outras classificações

Se considerarmos as outras classes de antibióticos listados nessa revisão, a variedade de enzimas capazes de inativa-las (Tabela 1.2) é muito menor quando comparada ao número de beta-lactamases e de enzimas modificadoras de aminoglicosídeos (Eliopoulos and Bush, 2001; Ramirez and Tolmasky, 2010).

Tabela 1.2 - Genes responsáveis pela resistência à diferentes classes de antibióticos, com exceção das beta-lactamases e enzimas modificadoras de aminoglicosídeos

Gene	Antibiótico	Variantes	Grupo
ere	Macrolídeo	*ere*(A), *ere*(B)	hidrolase
mph	Macrolídeo	*mph*(A), *mph*(B), *mph*(C)	transferase
lnu	Lincosamida	*lnu*(A), *lnu*(B), *lnu*(C), *lnu*(D), *lnu*(E), *lnu*(F), *lnu*(G), lin_{AN2}	transferase
saa	Estreptogramina A	*Saa*	transfease
vat	Estreptogramina A	*vat*A, *vat*B, *vatC, vat*D, *vat*E	transferase
vgb	Estreptogramina B	*vbg*(A), *vgb*(B)	liase
sbh	Estreptogramina B	*Sbh*	hidrolases
fos	Fosfomicina	*fos*X, *fos*A, *fos*B	transferase
nim	Nitroimidazol	*nim*A, *nim*B, *nim*C, *nim*D, *nim*E, *nim*F, *nim*G, *nim*H, *nim*I	oxido-redutases
arr	Rifampicina	*arr*-2, *arr*-ms, *arr*-sc, *arr*-cb	transferase
cat	Cloranfenicol	*cat*I, *cat*II	transferase
tetX	Tetraciclina	*tet*X	oxido-redutases

Apenas algumas dessas famílias de enzimas já foram analisadas quanto a critérios classificatórios. As fosforilases de macrolídeos (Mph) são divididas em duas classes, de acordo com diferenças entre suas sequências e especificidade de substrato. A classe Mph(2')I é codificada pelo gene *mph*(A), enquanto a classe Mph(2')II é codificada pelo gene *mph*(B). O gene *mph*(C) foi recentemente descrito, muito similar a *mph*(B). Mph(2')I e Mph(2')II apresentam enovelamento

[2] Enovelamento de uma proteína, sua forma tridimensional.

típico de proteínas quinases de eucariotos (Chesneau et al., 2007; Fong et al., 2017).

O trabalho de Roberts organiza os genes que conferem resistência aos MLS (macrolídeos, lincosamidas e estreptograminas). Nessa proposta, para serem classificados em um mesmo grupo e terem a mesma designação de letras, os genes precisariam apresentar identidade maior que 80% ao nível de aminoácidos (Roberts, 2008).

1.4 Distribuição e ambiente genético dos genes de resistência aos antibióticos

O estudo da diversidade e da distribuição dos determinantes de resistência entre populações bacterianas pode ajudar a entender melhor como a resistência aos antibióticos se desenvolve. Por exemplo, ainda não está claro por que alguns genes são encontrados em múltiplas espécies enquanto outros não. Novas fontes para monitorar mudanças nos padrões de resistência são muitos importantes (Roberts, 2008).

Genes de resistência funcionais estão presentes não apenas em espécies capazes de causar infecção, como também são encontrados em bactérias ambientais, comensais e não patogênicas. Essa prevalência pode ser alta, como mostrado em estudo utilizando genomas de bactérias ambientais e da microbiota humana, onde 84% deles codificavam pelo menos um gene de resistência (Gibson et al., 2014).

A maioria dos antibióticos utilizados no tratamento de infecções é produzida por microrganismos ambientais, e foi no ambiente que surgiram os primeiros mecanismos de resistência, onde permanece ocorrendo a evolução e seleção de novas estratégias de sobrevivência. Existem fortes indícios de que a resistência aos antibióticos surgiu bem antes da sua apropriação pelo homem (Martínez, 2008). Genes de resistência já foram encontrados na flora intestinal de pessoas que viveram em áreas isoladas aparentemente intocadas pela civilização moderna e não exposta ao uso de antibióticos (Bartoloni et al., 2009).

O conjunto de genes de resistência encontrado na natureza é chamado de resistoma (Davies and Davies, 2010). Embora amplamente distribuídos, os genes de resistência apresentam-se mais abundantes em determinados filos

bacterianos e ambientes (Gibson et al., 2014). Muitas espécies produtoras de antibióticos e consequentemente carreadoras de enzimas de resistência estão no filo *Actinobacteria* (Davies and Davies, 2010). O trabalho de Yongfei Hu e colaboadores mostrou que genes móveis de resistência são encontrados principalmente no filo *Proteobacteria*, seguido por *Firmicutes*, *Bacteroidetes* e *Actinobacteria*. Esses genes se dispersam entre espécies, gêneros e até filos diferentes. Eles observaram a transferência dos genes *tet*(C) (resistência a tetraciclina), *bla*$_{\text{TEM-116}}$ (resistência a beta-lactâmico), *aph*(3')-III (resistência a aminoglicosídeo) e *cat*A (resistência a cloranfenicol) entre cinco ou mais filos diferentes. Nesses casos, a filogenia se mostrou uma barreira determinante para a transferência horizontal de genes (*Horizontal Gene Transfer,* HGT), mais importante do que o fato das espécies compartilharem ou não o mesmo ambiente (Hu et al., 2016).

Avaliar a capacidade de organismos não patogênicos de trocar enzimas de resistência aos antibióticos com patógenos humanos pode levar a descoberta dos mais propensos a essa troca. Como a maioria dos patógenos humanos são *Proteobacteria*, pesquisar por genes de resistência compartilhados entre bactérias desse filo, isoladas no ambiente e no homem, poderia ser uma estratégia para avaliar essa capacidade (Forsberg et al., 2014).

Os mecanismos de resistência são classificados como intrínsecos ou adquiridos. Os mecanismos intrínsecos estão codificados no cromossomo de todas as cepas de determinado gênero ou espécie, sendo propagados verticalmente para seus descendentes. Já a aquisição de resistência ocorre por meio de mutações pontuais específicas ou por HGT (Courvalin, 2008). Já foi observada a troca de várias classes de genes de resistência entre bactérias ambientais não patogênicas e as de importância clínica (Forsberg et al., 2014).

A HGT tem ocorrido durante toda a história evolutiva das bactérias. No entanto, as mudanças que ocorreram durante a evolução das bactérias e outros organismos durante bilhões de anos foram muito mais lentas se comparadas ao fenômeno do desenvolvimento e transferência de resistência do último século. A atual pressão exercida pelo uso e pela disponibilidade de antibióticos é provavelmente muito mais intensa, pois as bactérias precisam sobreviver a

ambientes intensamente hostis, ao invés de evoluírem lentamente, adquirindo características que aumentem sua aptidão (Davies and Davies, 2010).

A resistência adquirida está geralmente associada a elementos genéticos móveis, responsáveis por sua disseminação. Esses elementos incluem sequências de inserção, integrons, transposons e plasmídios, que facilitam a persistência dos genes mesmo sem a presença da força seletiva dos antibióticos (Courvalin, 2008). O tipo de elemento varia de acordo com o gênero do patógeno, sendo a transmissão mediada por plasmídios, de longe, a forma mais comum de HGT (Norman et al., 2009).

Os plasmídios são moléculas extracromossômicas de DNA bacteriano. Essa estrutura é muito importante na evolução dos procariotos, pois pode ser transferida entre microrganismos e assim disseminar várias propriedades genéticas. Plasmídios então envolvidos com funções acessórias, que muitas vezes conferem vantagens evolutivas à cepa que o possuir. Determinantes de resistência são umas das informações que esses elementos podem carrear, exercendo um papel central na rede de disseminação (Tamminen et al., 2012).

Um antigo sistema de classificação separa os plasmídios em grupos de incompatibilidade (Inc). Dessa forma, plasmídios do mesmo grupo não são capazes de se manter no mesmo hospedeiro, devido à similaridade entre seus sistemas de replicação ou particionamento. A gama de hospedeiros plasmídicos pode variar bastante, em alguns casos ela é ampla e em outros, restrita. Os grupos IncF, IncH e IncI contém plasmídios com uma gama restrita de hospedeiros, enquanto plasmídios dos grupos IncN, IncP e IncW se transferem e replicam em várias espécies (Suzuki et al., 2010).

É difícil estabelecer uma correlação confiável entre a origem do gene e o organismo no qual ele foi encontrado. A proliferação de elementos transponíveis no genoma facilita a recombinação homóloga dentro dele, processo que pode resultar em rearranjos de cromossomos em larga escala, atrapalhando o estudo da ordenação dos genes por ancestralidade (Forsberg et al., 2014).

1.5 Bancos de Dados

As informações moleculares disponíveis sobre a resistência aos antimicrobianos facilitam o entendimento sobre evolução, dispersão e os mecanismos existentes, mas elas precisam estar integradas. Esforços para estabelecer a associação dessas informações vêm sendo aprimorados. Existem bancos de dados específicos para enzimas de resistência, sendo alguns exclusivos para as BLs. O primeiro deles, criado em 2001, é o *Lahey Clinical Database*, que atribuía números para novas variantes descritas em cada família de BLs. A partir de 2015, esse banco passou a ser mantido pelo NCBI, e está contido no *Bacterial Antimicrobial Resistance Reference Gene Database* (Naas et al., 2017).

Em 2008, outro banco de dados de BLs foi construído, o DLact, com 2.020 sequências similares às lactamases e suas respectivas informações de taxonomia, ecologia e tamanho de domínios. O número de sequências era maior que o do *Lahey*, mas apesar dos autores prometerem a atualização do banco, seu domínio não está mais disponível (Singh and Singh, 2008).

O *Lactamase Engineering Database* (LacED) quando foi publicado, em 2009, se destinava exclusivamente a família de BLs nomeadas como TEM, uma das mais difundidas e importantes clinicamente (Thai et al., 2009). Atualmente ele integra informações sobre estrutura primária e terciária, mutações e alinhamentos das duas principais famílias da classe A: TEM e SHV. Esse banco pode ser utilizado para a atribuição de novas sequências e a identificação de inconsistências em bancos de dados públicos, assim como facilitar a engenharia proteica dessas famílias de BLs.

Em 2012, foi criado um banco de dados curado e exclusivo para as MBLs, com o objetivo de facilitar sua classificação, nomenclatura e análise. O *Metallo-Beta-Lactamase Engineering Database* (MBLED) compreende 597 sequências, além dos perfis mutacionais das principais famílias de MBLs: IMP e VIM. O domínio do MBLED ainda está ativo, porém não foi atualizado desde sua publicação (Widmann et al., 2012). Outro banco de dados exclusivo para determinadas famílias de BLs, o *Institut Pasteur MLST Database*, pertence ao Instituto Pasteur. A nomenclatura das famílias OXY, OKP e LEN é curada e mantida por esse banco (Bialek-Davenet et al., 2014).

Disponibilizado pela primeira vez em 2013, o *Comprehensive Antibiotic Resistance Database* (CARD) fornece dados sobre várias classes de antibióticos, seus alvos e os respectivos mecanismos de resistência. A classificação dos genes de resistência é feita a partir de ontologia, um conjunto controlado de vocabulários que deve ser seguido. Desta forma, os genes e seus produtos podem ser conectados a suas atividades, permitindo uma investigação robusta de dados moleculares. Esse banco de dados curado foi atualizado em agosto de 2017, e possui mais de 2.300 genes conhecidos de resistência aos antibióticos. Inclui ainda ferramentas de bioinformática que permitem a identificação desses genes em genomas total ou parcialmente sequenciados, incluindo *contigs*[3] brutos ainda não anotados (McArthur et al., 2013).

Outro banco desenvolvido para mecanismos de resistência em geral é o Resfams. Ele foi criado para avaliar a relação entre resistomas ambientais e àqueles associados ao homem. Trata-se de um banco de dados curado, que também classifica os genes por ontologia, focado nas famílias de proteínas e seus perfis de Modelos Ocultos de Markov (*Hidden Markov Models*, HMM) associados. Os perfis HMM criados estão disponíveis, e a última atualização do banco foi feita em 2015 (Gibson et al., 2014). Na seção 1.7 os perfis HMM serão abordados com maior detalhamento.

O banco de dados mais completo sobre BLs atualmente disponível é o *Comprehensive Beta-lactamase Molecular Annotation Resource* (CBMAR). A arquitetura do banco é baseada no sistema de classificação de Ambler, e cada classe é dividida em famílias. Para cada família, é possível ter acesso a informações como origem do nome, gêneros onde já foi reportada, localização genômica, perfil de resistência, susceptibilidade a inibidores, sítios-ativos, *fingerprints*[4] específicos, perfis mutacionais, árvores filogenéticas, além de *links* para outros bancos de dados de estrutura e sequência. As buscas podem ser feitas por BLAST ou usando *fingerprints* específicos pelo MAST (Srivastava et al., 2014). O domínio disponibilizado no artigo não está mais ativo, o atual é http://proteininformatics.org/mkumar/lactamasedb, mas não houve atualização do banco.

[3] Conjunto de segmentos de DNA sobrepostos que juntos representam uma região de consenso
[4] Motivos proteicos

Além das fontes especializadas, informações sobre genes de resistência a antibióticos também podem ser recuperadas de bancos de dados não específicos e mais amplamente utilizados. Eles são fontes primárias para qualquer banco de dados especializado (Srivastava et al., 2014). O *National Center for Biotechnology Information* (NCBI) mantém o banco de dados de sequências de ácidos nucleicos, GenBank, que fornece muitos outros tipos de informações biológicas, assim como sistemas de recuperação e ferramentas computacionais para a análise de dados (Coordinators, 2015). Uma dessas ferramentas é o algoritmo BLAST, implementado em vários programas (Altschul et al., 1990).

O GenBank, assim como outras bases de dados biológicos, têm problemas de redundância, pois são estruturados de forma que uma entrada no banco reflete as descobertas de uma publicação, quando o ideal seria que uma entrada correspondesse a uma única entidade biológica. Dessa forma, a literatura desses bancos não é sintetizada e existem muitas declarações conflitantes (Karp, 2016).

O *Universal Protein Resource* (UniProt) é uma base de conhecimento abrangente para sequências de proteínas e dados de anotação associados a elas, que busca diminuir redundâncias. O Uniprot concentra informações do TrEMBL e do Swiss-Prot. Este último contém mais de meio milhão de sequências curadas manualmente, para as quais são disponibilizadas informações experimentais que foram extraídas da literatura. O TrEMBL (*Translated EMBL Nucleotide Sequence Data Library*) possui mais de 60 milhões de sequências anotadas automaticamente, oriundas principalmente de sequenciamento de alta vazão de DNA. A última revisão (*release* 2016_08) do UniProt procurou eliminar a redundância excessiva causada por proteomas muito similares, diminuindo o número total de sequências (Consortium, 2017).

Existem também bancos de dados específicos para estruturas tridimensionais. O *Protein Data Bank* (PDB) é o único repositório mundial de estruturas 3D de grandes moléculas biológicas, incluindo proteínas e ácidos nucleicos. Ele permite ao usuário fazer buscas simples ou complexas, visualizar e analisar o resultado. Atualmente ele armazena mais de 130.000 estruturas a

nível atômico, determinadas por cristalografia, espectroscopia RNM[5] e microscópios eletrônicos 3D (Burley et al., 2017).

A partir das estruturas tridimensionais disponíveis no PDB foi criado o CATH (*Protein Structure Classification Database*), um banco de dados que fornece informações sobre as relações evolutivas dos domínios proteicos. Depois de identificados, os domínios são separados em níveis estruturais. As "classes" se referem às estruturas secundárias, a "arquitetura" baseia-se no arranjo da estrutura secundária no espaço tridimensional, e a "topologia" usa informações sobre como os elementos secundários estão conectados e arranjados. Atribuições às "superfamílias homólogas" são feitas em último nível caso existam fortes evidências de que os domínios sejam relacionados evolutivamente (Sillitoe et al., 2015).

O "*Pfam - protein families database*" é uma grande coleção de famílias de proteínas, cada uma representada por um alinhamento múltiplo de sequências e um perfil HMM. Os dados disponíveis para cada entrada desse banco são baseados no proteoma do UniProt. Alinhamentos semente são utilizados para construção dos modelos probabilísticos, que são utilizados contra o proteoma, aplicando um *threshold*[6] curado mínimo (*Gathering Threshold*[7]) para as buscas (Finn et al., 2016). A última versão desse banco foi atualizada em março de 2017, e contém um total de 16.712 famílias de proteínas.

O PROSITE é uma fonte para identificação e anotaçao de regiões conservadas em sequências de proteínas. Ele se baseia em entradas para documentações que descrevem domínios, famílias e sítios funcionais, bem como padrões e perfis associados. Esse banco é usado para a anotação de domínios e características no UniProt. A ferramenta ScanProsite permite que os usuários pesquisem sequências de proteínas (incluindo proteomas inteiros) contra todas as assinaturas do PROSITE ou ainda busquem por correspondências para assinaturas definidas no PROSITE em bancos de dados como UniProt e PDB (Sigrist et al., 2013).

[5] Ressonância Magnética Nuclear
[6] Valor mínimo de alguma quantidade
[7] Valor mínimo de *bit score* estabelecido pelo Pfam para maximizar a cobertura dos perfis HMM e, ao mesmo tempo, excluir correspondências falso-positivas.

Todos os bancos citados aqui, se ainda ativos, possuem acesso gratuito para a comunidade global.

1.6 Bioinformática

A biologia computacional pode ser resumida como a aplicação de técnicas analíticas quantitativas à modelagem de sistemas biológicos. A bioinformática é parte da biologia computacional, uma ciência que usa as informações para entender a biologia. Os especialistas em bioinformática capturam, gerenciam e apresentam dados, numa interação entre biologia e uma ampla variedade de sistemas quantitativos, incluindo estatística, física, ciência da computação e engenharia (Gibas and Jambeck, 2001).

Nos primeiros anos da década de 1960, a biologia computacional emergiu impulsionada por três fatores principais: aumento na disponibilidade de sequências de aminoácidos, difusão da ideia de que macromoléculas carreiam informações, e o desenvolvimento de computadores digitais mais rápidos como consequência dos avanços da Segunda Guerra Mundial. Em 1970, diversas técnicas já haviam sido desenvolvidas para análise de estrutura molecular, função e evolução (Hagen, 2000).

Ao longo dos últimos anos, as informações fornecidas através de sequenciamento, biologia estrutural e bioinformática têm revolucionado a ciência das biomoléculas, através da disponibilidade de milhares de sequências e centenas de estruturas tridimensionais em banco de dados públicos. Além do grande volume de dados acessíveis, o desenvolvimento de ferramentas para sua manipulação vem permitindo que análises computacionais complexas sejam realizadas. Muitos *softwares*[8] são disponibilizados gratuitamente, e o conjunto de componentes físicos das máquinas (*hardware*) é constantemente aprimorado, tornando possíveis análises computacionais mais rápidas (Alderson et al., 2012). Porém, ainda que hoje a bioinformática tenha se revolucionado, ela permanece apoiada nas importantes bases intelectuais e técnicas estabelecidas em um período anterior a era do computador (Hagen, 2000), como o sequenciamento

[8] Todo programa armazenado em discos ou circuitos integrados de computador

da molécula de insulina por Frederick Sanger (Sanger, 1959) e a série de programas FORTRAN[9] escritos por Margaret Dayhoff para determinar a sequência de aminoácidos de moléculas proteicas (Dayhoff and Ledley, 1962).

Nas últimas duas décadas os avanços da bioinformática levaram a uma capacidade impressionante de sequenciar genomas, resultando numa quantidade de dados muito grande (Pallen, 2016). O potencial transformador da utilização do sequenciamento total do genoma bacteriano para diagnóstico clínico vem sendo muito discutido. As aplicações dessa tecnologia no diagnóstico molecular podem proporcionar mapas epidemiológicos mais precisos e informações sobre a história evolutiva e a composição genética de isolados específicos. Na clínica, seria possível fazer testes de susceptibilidade aos antimicrobianos sem a necessidade de cultivar o patógeno (Florian Fricke and Rasko, 2014).

O uso de sequenciamento de metagenomas é apontado como estratégia para detecção de doenças infecciosas, além de ter potencial para reduzir o tempo do diagnóstico (Florian Fricke and Rasko, 2014). Ao contrário do sequenciamento de cepas individualmente, a metagenômica permite a caracterização de uma comunidade inteira, incluindo bactérias não cultiváveis ou com crescimento muito trabalhoso em laboratório (Hugenholtz et al., 1998).

A viabilidade econômica da utilização de sequenciamento de genoma completo no diagnóstico molecular depende do custo e da facilidade da geração de sequências e das análises de bioinformática. As maiores dificuldades da bioinformática são procedimentos operacionais padronizados, uma infraestrutura adequada para a grande quantidade de dados, e o gerenciamento dos recursos computacionais. O primeiro passo seria a utilização dessa tecnologia pelos laboratórios nacionais de saúde, que trabalham em uma escala suficientemente grande para validar e otimizar os futuros testes moleculares (Florian Fricke and Rasko, 2014).

As informações genéticas microbianas derivam de milhares de patógenos, milhões de espécies comensais e até um bilhão de espécies bacterianas ambientais. Esse conjunto de dados tem uma magnitude maior que a de genes

[9] Linguagem de programação computacional de alto nível, com notação similar à da álgebra, concebida para aplicações matemáticas, científicas e técnicas

humanos. Problemas de *big data*[10] já são uma realidade para a bioinformática microbiana. Novas abordagens para o armazenamento e a análise de dados deverão ser desenvolvidas em um futuro próximo, como por exemplo, a construção de um banco de dados realmente não redundante. A aprendizagem de máquina e a inteligência artificial aos poucos irão substituir a curadoria manual para anotação de sequências ou metadados (Pallen, 2016).

Embora avanços significativos tenham sido feitos no processo de extrair informações de textos escritos usando inteligência artificial, ainda há um longo caminho a ser percorrido. Os programas de extração de informações ainda não são precisos ou abrangentes o suficiente para substituir a curadoria manual. Um dos problemas básicos, ainda com altas taxas de erro, é reconhecer o nome de entidades (genes, organismos, moléculas) em textos biológicos. De uma forma geral e lógica, o problema de automatizar a curadoria é proporcional à complexidade da mesma. A curadoria semi-automatizada parece ser o caminho para acelerar o processo, ainda que os primeiros resultados sejam limitados e a relação de custo-benefício não esteja clara (Karp, 2016).

1.7 Anotação de proteínas

Em bioinformática, anotação é um termo que se refere a todas às informações sobre uma proteína que não seja sua sequência. A elucidação da função de uma proteína abre o caminho para descrever completamente um organismo particular, através da caracterização de suas vias metabólicas e redes de regulação de transcrição (Rigoutsos et al., 2002).

Informações de função são normalmente extrapoladas a partir da identificação de sequências similares. No entanto, a complexidade funcional dentro das famílias e superfamílias de proteínas traz a necessidade de maior atenção no processo de transferência de anotação funcional. A maioria das superfamílias apresentam variações nos papéis exercidos por suas enzimas. A especificidade de substrato é comum, e para proteínas mais distantes, que compartilham menos de 30% de identidade, a variação funcional é significante.

[10] Grande conjunto de dados armazenados

A diversidade de função dentro de uma superfamília é gerada por variações locais nas sequências e reorganização (*shuffling*) dos domínios (Todd et al., 2001).

Avanços computacionais e de bioinformática vem permitindo melhoras substanciais na identificação de genes específicos. Algoritmos probabilísticos de anotação como os perfis HMM tem se mostrado superiores aos alinhamentos de sequências como o BLAST, em temos de precisão e recuperação de sequências homólogas (Gibson et al., 2014).

Perfis HMM são modelos probabilísticos, que especificam um sistema de pontuação para posições específicas, onde podem ocorrer substituições, deleções e inserções (Eddy, 1998). Na arquitetura desses perfis, para cada coluna do alinhamento múltiplo de sequências, um estado de correspondência modela a distribuição dos resíduos permitidos na coluna. Um estado de "inserção" e um estado de "deleção" em cada coluna permitem que um ou mais resíduos sejam inseridos ou deletados entre aquela coluna e a próxima (Krogh et al., 1994). Os parâmetros de probabilidade de um perfil são normalmente convertidos em pontuações aditivas *log-odds*[11] antes do alinhamento e da pontuação de uma sequência *query*[12]. O pacote HMMER implementa modelos baseados, ao menos em parte, nos perfis HMM originais de Krogh (Eddy, 1998).

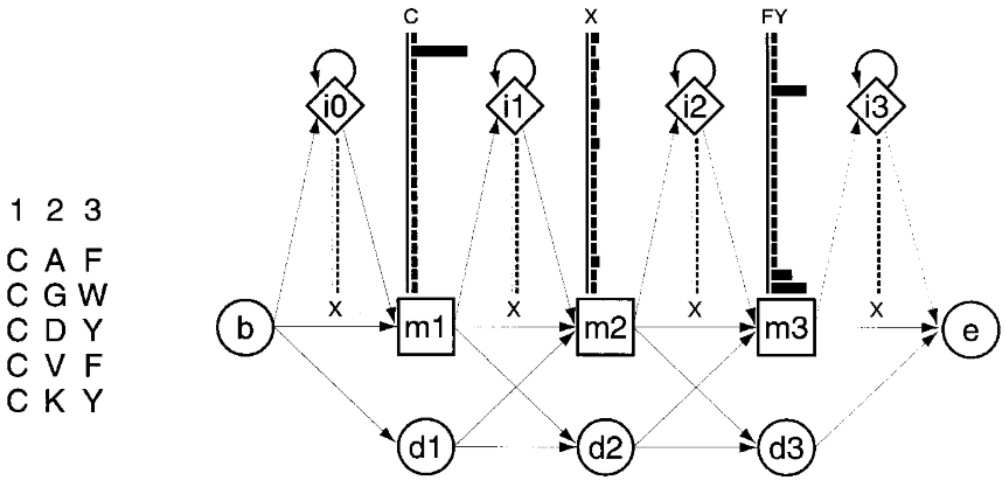


Figura 1.8: Um perfil HMM (direta) representando um alinhamento múltiplo de cinco sequências (esquerda) com três colunas consenso.

[11] Maneira alternativa de expressar probabilidades
[12] Comando de consulta

As três colunas são modeladas por três estados de correspondência (quadrados nomeados como m1, m2 e m3), cada um deles com 20 probabilidades de emissão de resíduos, mostrados como barras pretas. Estados de inserção (losangos nomeados como i0, i1, i3 e i3) também têm 20 probabilidades de emissão cada. Estados de deleção (círculos nomeados como d1, d2 e d3) são estados "mudos" que não tem probabilidades de emissão. Um estado inicial e outro final foram incluídos (b,e). Probabilidades dos estados de transição são mostradas como setas. (Fonte: Eddy, 1998).

Um único perfil HMM para uma determinada superfamília identifica proteínas com diferentes funções pois busca por sequências homólogas. São necessárias estratégias adicionais para a identificação de uma função específica (Brandt et al., 2017). Além disso, algumas famílias de proteínas distintas compartilham sinais de dobramentos, e com isso um perfil HMM construído para determinada família continua sendo pouco efetivo na tentativa de discriminá-la das demais (Sinha and Lynn, 2014).

Existem algumas alternativas para melhorar a precisão dos perfis HMM. Uma delas é usar *thresholds* curados junto ao *E-value*[13], como *Trusted Cutoff*[14] (TC), *Noise Cutoff*[15] (NC) e *Gathering threshold* (GA). Esses *thresholds*, introduzidos e utilizados pelo banco de dados Pfam, se baseiam em alinhamentos HMM. No entanto, esses critérios não se mantém uniformes quando aplicados em um conjunto de dados para treinamento, pré-classificados como sequências positivas e negativas. Isso acontece porque uma sequência negativa pode ter uma pontuação maior que uma sequência positiva. Uma opção seria usar essas sequências pré-classificadas para treinar os perfis HMM. O programa HMM-ModE gera perfis específicos para famílias de proteínas, através da otimização dos *thresholds* de discriminação utilizando sequências negativas de treinamento e o modo distribuição médio MCC (*Mathews correlation coefficient*[16]) (Sinha and Lynn, 2014).

Existem outras formas de atribuir função a uma sequência de aminoácidos. Uma estratégia é a atribuição de domínios proteicos, conservados ao longo da evolução. Como algumas proteínas possuem mais de um domínio,

[13] Número de alinhamentos que seriam esperados apresentando valores de pontuação iguais ou melhores que aos encontrados por acaso
[14] Pontuação da sequência de pontuação mais baixa no alinhamento do HMM
[15] Pontuação da sequência de pontuação mais alta que não está no alinhamento HMM
[16] Medida de qualidade para classificações binárias

a identificação dessas estruturas é mais informativa que utilizar apenas o melhor *hit*[17] fornecido pelos algoritmos de similaridade (Rigoutsos et al., 2002).

O papel evolucionário, estrutural e funcional desses domínios sugere que eles sejam “blocos de construção” indivisíveis, a partir dos quais proteínas modulares maiores são formadas. No entanto, fontes de anotações de domínios como o banco de dados Pfam, sugerem em alguns casos que apenas parte do domínio esteja presente em determinada sequência de aminoácidos. Na realidade, esses domínios parciais são resultado de alinhamentos locais dos perfis HMM, proteínas não funcionais ou montagens imprecisas do genoma (Triant and Pearson, 2015).

[17] Correspondência no banco de dados

2 OBJETIVOS

2.1 Objetivo Geral

Utilizar agrupamentos hierárquicos para caracterizar a diversidade entre as enzimas relacionadas à resistência aos antimicrobianos e sugerir um esquema de classificação padrão, além de analisar a distribuição e abundância de beta-lactamases entre diferentes táxons bacterianos.

2.2 Objetivos Específicos

1- Estabelecer um *pipeline* para clusterização de enzimas com mesma função enzimática usando como modelo as beta-lactamases;

2- Sugerir uma metodologia de classificação para a atividade enzimática em questão;

3- Construir modelos probabilísticos curados para cada *cluster* de beta-lactamases;

4- Avaliar abundância e distribuição de beta-lactamases entre diferentes táxons bacterianos;

5- Propagar o que foi desenvolvido para as beta-lactamases às outras atividades enzimáticas relacionadas à resistência aos antibióticos.

3 MATERIAL E MÉTODOS

O presente estudo está dividido em duas partes: a primeira foi realizada a partir de dados sobre BLs, uma vez que esse é o grupo de enzimas de resistência com maior disponibilidade de informações, e por isso foi utilizado para construir e validar a metodologia desenvolvida (Figura 3.1). A segunda parte consiste na extrapolação do processo para algumas famílias de enzimas modificadoras de aminoglicosídeos e acetiltransferases de cloranfenicol.

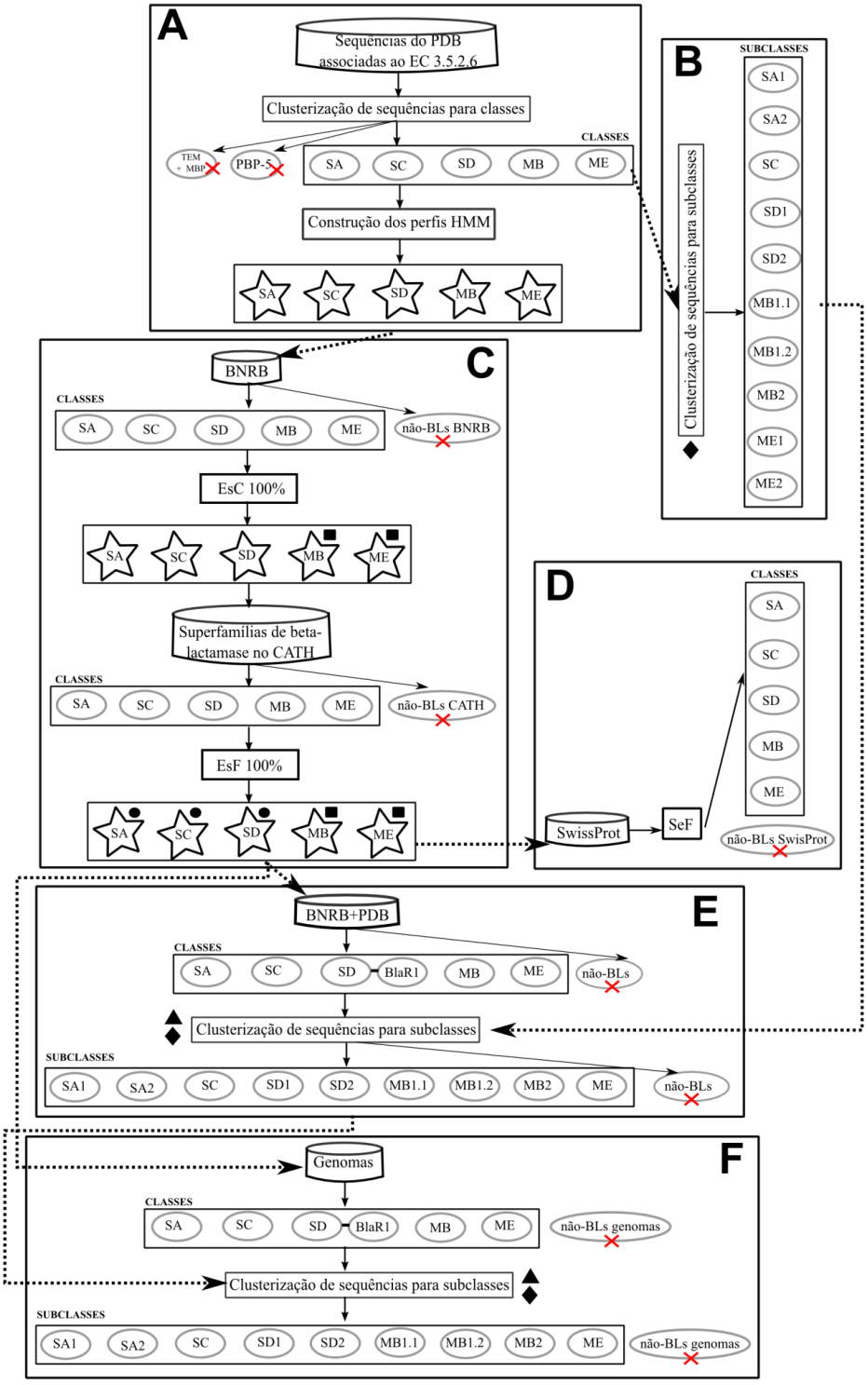


Figura 3.1 Fluxograma de construção da metodologia desenvolvida nesse estudo para identificação e classificação de sequências proteicas de BLs.
A:Construção de um perfil HMM para cada classe de BL; B:Calibração dos *thresholds* de similaridade para *clusterização* das sequências; C:Calibração dos *thresholds* para *hmmsearch*; D: Validação dos *thresholds* para *hmmsearch*; E:Calibração dos *thresholds* de cobertura para *clusterização* das sequências; F: Aplicação da metodologia em genomas. Estrela:Perfil HMM; Elipse: Agrupamento de sequências; X:Agrupamento desconsiderado. Os *thresholds* calibrados ao longo do metodologia estão representados por -> círculo preenchido:HMM *bit score*; quadrado preenchido:*Gathering threshold*; triângulo preenchido: Comprimento mínimo de cobertura; losango preenchido:Densidade de pontuação BLAST.

3.1 Beta-lactamases

3.1.1 Obtenção e preparação dos dados

Em março de 2016, realizamos uma busca avançada no banco de dados do PDB que está disponível no NCBI (Coordinators, 2015), utilizando como *query* o número de EC para BL (3.5.2.6), e excluindo as estruturas que possuíssem a palavra *mutant* na descrição do PDB. A lista de identificadores do PDB para todas as estruturas resultantes foi recuperada (Apêndice 9.1.1). Posteriormente, esses identificadores foram usados na plataforma RCSB PDB (Burley et al., 2017) para realizar o *download* dos seus respectivos arquivos, nos formatos FASTA e PDB. O formato FASTA representa a sequência de aminoácidos, enquanto o formato PDB é uma representação padronizada dos dados estruturais da macromolécula.

Os arquivos PDB passaram por uma triagem inicial onde átomos duplicados e arquivos com resolução maior que 3Å foram descartados (Apêndices 9.1.2 e 9.1.3). Além disso, apenas os monômeros ou a cadeia A dos homomultímeros foram considerados nas análises posteriores (Apêndices 9.1.4 e 9.1.5).

Uma “Base de dados Não Redundante de Beta-lactamases” (BNRB) foi construída para validar a metodologia desenvolvida. Sequências de aminoácidos foram recuperadas de sete bancos de dados específicos em resistência aos antibióticos. Cinco bancos são específicos para BLs: CBMAR (*download* feito em agosto de 2015), Instituto Pasteur (*download* feito em outubro de 2015), LacED (*download* feito em agosto de 2015), Dlact (os autores cederam as sequências em outubro de 2015) e MBLED v1.0 (Bialek-Davenet et al., 2014; Singh and Singh, 2008; Srivastava et al., 2014; Thai et al., 2009; Widmann et al., 2012). Os outros dois bancos, CARD v1.0.0 e Resfams v1.2, possuem proteínas relacionadas com diferentes mecanismos de resistência aos antibióticos (Gibson et al., 2014; McArthur et al., 2013). Nesses casos, para recuperar apenas sequências de BLs, buscamos pelos termos “bla” e “beta” no cabeçalho das sequências, nos arquivos FASTA. Após concatenar as sequências de BLs de todos os bancos em um arquivo multi-FASTA, sequências idênticas foram

removidas usando o programa CD-HIT (Huang et al., 2010), com um *threshold* de identidade igual a 100% e valor de cobertura padrão.

3.1.2 Clusterizações

Os testes de *clusterização*[18] têm como objetivo definir os parâmetros que agrupam as BLs nas classes e subclasses definidas na classificação hierárquica de Hall & Barlow (Hall and Barlow, 2005), onde são propostos quatro níveis distintos a partir de características estruturais (Figura 1.7). Foram feitos testes com os arquivos de estruturas e com as sequências de aminoácidos correspondentes.

O programa MaxCluster (http://www.sbg.bio.ic.ac.uk/maxcluster/index.html) foi utilizado localmente para *clusterizar* as estruturas. As cadeias A foram comparadas (todas contra todas) nas seguintes condições: alinhamento independente de sequência, comparação baseada em RMSD (*Root Mean Squared Deviation*) e *clusterização* hierárquica aplicando os testes de *single*, *average* e *maximum linkage*[19].

O alinhamento independente sequência é uma implementação do método de correspondência de similaridade local entre a cadeia principal de cada uma das proteínas comparadas, disponível no programa MAMMOTH. A cadeia principal é convertida em um conjunto de vetores (Ortiz et al., 2009). O RMSD estima a distância quadrada média entre carbonos *Alfa* equivalentes de duas estruturas sobrepostas (Armougom et al., 2006). Os testes de *single*, *average* e *maximum linkage* começam agrupando o par de entidades mais próximo; esse processo se repete até que todas as entidades estejam conectadas, formando uma estrutura de árvore, que depois é separada em clusters. A diferença entre eles é a maneira como calculam a distância entre os membros do par recém-formado e as demais entidades a serem comparadas. *Single* opta pelo membro com menor distância, *maximum* pelo maior valor, enquanto *average* usa a média entre esses dois valores (Yim and Ramdeen, 2015).

Para os testes com as sequências de aminoácidos utilizamos o programa BLASTClust v2.2.26 (Wei et al., 2012). Esse programa realiza *clusterizações*

[18] Agrupamento
[19] Diferentes métodos utilizados para calcular as distâncias entre membros dos *clusters*

hierárquicas do tipo *single linkage* baseadas em alinhamentos de duas sequências a partir do algoritmo BLAST. Foram testados diferentes *thresholds* de comprimento mínimo de cobertura (*minimum length coverage*) e de densidade de pontuação BLAST (BLAST *score density*), essa última definida como a pontuação do BLAST dividida pela extensão do alinhamento. Os valores finais que foram estipulados serão apresentados na seção "Resultados". O valor de *E-value* utilizado foi 1E-05. Para executar o programa utilizando os arquivos FASTA, foram realizadas as seguintes etapas: remoção das quebras de linhas das sequências de aminoácidos, seleção da sequência correspondente à cadeia A para cada homomultímero, inserção de um número GI hipotético no cabeçalho FASTA das sequências caso não exista (exigência do programa), e remoção de linhas vazias do arquivo final. Os *scripts*[20] em Perl estão disponibilizados nessa ordem nos Apêndices 9.1.6-9.

3.1.3 Construção dos perfis HMM

Para cada *cluster*[21] correspondente às classes de BLs definidas no esquema hierárquico (SA, SC, SD, ME e MB) foi construído um perfil HMM. Uma vez que as sequências de aminoácidos pertencem a proteínas com estrutura tridimensional resolvida, oriundas do RCSB PDB, elas foram consideradas confiáveis para serem utilizadas na construção dos modelos probabilísticos.

O resultado do programa BLASTClust apresenta apenas os identificadores das sequências, por isso foi necessário desenvolver um *script* em Perl para criar arquivos multi-FASTA com todas as sequências referentes a cada *cluster* (Apêndices 9.1.10). Sequências idênticas foram removidas usando o programa CD-HIT (Huang et al., 2010), com um *threshold* de identidade igual a 100% e valor de cobertura padrão. Cada *cluster* de sequências foi então alinhado usando o *software* MUSCLE v3.8.31 com parâmetros padrões (Edgar, 2004). Como o *output* do alinhamento é um arquivo FASTA e o programa para construção dos perfis HMM exige como *input* arquivos no formato Stockholm, essa conversão foi realizado com Bioconvert v0.4

[20] Conjunto de instruções para que uma função seja executada em determinado aplicativo
[21] Grupo de sequências proteicas ou estruturas terciárias

(http://www.agapow/net/software/bioscripts.convert). Os perfis HMM foram construídos com o programa *hmmbuild* do pacote HMMER v3.1b2 (Eddy, 2011).

3.1.4 Calibração e validação dos perfis HMM

O programa *hmmsearch* do pacote HMMER v3.1b2 (Eddy, 2011) foi empregado para realizar buscas por sequências utilizando os perfis HMM construídos. Foram feitos testes para calibrar e validar esses modelos probabilísticos. A base de dados BNRB e as superfamílias de SBL e MBL do banco de dados CATH (Sillitoe et al., 2015) (3.40.710.10 e 3.60.15.10, respectivamente) foram utilizadas nesses testes. O *download* das superfamílias foi realizado em março de 2016. O valor de corte do *E-value* foi 1E-05. Inicialmente, dois testes de calibração diferentes foram criados nesse estudo para avaliar os perfis HMM, a Especificidade de Classe (EsC) e a Especificidade de Função (EsF).

Considerando a possibilidade de dois ou mais perfis HMM recuperarem a mesma sequência no banco de dados, o índice de EsC avalia se cada perfil identifica um conjunto exclusivo de sequências. Ou seja, o objetivo é que não existam interseções entre as sequências recuperadas por cada perfil HMM, e assim seja possível classifica-las. A identificação de possíveis sobreposições foi feita utilizando o resultado de cada dois perfis (Apêndice 9.1.11).

Para calcular o índice EsC de um perfil HMM qualquer, o número de sequências identificadas exclusivamente por ele (SeqEx), é dividido pelo número total de sequências identificadas por esses perfil (T_P), incluindo as interseções caso existam [EsC = (SeqEx/T_P)*100]. A base de dados BNRB foi utilizada nas buscas.

Para o modelo probabilístico cujo índice EsC foi menor que 100%, a classe da(s) sequência(s) que estava sendo recuperada como falso-positivo foi usada como um conjunto negativo de treinamento, aplicando o protocolo do programa HMM-ModE (Sinha and Lynn, 2014), que gerou um novo perfil. Os perfis modificados são utilizados com um *threshold* de discriminação, *Gathering Threshold* (GA), substituindo o *E-value*.

Uma das formas de determinar a função de uma sequência é verificar sua proximidade filogenética com outras sequências de função conhecida. As 851

sequências de aminoácidos da superfamília CATH 3.40.710.10 (SBL) foram utilizadas na construção de uma árvore filogenética não enraizada com o programa MEGA-CC v7.0.18 (Kumar et al., 2012). Os parâmetros utilizados foram o algoritmo de *Neighbour-Joining*, o modelo de Jones-Taylor-Thornton (JTT) e as regiões de *gap* foram corrigidas com deleções par a par. A árvore resultante foi editada no programa Dendroscope v3.5.7 (Huson and Scornavacca, 2012), onde as sequências de BLs foram marcadas de acordo com a sua classe. Feito isso, *scripts* em Perl foram utilizados para identificar o nó dos clados correspondentes às classes de BLs, e em seguida arquivos multi-FASTA foram construídos com as sequências em cada um desses clados (Apêndices 9.1.12 e 9.1.13).

O índice EsF indica se todas as sequências recuperadas pelo perfil HMM de uma determinada classe possuem função de BL já descrita na literatura. O número de sequências da superfamília identificadas pelo perfil cuja atividade BL é conhecida (N_F), é dividido pelo número total de sequências recuperadas pelo perfil (T_P) [$EsF = (N_F/T_P)*100$]. Nos casos em que EsF foi inferior a 100%, um novo valor de *bit score threshold* foi estabelecido como parâmetro para as buscas com *hmmsearch*, substituindo o *E-value*. Os valores estipulados de *bit score threshold* serão apresentados na seção “Resultados”.

Esse novo valor foi estabelecido a partir do maior HMM *bit score* entre as sequências da superfamília que são incapazes de hidrolisar os antibióticos beta-lactâmicos. Gráficos foram criados com o programa GraphPad Prism v5.0.3 para Windows, GraphPad Software, La Jolla California USA, (www.graphpad.com). Todas sequências da superfamília recuperadas por cada um dos perfis HMM e seus respectivos *bit scores* foram plotados. Nessa representação puderam ser destacados três grupos distintos de sequências: i) com atividade BL, ii) dentro dos clados de BLs mas sem atividade BL, e iii) fora dos clados de BLs (Apêndice 9.1.14).

No banco de dados UniProt/Swiss-Prot (Consortium, 2017), 144 sequências são atribuídas à função molecular “atividade beta-lactamase” de acordo com o *Gene Ontology*[22] (GO). O teste de validação “Sensibilidade de Função” (SeF), criado nesse estudo, avalia a habilidade de todos os perfis HMM

[22] Representação unificada dos genes e seus atributos em todas as espécies

calibrados juntos recuperarem as sequências com função “atividade beta-lactamase”. O índice SeF foi calculado dividindo o número de sequências recuperadas pelos perfis HMM que estavam associadas a função BL no UniProt/Swiss-Prot (N_{SW}), por todas as sequências do banco atribuídas a essa atividade [$SeF = (N_{SW}/144)*100$]. O download do banco de dados foi feito em março de 2016.

O resultado do programa *hmmsearch* mostra apenas os identificadores das sequências. Para criar o arquivo multi-FASTA do resultado da busca utilizamos um *script* em Perl (Apêndice 9.1.15).

3.1.5 Validação dos thresholds usados para formar as subclasses de BLs

As sequências do BNRB foram *clusterizadas* a fim de reproduzir as subclasses observadas para as sequências curadas do PDB. Para isso, as BLs do PDB foram adicionadas às sequencias do BNRB. Esse conjunto de dados foi alvo de buscas usando os perfis HMM calibrados.

Para formar subclasses a partir das classes resultantes das buscas com os modelos probabilísticos, diferentes *thresholds* de “comprimento mínimo de cobertura” foram testados. Os *thresholds* de “densidade de pontuação BLAST” já haviam sido estabelecidos usando apenas as sequências do PDB e foram mantidos. Para a anotação das sequências em cada cluster, foi realizado BLASTP v2.2.28 (Altschul et al., 1997) conta o banco de dados não redundante de proteínas do NCBI. Como o *output* do programa BLASTClust apresenta apenas os identificadores das sequências, foi necessário rodar o *script* em Perl que cria arquivos multi-Fasta referentes a cada *cluster* (Apêndice 9.1.10). O melhor *hit* de cada sequência no banco de dados foi utilizado para sua anotação (Apêndice 9.1.16).

No final do processo de validação dos *thresholds* de subclasses, os *clusters* de sequências formados foram categorizados como: i) aqueles que correspondiam às subclasses já propostas por trabalhos anteriores (Brandt et al., 2017; Hall and Barlow, 2005; Philippon et al., 2016), e ii) *clusters* contendo proteínas que não se enquadram nos critérios de similaridade e cobertura aplicados (não-BL).

Segundo Triant & Pearson (Triant and Pearson, 2015), um domínio com menos de 50% do tamanho do modelo Pfam para sua família pode ser considerado um domínio parcial. Para melhor caracterizar as sequências não-BLs, utilizamos como referência os modelos Pfam para MBL (PF00753), classe SA (PF13354) e classe SC (PF00144), além do modelo para a classe SD descrito por Pratap e colaboradores (Finn et al., 2016; Pratap et al., 2016). Esses modelos possuem 197, 202, 330 e 214 resíduos, respectivamente.

3.1.6 Confrontando os Perfis HMM aprimorados x Perfis HMM do Pfam x Patterns de BLs

Para comprovar a eficiência dos perfis HMM construídos nesse estudo na identificação e classificação de sequências de BLs, eles foram confrontados contra perfis HMM para BLs do Pfam (Finn et al., 2016) e *patterns* [23]específicos para BLs descritos em artigos e bancos de dados.

Os perfis HMM do Pfam foram utilizados em buscas contra as sequências de BLs do PDB. O índice de EsC for calculado para cada perfil do Pfam. O *download* dos sete perfis foi feito em dezembro de 2016 (Tabela 3.1).

Tabela 3.1 - Perfis HMM para BLs do Pfam utilizados para buscar sequências entre as BLs do PDB

Identificador	Família	Alvo
PF00144.22	Beta-lactamase	SBL
PF13354.4	Beta-lactamase2	SBL
PF00753.25	Lactamase_B	MBL
PF12706.5	Lactamase_B_2	MBL
PF13483.4	Lactamase_B_3	MBL
PF14597.4	Lactamase_B_5	MBL
PF16661.3	Lactamase_B_6	MBL

BL: beta-lactamase; PDB: Proteins Data Bank; SBL: Serino-beta-lactamase; MBL: Metalo-beta-lactamase. O Identificador se refere ao banco Pfam.

Identificamos *patterns* específicos para BLs descritos em artigos e bancos de dados, e estes foram avaliados quanto a sua habilidade de distinguir entre as classes e subclasses dessa atividade enzimática. Um *script* em Perl foi criado

[23] Padrão de um grupo de sequências

com a função de buscar por esses *patterns* entre as sequências de BL do PDB (Apêndice 9.1.17).

3.1.7 Identificação e classificação de BLs em genomas completamente montados

Após a calibração e validação dos perfis HMM e dos *thresholds* de "densidade de pontuação BLAST" e "comprimento mínimo de cobertura" empregados com o BLASTClust (Wei et al., 2012), a aplicabilidade da metodologia foi exemplificada através de dados de sequenciamento disponíveis. Foram utilizados 2.774 genomas bacterianos completamente montados, depositados no NCBI (Coordinators, 2015) até junho de 2016 (a identificaçao das cepas está disponível em https://github.com/melisesilveira/betaLactamase-classification.git). Após o *download* dos proteomas preditos, o diretório de cada cepa continha um ou mais arquivos que correspondiam ao(s) cromossomo(s) e plasmídio(s). Através de *scripts* em Perl os arquivos foram discriminados como proteomas cromossomais ou plasmidiais (Apêndice 9.1.18). Os proteomas cromossomais de todas as cepas foram concatenados no mesmo multi-FASTA, e o mesmo foi feito para os proteomas plasmidiais.

Usando os perfis HMM buscamos pelas cinco classes de BLs, que em seguida foram separadas em suas respectivas subclasses aplicando o programa BLASTClust (Wei et al., 2012). O filo bacteriano de origem das sequências em cada *cluster* foi determinado a partir do gênero dos isolados (Apêndices 9.1.19-21). Para as sequências presentes em bactérias do filo *Proteobacteria*, foram especificadas as classes taxonômicas às quais elas pertenciam. Um arquivo tabular com múltiplas espécies e seus filos e classes correspondentes foi construído a partir das informações contidas no *Genome Online Database* (GOLD) (Mukherjee et al., 2017). Também foi verificado quantos cromossomos codificam para pelo menos uma sequência de BL em cada subclasse (Apêndice 9.1.22).

O fluxograma para a construção da metodologia apresentada está representado na Figura 3.1. Os *scripts*, perfis HMM e instruções necessários para identificar e classificar sequências proteicas de BL segundo o fluxo de

trabalho desenvolvido aqui, estão disponíveis no link https://github.com/melisesilveira/betaLactamase-classification.git. No mesmo endereço encontram-se também as bases de dados utilizadas nas buscas.

3.1.8 Identificação dos grupos de incompatibilidade plasmidial

Para determinar o grupo de incompatibilidade dos plasmídios que carreavam alguma sequência de BL, buscamos por homólogos das proteínas iniciadoras de replicação (Rep) de seis grupos de incompatibilidade: IncF, IncW, IncI, IncN, IncP e IncH (Suzuki et al., 2010). As proteínas Rep usadas estão na Tabela 3.2.

Tabela 3.2 - Proteínas iniciadoras de replicação (Rep) usadas para atribuir grupos de incompatibilidades aos plasmídios

Proteína	Identificador	Grupo de incompatibilidade	Origem
RepB	BAA97903.1	IncFI	Plasmídio F, *E. coli* K-12
RepE	BAA97915.1	IncFI	Plasmídio F, *E. coli* K-12
RepA4	BAA78895.1	IncFII	Plasmídio R100, *Shigella flexneri* 2b
RepA1	BAA78894.1	IncFII	Plasmídio R100, *Shigella flexneri* 2b
RepHIA	AAF69874.1	IncH	Plasmídio R27, *Salmonella enterica*
RepZ	NP_863360.1	IncI	Plasmídio R64, *Salmonella enterica*
RepA	AAL13416.1	IncN	Plasmídio R46, *Salmonella enterica*
TrfA	CAK02642.1	IncP	Plasmídio pKJK5, bactéria não cultivável
RepA	YP_009182140.1	IncW	Plasmídio R388, *E. coli*

Todos os arquivos correspondentes a plasmídios foram identificados e alocados no mesmo diretório e posteriormente realizamos BLASTP v2.2.28 (Altschul et al., 1997) contra as proteínas Rep, usando *E-value* menor ou igual a 1E-5 (Suzuki et al., 2010). Para selecionar o melhor *hit* para cada plasmídio, a prioridade foi dada para aquele com *E-value* igual à zero. Quando havia mais de um, foi escolhido aquele com maior comprimento do alinhamento. Caso não existisse nenhum *E-value* igual à zero, o *hit* com menor valor de *E-value* foi escolhido (Apêndices 9.1.23 e 9.1.24).

Os arquivos dos plasmídios que carreavam as sequências de BLs identificadas no item 3.1.7 foram detectados, e seus respectivos grupos de incompatibilidade foram registrados (Apêndices 9.1.25 e 9.1.26).

3.2 Outras atividades enzimáticas

3.2.1 Obtenção e preparação dos dados

A seleção das outras atividades enzimáticas envolvidas com a resistência aos antimicrobianos foi realizada por revisão da literatura. Foram selecionadas as seguintes atividades: N3'-acetiltransferase, N6'-acetiltransferase, 2''-nucleotidiltransferase, 3'-fosfotransferase e 3''-adenililtransferase de aminoglicosídeos; esterase (hidrolase) e fosfotransferase de macrolídeos; acetiltransferase e liase de estreptograminas; oxirredutases de tetraciclina; e O-acetiltransferase de cloranfenicol. Após a seleção, a primeira etapa foi verificar quais delas possuíam um número correspondente de EC específico e completo (quatro dígitos). A partir desse número, assim como para as BLs, realizamos uma busca avançada no banco de dados de estruturas do NCBI (Coordinators, 2015), utilizando como *query* o número de EC correspondente e excluindo as estruturas que possuíssem a palavra *mutant* na descrição do PDB.

Para àquelas atividades que apresentaram um número reduzido de estruturas tridimensionais disponíveis, também foram realizadas buscas no banco de dados UniProt/TrEMBL (Consortium, 2017), utilizando como *query* o número de EC correspondente. Além disso, também buscamos por sequências curadas dessas atividades no UniProt/Swiss-Prot.

O conjunto de sequências recuperadas do UniProt/TrEMBL é muito maior e mais variável quando comparado as sequências oriundas do PDB. Foram construídos histogramas no Excel usando o tamanho das sequências recuperadas com cada número de EC, na tentativa de normaliza-las. O *script* para determinar o comprimento das sequências está no Apêndice 9.1.27. Foram selecionados os intervalos de comprimento com maior frequência de sequências (Apêndice 9.1.28). As sequências nesses intervalos foram utilizadas na etapa de *clusterização.*

3.2.2 Clusterizações

Essa etapa é baseada no que foi realizado para as BLs (item 3.1.2), porém com alvos diferentes. Para as *clusterizações* das sequências foi utilizado o programa BLASTClust v2.2.26 (Wei et al., 2012) com as “densidades de

pontuação BLAST” estabelecidas a partir das sequências de BLs. O “comprimento mínimo de cobertura” foi mantido igual a zero em todas as análises, pois não existem sistemas de classificação estrutural bem estabelecidos na literatura para essas atividades. Esses sistemas serviriam de referência para o número de *clusters* que deveriam ser formados, como foi feito para BLs.

Para a anotação das sequências em cada cluster, foi realizado BLASTP v2.2.28 (Altschul et al., 1997) conta o banco de dados não redundante de proteínas do NCBI (Coordinators, 2015). Como *output* do programa BLASTClust apresenta apenas os identificadores das sequências, foi necessário rodar o *script* em Perl que cria arquivos multi-FASTA referentes a cada *clusters* (Apêndices 9.1.10). O melhor *hit* de cada sequência no banco de dados foi utilizado para sua anotação (Apêndice 9.1.16).

4 RESULTADOS

4.1 Beta-lactamases

4.1.1 Obtenção dos dados e clusterizações a partir do PDB

Um total de 516 estruturas de BLs foram recuperadas do RCSB PDB, e após a exclusão daquelas com resolução inferior, restaram 509. Destas, 208 eram monômeros e 301 possuíam duas ou mais cadeias idênticas.

As 509 estruturas e suas respectivas sequências de aminoácidos continuaram na análise. A *clusterização* de estruturas com os testes de *single*, *average* e *maximum linkage* resultaram em sete, nove e vinte *clusters*, respectivamente (Tabela 4.1). O método de *single linkage* formou *clusters* que corespondem as cinco classes de BL do esquema de Hall & Barlow (Hall and Barlow, 2005) (Figura 1.7). As proteínas ClbP e Pab87 se dissociaram do *cluster* SC quando usamos *average* e *maximum linkage*. O *cluster* MB se dividiu em MB1 e MB2 aplicando *average linkage,* enquanto que utilizando *maximum linkage* houve várias subdivisões em todos os *clusters,* com exceção do ME. Dois *clusters* formados em todos os testes foram desconsiderados. Um deles era composto por duas estruturas da BL TEM-1 fusionada a outra proteína (Ligadora de Maltose, MBP) (4DXB e 4DXC). O outro continha cinco estruturas de PBP5, isolada de *E. coli* (3MZF, 3MZE, 3MZD, 3BEC e 3BEB). É importante lembrar que todas essas estruturas estavam associadas ao número EC de BL no PDB (3.5.2.6).

Tabela 4.1 - *Clusterização* das estruturas proteicas de BLs do PDB com o programa MaxCluster utilizando os métodos de *single, average* e *maximum linkage*

Método	N do *cluster*	N de estruturas	Classe correspondente	TOTAL
Single	1	218	SA	509
	2	119	SC	
	3	60	SD	
	4	86	MB	
	5	19	ME	
	6	2	TEM+MBP *	
	7	5	PBP5 *	
Average	1	218	SA	509
	2	115	SC	
	3	4	ClbP, Pab87	
	4	60	SD	
	5	78	MB1	
	6	8	MB2	
	7	19	ME	
	8	2	TEM+MBP *	
	9	5	PBP5 *	
Maximum	1	56	SA	509
	2	76	SA	
	3	86	SA	
	4	107	SC	
	5	8	SC	
	6	3	ClbP	
	7	1	Pab87	
	8	1	SD	
	9	49	SD	
	10	3	SD	
	11	4	SD	
	12	3	SD	
	13	60	MB1	
	14	2	MB1	
	15	9	MB1	
	16	7	MB2	
	17	8	MB2	
	18	19	ME	
	19	2	TEM+MBP *	
	20	5	PBP5 *	

N: número; **clusters* desconsiderados por não incluírem beta-lactamases. S: Serino-BL; M: Metalo-BL; SA: serino classe SA, SC: Serino classe SC; SD: Serino classe SD; ME: Metalo classe ME; MB: Metalo classe MB.

As sequências de aminoácidos correspondentes às estruturas também foram *clusterizadas*. Foram testados os seguintes *thresholds* de “densidade de pontuação BLAST”: 0, 40%, 50%, 60% e 70%. O número de *clusters* formados foram seis, oito, nove, treze e dezesseis, respectivamente (Tabela 4.2). Quando nenhum valor mínimo para densidade de pontuação BLAST foi estipulado, os *clusters* correspondem às cinco classes de BLs do esquema de Hall & Barlow

(Hall and Barlow, 2005) (Figura 1.7). Usando 40%, as proteínas ClbP e Pab87 se separam do *cluster* SC, enquanto que 50% foi o *threshold* com o qual as subclasses MB1 e MB2 se dissociaram. A “densidade de pontuação BLAST” de 60% foi responsável pela divisão das classes SA (SA1 e SA2), SD (SD1 e SD2), MB1 (MB1.1 e MB1.2) e ME (ME1 e ME2), enquanto que com 70% houve mais três divisões da classe ME e duas da classe SD. Um *cluster* formado em todos os testes foi desconsiderado por se tratar de cinco estruturas de PBP5, isolada de *E. coli*. As duas estruturas da BL TEM-1 fusionada a MBP permaneceram no *cluster* SA.

Tabela 4.2 - *Clusterização* das sequências primárias de BLs do PDB com o programa BLASTClust utilizando diferentes *thresholds* de densidade de pontuação BLAST

Densidade de pontuação BLAST	N do *cluster*	N de seq.	Classe correspondente	TOTAL
0	1	220	SA	509
	2	86	MB	
	3	19	ME	
	4	119	SC	
	5	60	SD	
	6	5	PBP 5*	
40%	1	220	SA	509
	2	86	MB	
	3	19	ME	
	4	115	SC	
	5	60	SD	
	6	1	Pab87	
	7	3	ClbP	
	8	5	PBP 5*	
50%	1	220	SA	509
	2	79	MB1	
	3	7	MB2	
	4	19	ME	
	5	115	SC	
	6	60	SD	
	7	1	Pab87	
	8	3	ClbP	
	9	5	PBP 5*	
60%	1	218	SA1	509
	2	2	SA2	
	3	77	MB1.1	
	4	2	MB1.2	
	5	7	MB2	
	6	17	ME1	
	7	2	ME2	
	8	115	SC	
	9	56	SD1	
	10	4	SD2	
	11	1	Pab87	
	12	3	ClbP	
	13	5	PBP 5*	
70%	1	218	SA1	509
	2	2	SA2	
	3	77	MB1.1	
	4	2	MB1.2	
	5	7	MB2	
	6	12	ME1	
	7	4	ME2	
	8	2	ME3	
	9	1	ME4	
	10	115	SC	
	11	56	SD1	
	12	3	SD2	
	13	1	SD3	
	14	1	Pab87	
	15	3	ClbP	
	16	5	PBP 5*	

N: número; Seq:sequências; **clusters* desconsiderados por não incluírem BLs.S: Serino-BL; M: Metalo-BL; SA: serino classe SA, SC: Serino classe SC; SD: Serino classe SD; ME: Metalo classe ME; MB: Metalo classe MB.

Como a dissociação da classe MB em MB1 e MB2 corresponde as subclasses do esquema hierárquico de Hall & Barlow (Hall and Barlow, 2005) (Figura 1.7) e a divisão das classes SA e SD já foi sugerida por trabalhos anteriores (Brandt et al., 2017; Philippon et al., 2016), os *threshold* 50% e 60% de "densidade de pontuação BLAST" foram escolhidos para formar os *clusters* das subclasses de BLs, correspondendo ao quarto e quinto níveis da classificação hierárquica, respectivamente. Os identificadores do PDB (PDB IDs) das enzimas em cada subclasse estão no Apêndice 9.2.1. Na Figura 4.1 está representada a distribuição das sequências de BLs do PDB entre os cinco níveis classificatórios, dentre outros resultados que serão apresentados posteriormente. Destacamos que entre as classes e subclasses representadas existem aquelas que corroboram trabalhos anteriores (SA1, SA2, SD1, SD2, MB1 e MB2) (Brandt et al., 2017; Hall and Barlow, 2005; Philippon et al., 2016), além de novas divisões proposta pelo presente estudo (SCD e MB1.2).

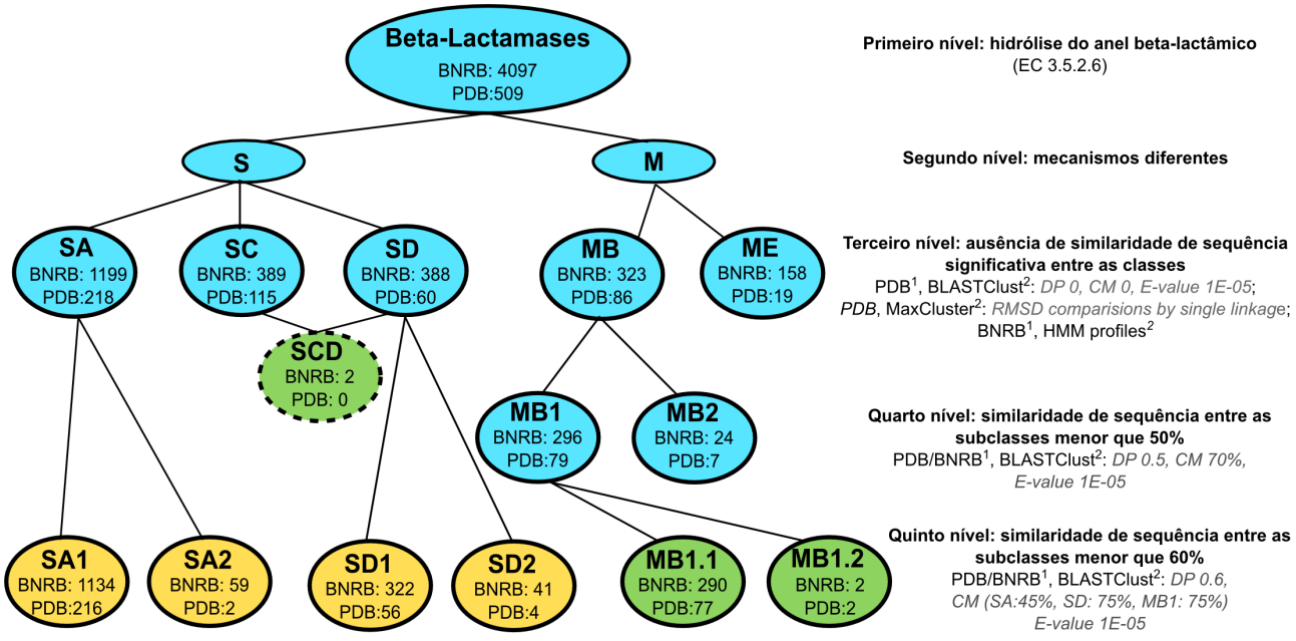


Figura 4.1: Classificação hierárquica de BLs.

Estão sendo descritos os cinco níveis da classificação hierárquica bem como o número de sequências de beta-lactamases (BL) do PDB e do BNRB em cada classe ou subclasse após as *clusterizações*. As sequências desconsideradas durante as análises não estão representadas. [1]conjunto de sequências *clusterizado*. [2] programa usado para *clusterizar*, seguido dos parâmetros aplicados. Azul: classificação de Ambler (Ambler, 1980) das BLs atualizada por Hall & Barlow (Hall and Barlow, 2005); Verde: novas classes ou subclasses propostas pela primeira vez nesse trabalho; Amarelo: subclasses descritas recentemente e confirmadas nesse trabalho (Brandt et al., 2017; Philippon et al., 2016); DP: Densidade de Pontuação BLAST; CM: Cobertura Mínima; S: Serino-BL; M: Metalo-BL; SA: serino classe SA, SC: Serino classe SC; SD: Serino classe SD; ME: Metalo classe ME; MB: Metalo classe MB.

4.1.2 Construção, calibração e validação dos perfis HMM

Os *clusters* utilizados para construção dos perfis HMM foram àqueles correspondentes as cinco classes de BLs do esquema de Hall & Barlow (Hall and Barlow, 2005) (Figura 4.1, terceiro nível). Como os resultados foram iguais tanto para o método *single linkage* utilizando estruturas como para a *clusterização* de sequências com "densidade de pontuação BLAST" igual à zero seguimos com os *clusters* de sequências, uma vez que os modelos probabilísticos são construídos a partir de um alinhamento de sequências. As duas estruturas de BLs fusionadas foram eliminadas do *cluster* SA. Os resultados de calibração e validação para os perfis de SBL (classes SA, SC e SD) e MBL (classes MB e ME) serão mostrados separadamente.

As buscas iniciais realizadas contra o BNRB resultaram em 1.292, 1.121 and 396 sequências, utilizando os perfis SA, SC e SD, respectivamente. O índice EsC (Especificidade de Classe) foi igual a 100% para o perfil de SA, e 99% para os perfis SC e SD. Haviam duas interseções entre as sequências identificadas pelos perfis SC e SD: a BL LRA-13 (ACH58991.1) e uma enzima anotada como "beta-lactamase classe C" codificada no genoma de *Janthinobacterium* sp. (WP_008451281.1), que serão comentadas na sessão 4.1.4.

Os modelos probabilísticos para as classes de MBL (MB e ME) identificaram 527 e 612 sequências no BNRB, respectivamente. O índice EsC foi 85% (MB) e 86% (ME), pois 84 sequências foram recuperadas por ambos. Novos perfis foram construídos com o protocolo do programa HMM-ModE, e esses últimos identificaram 323 (MB) e 158 (ME) sequências, sem interseções, com EsC igual a 100%.

O número de sequências identificadas com os perfis MB e ME iniciais foi maior que o número alcançado com os novos perfis calibrados. Então, os perfis HMM não calibrados foram comparados aos calibrados em buscas contra um

banco de dados exclusivo de MBLs (MBLED) (Widmann et al., 2012). Os perfis não calibrados identificaram 440 (MB, 85% EsCs) e 222 (ME, 70% EsCs) sequências, o que representa 99% do banco de dados de MBLs. Já os perfis calibrados recuperaram 424 (MB, 100% EsCs) and 164 (ME, 100% EsCs) sequências, representando 98,3% do MBLED. O 1,7% de sequências que os novos perfis calibrados não identificaram eram fragmentos proteicos de 75 a 131 aminoácidos. Isso mostra que os perfis calibrados de MBLs são capazes de identificar todas as sequências completas de MBL no banco de dados MBLED.

As relações filogenéticas entre as sequências pertencentes à superfamília de SBLs (CATH 3.40.710.10) mostraram algumas sequências incluídas nos clados de BLs sem a capacidade de hidrólise de beta-lactâmicos (Figura 4.2). PBP-A está em um ramo interno no clado da classe SA. Essa proteína possui estrutura muito similar à BL PER-1, subclasse SA2, mas com deleção de seis aminoácidos no *loop* conservado Ω e sem o ácido glutâmico 166, essencial e conhecido por estar envolvido no mecanismo de hidrólise da penicilina (Urbach et al., 2009). No clado da classe SD, a proteína regulatória BlaR1 (e sua cognata MecR1) está em um ramo mais interno que a BL OXA-1, subclasse SD2. O domínio extracelular desse regulador é fosforilado por beta-lactâmicos e consequentemente, essa proteína regula a resistência a esses antibióticos em *S. aureus* (Boudreau et al., 2016).

Quando os perfis HMM para as classes de SBLs foram testados contra sua respectiva superfamília, sequências dentro e fora dos clados de BLs foram recuperadas, incluindo as proteínas PBP-A e BlaR1. Devido a esses fatores, os índices de EsF (Especificidade de Função) foram iguais a 92,8%, 81,56% e 75,2% para as classes SA, SC e SD, respectivamente.

Foram estabelecidos *thresholds* de HMM *bit score* com os quais são recuperadas somente as sequências que realmente possuem atividade BL, excluindo 75 sequências. Na Figura 4.2 constam os gráficos com todas as sequências da superfamília identificadas por cada um dos perfis (Tabelas do Apêndice 9.2.2 a 9.2.4). Considerando as proteínas dentro dos clados de BLs, aquelas sem atividade BL possuem os menores valores de *bit score*.

Os valores de *bit score* das proteínas PBP-A, ClbP e Pab87 estavam bem distantes dos valores para as BLs (em todos com casos mais de 100 pontos). No

entanto, não foi possível separar as proteínas BlaR1/MecR1 e as BLs da subclasse SD2 nesse etapa da metodologia. O HMM *bit score* da única estrutura dessa subclasse disponível no banco de dados CATH foi muito próximo de algumas proteínas BlaR1 (227,6 para BlaR1 (1NRF) e 278,5 para OXA-1 (3ISG), por exemplo). Testes adicionais mostraram que, quando outras variantes dessa subclasse foram incluídas nas buscas, seus *bit scores* se confundiam com os de BlaR1/MecR1. Isso pode ser explicado pela similaridade estrutural do domínio extracelular de BlaR1 e das BLs da subclasse SD2 (Wilke et al., 2004).

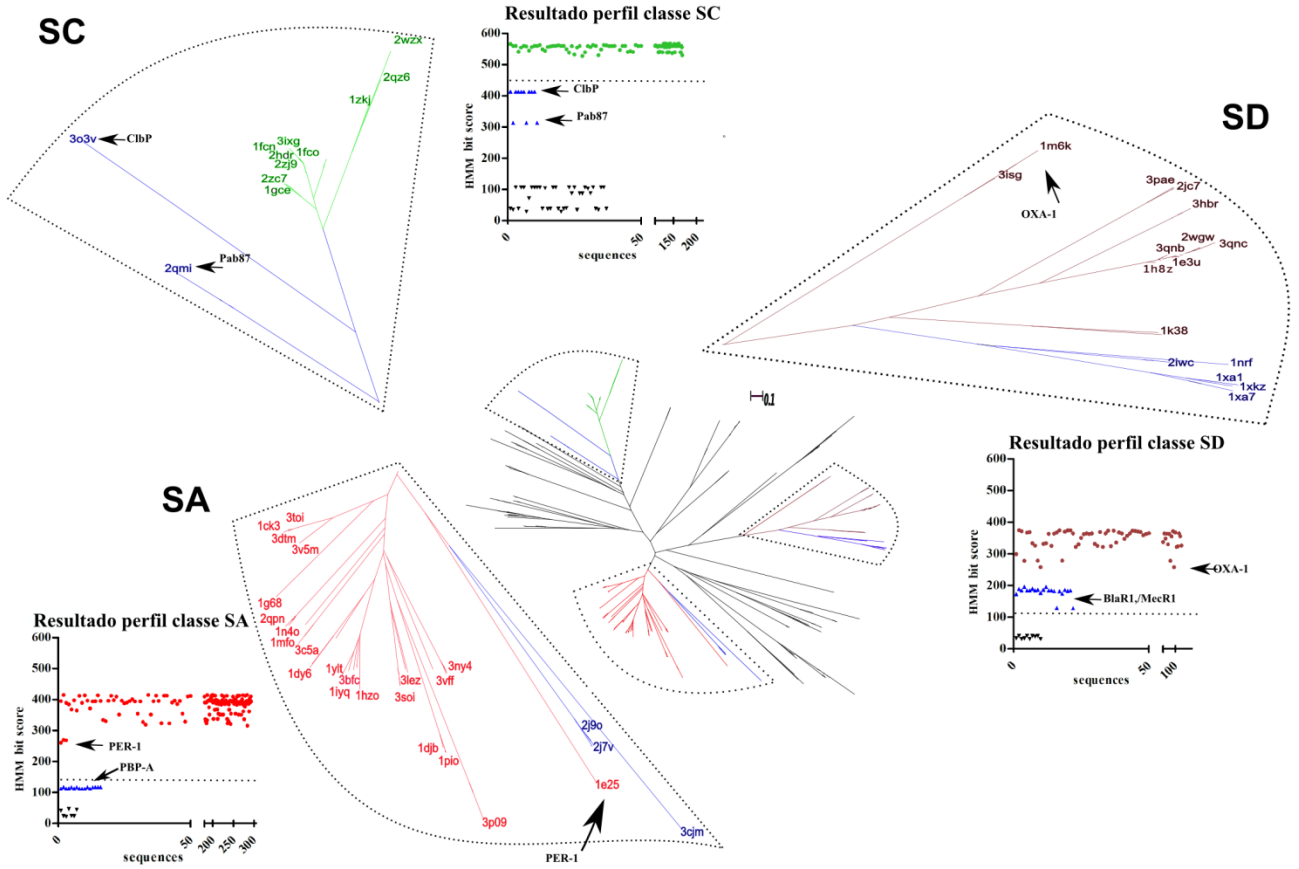


Figura 4.2: Árvore filogenética com as 851 sequências de proteína da superfamília das SBLs (CATH 3.40.710.10).

Na parte central está a árvore completa; SA: clado que inclui a classe SA; SC: clado que inclui a classe SC; SD: clado que inclui a classe SD. Nos gráficos que acompanham cada clado, os pontos representam o HMM *bit score* das sequências identificada pelo perfil da respectiva classe, e as linhas tracejadas mostram o *threshold* de HMM *bit score* estabelecido para cada perfil. X: sequências numeradas e ordenadas; Y: HMM *bit score*. Vermelho: 295 sequências de BL da classe SA; Verde: 177 sequências de BL da classe SC; Marrom: 103 sequências de BL da classe SD; Azul: Outras proteínas dentro do clado de BLs; Preto: outras proteínas da superfamília. As setas indicam proteínas específicas citadas no texto.

Os HMM *bit score thresholds* de 120, 430 e 120 foram definidos para as classes SA, SC e SD, respectivamente. Os modelos probabilísticos para as classes SA, SC e SD aliados ao novo valor de corte passaram a apresentar EsF igual a 100%, 100% e 81,2%, e recuperaram 1.199, 389 e 388 sequências do BNRB, respectivamente (Figura 4.1, terceiro nível).

Os perfis HMM para as classes MB e ME, calibrados anteriormente usando o índice EsC, recuperaram somente sequências de BL dentro da superfamília 3.60.15.10, por isso não foi necessário outro tipo de aperfeiçoamento.

Após os dois testes de calibração (EsC e EsF), os cincos perfis HMM identificaram 132 sequências de aminoácidos no UniProt/Swiss-Prot: 82, 10, 24, 15 e 1, para as classes SA, SC, SD, MB e ME, respectivamente. Dente essas, 125 são atribuídas ao termo GO “atividade beta-lactamase”, portanto o índice de SeF (Sensibilidade de Função) foi 87%. As anotações das outras 19 sequências que estão atribuídas a esse termo, mas não foram identificadas pelos perfis HMM, são: fragmentos de BLs, proteínas BL-*like*, carboxipeptidade DacA, hidroxiglutationa hidrolases, proteínas da família Hcp, PenA BL isolada de *Burkholderia cepacia* e ribonuclease. Nenhuma dessas anotações representa BLs verdadeiras ou sequências inteiras. Outras sete proteínas anotadas como BlaR1/MecR1, que não estão associadas ao termo GO para BL, foram recuperadas pelos perfis HMM da classe SD.

4.1.3 Validação dos thresholds usados para formar as subclasses de BLs

Os valores de “densidade de pontuação BLAST” para formar as subclasses de BLs, padronizados usando as sequências curadas do PDB, foram aplicados ao conjunto de dados PDB+BNRB. Foi observada uma variação importante de comprimento entre sequências do PDB e aquelas no BNRB. O tamanho das sequências de BLs do PDB varia entre 219 to 447, enquanto as sequências do BNRB possuem entre 96 e 619 aminoácidos. Como citado no item 3.1.5, o número de aminoácidos de um domínio de BL costuma variar entre 197 e 303 resíduos (Finn et al., 2016; Pratap et al., 2016). Por isso foi necessário

estabelecer *thresholds* de “tamanho de cobertura mínimo” para formar as subclasses a partir das sequências do BNRB (Figura 4.1).

Os *clusters* formados estão presentes na Tabela 4.3, com a anotação das famílias de BLs e suas variantes contidas em cada um deles. Alguns *clusters* correspondem às subclasses já propostas por trabalhos anteriores (Brandt et al., 2017; Hall and Barlow, 2005; Philippon et al., 2016), e outros *clusters* abrigam sequências chamadas de não-BLs.

As sequências não-BLs não se encaixaram nos cortes de similaridade (densidade de pontuação BLAST) e cobertura (comprimento mínimo de cobertura) estipulados para formar as subclasses do sistema hierárquico. No caso da classe SA, a maioria das sequências não-BL tem tamanho médio maior que os das subclasses SA1 e SA2 (*cluster*s não-BL1, 2 e 3). A enzima LRA-5, com atividade BL já descrita (Allen et al., 2009), foi isolada das demais sequências (*cluster* não-BL4). Não foi identificada nenhuma sequência não-BL para a classe SC. Já na classe SD, cada sequência não-BL foi separada em um *cluster* diferente, e a maioria apresenta tamanho similar às sequências de BlaR1/MecR1 (*cluster*s não-BL1, 2 e 3), ou parece se tratar de domínios parciais (próximo a 50% de 214 aminoácidos) (não-BL5 e não-BL6). Apenas uma sequência da classe SD tem tamanho semelhante ao da subclasse SD1 (não-BL4, YP_612206.1). O melhor *hit* dessa sequência no banco de dados não redundante de proteínas do NCBI apresenta 43% de identidade (*E-value* of 1E-67) com uma sequência anotada como “beta-lactamase clase D” isolada de *Oceanicaulis alexandrii*. Os *clusters* não-BL1, 2 e 3 da classe MB possuem domínios parciais (próximo a 50% de 197 aminoácidos). Dentre os dois *clusters* contendo sequências não-BL que foram formados a partir da subclasse MB1, um possui domínios parcial (não-BL 2) e o outro apresenta uma sequência de 340 aminoácidos isolada de *Stigmatella aurantiaca* (nãoBL 1, tamanho bem maior que a média da subclasse MB1, 249 aminoácidos) (Tabela 4.3). Não houve formação de subclasses para a classe ME após a *clusterização* das sequências do BNRB, mesmo com altos *thresholds* de cobertura (90%), fato que não corrobora a divisão observada quando apenas as BLs curadas do PDB foram *clusterizadas*.

Tabela 4.3 - *Clusterização* das sequências de BLs do PDB e BNRB para formar subclasses

Classe	DP	CM	Subclasses	N de enzimas (%)	Anotação segundo BLASTP	Comprimento médio
ME (177)	-	-	ME	177 (100%)	AIM-1 Asp-120 BJP-1 CAU-1 FEZ-1 GOB-(1,7,8,9,10,11,12,13,14,15,16,17,18) LRA-(2,3,8,9,12,17,19) POM-1 SMB-1	288
MB (410)	0.5	70%	MB1	376 (92%)	BlaB-(1-14) CGB-1 CcrA DIM-1 EBR-1 GIM-(1,2) IMP-(1-55) IND-(1-15) JOHN-1 KHM-1 MUS-1 NDM-(1-13) SIM-1 SLB-1 SPM-1 TMB-(1,2) TUS-1 VIM-(1-43)	249
			MB2	31 (7,4%)	ChpA ImiH ImiS Sfh-I	245
			não-BL1	1 (0,2%)	VIM-11	160
			não-BL2	1 (0,2%)	“metallo-beta-lactamase”	104
			não-BL3	1 (0,2%)	“hypothetical protein”	95
MB1 (376)	0.6	75%	MB1.1	368 (98%)	BlaB-(1,2,3,5,6,7,8,9,10,11,12,13,14) CGB-1 CcrA DIM-1 EBR-1 GIM-(1,2) IMP-(1-55) IND-(1-15) JOHN-1 KHM-1 MUS-1 NDM-(1-13) SIM-1 SLB-1 TMB-(1,2) TUS-1 VIM-(1-43)	249
			MB1.2	4 (1%)	SPM-1	254
			não-BL1	1 (0,25%)	“CcrA/NDM Family”	340
			não-BL2	3 (0,25%)	VIM-2*	111
SC (504)	0.5	45%	SC	504 (100%)	ACC-(1,2,4,5) ACT-(1-37) ADC-(1,7,8,25,26,31,67,81) AmpC CFE-1 CMY-(1 - 119) DHA-(1-22) FOX-(1-10) LAT-(1,3,4) LRA-13 MIR-(1-17) MOX-(1-8) OCH-(1-8) PDC-(1,2,3,4,5,6,7,8,9,10,16,22,23,3,35) SRT-(1,2) SST-1	383
SA (1419)	0.6	45%	SA1	1350 (95,1%)	ACI-1 AER-1 AST-1 BEL-(1,2,3) BES-1 BlaC BlaZ CARB-(1-22) CKO-1 DES-1 ERP-1 FEC-1 FONA-(1,2,3,4,5,6) FTU-1 GES-(1-26) HER-(1,2,3) IMI-(1,2,3,4,7) KLUC-(2,3,4) KLUG-(1,2,3) KPC-(1-17,22) LAP-(1,2) LEN-(1-37) LRA-1 LUT-(1,2,3,4,5,6) MAL-(1,2) OHIO-1 OIH-1 ORN-(1,2,3,4,5,6) OXY-(2,5) PLA-(3,6) PenA Pse-(1,4) RAHN-(1,2) ROB-1 SCO-1 SED-1 SFC-1 SFO-1 SGM-4 SHV-(1-189) SME-(1-5) TEM-(1-220) TER-(1,2) Toho-1 VHH-1 VHW-1	285
			SA2	61 (4,3%)	CME1 ,2 CSP-1 CblA CepA,29,44,49 CfxA PER-1-7 TLA-1 VEB-1-9,16	300
			não-BL1	3 (0,2%)	“serine hydrolase”	366
			não-BL2	2 (0,15%)	TEM-1 fusionada	637

Classe	DP	CM	Subclasses	N de enzimas (%)	Anotação segundo BLASTP	Comprimento médio
SA (1419)	0.6	45%	não-BL3	2 (0,15%)	“serine hydrolase”	345
			não-BL4	1 (0,1%)	LRA-5	326
SD (448)	0.6	75%	SD1	378 (85%)	BPU-1 LCR-1 OXA-(2,3,5-11,13-15,17,19-21,23-28,32,35,37,46,48,49,51,53-56,58,61-80,83-101,103,104,106-113,115,117-121,128-134,136-139,141-150,160-185,192-217,219,223,225,226,228-233,235-237,239-242,244-251,253-257,259,282,283,285,309,312-317,322-335,338,347-363,365,366,370,371,374-391,397,398,415,418,420-426,435,454,460,471,488,505	272
			SD2	45 (10%)	OXA-1,4,12,18,29-31,42,43,45,47,57,59,224,243,258,320	269
			BlaR1	14 (3%)	BlaR1	585
			BlaR1	3 (1,7%)	BlaR1	590
			não-BL1	1 (0,2%)	“class D beta-lactamase”	473
			não-BL2	1 (0,2%)	“class D beta-lactamase”	472
			não-BL3	1 (0,2%)	“class D beta-lactamase”	427
			não-BL4	1 (0,2%)	Beta-lactamase	274
			não-BL5	1 (0,2%)	OXA1*	172
			não-BL6	1 (0,2%)	OXA-23*	149
			SCD	2 (0,4%)	LRA-13 (classe SCD)	619

DP: Densidade de Pontuação BLAST; CM: comprimento mínimo de cobertura; coluna A: as classes que foram subdivididas; coluna D: subclasses formadas após *clusterização*; coluna E: número de enzimas em cada subclasse; coluna F: anotações encontradas para as sequências em cada subclasse; coluna H: comprimento médio das sequências em cada subclasse, em aminoácidos. O comprimento médio diz respeito ao número de aminoácidos. O símbolo “-” indica o intervalo de variantes encontrado para cada família. BL: Beta-lactamase. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SC: Serino-BL classe SC; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1.1: Metalo-BL subclasse MB1.1; MB1.2: Metalo-BL subclasse MB1.2; MB2: Metalo-BL subclasse MB2; ME: Metalo-BL classe ME. *sequências parciais.

4.1.4 Nova classe de BLs com dois domínios

A BL LRA-13 (ACH58991.1), identificada pelos perfis HMM das classes SC e SD (sem aplicação dos *thresholds* de HMM *bit score*), possui domínios de ambas as classes e um espectro de ação ampliado (Allen et al., 2009). Como recuperamos do BNRB uma proteína semelhante a essa BL (WP_008451281.1), buscamos por outros homólogos putativos no banco de dados não-redundante de proteínas do NCBI (Junho, 2016).

Oito outras proteínas foram identificadas, codificadas nos genomas de diferentes espécies bacterianas (Tabela 4.4). Essas proteínas apresentaram cobertura igual ou superior a 94% e similaridade igual ou superior a 65% em relação ao LRA-13. Todas as nove proteínas têm domínios característicos completos das classes SC (COG1680) e SD (COG2602) de acordo com o *Conserved Domain Database* (CDD, Batch CD-search tool) (Marchler-Bauer et al., 2017). Além disso, essas proteínas exibem os padrões característicos de sítio-ativo dos domínios das classes SC (PS00336) e SD (PS00337) de acordo com PROSITE (Sigrist et al., 2013), incluindo o resíduo catalítico de serina. Com isso foi sugerida uma nova classe de BLs bifuncionais, classe SCD (Figura 4.1).

As anotações originais do produto dos genes que codificam essas BLs bifuncionais putativas são "beta-lactamase classe C" ou "beta-lactamase classe D" (Tabela 4.4). Em todos os casos, o gene que codifica a proteína a montante é anotado como "beta-lactamase classe D" e seus produtos exibem um domínio característico completo da classe SD (COG2602), bem como domínios completos ou incompletos da proteína reguladora BlaR1 (COG4219 , cd07341), o que corresponde à estrutura da proteína integral de membrana transdutora de sinal que regula a resistência a beta-lactâmicos na espécie Gram-positiva *S. aureus* (Wilke et al., 2004). Os genes localizados a jusante da BL bifuncional são variáveis, dependendo da espécie: "diguanilato ciclase", "transaldolase", "glicosídeo hidrolase", “nitrito redutase” ou "proteína hipotética". Todas as cepas de *Massilia* sp. abrigam um gene que codifica para um regulador transcricional a montante do gene da BL bifuncional, mas em uma orientação invertida, que está envolvido na regulação da expressão do regulador BlaR1, também invertido em relação à BL bifuncional (Figura 4.3).

Tabela 4.4 - Genomas, anotação original e contexto genômico dos genes que codificam as BL da classe SCD

		BL bifuncional		Produto do gene a montante		Produto do gene a jusante	
Cepa	**Acesso**	**Anotação**	**Acesso**	**Anotação**	**Acesso**	**Anotação**	**Acesso**
Bactéria não-cult BLR13	EU408352.1	BL bifunctional	ACH58991.1	Regulador	ACH58992.1	glicosídeo hidrolase	ACN58887.1
Janthinobacterium sp. HH01	NZ_AMWD01000002.1	BL classe C	WP_008451281.1	BL classe D	WP_008451277.1	transaldolase	WP_008451283.1
Massilia sp. Root418	LMEC01000020.1	Proteína hipot.	KQW93884.1	BL classe D	KQW93885.1	diguanilato ciclase	KQW93883.1
Massilia sp. Root351	NZ_LMDJ01000033.1	BL classe C	WP_082552146.1	BL classe D	WP_057157847.1	diguanilato ciclase	WP_057157849.1
Massilia sp. CF038	FQWU01000002.1	BL classe C	SHH20105.1	BL classe D	SHH20059.1	proteína hipotética	SHH20125.1
Duganella sp. HH105	LRHV01000029.1	BL classe C	OEZ55387.1	BL classe D	OEZ55388.1	transaldolase	OEZ55386.1
Duganella sp. CF458	FOOF01000012.1	BL classe D	SFG43659.1	BL classe D	SFG43677.1	nitrito redutase	SFG43637.1
Duganella sp. Root198D2	NZ_LMIC01000034.1	BL classe C	WP_082591432.1	BL classe D	WP_082591444.1	nitrito redutase	WP_082507115.1
Duganella sp. Root336D2	NZ_LMDB01000002.1	BL classe C	WP_082507116.1	BL classe D	WP_082507139.1	nitrito redutase	WP_082507115.1
Duganella sp. Root1480D1	NZ_LMFZ01000003.1	BL classe C	WP_082565248.1	BL classe D	WP_082565235.1	nitrito redutase	WP_082565234.1

Não-cult: não-cultivável; Hipot: hipotética. A anotação original é referente ao GenBank.

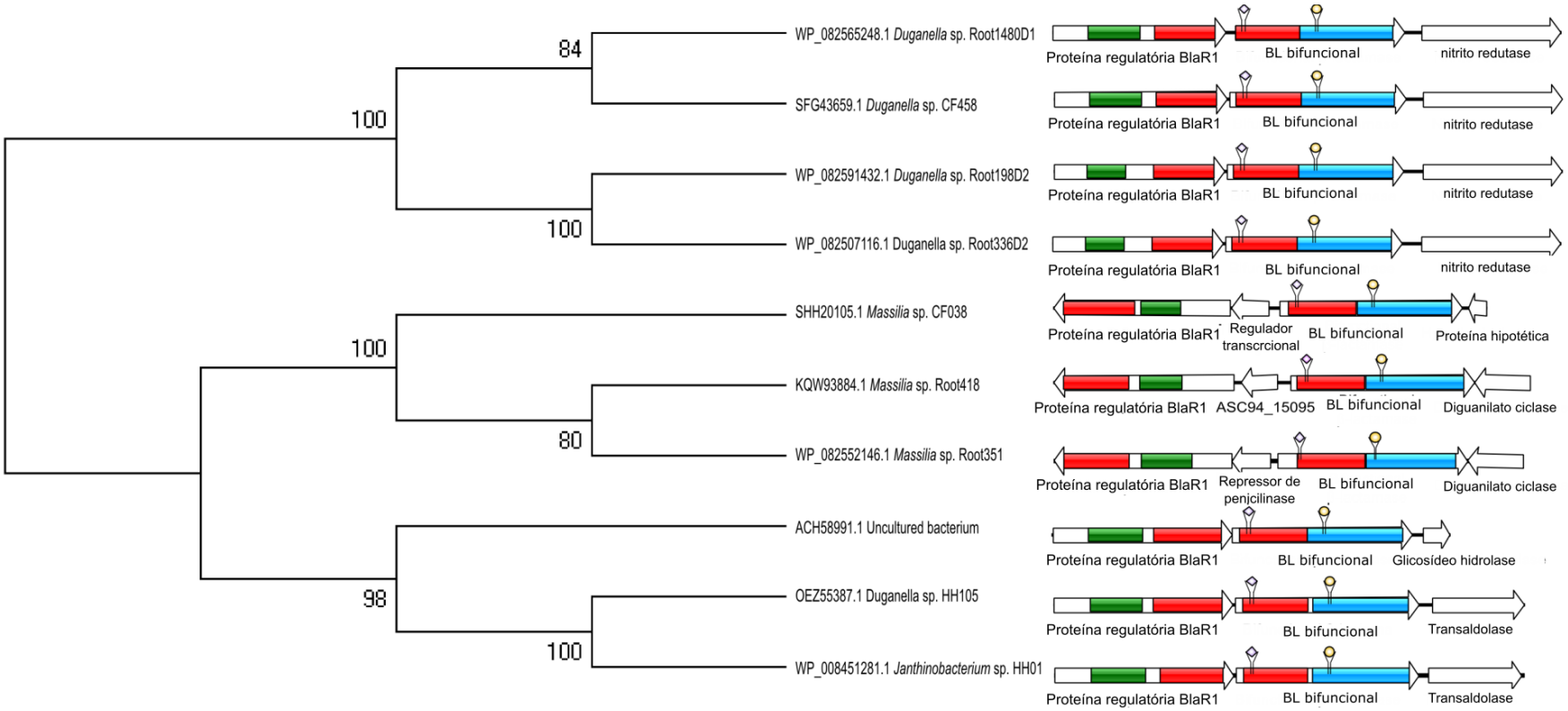


Figura 4.3: Padrão filético do genes codificadores de BLs bifuncionais da classe SCD e seu contexto genômico.

Esquerda: Dendrograma representando as relações de similaridade entre as sequências proteicas das BLs bifuncionais. Direita: representação da ordem e orientação dos genes codificantes das BLs bifuncionais e os genes ao entorno. Retângulos: domínios; vermelho: classe SD, verde: BlaR1; Azul: classe SC. Setas indicam a orientação dos genes. Losangos e círculos representam os sítios-ativos das classes SD e SC, respectivamente. As sequências foram alinhadas globalmente utilizando a ferramenta MAFFT v7 (Katoh et al., 2017). O dendrograma foi construído com o programa MEGA v7 (Kumar et al., 2012), aplicando o algoritmo de *Neighbour-Joining* e 500 réplicas de bootstrap.

4.1.5 Confrontando: Perfis HMM desse estudo x Perfis HMM do Pfam x Patterns de BLs

Os perfis HMM disponíveis no Pfam para BLs são destinados às SBLs ou MBLs. Assim, nenhum deles foi capaz de distinguir entre classes (SA, SC, SD, MB e ME). Os perfis para SBLs (PF00144.22 e PF13354.4) identificaram 339 (EsC 35%) e 227 (EsC 3%) sequências, respectivamente, de um total de 399 disponíveis no conjunto de BLs do PDB (Figura 4.4 A). Os três perfis para MBLs (PF00753.25, PF12706.5, e PF16661.3) recuperaram 98, 35 e 12 sequências de um total de 105 MBLs disponíveis (EsC igual a 64%, 0% e 0%, respectivamente) (Figura 4.4 B). Dois outros perfis HMM para MBLs não identificaram nenhuma enzima entre as BLs curadas do PDB (PF13483.4 e PF14597.4).

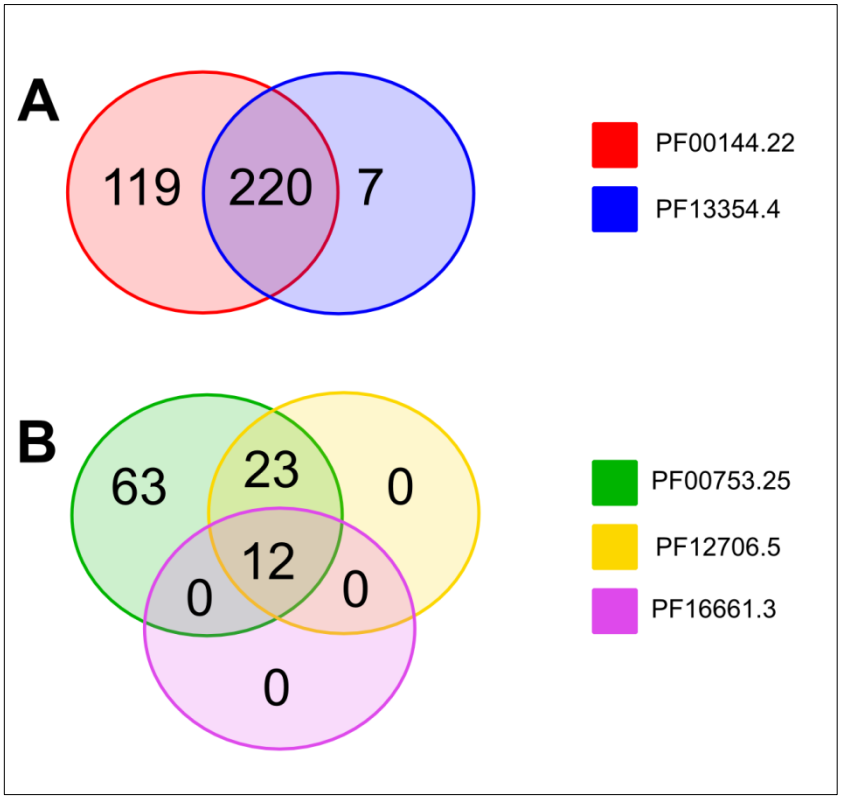


Figura 4.4: Diagrama de Venn representando o número de sequências recuperadas pelos os perfis HMM do Pfam para SBLs (A) e MBLs (B).
Nome dos perfis HMM no Pfam -> PF00144.22: Beta-lactamase; PF13354.4: Beta-lactamase2; PF00753.25: Lactamase_B; PF12706.5: Lactamase_B_2; PF16661.3: Lactamase_B_6.

A partir de fontes variadas, 14 *patterns* referentes às diferentes classes de BLs foram selecionados. Buscamos por eles entre as 509 sequências do PDB. Doze *patterns* são específicos para o terceiro nível da classificação hierárquica (classes SA,

SC e SD), enquanto outros dois *patterns* são específicos apenas para o segundo nível (MBLs). Todos foram testados quanto à capacidade de distinguir entre as classes e subclasses de BLs. O número de sequências que possuem o *pattern* foi dividido pelo total de sequências já classificadas anteriormente pelos perfis HMM calibrados na classe/subclasse respectiva (Tabelas 4.5 e 4.6).

Tabela 4.5 - *Patterns* para as classes de SBL presentes entre as sequências de BLs do PDB

Patterns	Alvo	Classe	Subclasse 1*	Subclasse 2**
ExxLN[a]	SA	174/218 (80%)	174/216 (81%)	0/2 (0%)
SDN[a]	SA	183/218 (84%)	181/216 (83%)	2/2 (100%)
KTG[a]	SA	169/218 (78%)	167/216 (77%)	2/2 (100%)
KSG[a]	SA	44/218 (20%)	44/216 (20%)	0/2 (0%)
S-[DG]-N-x(1,2)-A-[ACGNST]-x(2)-[ILMV]-x(4)-[AGSTV][b]	SA	107/218 (49%)	105/216 (48%)	2/2 (100%)
[FY]-x-[LIVMFY]-{E}-S-[TV]-x-K-x(3)-{T}-[AGLM]-{D}-{KA}-[LC][c]	SA	192/218 (88%)	190/216 (87%)	2/2 (100%)
YxN[d]	SC	116/119 (97%)	-	-
KxxS[e]	SC	119/119 (100%)	-	-
[FY]-E-[LIVM]-G-S-[LIVMG]-[SA]-K[c]	SC	118/119 (99%)	-	-
SxV[d]	SD	60/60 (100%)	56/56 (100%)	4/4 (100%)
SxxxxS[e]	SD	50/60 (83%)	46/56 (82%)	4/4 (100%)
[PA]-x-S-[ST]-F-K-[LIV]-[PALV]-x-[STA]-[LI][c]	SD	43/60 (72%)	41/56 (73%)	2/4 (50%)

a: (Ambler, 1980); b: (Singh et al., 2009); c: (Sigrist et al., 2013); d: (Bush, 2013); e: (Holliday et al., 2012). Alvo: classe de BL para qual o *pattern* foi feito.*: subclasses SA1 e SD1. **: subclasses SA2 e SD2. – representa as classes que não têm subclasses.

Tabela 4.6 - *Patterns* para MBL presentes entre as sequências de BLs do PDB

Patterns	Alvo	MBL	ME	MB	MB1	MB2
[LI]-x-[STN]-[HN]-x-H-[GSTAD]-D-x(2)-G-[GP]-x(7,8)-[GS][c]	MBL	55/105 (52%)	12/19 (63%)	43/86 (51%)	36/79 (46%)	7/7 (100%)
P-x(3)-[LIVM](2)-x-G-x-C-[LIVMF](2)-K[c]	MBL	46/105 (44%)	0	46/86 (54%)	39/79 (50%)	7/7 (100%)

c: (Sigrist et al., 2013). Alvo: classe de BL para qual o *pattern* foi feito. MBL: Metalo-beta-lactamase; ME: Metalo-BL classe ME; MB: Metalo-BL classe MB; MB1: Metalo-BL subclasse MB1; MB2: Meta-BL subclasse MB2.

Com esses resultados é possível notar que nenhum *pattern* para MBLs ou para a classe SA está presente em todas as sequências do seu respectivo grupo. O *pattern* mais comum para as classes SA, ME e MB estão presentes em 88% ([FY]-x-[LIVMFY]-{E}-S-[TV]-x-K-x(3)-{T}-[AGLM]-{D}-{KA}-[LC]), 63% ([LI]-x-[STN]-[HN]-x-H-[GSTAD]-D-x(2)-G-[GP]-x(7,8)-[GS]) e 54% (P-x(3)-[LIVM](2)-x-G-x-C-[LIVMF](2)-K) de suas respectivas sequências. O *pattern* KxxS (Holliday et al., 2012) foi encontrado em todas as sequências da classe SC, enquanto o *patter* SxV (Bush, 2013) está presente em todos os membros da classe SD. Para as classes SA, SC, SD, ME e MB, os *patterns* menos frequentes

estão em 20% (KSG), 97% (YxN), 72% ([PA]-x-S-[ST]-F-K-[LIV]-[PALV]-x-[STA]-[LI]), 0% (P-x(3)-[LIVM](2)-x-G-x-C-[LIVMF](2)-K) e 51% ([LI]-x-[STN]-[HN]-x-H-[GSTAD]-D-x(2)-G-[GP]-x(7,8)-[GS]) de suas respectivas sequências. Nenhum dos *patters* analisados foi caaz de distinguir entre as subclasses de BLs, pois não estavam presentes em todas as sequências de apenas uma das subclasses (MB1/MB2, SA1/SA2, SD1/SD2).

4.1.6 Identificação e classificação de BLs em genomas completamente montados

Foram identificadas 1.476 sequências de BLs entre as 2.774 cepas de bactérias estudadas. Os perfis HMM para as classes SA, SC, SD, MB e ME recuperaram 616, 280, 366, 103 e 111 sequências, respectivamente. Destas, 1.362 são cromossômicas e 114 são plasmidiais. Nenhuma sequência foi identificada pelos perfis das classes SC e SD simultaneamente, portanto não foi encontrado nenhum membro da classe SCD.

Depois do processo de agrupamentos, 8,3% das sequências foram consideradas não-BLs e por isso descartadas (Tabela 4.7). Um total de 91% (58 sequências) das não-BLs da classe SD foram identificadas em cepas do filo *Firmicutes*.

Tabela 4.7 - Sequências identificadas pelos perfis HMM e consideradas não-BL após os processos de *clusterização* e consequente formação das subclasses

Filo	Localização	SC	SD	MB	Total (C)	Total (P)
		não-BL	não-BL	não-BL		
Proteobacteria	C	21	6	2	29	
	P	4	1			5
Firmicutes	C		58	5	63	
	P		22			22
Bacteroidetes	C	1			1	
Cyanobacteria	C	2			2	
Planctomycetes	C					
Fusobacteria	C		1		1	
		28	88	7	96	27

C: cromossomal; P: plasmidial; Total: número de sequencias de beta-lactamases (BLs) identificadas. SC: Serino-BL classe SC; SD: Serino-BL classe SD; MB: Metalo-BL classe MB.

As demais sequências foram classificadas de acordo com as subclasses de BLs à qual pertencem. Para aquelas localizadas em cromossomos (Tabela 4.8), *Proteobacteria* foi o único filo onde encontramos todas as diferentes subclasses, com exceção de MB1.2, encontrada apenas em *Spirochaetes*. A subclasse SD1 e classe ME foram as mais disseminadas entre filos diferentes. Já a classe SC e as subclasses

SD2 e MB2 estão praticamente restritas às *Proteobacteria*. Mais de 99% das sequências na subclasse SA1 se distribui entre os filos *Proteobacteria*, *Firmicutes* e *Actinobacteria*, enquanto a distribuição de sequências da subclasse SA2 tem destaque para o filo *Bacteroidetes* (61%). Para as MBLs, a subclasse MB1 está presente principalmente no filo *Firmicutes* (52%), enquanto para a classe ME o filo *Proteobacteria* é majoritário (82%). Ao todo, 66% das sequências de BLs foram encontradas em *Proteobacteria*. Entre todos os filos onde foi identificada alguma BL, 49% dos genomas analisados pertencem às *Proteobacteria*. Apenas um genoma dos 106 analisados para o filo *Chlamydiae* carreia uma sequência de BL, subclasse SD2. Apesar de terem sido estudados apenas cinco genomas do filo *Acidobacteria*, foram identificadas oito sequências BLs, subclasses SA1 e SA2 e classe ME.

Tabela 4.8 - Distribuição das sequências cromossômicas de BLs entre filos bacterianos

Filo	N genom.	SA1	SA2	SC	SD1	SD2	MB1.1	MB1.2	MB2	ME	Total
Proteobacteria	1176	263	4	261	129	61	22		5	91	836
Firmicutes	583	127			30		46			3	206
Actinobacteria	283	107		5	4					1	117
Bacteroidetes	88		17		14		18			3	52
Cyanobacteria	73		2		17						19
Spirochaetes	60		1		3		2	1		7	14
Chlamydiae	106					1					1
Chlorobi	11				7						7
Fusobacteria	8				2					1	3
Acidobacteria	5	2	3							3	8
Verrucomicrobia	4		1							1	2
Gemmatimonadetes	1									1	1
Total	2398	499	28	266	206	62	88	1	5	111	1266

N genom.: número de genomas analisados; Total: número de sequencias de beta-lactamases (BLs) identificadas. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SC: Serino-BL classe SC; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1.1: Metalo-BL subclasse MB1.1; MB1.2: Metalo-BL subclasse MB1.2; MB2: Metalo-BL subclasse MB2; ME: Metalo-BL classe ME.

A distribuição de BLs entre as classes de *Proteobacteria* foi analisada devido ao elevado número de sequências encontradas nesse filo (Tabela 4.9). A classe *Gammaproteobacteria* concentra 60% das BLs, divididas entre todas as subclasses. SD1 é a única subclasse de BL encontrada em todas das classes de *Proteobacteria*, incluindo as *Epsilonproteobacteria*. A classe *Deltaproteobacteria* parece ter mais importância para a subclasse MB1.

Tabela 4.9 - Distribuição das sequências cromossômicas de BLs entre classes de Proteobacteria

Classe	N genom.	SA1	SA2	SC	SD1	SD2	MB1.1	MB2	ME	TOTAL
Gammaproteobacteria	549	116	2	211	74	17	11	4	64	499
Alphaproteobacteria	320	68	2	20	15	15	2		22	144
Betaproteobacteria	145	76		30	13	25		1	4	149
Epsilonproteobacteria	103				21					21
Deltaproteobacteria	59	3			6	4	9		1	23
Total	1176	263	4	261	129	61	22	5	91	836

N genom.: número de genomas analisados; Total: número de sequencias de beta-lactamases (BLs) identificadas. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SC: Serino-BL classe SC; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1.1: Metalo-BL subclasse MB1.1; MB2: Metalo-BL subclasse MB2; ME: Metalo-BL classe ME

Uma única cepa bacteriana pode carrear várias BLs. Sabendo disso, analisamos quantos cromossomos codificam para pelo menos uma sequência de BL em cada subclasse. Dessa forma cada cromossomo é considerado uma única vez na contagem, mesmo que ele tenha várias sequências da mesma subclasse (Tabela 4.10). O padrão da distribuição das subclasses entre os filos continua o mesmo, assim como os filos onde mais se encontram BLs. Porém dessa forma é possível calcular a porcentagem de genomas de cada filo que carreiam BLs das diferentes subclasses. As maiores porcentagens para cada filo então assinaladas na Tabela 4.11. Em *Firmicutes* e *Actinobacteria* a subclasse SA1 é a mais comum (17,8% e 33,8%, respectivamente), enquanto em *Proteobacteria* a classe SC é a mais encontrada (22,6%). Para os filos *Chlorobi* e *Cyanobacteria*, 63,6% e 20,5% dos genomas possuem no mínimo uma BL da subclasse SD1, respectivamente.

Tabela 4.10 - Número de genomas codificando no mínimo uma sequência de BL em cada subclasse

Filo	N genomas	SA1	SA2	SC	SD1	SD2	MB1.1	MB1.2	MB2	ME	Total
Proteobacteria	1176	222	4	255	119	59	21		5	88	773
Firmicutes	583	104			28		46			2	180
Actinobacteria	283	96		3	4					1	104
Bacteroidetes	88		16		11		18			3	48
Cyanobacteria	73		2		15						17
Spirochaetes	60		1		3		2	1		7	14
Chlamydiae	106					1					1
Chlorobi	11				7						7
Fusobacteria	8				1					1	2
Acidobacteria	5	2	3							3	8
Verrucomicrobia	4		1							1	2
Gemmatimonadetes	1									1	1
Total	2398	424	27	258	188	60	87	1	5	107	1157

N genom.: número de genomas analisados; Total: número de sequencias de beta-lactamases (BLs) identificadas. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SC: Serino-BL classe SC; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1.1: Metalo-BL subclasse MB1.1; MB1.2: Metalo-BL subclasse MB1.2; MB2: Metalo-BL subclasse MB2; ME: Metalo-BL classe ME.

Tabela 4.11 - Porcentagem de genomas por filo codificando no mínimo uma sequência de BL em cada subclasse

	SA1	**SA2**	**SC**	**SD1**	**SD2**	**MB1.1**	**MB1.2**	**MB2**	**ME**
Proteobacteria	18,9%	0,3%	**21,7%**	10,1%	5%	1,8%		0,4%	7,5%
Firmicutes	**17,8%**			4,8%		7,9%			0,3%
Actinobacteria	**33,9%**		1,1%	1,4%					0,3%
Bacteroidetes		18,2%		12,5%		**20,4%**			3,4%
Cyanobacteria		2,7%		**20,5%**					-
Spirochaetes		1,7%		5%		3,3%	1,7%		**11,7%**
Chlamydiae					**0,9%**				-
Chlorobi				**63,7%**					-
Fusobacteria				**12,5%**					**12,5%**
Acidobacteria	**40%**	**60%**							**60%**
Verrucomicrobia		**25%**							**25%**
Gemmatimonadetes									**100%**

BL: Beta-lactamases. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SC: Serino-BL classe SC; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1.1: Metalo-BL subclasse MB1.1; MB1.2: Metalo-BL subclasse MB1.2; MB2: Metalo-BL subclasse MB2; ME: Metalo-BL classe ME

Entre os 2.162 plasmídios analisados foram encontradas 87 sequências de BLs, distribuídas ente os filos *Proteobacteria* e *Firmicutes* (Tabela 4.12). A subclasse SA1 foi a mais encontrada e a única presente em *Firmicutes*. A classe SC foi a segunda mais prevalente entre os plasmídios. Não foram encontradas sequências plasmidiais de BLs de MB1.2, MB2 e ME.

Tabela 4.12 - Distribuição das sequências plasmidiais de BLs entre filos bacterianos

Filo	**N plasmídios**	**SA1**	**SA2**	**SC**	**SD1**	**SD2**	**MB1.1**	**Total**
Proteobacteria	976	39	1	14	3	7	2	66
Firmicutes	472	21						21
Total	1448	60	1	14	3	7	2	87

N plasmídios: número de plasmídios analisados; Total: número de sequencias de beta-lactamases (BLs) identificadas. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SC: Serino-BL classe SC; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1.1: Metalo-BL subclasse MB1.1.

4.1.7 Anotação das sequências de BLs identificadas nos genomas

Todas as sequências recuperadas pelos perfis HMM foram atribuídas a uma classe ou subclasse de BL, após a exclusão das não-BL. Considerando suas anotações originais presentes no banco de dados do NCBI, 70% eram designadas apenas como “beta-lactamases”; 24% possuíam informação sobre o segundo (SBL e MBL) ou terceiro (classes) nível da classificação hierárquica, ou ainda o nome do gene; e também existiam 6% com anotação errada ou imprecisa, como por exemplo, “proteína hipotética” (Figura 4.5).

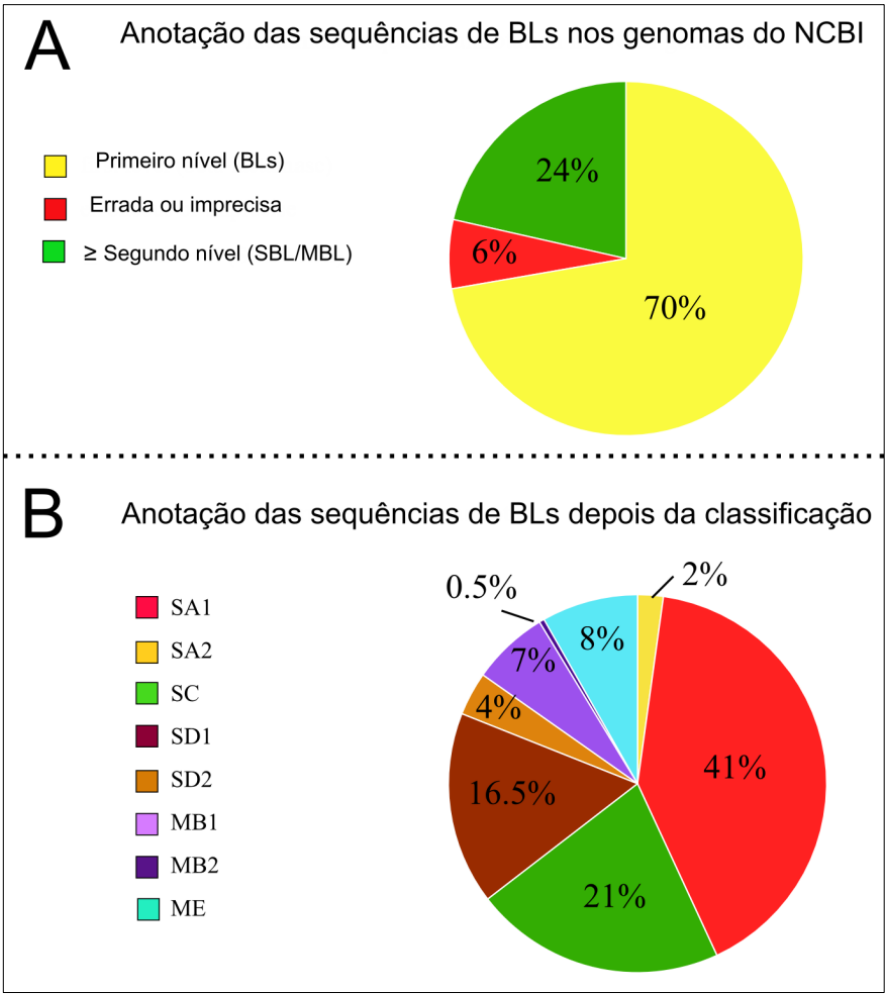


Figura 4.5 - Anotação original (A) e reanotação (B) das 1.363 sequências classificadas nesse estudo seguindo a classificação hierárquica das BLs.
*0,5% na parte B se refere à porcentagem de enzimas da subclasse MB2. BL: Beta-lactamase. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SC: Serino-BL classe SC; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1: Metalo-BL subclasse MB1; MB2: Metalo-BL subclasse MB2; ME: Metalo-BL classe ME

4.1.8 Identificação dos grupos de incompatibilidade plasmidial

Dos 2.162 plasmídios analisados nesse estudo, atribuímos o grupo de incompatibilidade plasmidial à 457 deles. Das 87 sequências plasmidiais de BLs

(Tabela 4.12), foi possível inferir o grupo de incompatibilidade de 30 (35%) plasmídios carreadores (Tabela 4.13). O grupo de incompatibilidade mais encontrado foi IncF. Também foram encontradas BLs da subclasse SA1 em plasmídeos IncH, IncI, IncN e IncW. O grupo de incompatibilidade plasmidial mais encontrado carreando sequências da subclasse SD2 foi IncW. Não foi possível atribuir o grupo de incompatibilidade para nenhum dos plasmídios carreadores de BLs da classe SC.

Tabela 4.13 – Grupo de incompatibilidade dos plasmídios carreadores de BLs

Subclasse	IncF	IncH	IncI	IncN	IncW	Total
SA1	16	2	1	1	2	22
SA2	1					1
SD1	1					1
SD2	1				4	5
MB1	1					1
Total	20	2	1	1	6	30

Inc: Grupo de Incompatibilidade Plasmidial. Total: número de plasmídios identificados. SA1: Serino-BL subclasse SA1; . SA2: Serino-BL subclasse SA2; SD1: Serino-BL subclasse D1; SD2: Serino-BL subclasse SD2; MB1: Metalo-BL subclasse MB1.

4.2 Outras atividades enzimáticas

4.2.1 Obtenção e preparação dos dados

O processo de *clusterização* foi utilizado para estudar outros grupos de enzimas responsáveis pela resistência aos antibióticos, com os mesmos critérios de similaridade aplicados às BLs. Com isso foi possível analisar a diversidade das enzimas de cada atividade sob uma perspectiva comum.

Após revisão da literatura, 14 atividades enzimáticas envolvidas com a resistência aos antimicrobianos foram selecionadas. Entre elas estão transferases de aminoglicosídeos, macrolídeos, lincosamidas, streptogramíneas A, rifampicina e cloranfenicol; hidrolase de macrolídeos; liase de streptogramíneas B; e oxirredutases de tetraciclinas e nitroimidazol (Tabela 4.14).

Tabela 4.14 - Atividades enzimáticas envolvidas com a resistências a diferentes classes de antibióticos selecionadas após revisão da literatura

	Atividade enzimática
1	nucleotidiltransferases de lincosamidas
2	oxirredutases de nitroimidazol
3	ADP-ribosil transferase de rifampicina
4	N3'-acetiltransferase de aminoglicosídeos
5	N6'-acetiltransferase de aminoglicosídeos
6	nucleotidiltransferase de aminoglicosídeos
7	3'-fosfotransferase de aminoglicosídeos
8	3''-adenililtransferase de aminoglicosídeos
9	esterase (hidrolase) de macrolídeos (Ere)
10	fosfotransferase de macrolídeos (Mph)
11	acetiltransferase de estreptograminas A
12	liase de estreptograminas A (Vgb)
13	transfease de fosfomicina (FosA)
14	acetiltransferase de cloranfenicol (CAT)

Não foi possível identificar o número de EC específicos para as nucleotidiltransferases de lincosamidas (Lnu), para as oxirredutases de nitroimidazol (Nim), nem para a ADP-ribosil transferase de rifampicina (ARR).

Entre as 11 atividades enzimáticas restantes, três possuíam número de EC incompleto: esterase de aminoglicosídeos (Ere), acetiltransferase de streptogramíneas A, e liase de streptogramíneas B (Vbg). Utilizando os 8 números de EC completos, foram realizadas buscas nos bancos de dados RCSB PDB, UniProt/TrEMBL e UniProt/Swiss-Prot. As buscas por estruturas teve como resultado um número muito menor que aquele observado para BLs (516). A atividade com maior número de estruturas foi a 3'-fosfotransferase de aminoglicosídeos (17). Por outro lado, a busca por sequências no UniProt/TrEMBL mostrou mais resultados, variando de 11 a 1.705 sequências disponíveis para cada atividade. No banco de dados curado do UniProt/Swiss-Prot a disponibilidade de sequências foi menor, variando de 3 a 29 sequências por atividade. A oxirredutase de tetraciclinas TetX e as fosforilases de macrolídeos Mph, números de EC 1.14.13.231 e 2.7.1.136, não tiveram resultado em nenhum dos bancos pesquisados (Tabela 4.15).

Tabela 4.15 - Outras atividades enzimáticas envolvidas com a resistência aos antimicrobianos com número de EC associado

	Atividade	Classe de antibiótico	EC	PDB	UniProt	Swiss-Prot
1	N3'-acetiltransferase	aminoglicosídeos	2.3.1.81	2	173	10

2	N6'-acetiltransferase	aminoglicosídeos	2.3.1.82	13	687	12
3	2''-nucleotidiltransferase	aminoglicosídeos	2.7.7.46	2	11	3
4	3'-fosfotransferase	aminoglicosídeos	2.7.1.95	17	440	12
5	3''-adenililtransferase	aminoglicosídeos	2.7.7.47	0	1149	8
6	Hidrolase, Esterase Ere	macrolideos	3.1.1.-	-	-	-
7	Transferase, Fosforilases Mph	macrolídeos	2.7.1.136	0	0	0
8	Acetiltransferase	streptogramíneas A	2.3.1.-	-	-	-
9	Liase Vgb	streptogramíneas B	4.2.99.-	-	-	-
10	TetX	tetraciclina	1.14.13.231	0	0	0
11	O-acetiltransferase CAT	cloranfenicol	2.3.1.28	9	1705	29

Devido a quantidade de dados do PDB e do UniProt/Swiss-Prot ser muito pequena, seguimos com as sequências do Uniprot/TrEMBL. Foram construídos histogramas com o comprimento das sequências recuperadas do UniProt/TrEMBL para cada atividade enzimática (Figuras 4.6 - 4.11). A partir desses resultados, foram selecionados os intervalos de comprimento com maior frequência de sequências. Em todos os casos, o número de sequências no intervalo selecionado foi superior a 90% do número inicial de sequências baixadas (Tabela 4.16).

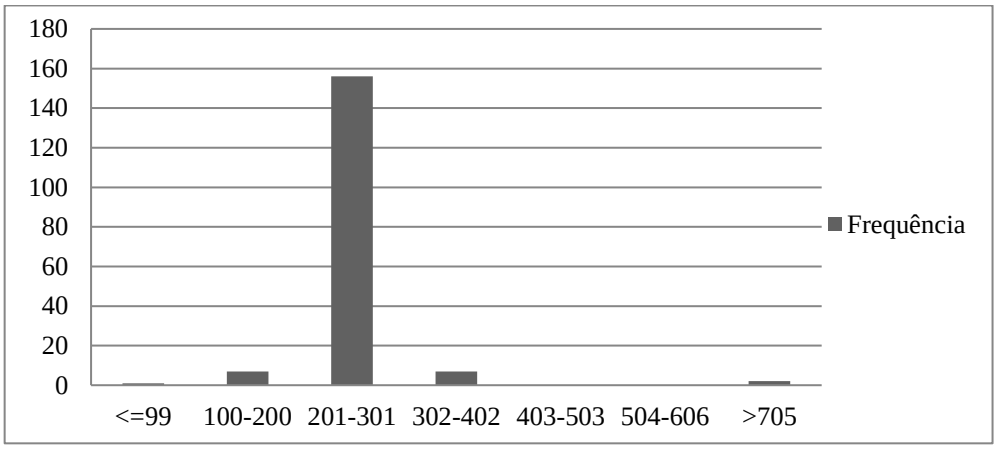


Figura 4.6: : Histograma do comprimento em aminoácidos das sequências disponíveis no UniProt/TrEMBL para o EC 2.3.1.81 (N3'-acetiltransferase de aminoglicosídeos).

X: número de sequências; Y: comprimento das sequências.

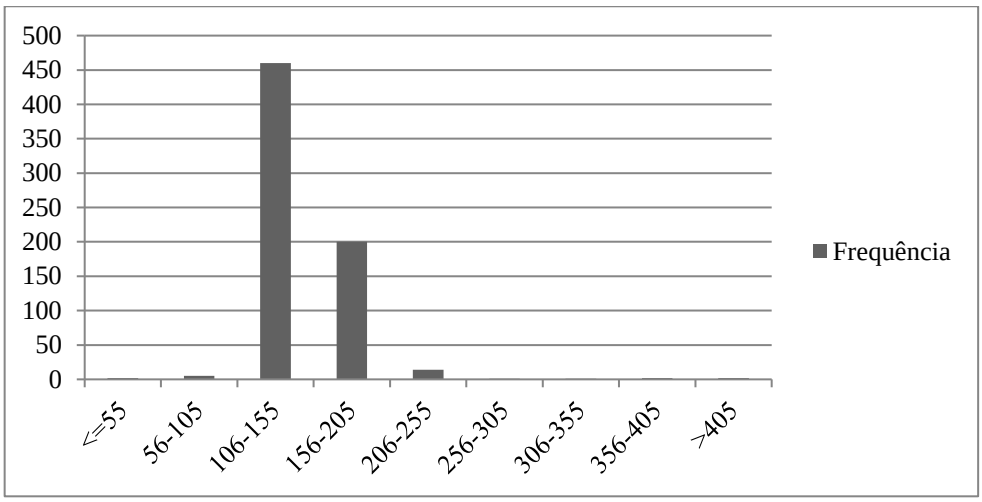


Figura 4.7: Histograma do comprimento em aminoácidos das sequências disponíveis no UniProt/TrEMBL para o EC 2.3.1.82 (N6'-acetiltransferase de aminoglicosídeos).

X: número de sequências; Y: comprimento das sequências.

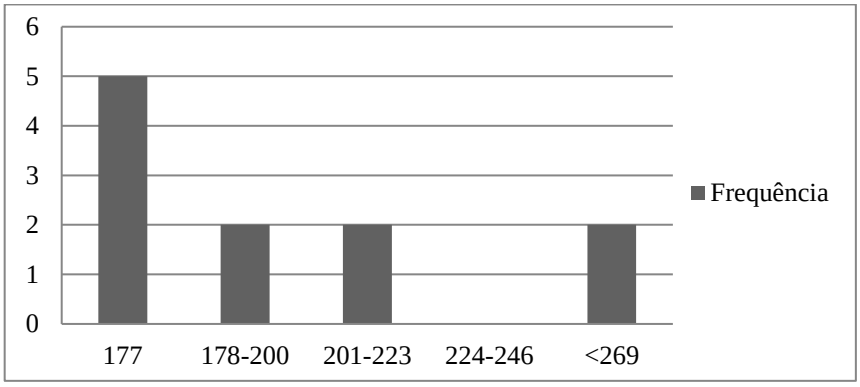


Figura 4.8: Histograma do comprimento em aminoácidos das sequências disponíveis no UniProt/TrEMBL para o EC 2.7.7.46 (2"-nucleotidiltransferase de aminoglicosídeos).

X: número de sequências; Y: comprimento das sequências.

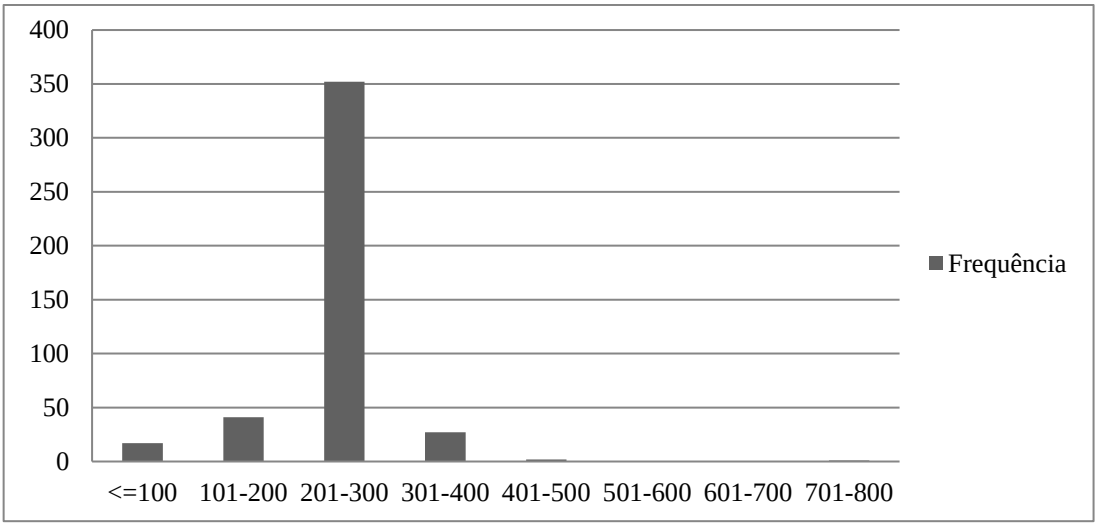


Figura 4.9: Histograma do comprimento em aminoácidos das sequências disponíveis no UniProt/TrEMBL para o EC 2.7.7.47 (3"-adenililtransferase de aminoglicosídeos).

X: número de sequências; Y: comprimento das sequências.

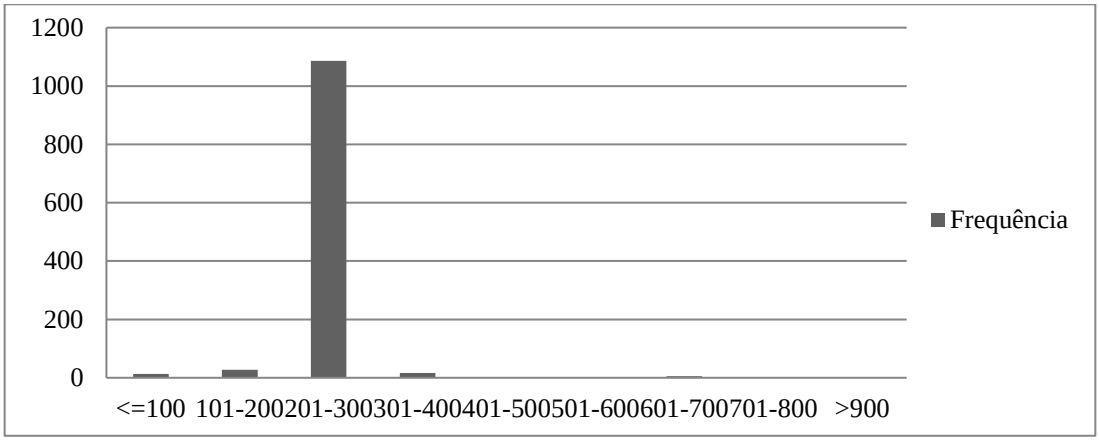


Figura 4.10: Histograma do comprimento em aminoácidos das sequências disponíveis no UniProt/TrEMBL para o EC 2.7.1.95 (3'-fosfotransferase de aminoglicosídeos).

X: número de sequências; Y: comprimento das sequências.

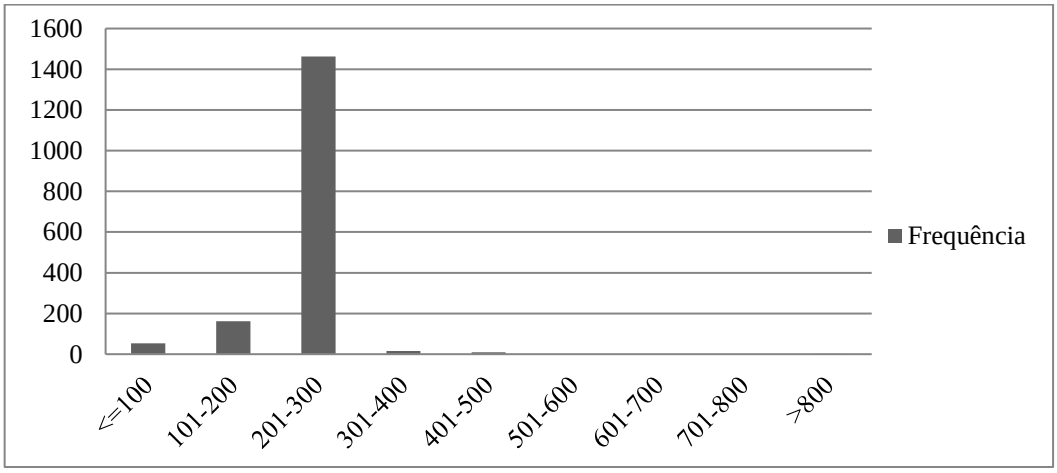


Figura 4.11: Histograma do comprimento em aminoácidos das sequências disponíveis no UniProt/TrEMBL para o EC 2.3.1.28 (O-acetiltransferase de cloranfenicol).

X: número de sequências; Y: comprimento das sequências.

Tabela 4.16 - Intervalos de comprimento de sequência selecionados para a etapa de clusterizações utilizando dados do UniProt

Atividade	Alvo	EC	Comprimento (aa)	N de sequências	% do total inicial
N3'-acetiltransferase	aminoglicosídeos	2.3.1.81	200-300	156	90%
N6'-acetiltransferase	aminoglicosídeos	2.3.1.82	100-250	674	98%
2''-nucleotidiltransferase	aminoglicosídeos	2.7.7.46	100-300	11	100%
3'-fosfotransferase	aminoglicosídeos	2.7.1.95	100-400	421	96%
3''-adenililtransferase	aminoglicosídeos	2.7.7.47	200-300	1.086	95%
O-acetiltransferase	cloranfenicol	2.3.1.28	100-300	1.628	95%

aa: aminoácidos; N: número. EC: *Enzyme Comission Number*.

4.2.2 Clusterizações

As sequências cujo comprimento estava incluído nesses intervalos foram utilizadas na etapa de *clusterização*. Apenas os *clusters* com um número significativo de sequências em relação ao total inicial foram avaliados quanto a anotação das proteínas (eclipse cinza nas Figuras 4.12-17).

A atividade aminoglicosídeo N3'-acetiltransferase parece não apresentar classes ou subclasses. Com os *thresholds* de “densidade de pontuação BLAST” iguais a 50% e 60%, apenas quatro sequências se separaram do grupo principal (Figura 4.12).

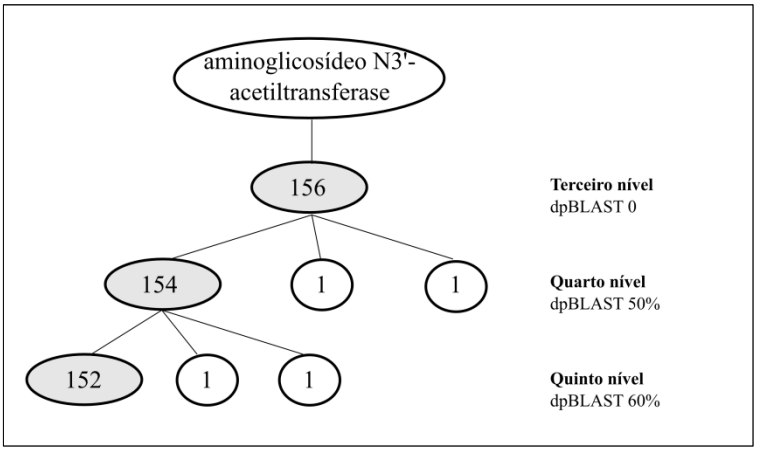


Figura 4.12: *Clusterizações* de 156 sequências do UniProt/TrEMBL para o EC 2.3.1.81 (N3'-acetiltransferase de aminoglicosídeos) com comprimento entre 200 e 300 aminiácidos.

dpBLAST: Densidade de pontuação BLAST.

Para as aminoglicosídeo N6'-acetiltransferases, existem duas classes principais, que apresentam baixa similaridade ente suas sequências primárias (Figura 4.13). Entre as anotações mais comuns para as sequências na classe 1 estão os genes “aacA7” (AAC(6')-Il) e “aacA4” (AAC(6')-Ib). Já as anotações das sequências na classe 2 são “aacA1”, “aac(6`)-Ii”, “aac(6`)-Iae” e “aac(6`)-Iaf”. A classe 1 é dividida em duas subclasses principais (1.1 e 1.2) que apresentam similaridade menor que 50% entre seus representantes, e parecem corresponder a separação entre “aacA7” e “aacA4”, respectivamente. Os demais *clusters*, todos com 11 ou menos sequências, não foram considerados como classes ou subclasses de N6'-acetiltransferases (círculos brancos). Suas sequências possuem domínios da família N'-acetiltransferase (pfam00583) ou se tratam se sequências de outras famílias como RimJ/RimL ou RimI.

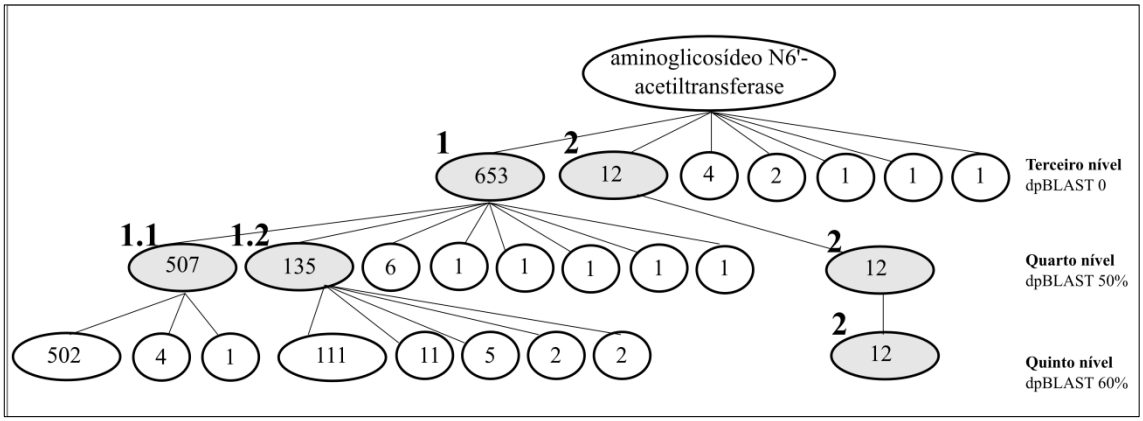


Figura 4.13: *Clusterizações* de 674 sequências do UniProt/TrEMBL para o EC 2.3.1.82 (N6'-acetiltransferase de aminoglicosídeos) com comprimento ente 100 e 250 aminoácidos.

dpBLAST: Densidade de pontuação BLAST.

As 2''-nucleotidiltransferase de aminoglicosídeos (aadB) não possuem classes e subclasses. O segundo *cluster* é formado por nucleotidiltransferase de kanamicina, que constituem uma atividade enzimática distinta de 2.7.7.46 (Figura 4.14).

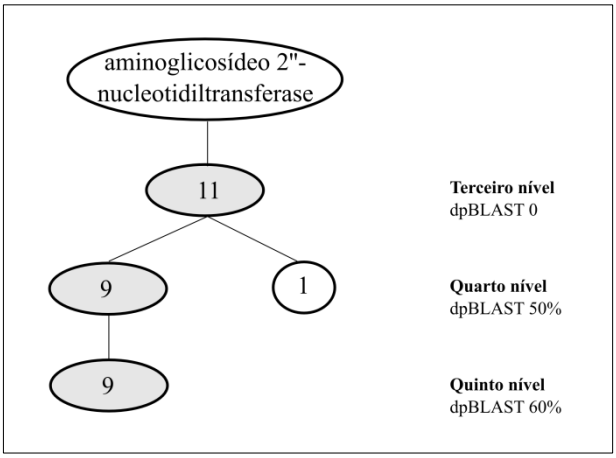


Figura 4.14: *Clusterizações* de 11 sequências do UniProt/TrEMBL para o EC 2.7.7.46 (2''-nucleotidiltransferase de aminoglicosídeos) com comprimento ente 100 e 300 aminoácidos.

dpBLAST: Densidade de pontuação BLAST.

A atividade aminoglicosídeo 3''-adenililtransferase parece não apresentar classes ou subclasses, uma vez que a sequências se mantiveram agrupadas, com apenas seis se separando do grupo principal (Figura 4.15).

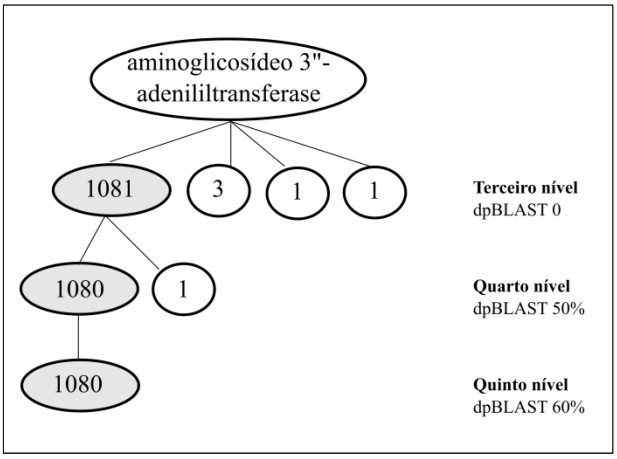


Figura 4.15: *Clusterizações* de 1.086 sequências do UniProt/TrEMBL para o EC 2.7.7.47 (3"-adenililtransferase de aminoglicosídeos) com comprimento entre 200 e 300 aminioácidos.

dpBLAST: Densidade de pontuação BLAST.

Para a atividade 3'-fosfotransferase de aminoglicosídeos, os resultados indicam a existência de uma única classe principal (Figura 4.16). As anotações mais frequentes para as sequências nessa classe são “aphA1” “aphA-3”, “aphA-4”, “aphA-6” e “aphA-7”. Foram descartados um *cluster* com 12 sequências de 3"-quinase de estreptomicina por pertencerem ao EC 2.7.1.87, além de outro *clusters* menores com anotações imprecisas. O *cluster* principal foi separado em vários outros, cujas sequencias apresentam similaridade menor que 50% entre si. As sequências no maior *cluster* apresentam o domínio COG3231 (fosfotransferase de aminoglicosídeo), enquanto aquelas nos demais *clusters* apresentam o domínio COG3173 (menos específico).

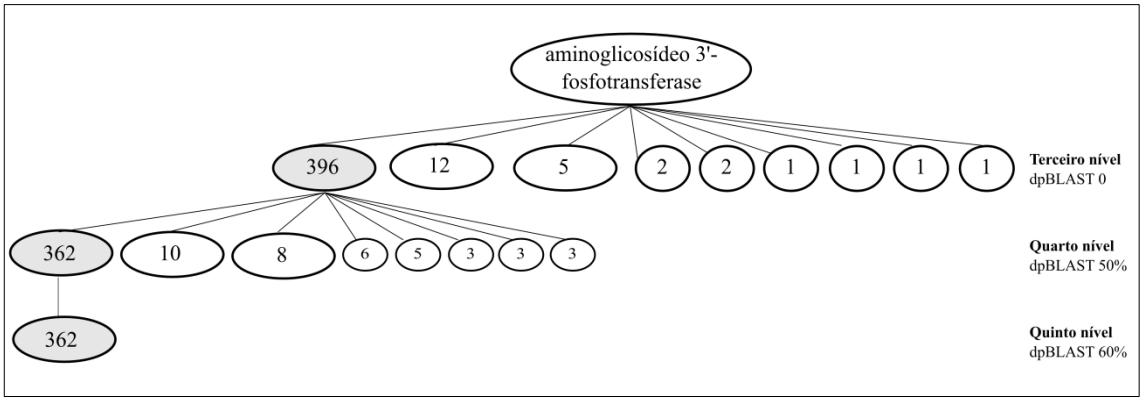


Figura 4.16: *Clusterizações* de 421 sequências do UniProt/TrEMBL para o EC 2.7.1.95 (3'-fosfotransferase de aminoglicosídeos) com comprimento entre 100 e 400 aminoácidos.

dpBLAST: Densidade de pontuação BLAST.

As O-acetiltransferases de cloranfenicol se separam em duas classes (Figura 4.17). A primeira é formada pelas sequências anotadas como cloranfenicol O-

acetiltransferase do tipo A, cujo domínio presente é o COG4845. Já a segunda classe é constituída de sequências de cloranfenicol O-acetiltransferase do tipo B, cujo domínio correspondente é o COG0110.

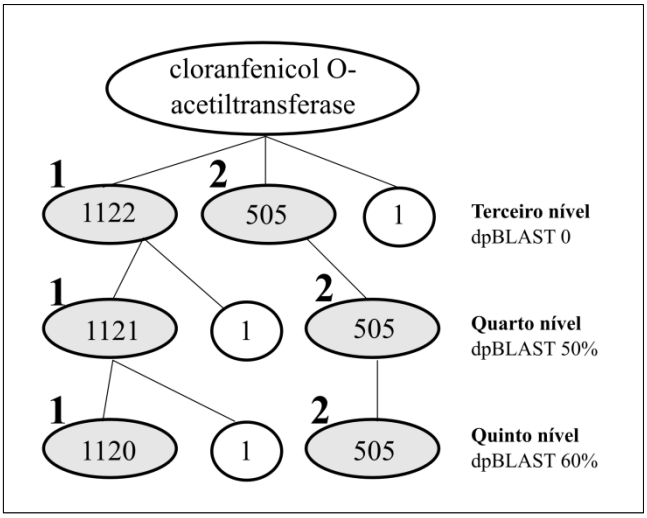


Figura 4.17: *Clusterizações* de 1.628 sequências do UniProt/TrEMBL para o EC 2.3.1.28 (O-acetiltransferase de cloranfenicol) com comprimento entre 100 e 300 aminiácidos.

dpBLAST: Densidade de pontuação BLAST.

5 DISCUSSÃO

5.1 Aspectos metodológicos

Um grande número de sequências de DNA vem sendo disponibilizado nas últimas décadas com as facilidades proporcionadas pelas novas tecnologias de sequenciamento. Esse volume de dados torna cada vez mais necessária à utilização de métodos automatizados de análises (Pallen, 2016). No entanto, a automatização é acompanhada de falhas, em maior ou menor grau, que podem ser percebidas apenas através de curadorias realizadas por um grupo especializado de pessoas (Karp, 2016).

O presente trabalho se iniciou com a ideia de propor um esquema de classificação estrutural hierárquica para enzimas envolvidas com a resistência aos antimicrobianos. Esse estudo requer análises dos esquemas atuais e dos dados disponíveis, tanto para entender as limitações quanto para estabelecer o número de diferentes grupos existentes. O processo foi pensado para ser inicialmente curado, formando grupos de alta qualidade, e depois estendido para um sistema automatizado, reprodutível em conjuntos de dados maiores e mais diversos.

A utilização do sistema de classificação funcional de enzimas (*Enzyme Commission*) na recuperação de proteínas envolvidas com a resistência aos antibióticos permitiu uma seleção baseada nas reações químicas catalisadas por elas, e não por suas sequências (Omelchenko et al., 2010). Dessa forma, puderam ser selecionadas inclusive NISEs, como as SBLs e MBLs, e as acetiltransferases de cloranfenicol tipos A e B, discutidas posteriormente.

A metodologia foi estabelecida a partir das BLs, um grupo de enzimas que hidrolisam antibióticos beta-lactâmicos, sendo o principal mecanismo de resistência em bactérias Gram-negativas (Eliopoulos and Bush, 2001). Existem milhares de BLs diferentes, amplamente distribuídas, e codificadas tanto em cromossomos como em elementos genéticos móveis (Eliopoulos and Bush, 2001; Srivastava et al., 2014). Vários autores vêm estudando esquemas estruturais de classificação para BLs, baseados principalmente no trabalho de Ambler em 1980 (Brandt et al., 2017; Hall and Barlow, 2005; Philippon et al., 2016; Rasmussen and Bush, 1997).

Além disso, as crescentes investigações em metagenomas por novas famílias de BLs aumentou em muito a variabilidade desse grupo de enzimas (Allen et al., 2009; Berglund et al., 2017). Essas famílias podem inclusive representar novas classes e subclasses, que precisam ser definidas.

Em 2005, Hall & Barlow questionaram a organização das classes e subclasses de BLs adotada pela comunidade científica, propondo pequenas modificações que adequariam essa classificação às informações disponíveis (Hall and Barlow, 2005). Dessa forma, as BLs puderam ser claramente separadas em quatro níveis estruturais hierárquicos distintos. Algum tempo depois, novas subclasses foram apontadas, usando informações de filogenia e similaridade de sequência (Brandt et al., 2017; Philippon et al., 2016). No presente estudo, a metodologia desenvolvida permite a identificação e classificação das BLs em cinco níveis hierárquicos que corroboram os trabalhos anteriores, adotando critérios de similaridade de sequência.

A abordagem de agrupamento possibilitou a construção de *clusters* que correspondem às classes e subclasses de BLs. Nesse sentido, a clusterização hierárquica foi utilizada tanto para as estruturas primárias como para as terciárias. C*lusters* formados com *single linkage* são mais heterogêneos em relação aos demais métodos (Yim and Ramdeen, 2015). No presente trabalho eles corresponderam às cinco classes de BLs do esquema de classificação sugerido por Hall & Barlow (Hall and Barlow, 2005), utilizando tanto estruturas quanto sequências.

Ainda que as MBLs sejam separadas em três subclasses (Rasmussen and Bush, 1997), estas não são equivalentes. Considerando suas estruturas primárias, as subclasses B1 e B2 são similares o suficiente para serem reconhecidas como homólogas, porém o mesmo não é observado para a subclasse B3 (Hall et al., 2003). Existem inclusive indícios de que a origem evolutiva desses dois grupos seja distinta (Alderson et al., 2014). Por essa razão, adotamos as classes MB (B1+B2) e ME (ME), assim como o sugerido por Hall & Barlow (Hall and Barlow, 2005).

Os perfis HMM construídos para as cinco classes de BLs (SA, SC, SD, MB e ME) são a primeira etapa no processo de identificação e anotação dessas enzimas. Métodos de comparação de sequências, como a ferramenta BLAST (Altschul et al., 1990), são normalmente os mais utilizados para identificar uma nova proteína (Feuermann et al., 2016). No entanto, os modelos probabilísticos de Markov e outros métodos de comparação perfil-sequência aumentam significativamente a sensibilidade do processo. Isso acontece porque os perfis são construídos a partir do alinhamento de várias sequências, e logicamente contém mais informações que uma sequência sozinha. Os perfis descrevem exatamente quais variações de aminoácidos são possíveis em cada posição (Söding, 2005). Essa abordagem já vem sendo empregada na identificação de proteínas que promovem a resistência aos antibióticos (Fróes et al., 2016; Gibson et al., 2014).

Entretanto, perfis HMM são pouco específicos para diferenciar entre famílias de proteínas, que costumam compartilhar sinais de dobramentos que tornam a discriminação entre elas mais difícil, diminuindo a precisão das anotações funcionais. Existem estratégias para tornar os perfis mais específicos, como o protocolo HMM-ModE adotado nesse trabalho (Sinha and Lynn, 2014). Essa abordagem envolve o treinamento do perfil com um grupo de sequências negativas, que não devem ser identificadas por ele. Após essa etapa, o índice EsC dos perfis de MBLs foi aperfeiçoado e chegou a 100%.

Outra abordagem para anotação de proteínas é a utilização das relações evolutivas entre suas sequências para predizer função, se baseando nos membros do mesmo grupo filogenético que têm papel conhecido (Feuermann et al., 2016). A árvore filogenética construída com todos os membros da superfamília de SBL incluiu as classes SA, SC, SD, além de outras proteínas com funções variadas. No entanto a filogenia ainda não foi suficiente para separar algumas proteínas que estavam incluídas no clado das classes de BLs apesar destas não hidrolisarem antibióticos beta-lactâmicos (PBP-A no clado da classe SA, e BlaR1 no clado da classe SD).

Foi necessário estabelecer um *threshold* de HMM *bit score* para as buscas com cada um dos perfis HMM para que estes passassem a recuperar apenas BLs. Escolhemos o *bit score* por se tratar de um indicador estatístico constante para pesquisar diferentes bancos de dados com tamanhos variados. Assim foi possível aprimorar a anotação funcional das SBLs fornecida pelas buscas utilizando modelos probabilísticos de Markov.

As proteínas PBP-A e BlaR1 apresentam alta similaridade filogenética e estrutural com as BLs. A enzima PBP-A possuí estrutura terciária muito similar as BLs do tipo PER (subclasse SA2), e não apresenta similaridade detectável à nível de sequência primária com outras PBPs. Além disso, não existe nenhuma explicação aparente do por que essa proteína não possui atividade BL (Urbach et al., 2009).

A proteína BlaR1 (ou sua cognata MecR1) regula a expressão de genes que causam resistência aos beta-lactâmicos em *S. aureus*. Essa proteína transmembrana possui dois domínios, um extracelular que é fosforilado pelo antibiótico beta-lactâmico e comunica a presença deste, e outro domínio citoplasmático, que é em seguida ativado levando a expressão de determinantes de resistência (Boudreau et al., 2016). Quando BlaR1 é inibida, a sensibilidade à meticilina aumenta (Hou et al., 2011). A estrutura de BlaR1 mostra que o domínio sensor da parte extracelular se assemelha as enzimas da subclasse SD2 de BLs, mas que se tornou uma proteína de ligação à

beta-lactâmicos devido a formação de uma acil-enzima muito estável (Wilke et al., 2004). Por isso, os valores de HMM *bit score* dessas proteínas nas buscas utilizando o perfil da classe SD são muito próximos aos das BLs. Portanto, a discriminação entre BlaR1/MecR1 e as BLs da classe SD, especificamente da subclasse SD2, não foi possível nem por filogenia nem pelos valores de HMM *bit score*. Esses dois grupos de proteínas foram separados apenas com o processo de *clusterização* que define as subclasses, discutido mais a diante.

As enzimas ClbP e Pab87 estão associadas ao número de EC das BLs no PDB, e permaneceram agrupadas com as BLs da classe SC tanto na *clusterização* de estruturas usando *single linkage,* quanto na *clusterização* de sequências aplicando "densidade de pontuação BLAST" igual à zero. Pab87 é uma serino protease octamérica, homóloga às PBPs, que faz parte de uma família de proteínas de auto-compartimentalização, CubicO (Delfosse et al., 2009). Já a ClbP é uma enzima essencial na maquinaria que codifica policetídeos (PK), que quando combinados com peptídeos não-ribossomais (NRP) induzem a quebra do DNA fita dupla de células eucarióticas. ClbP é destituída de atividade BL significativa, apresentando K_{cat}[24] um milhão de vezes menor que enzimas típicas da classe SC. Essas duas proteínas compartilham similaridade significativa entre seus sítios ativos e aqueles das BLs da classe SC (Dubois et al., 2011). No entanto, mostramos aqui que elas podem ser separadas das demais BLs da classe SC através de filogenia e da utilização de um *threshold* de HMM *bit score*.

A utilização dos perfis HMM e seus *thresholds* associados para identificação de sequências BLs poderia ser utilizada em metagenomas. Porém o número de sequências recuperadas a partir de *reads* brutas seria muito pequeno. Quando se trata de dados fragmentados, novos *thresholds* devem ser estabelecidos, e para isso outros testes de validação devem ser realizados (Berglund et al., 2017). Uma das alternativas possíveis é a utilização de metagenomas gerados por tecnologias que fornecem *reads* maiores ou ainda buscar por BLs entre metagenomas montados.

Os primeiros indícios de que existiriam dois grupos distintos dentro da classe SA surgiram através de análises filogenéticas. O grupo principal seria composto pelos tipos mais disseminados, enquanto o segundo seria formado por BLs isoladas principalmente em *Cytophagales-Flavobacteriales-Bacteroidales* (Hall and Barlow, 2004). Mais tarde, esses dois grupos foram chamados de subclasses A1 (SA1) e A2

[24] constante de velocidade utilizada para determinar a velocidade limitante de qualquer reação catalisada por uma enzima em condições de saturação

(SA2), que apresentam resíduos conservados muito diferentes entre elas, além da subclasse SA2 possuir blocos de inserções em relação à SA1 (Philippon et al., 2016). A subclasse SA2 mostra pouca relação filogenética com os membros de SA1, e não possui enzimas com amplo espectro de ação (Brandt et al., 2017). Essas subclasses foram identificadas quando adotamos o *threshold* de “densidade de pontuação BLAST” igual à 60% e “tamanho de cobertura mínimo” de 45% entre as sequências de BLs da classe SA.

Até o momento da coleta de dados para esse trabalho a única BL da subclasse SA2 com estrutura determinada, e portanto disponível no PDB, era a enzima PER-1 (família PER). Quando foi realizada a separação das sequências do BNRB em subclasses, outras famílias de BLs da subclasse SA2 foram anotadas, corroborando com aquelas citadas no trabalho de Philippon e colaboradores (CblA, CfxA, CEF, CepA, CGA, CIA, CME, CSP, PER, SPU, TLA, TLA e VEB) (Philippon et al., 2016).

Atualmente, a distinção entre as enzimas da classe SD é baseada somente em números, o que não reflete a diversidade dessa classe (Brandt et al., 2017). Na página atualizada do banco de dados CBMAR (Srivastava et al., 2014) é possível ter acesso à árvore filogenética da classe SD, onde dois clados principais podem ser observados. Um deles contém a maioria das enzimas, enquanto o segundo é composto das variantes OXA-1, OXA-4, OXA-9, OXA-18, OXA-22, OXA-30, OXA-31, OXA-45, OXA-57, OXA-59, OXA-114a, OXA-224, OXA-243, OXA-258 e OXA-320 (http://proteininformatics.org/mkumar/lactamasedb/OXA_phylogeny.pdf).

Recentemente, a subclasse D2 (SD2) foi proposta como um clado distante e distinto das demais enzimas da classe SD, com variantes de espectro de ação ampliado (OXA-1, OXA-18 e OXA-45) e restrito (OXA-9 e OXA-22), mas nenhuma carbapenemase (Brandt et al., 2017).

A anotação das enzimas do BNRB nas subclasses SD1 e SD2 coincidem com aquelas variantes apontadas nos trabalhos mencionados acima (Brandt et al., 2017; Srivastava et al., 2014). Portanto, conseguimos reproduzir os resultados alcançados através de análises filogenéticas utilizando agrupamentos de sequências por similaridade, com a formação das subclasses da classe SD quando aplicada “densidade de pontuação BLAST” igual à 60% e “tamanho de cobertura mínimo” de 75%.

Também utilizando *threshold* de “densidade de pontuação BLAST” de 60%, a subclasse MB1 é dividida em MB1.1 e MB1.2. A única família de BLs da subclasse MB1.2 é a SPM. Essa família possui apenas uma variante já descrita, SPM-1, encontra

somente em isolados da espécie *Pseudomonas aeruginosa*. Sua localização cromossômica pode estar contribuindo para o isolamento dessa enzima em relação às demais BLs da subclasse MB1 (Silveira et al., 2016). A distância de SPM-1 em relação às outras representantes de MB1 pode ser corroborada inclusive por filogenia. Um estudo recente, que descreve 56 novas familias da subclasse MB1, mostra SPM-1 no clado mais ancestral em relação às demais enzimas conhecidas dessa subclasse (Berglund et al., 2017). Além disso, a análise da estrutura de SPM-1 mostrou algumas inserções e deleções em relação aos outros membros da subclasse MB1, sugerindo que essa enzima seja um híbrido estrutural entre MB1 e MB2 (Bebrone, 2007).

Bush e colaboradores observaram baixa similaridade entre a família NDM de BLs e as outras enzimas da classe B1 (MB1), sugerindo por isso a subdivisão em B1a e B1b (Bush, 2013). No entanto, as sequências dessa família permaneceram agrupadas no grupo principal, MB1.1, de acordo com o presente estudo.

A separação de SBM-1 das demais sequências da classe ME observada aplicando *threshold* de “densidade de pontuação BLAST” de 60% para as sequências do PDB, não foi mantida quando esse mesmo limiar foi utilizado para *clusterizar* as sequências do BNRB. Isso aconteceu porque o BNRB inclui outras famílias de BLs da classe ME que não estavam presentes no PDB. Entre elas esta a enzima AIM-1, com 44% dos aminoácidos idênticos à SMB-1, a maior porcentagem entre as sequências da classe ME (Wachino et al., 2013).

Devido a grande heterogeneidade no tamanho das sequências presente no BNRB, a formação das subclasses utilizando sequências dessa base exigiu que fossem estipulados *thresholds* de “tamanho de cobertura mínimo”, como 45%, 70% e 75% para as classes SA, MB e SD, respectivamente. Como consequência dessa abordagem, além de obtermos as subclasses de BLs como esperado, alguns *clusters* se formaram correspondendo à BlaR1 e outros possuíam sequências com anotações imprecisas e/ou tamanhos reduzidos, que optamos por chamar de não-BLs.

Os fragmentos de proteínas não puderam ser classificados em subclasses de acordo com os critérios estabelecidos na metodologia apresentada aqui. A partir do tamanho dos modelos do Pfam para BLs, nós inferimos quais fragmentos se tratavam de domínios parciais. A maioria das sequências nos *clusters* não-BL foram categorizadas como domínios parciais. Os papeis evolucionário, estrutural e funcional dos domínios proteicos sugerem que eles sejam “blocos de construção” indivisíveis, a partir dos quais proteínas modulares maiores são formadas (Triant and Pearson, 2015). Esses domínios parciais são na verdade resultado de alinhamentos locais dos

perfis HMM, proteínas não funcionais ou montagens imprecisas/incompletas do genoma de origem (Triant and Pearson, 2015).

Existem algumas sequências em *clusters* isolados com tamanho semelhantes às BLs nas subclasses, mas essas exceções são justificáveis. A BL LRA-5 da classe SA não agrupou nas subclasses. Essa enzima apresenta baixa similaridade e é considerada distante evolutivamente tanto das BLs da classe SA caracterizadas funcionalmente, quanto de seus ancestrais (Allen et al., 2009). Para a classe SD, a sequência de tamanho compatível possui 43% de identidade com uma "beta-lactamase classe D", no banco de dados de proteínas do NCBI, isolada da bactéria dimórfica *Oceanicaulis alexandrii*. No entanto, a atividade da possível BL de *O. alexandrii* ainda não foi demonstrada (Oh et al., 2011).

A utilização de perfis HMM aprimorados aliados as *clusterizações* de sequências para identificar e classificar de BLs se mostrou superior às outras abordagens comparadas. Já foi demonstrado que os perfis do Pfam para sequências categorizadas na família "Serino beta-lactamase *like*" capturam proteínas não relacionadas com as BLs conhecidas (Brandt et al., 2017). Corroboramos esses resultados mostrando que esses perfis são inespecíficos para as classes de BL. Da mesma forma, os *patterns* já descritos para diferentes classes de BLs testados aqui não estão presentes em todos os membros da sua respectiva classe, além de não serem capazes de diferenciar entre subclasses. Por exemplo, a sensibilidade do *pattern* representado pela expressão regular "S-[DG]-N-x(1,2)-A-[ACGNST]-x(2)-[ILMV]-x(4)-[AGSTV]", desenvolvido para a classe SA, é descrita como aproximadamente 70% (Singh et al., 2009), e quando testado aqui a sensibilidade do mesmo se mostrou ainda menor (49%).

5.2 Nova classe de BLs com domínios fusionados

Foi proposta aqui uma nova classe de enzimas com dois domínios BL, classe SCD. Fazem parte dessa classe a BL LRA-13 e outras nove sequências homólogas, considerando as buscas que realizamos no banco de dados não redundante de proteínas do NCBI. A enzima LRA-13 foi identificada no metagenoma de um solo remoto do Alasca, oriunda de uma bactéria não cultivável. A fusão de dois domínios BL expandiu a capacidade hidrolítica dessa proteína além do que qualquer um dos domínios poderia exibir sozinho, causando resistência à amoxilina, ampicilina,

carbenicilina (classe SD) e cefalexina (classe SC), demonstrado experimentalmente (Allen et al., 2009).

As nove proteínas homólogas à LRA-13 foram identificadas nos gêneros *Janthinobacterium*, *Duganella* e *Massilia*. As cepas *Janthinobacterium* sp. HH01 e *Duganella* sp. HH105 foram isoladas do ambiente aquático e exibem resistência à ampicilina (Hornung et al., 2013). As cepas *Duganella* sp. CF458 (Gp0136797) e *Massilia* sp. CF038 (Gp0136806) foram isoladas da raíz da árvore *Populus* (NCBI BioProject PRJEB18228), enquanto que as demais cepas de *Duganella* sp. e *Massilia* sp. foram isoladas da microbiota da raiz da planta *Arabidopsis* (Bai et al., 2015). Esses três gêneros pertencem à família *Oxalobacteraceae* (classe *Betaproteobacteria*), são não-patogênicas aos homens, animais e plantas, e são conhecidas pelo seu efeito antifúngico (Haack et al., 2016; Yin et al., 2013). Bactérias dessa família apresentam diferenças fenotípicas mínimas e a distinção entre os gêneros é feita principalmente através da sequência gene de rRNA 16S (Kämpfer et al., 2007).

Nenhum dos domínios de BLs (SC ou SD) apresentam deleções ou inserções significativas em LRA-13, além do conteúdo GC do gene *bla*$_{\text{LRA-13}}$ ser semelhante ao do DNA flanqueador. Por isso, essa BL parece ser o resultado de uma fusão natural antiga de genes que codificavam enzimas completas, e não devido à pressão seletiva recente causada pelo uso extensivo de antibióticos (Allen et al., 2009). A sequência e o contexto genético das BLs bifuncionais nos genomas analisados são muito semelhantes, sugerindo que, como para a enzima LRA-13, a fusão dos domínios SC e SD deva ter ocorrido naturalmente, há muito tempo.

Embora não exista um significado clínico para as BLs bifuncionais putativas apresentadas aqui, essa possibilidade não pode ser ignorada. Os genomas que carreiam essas enzimas podem ser beneficiados de várias formas. Por exemplo, a mobilização concomitante de duas funções diferentes, o potencial de resistência complementar e ampliado e a seleção simultânea de duas atividades enzimáticas pela pressão seletiva exercida por um único antibiótico (Zhang et al., 2009). É preciso estar alerta para a ameaça da disseminação dessa enzima para espécies bacterianas de importância clínica, uma vez que já é possível observar sua presença na natureza em cepas ambientais dos gêneros *Duganella*, *Massilia* e *Janthinobacterium*.

5.3 Beta-lactamases em genomas bacterianos

O aperfeiçoamento na anotação das sequências de BLs alcançado nesse estudo possibilitou a expansão de conhecimento sobre a dispersão e a prevalência dessas enzimas, capazes de hidrolisar antibióticos beta-lactâmicos, um importante mecanismo de resistência. A análise da distribuição das classes e subclasses de BLs entre diferentes genomas bacterianos confirmou algumas tendências já apontadas por estudos anteriores, além de acrescentar informações novas.

Constamos que o enriquecimento já descrito de BLs em *Actinobacteria* em relação aos demais (Gibson et al., 2014) é principalmente causado por enzimas da subclasse SA1. O filo *Actinobacteria* é apontado como um importante reservatório de genes de resistência à antibióticos para as *Gammaproteobacteria,* devido a grande quantidade desses genes que são compartilhados entre os plasmídios desses filos (Tamminen et al., 2012).

BLs da subclasse SA2, predominantemente associadas ao filo *Bacteroidetes*, também foram encontras em outros filos, alguns deles cuja presença ainda não havia sido descrita (*Cyanobacteria*, *Spirochaetes*, *Acidobacteria* e *Verrucomicrobia*). A identificação de BLs da subclasse SA2 em *Proteobactéria,* fato já reportado em trabalhos anteriores, reforça a ideia de uma transferência antiga entre filos, uma vez que *Bacteroidetes* e *Proteobactéria* podem habitar o intestino (Brandt et al., 2017). O único plasmídio identificado aqui carreando uma enzima dessa subclasse pertence ao filo *Proteobacteria,* corroborando a relação entre BLs móveis de SA2 com esse filo específico (Philippon et al., 2016).

A presença majoritária da classe SC em *Proteobacteria* é um consenso*,* destacando as classes *Gamma*, *Alpha* e *Betaproteobacteria*. Enquanto isso, as poucas sequências identificadas em *Actinobacteria* confirmam o ocasional isolamento da classe SC (Brandt et al., 2017). A maioria das BLs da classe SC são espécie-específicas e localizadas em cromossomos (Jacoby, 2009), no entanto, existem algumas enzimas características de plasmídios, restritas à família *Enterobacteriaceae* e ao gênero *Aeromonas*, ambas do filo *Proteobacteria* (Brandt et al., 2017). No nosso estudo as BLs da classe SC encontradas em plasmídios estão presentes apenas em *Proteobacteria*.

A classe SD possui quase 500 membros descritos, considerada a classe de BLs que cresce mais rápido (Toth et al., 2016). Essas enzimas podem apresentar espectro de ação restrito, estendido ou ainda serem capazes de hidrolisar

carbapenêmicos (Leonard et al., 2013). A presença de BLs da classe SD em bactérias Gram-negativas é bastante comum, no entanto, apenas recentemente elas foram descritas em Gram-positivas (Toth et al., 2016). Nós também encontramos sequências da subclasse SD1 em 4,8% das cepas de *Firmicutes* analisadas (Gram-positivo).

Nesse trabalho, a subclasse SD1 foi a mais distribuída entre diferentes filos e entre as classes de *Proteobacteria*. Há bem pouco tempo, um número enorme de variantes não caracterizadas da classe SD1 foi reportado, espalhado por diferentes filos, indicando que esse grupo de enzimas venha sendo subestimado (Brandt et al., 2017). Destacamos a alta porcentagem de genomas do filo *Chlorobi* codificando BL da subclasse SD1 (63,6%), filo de bactérias obrigatoriamente anaeróbicas e fotoautotróficas e intimamente relacionado ao filo *Bacteroidetes* (Gupta, 2004). A presença de genes de BLs ainda não havia sido descrita para *Chlorobi,* de acordo com nossas pesquisas.

A subclasse SD1 foi a única encontrada em *Epsilonproteobacteria*, uma classe de *Proteobacteria* amplamente distribuída e com algumas espécies patogênicas ao homem, como *Campylobacter* e *Helicobacter*. A produção de duas BLs em particular (OXA-61 e OXA-184) já foi reportada em 85% das cepas de *Campylobacter*, agente causador de diarreias severas em humanos (Weis et al., 2016).

As BLs da subclasse SD2 estão praticamente restritas a *Proteobacteria*, fato que deve estar relacionado com a característica cromossomal e intrínseca desses genes (Brandt et al., 2017).

Já foram reportadas enzimas cromossômicas da subclasse MB1 nos filos *Bacteroidetes*, *Firmicutes* e *Proteobacteria*, enquanto as BLs móveis dessa subclasse são características de *Proteobacteria* (Berglund et al., 2017). Essas informações corroboram os resultados apresentados aqui. Além disso, Berglund e colaboradores destacam uma forte sobrerrepresentação de bactérias carreando BLs de MB1 no filo *Bacteroidetes* (22,4%), porcentagem próxima à encontrada no presente estudo (19,8%) (Berglund et al., 2017). Destacamos também a ocorrência dessa subclasse em *Firmicute*s, lembrando que esse filo e *Bacteroidetes* são os principais componentes na microbiota normal do intestino (Jandhyala et al., 2015), o que pode favorecer a troca de informações genéticas entre eles.

Poucos representantes da subclasse MB2 foram encontrados nos cromossomos analisados, e nenhum em plasmídio. Essa subclasse inclui enzimas produzidas por diferentes espécies de *Aeromonas* (CphA e ImiS), assim como Sfh-I codificada no genoma de *Serratia fonticola*. Todos esses genes são cromossômicos,

e codificam enzimas com um espectro de ação bastante limitado em relação às demais MBLs (Bebrone, 2007).

A classe ME de BLs foi descrita aqui em nove filos diferentes. A associação dessa subclasse a bactérias do solo (Gibson et al., 2014) pode estar contribuindo para essa diversidade. O trabalho de Allen e colaboradores descreve e caracteriza funcionalmente oito novas BLs ME em bactérias isoladas de um solo remoto no Alaska, a maioria delas muito divergente em relação às outras BLs ME já conhecidas (Allen et al., 2009). Esses achados expandem o conhecimento sobre reservatórios de enzimas dessa classe, e podem explicar a dispersão dessas MBLs entre bactérias pertencentes a filos variados.

Acidobacteria é um filo descrito recentemente e já reconhecido como um dos mais abundantes e diversos da Terra, principalmente no solo. No entanto, a maioria das espécies é de difícil cultivo, e por isso costumam ser identificadas principalmente em estudos de metagenoma (Kielak et al., 2016). Apesar de poucos genomas disponíveis, esse filo parece ser importante na produção de BLs das classes SA e ME. Já foi sugerido que as *Acidobacteria* sejam capazes produzir novos compostos antibióticos (Ward et al., 2009). A ocorrência de BLs nesse filo pode servir de autoproteção, como ocorre com as *Actinobacteria* (Gibson et al., 2014).

Identificamos 14 sequências de BLs entre as 60 cepas analisadas do filo *Spirochaetes*. A presença de BLs da classe SD em *Brachyspira pilosicoli* vem sendo reportada desde 2008, espécie essa que pertence ao filo *Spirochaetes* e é agente de uma zoonose intestinal (Jansson and Pringle, 2011; Meziane-Cherif et al., 2008). Também identificamos MBLs da classe ME nesse gênero bacteriano. Além disso, reportamos pela primeira vez BLs (subclasse MB1.1) no gênero *Leptospira*, espécie *biflexa*, uma espiroqueta saprofítica, de vida livre (Picardeau et al., 2008).

A classe *Gammaproteobacteria* inclui vários patógenos comuns para os homens, como as famílias *Enterobacteriaceae* e *Pseudomonadacea*. Esses patógenos estão continuamente expostos a pressão seletiva dos antibióticos, o que favorece a aquisição de resistência, e pode justificar as 499 sequências de BLs encontradas nesse grupo de bactérias, distribuídas entre as classes e subclasses SA1, SA2, SC, SD1, SD2, MB1.1, MB1 e ME.

Um total de 35% dos plasmídios carreadores de BLs foi classificado nos grupos de incompatibilidade pesquisados. A identificação do grupo de incompatibilidade plasmidial não é sempre possível, por exemplo, Suzuki e colaboradores anotaram o grupo de apenas 55 plasmídios em 1.945 analisados (Suzuki et al., 2010). O grupo

IncF, apesar de ter sido o mais identificado no presente estudo, possui uma gama de hospedeiros bastante limitada, onde se destacam as *Gammaproteobacteria* (Suzuki et al., 2010). Importantes famílias de BLs da classe SA são encontradas em plasmídios IncF, como CTX-M, TEM e SHV. Além disso, IncF é o grupo mais descrito entre os plasmídios carreadores de genes de resistência em *Enterobacteriaceae* (Carattoli, 2009). Plasmídios do grupo IncW foram identificados como principais carreadores de BLs da subclasse SD2. Esse grupo já foi encontrado em uma grande variedade de bactérias hospedeiras, incluindo *Alfa*, *Beta*, *Gamma*, *Deltaproteobacteria* e *Bacteroidetes* (Fernández-López et al., 2006).

5.4 Outras atividades enzimáticas envolvidas na resistência a antibióticos

Baseado no que foi realizado para BLs, tentamos identificar grupos de sequências similares e estabelecer uma classificação hierárquica para outras atividades enzimáticas envolvidas com a resistência aos antibióticos em bactérias. Dentre elas, uma das mais heterogêneas foi a N6'-acetiltransferase de aminoglicosídeos. Duas famílias das enzimas AAC(6') são conhecidas, mas se baseiam na especificidade de substrato (AAC(6')-I e AAC(6')-II) (Shaw et al., 1993). A partir de critérios filogenéticos, AAC(6') é separada em três clados distintos, cujos as estruturas primárias não são relacionadas entre si (Salipante and Hall, 2003). Esses clados foram chamados de subfamílias, e correspondem as subclasses 1.1 e 1.2, e a classe 2 presentes nos nossos resultados (AAC(6')[A], AAC(6')[C]) e AAC(6')[B]), respectivamente). Com a resolução de novas estruturas tridimensionais de AAC(6'), Stogios e colaboradores mostram que essas três subfamílias apresentam sítios-ativos com arquiteturas muito diferentes, e por isso podem ser resultado de evolução convergente, ainda que elas apresentem o mesmo *fold* (Stogios et al., 2017). No entanto, observamos aqui que as suclasses 1.1 e 1.2 compartilham similaridade entre suas sequências, e por isso acreditamos que a convergência evolutiva seja uma possibilidade somente entre a família AAC(6')[B] e o grupo que inclui AAC(6')[A] e AAC(6')[C].

A maior família de fosfotransferases de aminoglicosídeos modificam o grupo 3'-OH desses antibióticos (Wright and Thompson, 1999). As APH(3') possuem oito tipos de enzimas, que se diferenciam em relação à sequência e a especificidade de substrato. Esse grupo de enzimas compartilha o mesmo *fold*, com dois domínios estruturais (Boyko et al., 2016). Dados filogenéticos mostram que os tipos III, IV, VI e

VII são mais relacionados entre si, assim como os tipos I, II, V e VIII (Hächler et al., 1996). Segundo os nossos resultados, considerando o critério de similaridade de sequências, as enzimas que pertencem a esse EC são bastante similares, e apenas um *cluster* principal pôde ser observado.

As acetiltransferases de cloranfenicol (CAT) pertencem a dois tipos bem definidos (A e B), com sequências e estruturas distintas, podendo ser considerados NISE. Esses grupos utilizam uma gama estruturalmente diferente de compostos hidroxilados como aceptores de acetil (Murray and Shaw, 1997). Os dois *clusters* foram identificados em nosso estudo, e não apresentaram novas subdivisões. Os tipos principais de CAT já foram separados em outros grupos, no entanto essa separação é baseada em identidades maiores que 80% (Schwarz et al., 2004), muito superior aos *thresholds* de similaridade estabelecidos aqui.

As demais atividades *clusterizadas* apresentaram apenas um grupo principal de enzimas de acordo com os critérios de similaridade aplicados. Os resultados apresentados nesse estudo mostram que as BLs são o grupo de enzimas mais diverso em relação às demais atividades enzimáticas relacionadas com a resistência bacteriana aos antimicrobianos estudadas aqui. As BLs possuem um grande número de estruturas e sequencias disponíveis, além de ampla distribuição entre os filos bacterianos. Alguns fatores poderiam justificar essa diversidade, como o longo período desde o aparecimento das primeiras BLs (Hall and Barlow, 2004) e o fato dos beta-lactâmicos serem, continuamente, uma classe de antibióticos amplamente utilizados (Bush, 2013).

6 CONCLUSÕES

A classificação hierárquica de BLs mostrou a presença de cinco níveis distintos, segundo critérios de similaridade de sequência, que corroboram estudos filogenéticos realizados anteriormente para essa atividade enzimática.

A metodologia apresentada permitiu a identificação e classificação de sequências proteicas de BLs com alta confiabilidade, e apontou ainda alguns problemas na anotação dessas enzimas em bancos de dados públicos.

Os perfis HMM para as classes de BLs desenvolvidos aqui apresentaram especificidade superior aos perfis disponíveis no Pfam e *patterns* descritos na literatura.

Uma nova classe de BLs bifuncionais foi proposta, além da descrição de uma nova subclasse de MBLs, MB1.2, que inclui a família SPM.

A análise da distribuição das classes e subclasses de BLs entre diferentes genomas bacterianos mostrou que todas estão presentes no filo *Proteobacteria*, e que a classe ME e a subclasse SD1 são as mais distribuídas entre diferentes filos.

O principal grupo de incompatibilidade para os plasmídios carregadores de BLs foi o IncF, mas a maioria dos plasmídios não teve seu grupo definido.

As outras atividades enzimáticas de resistência aos antimicrobianos estudadas se mostraram menos diversas que as BLs, e as classes e subclasses observadas coincidem com estudos anteriores.

7 PERSPECTIVAS

O estudo apresentado abre novas possibilidades, dentre as quais estão o desenvolvimento de perfis HMM para as enzimas modificadoras de aminoglicosídeos e para as acetiltransferases de cloranfenicol, permitindo que a distribuição e a abundância dessas enzimas também sejam inferidas. Os perfis HMM finais das atividades enzimáticas relacionadas à resistência aos antibióticos podem ainda ser aplicados em metagenomas montados. A associação entre as subclasses e os filos bacterianos pode ser mais explorada, incluindo, por exemplo, a relação filogenética entre os filos onde essas sequências são identificadas. Além disso, pode ser criado a partir do BNRB um banco de dados públicos com todas as informações apresentadas nessa tese. A relevância do tema e a quantidade de dados disponíveis faz da resistência aos antimicrobianos uma fonte rica para estudos que contribuam para sua prevenção e controle.

8 REFERÊNCIAS BIBLIOGRÁFICAS

Abraham, E.P., Chain, E., 1940. An Enzyme from Bacteria able to Destroy Penicillin. Nature. https://doi.org/10.1038/146837a0

Alderson, R.G., Barker, D., Mitchell, J.B.O., 2014. One origin for metallo-β-lactamase activity, or two? An investigation assessing a diverse set of reconstructed ancestral sequences based on a sample of phylogenetic trees. J. Mol. Evol. 79, 117–129. https://doi.org/10.1007/s00239-014-9639-7

Alderson, R.G., De Ferrari, L., Mavridis, L., McDonagh, J.L., Mitchell, J.B.O., Nath, N., 2012. Enzyme informatics. Curr. Top. Med. Chem. 12, 1911–23. https://doi.org/CTMC-EPUB-20121030-7 [pii]

Allen, H.K., Moe, L. a, Rodbumrer, J., Gaarder, A., Handelsman, J., 2009. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. ISME J. 3, 243–251. https://doi.org/10.1038/ismej.2008.86

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Ambler, R.P., 1980. The structure of beta-lactamases. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 289, 321–331. https://doi.org/10.1098/rstb.1980.0049

Armougom, F., Moretti, S., Keduas, V., Notredame, C., 2006. The iRMSD: A local measure of sequence alignment accuracy using structural information. Bioinformatics 22, 35–39. https://doi.org/10.1093/bioinformatics/btl218

Bai, Y., Müller, D.B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P.C., Spaepen, S., Remus-Emsermann, M., Hüttel, B., McHardy, A.C., Vorholt, J.A., Schulze-Lefert, P., 2015. Functional overlap of the Arabidopsis leaf and root microbiota. Nature 528, 364–369. https://doi.org/10.1038/nature16192

Bartoloni, A., Pallecchi, L., Rodríguez, H., Fernandez, C., Mantella, A., Bartalesi, F., Strohmeyer, M., Kristiansson, C., Gotuzzo, E., Paradisi, F., Rossolini, G.M., 2009. Antibiotic resistance in a very remote Amazonas community. Int. J. Antimicrob. Agents 33, 125–129. https://doi.org/10.1016/j.ijantimicag.2008.07.029

Bebrone, C., 2007. Metallo-beta-lactamases (classification, activity, genetic

organization, structure, zinc coordination) and their superfamily. Biochem. Pharmacol. 74, 1686–1701. https://doi.org/10.1016/j.bcp.2007.05.021

Berglund, F., Marathe, N.P., Österlund, T., Bengtsson-Palme, J., Kotsakis, S., Flach, C.-F., Larsson, D.G.J., Kristiansson, E., 2017. Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. Microbiome 5, 134. https://doi.org/10.1186/s40168-017-0353-8

Bialek-Davenet, S., Criscuolo, A., Ailloud, F., Passet, V., Jones, L., Delannoy-Vieillard, A.S., Garin, B., Hello, S. Le, Arlet, G., Nicolas-Chanoine, M.H., Decré, D., Brisse, S., 2014. Genomic definition of hypervirulent and multidrug-resistant klebsiella pneumoniae clonal groups. Emerg. Infect. Dis. 20, 1812–1820. https://doi.org/10.3201/eid2011.140206

Boudreau, M.A., Fishovitz, J., Llarrull, L.I., Xiao, Q., Mobashery, S., 2016. Phosphorylation of BlaR1 in Manifestation of Antibiotic Resistance in Methicillin-Resistant Staphylococcus aureus and Its Abrogation by Small Molecules. ACS Infect. Dis. https://doi.org/10.1021/acsinfecdis.5b00086

Boyko, K.M., Gorbacheva, M.A., Korzhenevskiy, D.A., Alekseeva, M.G., Mavletova, D.A., Zakharevich, N. V., Elizarov, S.M., Rudakova, N.N., Danilenko, V.N., Popov, V.O., 2016. Structural characterization of the novel aminoglycoside phosphotransferase AphVIII from Streptomyces rimosus with enzymatic activity modulated by phosphorylation. Biochem. Biophys. Res. Commun. https://doi.org/10.1016/j.bbrc.2016.06.097

Brandt, C., Braun, S.D., Stein, C., Slickers, P., Ehricht, R., Pletz, M.W., Makarewicz, O., 2017. In silico serine β-lactamases analysis reveals a huge potential resistome in environmental and pathogenic species. Sci. Rep. 7, 43232. https://doi.org/10.1038/srep43232

Burley, S.K., Berman, H.M., Christie, C., Duarte, J.M., Feng, Z., Westbrook, J., Young, J., Zardecki, C., 2017. TOOLS FOR PROTEIN SCIENCE RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. https://doi.org/10.1002/pro.3331

Bush, K., 2013. The ABCD's of Beta-lactamase nomenclature. J. Infect. Chemother. 19, 549–559. https://doi.org/10.1007/s10156-013-0640-7

Bush, K., Jacoby, G. a, Medeiros, a a, 1995. A functional classification scheme for beta-lactamases and its correlation with molecular structure. Antimicrob. Agents Chemother. 39, 1211–1233. https://doi.org/10.1128/AAC.39.6.1211

Carattoli, A., 2009. Resistance plasmid families in Enterobacteriaceae. Antimicrob. Agents Chemother. https://doi.org/10.1128/AAC.01707-08

Castilho, S.R.A., Godoy, C.S.D.M., Guilarde, A.O., Cardoso, J.L., André, M.C.P., Junqueira-Kipnis, A.P., Kipnis, A., 2017. Acinetobacter baumannii strains isolated from patients in intensive care units in Goiânia, Brazil: Molecular and drug susceptibility profiles. PLoS One 12, 1–13. https://doi.org/10.1371/journal.pone.0176790

Chesneau, O., Tsvetkova, K., Courvalin, P., 2007. Resistance phenotypes conferred by macrolide phosphotransferases. FEMS Microbiol. Lett. 269, 317–322. https://doi.org/10.1111/j.1574-6968.2007.00643.x

Consortium, T.U., 2017. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169. https://doi.org/10.1093/nar/gkw1099

Coordinators, N.R., 2015. Database resources of the National Center for Biotechnology Information NCBI Resource Coordinators. Nucleic Acids Res. 43. https://doi.org/10.1093/nar/gku1130

Courvalin, P., 2008. Predictable and unpredictable evolution of antibiotic resistance. J. Intern. Med. 264, 4–16. https://doi.org/10.1111/j.1365-2796.2008.01940.x

Cox, G., Stogios, P.J., Savchenko, A., Wright, G.D., 2015. Structural and molecular basis for resistance to aminoglycoside antibiotics by the adenylyltransferase ANT(2″)-Ia. MBio. https://doi.org/10.1128/mBio.02180-14

Dayhoff, M.O., Ledley, R.S.,1963. Comprotein: a computer program to aid primary protein structure determination. Proc. Fall Joint Comp. Conf. 22, 262–274.

Davies, J., Davies, D., 2010. Origins and evolution of antibiotic resistance. Microbiol. Mol. Biol. Rev. 74, 417–433. https://doi.org/10.1128/MMBR.00016-10

Delfosse, V., Girard, E., Birck, C., Delmarcelle, M., Delarue, M., Poch, O., Schultz, P., Mayer, C., 2009. Structure of the Archaeal Pab87 Peptidase Reveals a Novel Self-Compartmentalizing Protease Family. PLoS One 4, 1–9. https://doi.org/10.1371/journal.pone.0004712

Diaz-Torres, M.L., Villedieu, A., Hunt, N., McNab, R., Spratt, D.A., Allan, E., Mullany, P., Wilson, M., 2006. Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using ametagenomic approach. FEMS Microbiol. Lett. 258, 257–262. https://doi.org/10.1111/j.1574-6968.2006.00221.x

Diniz, C.G., Farias, L.M., Carvalho, M.A.R., Rocha, E.R., Smith, C.J., 2004. Differential gene expression in a Bacteroides fragilis metronidazole-resistant mutant. J. Antimicrob. Chemother. 54, 100–108. https://doi.org/10.1093/jac/dkh256

Dubois, D., Baron, O., Cougnoux, A., Delmas, J., Pradel, N., Boury, M., Bouchon, B., Bringer, M.A., Nougayr??de, J.P., Oswald, E., Bonnet, R., 2011. ClbP is a prototype of a peptidase subgroup involved in biosynthesis of nonribosomal peptides. J. Biol. Chem. 286, 35562–35570. https://doi.org/10.1074/jbc.M111.221960

Eddy, S.R., 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7. https://doi.org/10.1371/journal.pcbi.1002195

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Eliopoulos, G.M., Bush, K., 2001. New -Lactamases in Gram-Negative Bacteria: Diversity and Impact on the Selection of Antimicrobial Therapy. Clin. Infect. Dis. https://doi.org/10.1086/319610

Fabiane, S.M., Sohi, M.K., Wan, T., Payne, D.J., Bateson, J.H., Mitchell, T., Sutton, B.J., 1998. Crystal structure of the zinc-dependent ??-lactamase from Bacillus cereus at 1.9 ?? resolution: Binuclear active site with features of a mononuclear enzyme. Biochemistry 37, 12404–12411. https://doi.org/10.1021/bi980506i

Fernández-López, R., Pilar Garcillán-Barcia, M., Revilla, C., Lázaro, M., Vielva, L., De La Cruz, F., 2006. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. FEMS Microbiol. Rev. 30, 942–966. https://doi.org/10.1111/j.1574-6976.2006.00042.x

Feuermann, M., Gaudet, P., Mi, H., Lewis, S.E., Thomas, P.D., 2016. Large-scale inference of gene function through phylogenetic annotation of Gene Ontology terms: Case study of the apoptosis and autophagy cellular processes. Database 2016, 1–11. https://doi.org/10.1093/database/baw155

Fillgrove, K.L., Pakhomova, S., Schaab, M.R., Newcomer, M.E., Armstrong, R.N., 2007. Structure and mechanism of the genomically encoded fosfomycin resistance protein, FosX, from Listeria monocytogenes. Biochemistry 46, 8110–8120. https://doi.org/10.1021/bi700625p

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv1344

Fleiss J.L., Zubin J., 1969. On the methods and theory of clustering. Multivariate Behav Res. 4(2):235-50. doi: 10.1207/s15327906mbr0402_8.

Florian Fricke, W., Rasko, D.A., 2014. Bacterial genome sequencing in the clinic: Bioinformatic challenges and solutions. Nat. Rev. Genet. https://doi.org/10.1038/nrg3624

Fong, D.H., Burk, D.L., Blanchet, J., Yan, A.Y., Berghuis, A.M., 2017. Structural Basis for Kinase-Mediated Macrolide Antibiotic Resistance. Structure 25, 750–761.e5. https://doi.org/10.1016/j.str.2017.03.007

Forsberg, K., Patel, S., Gibson, M.K., Lauber, C.L., Knight, R., Fierer, N., Dantas, G., 2014. Bacterial phylogeny structures soil resistomes across habitats. Nature 509, 612–616. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Frère, J.M., Galleni, M., Bush, K., Dideberg, O., 2005. Is it necessary to change the classification of β-lactamases? J. Antimicrob. Chemother. 55, 1051–1053. https://doi.org/10.1093/jac/dki155

Fróes, A.M., da Mota, F.F., Cuadrat, R.R.C., D'Avila, A.M.R., 2016. Distribution and classification of serine Beta-lactamases in Brazilian hospital sewage and other environmental metagenomes deposited in public databases. Front. Microbiol. 7, 1–15. https://doi.org/10.3389/fmicb.2016.01790

Gales, A.C., Castanheira, M., Jones, R.N., Sader, H.S., 2012. Antimicrobial resistance among Gram-negative bacilli isolated from Latin America: Results from SENTRY Antimicrobial Surveillance Program (Latin America, 2008-2010). Diagn. Microbiol. Infect. Dis. 73, 354–360. https://doi.org/10.1016/j.diagmicrobio.2012.04.007

Galleni, M., Lamotte-brasseur, J., Rossolini, G.M., 2001. Standard Numbering Scheme for Class B Beta-Lactamases. Society 45, 660–663. https://doi.org/10.1128/AAC.45.3.660

Galperin, M.Y., Koonin, E. V., 2012. Divergence and convergence in enzyme evolution. J. Biol. Chem. https://doi.org/10.1074/jbc.R111.241976

Galperin, M.Y., Koonin, E. V, 2000. FUNCTIONAL GENOMICS AND ENZYME EVOLUTION. Library (Lond). 1–21.

Gherardini, P.F., Wass, M.N., Helmer-Citterich, M., Sternberg, M.J.E., 2007. Convergent Evolution of Enzyme Active Sites Is not a Rare Phenomenon. J. Mol. Biol. 372, 817–845. https://doi.org/10.1016/j.jmb.2007.06.017

Gibas C., Jambeck P., 2001. Desenvolvendo a bioinformática. Editora Campus e O'Reilly.

Gibson, M.K., Forsberg, K.J., Dantas, G., 2014. Improved annotation of antibiotic

resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 9, 1–10. https://doi.org/10.1038/ismej.2014.106

Gupta, R.S., 2004. The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. Crit. Rev. Microbiol. 30, 123–143. https://doi.org/10.1080/10408410490435133

Haack, F.S., Poehlein, A., Kr??ger, C., Voigt, C.A., Piepenbring, M., Bode, H.B., Daniel, R., Sch??fer, W., Streit, W.R., 2016. Molecular keys to the Janthinobacterium and Duganella spp. Interaction with the plant pathogen Fusarium graminearum. Front. Microbiol. 7. https://doi.org/10.3389/fmicb.2016.01668

Hächler, H.H., Santanam, P., Kayser, F.H., 1996. Sequence and Characterization of a Novel Chromosomal Aminoglycoside Phosphotransferase Gene, aph(3J)-IIb, in Pseudomonas aeruginosa. Antimicrob. Agents Chemother. 40, 1254–1256.

Hagen, J., 2000. The origins of bioinformatics. Nat. Rev. 1, 231–236.

Hall, B.G., Barlow, M., 2005. Revised Ambler classification of beta-lactamases. J. Antimicrob. Chemother. 55, 1050–1051. https://doi.org/10.1093/jac/dki130

Hall, B.G., Barlow, M., 2004. Evolution of the serine beta-lactamases: past, present and future. Drug Resist. Updat. 7, 111–23. https://doi.org/10.1016/j.drup.2004.02.003

Hall, B.G., Salipante, S.J., Barlow, M., 2004. Independent origins of subgroup Bl + B2 and subgroup B3 metallo-beta-lactamases. J. Mol. Evol. 59, 133–141. https://doi.org/10.1007/s00239-003-2572-9

Hall, B.G., Salipante, S.J., Barlow, M., 2003. The metallo-beta-lactamases fall into two distinct phylogenetic groups. J. Mol. Evol. 57, 249–254. https://doi.org/10.1007/s00239-003-2471-0

Holliday, G.L., Andreini, C., Fischer, J.D., Rahman, S.A., Almonacid, D.E., Williams, S.T., Pearson, W.R., 2012. MACiE: Exploring the diversity of biochemical reactions. Nucleic Acids Res. 40, 783–789. https://doi.org/10.1093/nar/gkr799

Hornung, C., Poehlein, A., Haack, F.S., Schmidt, M., Dierking, K., Pohlen, A., Schulenburg, H., Blokesch, M., Plener, L., Jung, K., Bonge, A., Krohn-Molt, I., Utpatel, C., Timmermann, G., Spieck, E., Pommerening-R??ser, A., Bode, E., Bode, H.B., Daniel, R., Schmeisser, C., Streit, W.R., 2013. The Janthinobacterium sp. HH01 Genome Encodes a Homologue of the V. cholerae CqsA and L. pneumophila LqsA Autoinducer Synthases. PLoS One 8. https://doi.org/10.1371/journal.pone.0055045

Hou, Z., Zhou, Y., Wang, H., Bai, H., Meng, J., Xue, X., Luo, X., 2011. Co-blockade of mecR1/blaR1 signal pathway to restore antibiotic susceptibility in clinical isolates of methicillin-resistant Staphylococcus aureus. Arch. Med. Sci. 7, 414–422. https://doi.org/10.5114/aoms.2011.23404

Hu, Y., Yang, X., Li, J., Lv, N., Liu, F., Wu, J., Lin, I.Y.C., Wu, N., Weimer, B.C., Gao, G.F., Liu, Y., Zhu, B., 2016. The bacterial mobile resistome transfer network connecting the animal and human microbiomes. Appl. Environ. Microbiol. https://doi.org/10.1128/AEM.01802-16

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682. https://doi.org/10.1093/bioinformatics/btq003

Hugenholtz, P., Goebel, B.M., Pace, N.R., 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. 180, 6793. https://doi.org/0021-9193/98/$04.00+0

Husain, F., Veeranagouda, Y., Hsi, J., Meggersee, R., Abratt, V., Wexler, H.M., 2013. Two multidrug-resistant clinical isolates of bacteroides fragilis carry a novel metronidazole resistance nim gene (nimJ). Antimicrob. Agents Chemother. https://doi.org/10.1128/AAC.00386-13

Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61, 1061–1067. https://doi.org/10.1093/sysbio/sys062

Jacoby, G.A., 2009. AmpC Beta-Lactamases. Clin. Microbiol. Rev. 22, 161–182. https://doi.org/10.1128/CMR.00036-08

Jacoby, G.A., Baquero, F., Cantón, R., Drusano, G.L., Shlaes, D.M., Projan, S.J., Roy, P.H., Babic, M., Bonomo, R.A., Rice, L.B., D'Costa, V., Wright, G.D., McPhee, J.B., Tamber, S., Brazas, M.D., Lewenza, S., Hancock, R.E.W., Walmsley, A.R., Rosen, B.P., Fux, C.A., Stoodley, P., Shirtliff, M., Costerton, J.W., Bush, K., Zapun, A., Macheboeuf, P., Vernet, T., Magalhães, M.L., Blanchard, J.S., Roberts, M.C., Schwarz, S., Moudgal, V. V., Kaatz, G.W., Canu, A., Leclercq, R., Dhand, A., Snydman, D.R., Périchon, B., Courvalin, P., Leuthner, K.D., Rybak, M.J., Shinabarger, D., Eliopoulos, G.M., Sköld, O., Karakousis, P.C., O'Shaughnessy, E.M., Lyman, C.A., Walsh, T.J., Lopez-Ribot, J.L., Patterson, T.F., Chandra, J., Mohammad, S., Ghannoum, M.A. et. al., 2009. Antimicrobial Drug Resistance. https://doi.org/10.1007/978-1-59745-180-2

Jandhyala, S.M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., Reddy,

D.N., 2015. Role of the normal gut microbiota. World J. Gastroenterol. https://doi.org/10.3748/wjg.v21.i29.8787

Jansson, D.S., Pringle, M., 2011. Antimicrobial susceptibility of Brachyspira spp. Isolated from commercial laying hens and free-living wild mallards (Anas platyrhynchos). Avian Pathol. 40, 387–393. https://doi.org/10.1080/03079457.2011.588197

Jaurin, B., Grundström, T., 1981. ampC cephalosporinase of Escherichia coli K-12 has a different evolutionary origin from that of beta-lactamases of the penicillinase type. Proc. Natl. Acad. Sci. U. S. A. 78, 4897–901. https://doi.org/10.1073/pnas.78.8.4897

Kämpfer, P., Rosselló-Mora, R., Hermansson, M., Persson, F., Huber, B., Falsen, E., Busse, H.J., 2007. Undibacterium pigrum gen. nov., sp. nov., isolated from drinking water. Int. J. Syst. Evol. Microbiol. 57, 1510–1515. https://doi.org/10.1099/ijs.0.64785-0

Karp, P.D., 2016. Can we replace curation with information extraction software? Database. https://doi.org/10.1093/database/baw150

Katoh, K., Rozewicki, J., Yamada, K.D., 2017. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 1–7. https://doi.org/10.1093/bib/bbx108

Kielak, A.M., Barreto, C.C., Kowalchuk, G.A., van Veen, J.A., Kuramae, E.E., 2016. The ecology of Acidobacteria: Moving beyond genes and genomes. Front. Microbiol. https://doi.org/10.3389/fmicb.2016.00744

Koonin, E. V., Tatusov, R.L., Galperin, M.Y., 1998. Beyond complete genomes: From sequence to structure and function. Curr. Opin. Struct. Biol. 8, 355–363. https://doi.org/10.1016/S0959-440X(98)80070-5

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., Haussler, D., 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J. Mol. Biol. 235, 1501–1531. https://doi.org/10.1006/jmbi.1994.1104

Kumar, S., Stecher, G., Peterson, D., Tamura, K., 2012. MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28, 2685–2686. https://doi.org/10.1093/bioinformatics/bts507

Laehnemann, D., Peña-Miller, R., Rosenstiel, P., Beardmore, R., Jansen, G., Schulenburg, H., 2014. Genomics of rapid adaptation to antibiotics: Convergent evolution and scalable sequence amplification. Genome Biol. Evol. 6, 1287–1301.

https://doi.org/10.1093/gbe/evu106

Lee, D., Das, S., Dawson, N.L., Dobrijevic, D., Ward, J., Orengo, C., 2016. Novel Computational Protocols for Functionally Classifying and Characterising Serine Beta-Lactamases. PLOS Comput. Biol. 12, e1004926. https://doi.org/10.1371/journal.pcbi.1004926

Leonard, D.A., Bonomo, R.A., Powers, R.A., 2013. Class D β-lactamases: A reappraisal after five decades. Acc. Chem. Res. https://doi.org/10.1021/ar300327a

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., Bryant, S.H., 2017. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45, D200–D203. https://doi.org/10.1093/nar/gkw1129

Martínez, J.L., 2008. Antibiotics and antibiotic resistance genes in natural environments. Science 321, 365–367. https://doi.org/10.1126/science.1159483

Marvaud, J.C., Lambert, T., 2017. Arr-cb is a rifampin resistance determinant found active or cryptic in Clostridium bolteae strains. Antimicrob. Agents Chemother. https://doi.org/10.1128/AAC.00301-17

Marvig, R.L., Sommer, L.M., Molin, S., Johansen, H.K., 2015. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat. Genet. 47, 57–65. https://doi.org/10.1038/ng.3148

Massova I., Mobashery S., 1999. Structural and mechanistic aspects of evolution of beta-lactamases and penicillin-binding proteins. Curr Pharm Des. 5(11):929-37.

McArthur, A.G., Waglechner, N., Nizam, F., Yan, A., Azad, M.A., Baylay, A.J., Bhullar, K., Canova, M.J., De Pascale, G., Ejim, L., Kalan, L., King, A.M., Koteva, K., Morar, M., Mulvey, M.R., O'Brien, J.S., Pawlowski, A.C., Piddock, L.J. V, Spanogiannopoulos, P., Sutherland, A.D., Tang, I., Taylor, P.L., Thaker, M., Wang, W., Yan, M., Yu, T., Wright, G.D., 2013. The comprehensive antibiotic resistance database. Antimicrob. Agents Chemother. 57, 3348–3357. https://doi.org/10.1128/AAC.00419-13

Medeiros A.A., 1997. Evolution and dissemination of beta-lactamases accelerated by generations of beta-lactam antibiotics. Clin Infect Dis. 24 Suppl 1:S19-45.

Meziane-Cherif, D., Lambert, T., Dupêchez, M., Courvalin, P., Galimand, M., 2008. Genetic and biochemical characterization of OXA-63, a new class D β-lactamase

from Brachyspira pilosicoli BM4442. Antimicrob. Agents Chemother. 52, 1264–1268. https://doi.org/10.1128/AAC.00684-07

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., Thomas, A.D., Ali, R., Sharma, K., Kyrpides, N.C., Reddy, T.B.K., 2017. Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancements. Nucleic Acids Res. 45, D446–D456. https://doi.org/10.1093/nar/gkw992

Mukhtar, T.A., Koteva, K.P., Hughes, D.W., Wright, G.D., 2001. Vgb from Staphylococcus aureus inactivates streptogramin B antibiotics by an elimination mechanism not hydrolysis. Biochemistry 40, 8877–8886. https://doi.org/10.1021/bi0106787

Murray, I. a, Shaw, W. V, 1997. MINIREVIEW O -Acetyltransferases for Chloramphenicol and Other Natural Products. Antimiro 41, 1–6.

Naas, T., Oueslati, S., Bonnin, R.A., Dabos, M.L., Zavala, A., Dortet, L., Retailleau, P., Iorga, B.I., 2017. Beta-lactamase database (BLDB) – structure and function. J. Enzyme Inhib. Med. Chem. 32, 917–919. https://doi.org/10.1080/14756366.2017.1344235

Norman, A., Hansen, L.H., Sorensen, S.J., 2009. Conjugative plasmids: vessels of the communal gene pool. Philos. Trans. R. Soc. B Biol. Sci. 364, 2275–2289. https://doi.org/10.1098/rstb.2009.0037

Novick, R.P., Clowes, R.C., Cohen, S.N., Iii, R.O.Y.C., Falkow, S., 1976. Uniform nomenclature for bacterial plasmids: a proposal. Bacteriol. Rev. 40, 525.

Nelson, D., Cox, M., 2006. Lehninger principles of biochemistry. 4th ed. https://doi.org/10.1002/bmb.2005.494033010419.

Oh, H.M., Kang, I., Vergin, K.L., Lee, K., Giovannoni, S.J., Cho, J.C., 2011. Genome sequence of Oceanicaulis sp. strain HTCC2633, isolated from the western Sargasso Sea. J. Bacteriol. 193, 317–318. https://doi.org/10.1128/JB.01267-10

Omelchenko, M. V, Galperin, M.Y., Wolf, Y.I., Koonin, E. V, 2010. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol. Direct 5, 31. https://doi.org/10.1186/1745-6150-5-31

Ortiz, A.R., Strauss, C.E.M., Olmea, O., 2009. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Sci. 11, 2606–2621. https://doi.org/10.1110/ps.0215902

Ouellette, M., Bissonnette, L., Roy, P.H., 1987. Precise insertion of antibiotic resistance determinants into Tn21-like transposons: nucleotide sequence of the

OXA-1 beta-lactamase gene. Proc. Natl. Acad. Sci. U. S. A. 84, 7378–7382. https://doi.org/10.1073/pnas.84.21.7378

Ouzounis, C. a, Coulson, R.M.R., Enright, A.J., Kunin, V., Pereira-Leal, J.B., 2003. Classification schemes for protein structure and function. Nat. Rev. Genet. 4, 508–519. https://doi.org/10.1038/nrg1113

Pallen, M.J., 2016. Microbial bioinformatics 2020. Microb. Biotechnol. https://doi.org/10.1111/1751-7915.12389

Philippon, A., Slama, P., Dény, P., Labia, R., 2016. A Structure-Based Classification of Class A beta-Lactamases , a Broadly. Clin. Microbiol. Rev. 29, 29–57. https://doi.org/10.1128/CMR.00019-15.Address

Picardeau, M., Bulach, D.M., Bouchier, C., Zuerner, R.L., Zidane, N., Wilson, P.J., Creno, S., Kuczek, E.S., Bommezzadri, S., Davis, J.C., McGrath, A., Johnson, M.J., Boursaux-Eude, C., Seemann, T., Rouy, Z., Coppel, R.L., Rood, J.I., Lajus, A., Davies, J.K., Médigue, C., Adler, B., 2008. Genome sequence of the saprophyte Leptospira biflexa provides insights into the evolution of Leptospira and the pathogenesis of leptospirosis. PLoS One 3, 1–9. https://doi.org/10.1371/journal.pone.0001607

Pratap, S., Katiki, M., Gill, P., Kumar, P., Golemi-Kotra, D., 2016. Active-site plasticity is essential to carbapenem hydrolysis by OXA-58 class D ??-lactamase of Acinetobacter baumannii. Antimicrob. Agents Chemother. 60, 75–86. https://doi.org/10.1128/AAC.01393-15

Ramirez, M.S., Tolmasky, M.E., 2010. Aminoglycoside Modifying Enzymes. Drug Resist Updat 13, 151–171. https://doi.org/10.1016/j.drup.2010.08.003.Aminoglycoside

Rasmussen, B.A., Bush, K., 1997. Carbapenem-hydrolyzing ??-lactamases. Antimicrob. Agents Chemother. 41, 223–232.

Rigoutsos, I., Huynh, T., Floratos, A., Parida, L., Platt, D., 2002. Dictionary-driven protein annotation. Nucleic Acids Res. 30, 3901–3916.

Roberts, M.C., 2008. Update on macrolide-lincosamide-streptogramin, ketolide, and oxazolidinone resistance genes. FEMS Microbiol. Lett. https://doi.org/10.1111/j.1574-6968.2008.01145.x

Robicsek, A., Jacoby, G.A., Hooper, D.C., 2006. The worldwide emergence of plasmid-mediated quinolone resistance. Lancet Infect. Dis. 6, 629–640. https://doi.org/10.1016/S1473-3099(06)70599-0

Salipante, S.J., Hall, B.G., 2003. Determining the limits of the evolutionary potential of

an antibiotic resistance gene. Mol. Biol. Evol. 20, 653–659. https://doi.org/10.1093/molbev/msg074

Sanger F., 1959. Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes. Science. 15;129(3359):1340-4.

Savjani, J., Gajjar, A., Savjani, K., 2009. Mechanisms of Resistance: useful tool to design Antibacterial Agents for Drug-Resistant bacteria. Mini-Reviews Med. Chem. 9, 194–205.

Schwarz, S., Kehrenberg, C., Doublet, B., Cloeckaert, A., 2004. Molecular basis of bacterial resistance to chloramphenicol and florfenicol. FEMS Microbiol. Rev. https://doi.org/10.1016/j.femsre.2004.04.001

Shaw, K.J., Rather, P.N., Hare, R.S., Miller, G.H., 1993. Molecular Genetics of Aminoglycoside Resistance Genes and Familial Relationships of the Aminoglycoside-Modifying Enzymes. Microbiol. Rev. 57, 138–163.

Sigrist, C.J.A., De Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I., 2013. New and continuing developments at PROSITE. Nucleic Acids Res. 41, 344–347. https://doi.org/10.1093/nar/gks1067

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., Lehtinen, S., Studer, R.A., Thornton, J., Orengo, C.A., 2015. CATH: Comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. https://doi.org/10.1093/nar/gku947

Silveira, M.C., Albano, R.M., Asensi, M.D., Carvalho-Assef, A.P.D.A., 2016. Description of genomic islands associated to the multidrug-resistant Pseudomonas aeruginosa clone ST277. Infect. Genet. Evol. 42, 60–65. https://doi.org/10.1016/j.meegid.2016.04.024

Singh, R., Saxena, A., Singh, H., 2009. Identification of group specific motifs in beta-lactamase family of proteins. J. Biomed. Sci. 16, 109. https://doi.org/10.1186/1423-0127-16-109

Singh, R., Singh, H., 2008. DLact : An Antimicrobial Resistance Gene Database. J. Comput. Intell. Bioinforma. 1, 93–108.

Sinha, S., Lynn, A.M., 2014. HMM-ModE: implementation, benchmarking and validation with HMMER3. BMC Res. Notes 7, 483. https://doi.org/10.1186/1756-0500-7-483

Söding, J., 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951–960. https://doi.org/10.1093/bioinformatics/bti125

Speer, B.S., Shoemaker, N.B., Salyers, A.A., 1992. Bacterial resistance to tetracycline: Mechanisms, transfer, and clinical significance. Clin. Microbiol. Rev. https://doi.org/10.1128/CMR.5.4.387

Srivastava, a., Singhal, N., Goel, M., Virdi, J.S., Kumar, M., 2014. CBMAR: a comprehensive -lactamase molecular annotation resource. Database 2014, bau111-bau111. https://doi.org/10.1093/database/bau111

Stogios, P.J., Evdokimova, E., Morar, M., Koteva, K., Wright, G.D., Courvalin, P., Savchenko, A., 2015. Structural and functional plasticity of antibiotic resistance nucleotidylyltransferases revealed by molecular characterization of lincosamide nucleotidylyltransferases Lnu(A) and Lnu(D). J. Mol. Biol. https://doi.org/10.1016/j.jmb.2015.04.008

Stogios, P.J., Kuhn, M.L., Evdokimova, E., Law, M., Courvalin, P., Savchenko, A., 2017. Structural and Biochemical Characterization of Acinetobacter spp. Aminoglycoside Acetyltransferases Highlights Functional and Evolutionary Variation among Antibiotic Resistance Enzymes. ACS Infect. Dis. https://doi.org/10.1021/acsinfecdis.6b00058

Suzuki, H., Yano, H., Brown, C.J., Top, E.M., 2010. Predicting plasmid promiscuity based on genomic signature. J. Bacteriol. 192, 6045–6055. https://doi.org/10.1128/JB.00277-10

Tamminen, M., Virta, M., Fani, R., Fondi, M., 2012. Large-scale analysis of plasmid relationships through gene-sharing networks. Mol. Biol. Evol. 29, 1225–1240. https://doi.org/10.1093/molbev/msr292

Thai, Q.K., Bös, F., Pleiss, J., 2009. The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. BMC Genomics 10, 390. https://doi.org/10.1186/1471-2164-10-390

Todd, a E., Orengo, C. a, Thornton, J.M., 2001. Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307, 1113–1143. https://doi.org/10.1006/jmbi.2001.4513

Toth, M., Antunes, N.T., Stewart, N.K., Frase, H., Bhattacharya, M., Smith, C.A., Vakulenko, S.B., 2016. Class D β-lactamases do exist in Gram-positive bacteria. Nat. Chem. Biol. https://doi.org/10.1038/nchembio.1950

Triant, D.A., Pearson, W.R., 2015. Most partial domains in proteins are alignment and annotation artifacts. Genome Biol. 16, 99. https://doi.org/10.1186/s13059-015-0656-7

Urbach, C., Evrard, C., Pudzaitis, V., Fastrez, J., Soumillion, P., Declercq, J.P., 2009.

Structure of PBP-A from Thermosynechococcus elongatus, a Penicillin-Binding Protein Closely Related to Class A ??-Lactamases. J. Mol. Biol. 386, 109–120. https://doi.org/10.1016/j.jmb.2008.12.001

Van Boeckel, T.P., Gandra, S., Ashok, A., Caudron, Q., Grenfell, B.T., Levin, S.A., Laxminarayan, R., 2014. Global antibiotic consumption 2000 to 2010: An analysis of national pharmaceutical sales data. Lancet Infect. Dis. 14, 742–750. https://doi.org/10.1016/S1473-3099(14)70780-7

Wachino, J.I., Yamaguchi, Y., Mori, S., Kurosaki, H., Arakawa, Y., Shibayama, K., 2013. Structural insights into the subclass B3 metallo-??-lactamase SMB-1 and the mode of inhibition by the common metallo-??-lactamase inhibitor mercaptoacetate. Antimicrob. Agents Chemother. 57, 101–109. https://doi.org/10.1128/AAC.01264-12

Ward, N.L., Challacombe, J.F., Janssen, P.H., Henrissat, B., Coutinho, P.M., Wu, M., Xie, G., Haft, D.H., Sait, M., Badger, J., Barabote, R.D., Bradley, B., Brettin, T.S., Brinkac, L.M., Bruce, D., Creasy, T., Daugherty, S.C., Davidsen, T.M., DeBoy, R.T., Detter, J.C., Dodson, R.J., Durkin, A.S., Ganapathy, A., Gwinn-Giglio, M., Han, C.S., Khouri, H., Kiss, H., Kothari, S.P., Madupu, R., Nelson, K.E., Nelson, W.C., Paulsen, I., Penn, K., Ren, Q., Rosovitz, M.J., Selengut, J.D., Shrivastava, S., Sullivan, S.A., Tapia, R., Thompson, S., Watkins, K.L., Yang, Q., Yu, C., Zafar, N., Zhou, L., Kuske, C.R., 2009. Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. Appl. Environ. Microbiol. https://doi.org/10.1128/AEM.02294-08

Wei, D., Jiang, Q., Wei, Y., Wang, S., 2012. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics 13, 174. https://doi.org/10.1186/1471-2105-13-174

Weis, A.M., Storey, D.B., Taff, C.C., Andrea K. Townsend, Bihua C. Huang, Nguyet T. Kong, Kristin A. Clothier, Abigail Spinner, B.A.B., Weimera, B.C., 2016. Genomic Comparison of Campylobacter spp. and Their Potential for Zoonotic Transmission between Birds, Primates, and Livestock. Appl. Environ. Microbiol. 82, 7165–7175. https://doi.org/10.1128/JB.185.2.553

Widmann, M., Pleiss, J., Oelschlaeger, P., 2012. Systematic analysis of metallo-β-lactamases using an automated database. Antimicrob. Agents Chemother. 56, 3481–3491. https://doi.org/10.1128/AAC.00255-12

Wilke, M.S., Hills, T.L., Zhang, H.Z., Chambers, H.F., Strymadka, N.C.J., 2004. Crystal structures of the apo and penicillin-acylated forms of the BlaR1 β-lactam sensor

of Staphylococcus aureus. J. Biol. Chem. https://doi.org/10.1074/jbc.M407054200

Wright, G.D., 2005. Bacterial resistance to antibiotics: Enzymatic degradation and modification. Adv. Drug Deliv. Rev. 57, 1451–1470. https://doi.org/10.1016/j.addr.2005.04.002

Wright, G.D., Thompson, P.R., 1999. Aminoglycoside Phospotransferases: proteins, structure and mechanisms. Front. Biosci. 1–17.

Yim, O., Ramdeen, K.T., 2015. Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. Quant. Methods Psychol. 11, 8–21. https://doi.org/10.20982/tqmp.11.1.p008

Yin, C., Hulbert, S.H., Schroeder, K.L., Mavrodi, O., Mavrodi, D., Dhingra, A., Schillinger, W.F., Paulitz, T.C., 2013. Role of bacterial communities in the natural suppression of Rhizoctonia solani bare patch disease of wheat (Triticum aestivum L.). Appl. Environ. Microbiol. 79, 7428–7438. https://doi.org/10.1128/AEM.01610-13

Yong, D., Toleman, M.A., Giske, C.G., Cho, H.S., Sundman, K., Lee, K., Walsh, T.R., 2009. Characterization of a new metallo-β-lactamase gene, blaNDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. Antimicrob. Agents Chemother. 53, 5046–5054. https://doi.org/10.1128/AAC.00774-09

Zhang, W., Fisher, J.F., Mobashery, S., 2009. The bifunctional enzymes of antibiotic resistance. Curr. Opin. Microbiol. 12, 505–511. https://doi.org/10.1016/j.mib.2009.06.013

9 APÊNDICES

9.1 Scripts

9.1.1 Obtenção do PDB ID

```
#!/usr/bin/env perl
#use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [PDB summary from GenBank]  >> pdb_ids
#Program to extract the PDB ID from file downloaded from GenBank
#Input    PDB summary downloaded from GenBank
#Output PDB list with PDB ID

my $arquivo= $ARGV[0];
# Open the PDB summary downloaded from GenBank
open (IN, "<", $arquivo) || die "File not open\n";
while (my $row = <IN>) {
        # remove \n of the line end
        chomp $row;
        if ($row=~m/^MMDB\sID:.*PDB\sID:\s(.*)/gi){
        print "$1\n";
        }
}
close (IN);
exit;
```

9.1.2 Remoção de átomos duplicados do arquivo .pdb

```
#!/usr/bin/perl –w
# Author: Alex Herbert
use Getopt::Long;
use File::Basename;
use File::Copy;

my $prog = basename($0);
my $bak = 'bak';

my $usage = "
  Program to remove all but the first alternative position from ATOM records
  in PDB files

Usage:
  $prog input [...]

Options:
  input     PDB files
  -help     Print this help and exit
  -bak=%s   Save the original file to [filename].[ext] (default $bak)
  -silent   Do not print the positions that are ignored
";

my $help;
GetOptions(
        "help" => \$help,
```

```
        "bak" => \$bak,
        "silent" => \$silent,
);
die $usage if $help;
@ARGV or die $usage;

# Get the input files
my @files;
for (@ARGV) {
        if (m/\*/) {
                push @files, glob "$_";
        } else {
                push @files, $_;
        }
}

for $input (@files) {
        open (IN, $input) or die "Failed to open '$input': $!\n";
        @pdb = <IN>;
        close IN;

        if ($bak) {
                move($input, "$input.$bak") or
                        die "Failed to create backup '$input.$bak': $!\n";
        }

        open (OUT, ">$input") or die "Failed to open '$input': $!\n";
        my %alt;
        my $ignore;
        for (@pdb) {
                if (m/^ATOM/) {
                        $alt = substr($_,16,1);
                        $code = substr($_,21,6);

                        # Ignore all but the first ALT code
                        $alt2 = $alt{$code};
                        if (defined $alt2) {
                                if ($alt ne $alt2) {
                                        unless ($silent) {
                                                warn "Ignoring $input '$code' ALT '$alt'\n"
                                                        unless $ignore{$code}{$alt};
                                                $ignore{$code}{$alt} = 1;
                                        }
                                        next;
                                };
                        }
                        $alt{$code} = $alt;
                }
                print OUT $_;
        }
        close OUT
}
```

9.1.3 Selecionar resoluções maiores que 3Å

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [.pdb file] >> goodResolution
```

```
#Program to remove pdb files with resolution less than 3
#Input    file .pdb
#Output list of pdb files with resolution less than 3, print ih the terminal, goodResolution

my $arquivo= $ARGV[0];
# Open the file .pdb
open (IN, "<", $arquivo) || die "File not open\n";
while (my $row = <IN>) {
        # remove \n of the line end
        chomp $row;
        if ($row=~m/^REMARK\s{3}\d\sRESOLUTION\.\s{4}(\d)\..*ANGSTROMS\./gi){
                if ($1 < 3) {
                print "$ARGV[0]\n";
                exit;
                }
        }
}
close (IN);
exit;
```

9.1.4 Fazer uma lista indicando os arquivos PDB com monômeros e homomultímeros

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [LIST] >> [OUTPUT]
#Program to make a list indicating the monomers and homodimers, that can be use by MaxCluster program
#Input    PDB list
#Output PDB list indicating homodimers and monomers
#Run in directory with .fasta files

#Open the pdb list with good resolution
my $resolution= $ARGV[0];
open (RES, "<", $resolution) || die "File resolution not open\n";
#make a list indicating the monomers and homodimers
#open the pdb list
while (my $pdb = <RES>) {
        # remove \n of the line end
        chomp $pdb;
        if ($pdb=~m/(.+)\.pdb/gi){

        my $arquivo= "fasta_files/$1.fasta";
        # Open the fasta file correspoding to the pdb on the list
        open (IN, "<", $arquivo) || die "File fasta not open\n";
        my $chains = 0;
        #Check the numbers of sequences in fasta file
        while (my $row = <IN>) {
                # remove \n of the line end
                chomp $row;
                if ($row=~m/^>/gi){
                        $chains++;
                }
        }
        #If there is more than one sequence run the align on terminal
        if ($chains > 1){
                system("cd fasta_files/ && clustalw $1.fasta > $1_score");
```

```
	} else { #if there is just one sequence print on terminal the file name indicating that it is a monomer

		print "mono_$1\n";
		next;
	}
	#after the align, access the file output contaning the score
	my $file= "fasta_files/$1_score";

	open (IN2, "<", $file) || die "File score not open\n";
	#print at terminal the file name of homodimers, with score iqual 100, identical sequences
	while (my $row2 = <IN2>) {
		chomp $row2;
		if ($row2=~m/^Sequences.+Score:\s+(\d+)/gi){
		my $score = $1;
			if ($score != 100) {
			next;
			}
		}
	}
	print "homo_$1\n";
	#remove the output file from align
	system("cd fasta_files/ && rm $1_score");
	close IN2;
	close IN;
	}
}
exit;
```

9.1.5 Extrair a cadeia A dos homodímeros nos arquivos PDB

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script homo_mono_list >> pbd_chainA
#Program to extraxt the chain A from the pdb file of homodimers
#Input    list of pdb files indicating homodimers and monomers
#Output  pdb files with one chain AND a list od pdb that will be used by Maxcluster program
#Run at a directtory containing the pdb files

#open the list with monomers and homodimers
my $arquivo= $ARGV[0];
open (IN, "<", $arquivo) || die "File homo_mono_list not open\n";
#read the files that are homodimers and extract chain A
while (my $row = <IN>) {
	# remove \n f the line end
	chomp $row;
	if ($row=~m/homo_(.+)/gi){

	#acess the file pdb correspoding to the homodimer
	my $arquivo2= "pdb_files/$1.pdb";
	open (IN2, "<", $arquivo2) || die "File .pdb not open\n";
	#creat a new file with the chain A
		while (my $row2 = <IN2>) {
			# remove \n of the line end
			chomp $row2;
			#file name
			my $arquivo3= "pdb_files/$1_A.pdb";
			open (IN3, ">>", $arquivo3) || die "File 3 not open\n";
			#print the rows corresponding to ATOM
```

```
                        if ($row2=~m/^ATOM\s+\d+\s+\w+\s+\w{3}\s+(\w)\s+\d+\s+.+/gi){
                                if ($1 eq "A") {
                                print IN3 "$row2\n";
                                }
                        }
                close IN3;
                }
        #print in the terminal the name of pdb files which will be used by MaxCluster program, pbd_chainA
        print "$1_A.pdb\n";
        close IN2;
        } elsif ($row=~m/mono_(.+)/gi) {
                print "$1.pdb\n";
        }
}
close IN;
exit;
```

9.1.6 Remoção das quebras de linha dos arquivos FASTA

```
#!/usr/bin/perl -w
use strict;
# Author: Melise Chaves Silveira
# > perl script [FASTA] >> single_[FASTA]
#Program to extraxt \n of line containing amino acids sequences
#Input    fasta files
#Output fasta file without \n
#Run at a directtory containing the fasta files

#open the fasta file
my $input_fasta=$ARGV[0];
open(IN,"<$input_fasta") || die ("Error opening $input_fasta $!");
while (my $line = <IN>){
        chomp $line;
        if ($line=~m/^>/) {
                print "\n",$line,"\n"; }
        else { print $line; }
}
print "\n";
exit;
```

9.1.7 Seleção da sequência correspondente à cadeia A para cada homodímeros nos arquivos FASTA

```
#!/usr/bin/perl -w
use strict;
# Author: Melise Chaves Silveira
# > perl script [FASTA] >> chainA_bcl
#Program to print in a multifasta file all chains A
#Input    fasta files
#Output multifastafile with chains A
#Run at a directtory containing the fasta files

#open the fasta file
my $input_fasta=$ARGV[0];
open(IN,"<$input_fasta") || die ("Error opening $input_fasta $!");
while (my $line = <IN>){
```

```
        chomp $line;
        if ($line=~m/^>.+:B.*/gi) {
        exit;
        } elsif ($line=~m/(^>)(.*)/gi) {
                print "$line\n"; }
        elsif ($line=~m/^$/gi){
        next;}
        else { print "$line\n"; }
}
print "\n";
exit;
```

9.1.8 Inserção de um número GI hipotético no cabeçário FASTA

```
#!/usr/bin/perl -w
use strict;
# Author: Melise Chaves Silveira
# > perl script chainA_bcl >> chainAgi_bcl
#Program to insert a imaginary GI
#Input    multifasta file
#Output multifastafile with GI

#open the multifasta file
my $input_fasta=$ARGV[0];
open(IN,"<$input_fasta") || die ("Error opening $input_fasta $!");
my $x= 0;

while (my $line = <IN>){
        chomp $line;
        if ($line=~m/(^>)(.*)/) {
                $x= $x+1;
                print "\n",$1,"gi|$x|",$2,"\n"; }
        else { print $line; }
}
print "\n";
exit;
```

9.1.9 Remoção de linhas vazias do arquivo final

```
awk 'NF' chainAgi_bcl > seqs_gi_chainA
```

9.1.10 Criar arquivos multi-FASTA com todas as sequências referentes a cada cluster formados pelo BLASTClust

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [BLASTClust output]
#Program to creat multi-FASTA files with cluster's sequences
#Input BLASTClust output
#Output multi-FASTA file
```

```
my $file = $ARGV[0];
	#open the file with BlastClust cluster result
	open (IN, "<", $file) || die "File cluster result not open\n";
#if (-z $file){ #remove empty file
#		system ("rm $file");
#	}else{
	my $z = 0;
		while (my $row = <IN>) {
			# remove \n of line end
			chomp $row;
			#print "$row\n";
			my @cluster = split ('\s', $row);
			#print in a multifasta file the sequences
			#clusters number
			$z++;
			#print "$z\n";
			my $file2= $ARGV[1]; #"results/seqs_gi_chainA";
			#open de multifasta file with all sequences
			open (IN2, "<", $file2) || die "File sequences not open\n";
				while (my $row2 = <IN2>) {
					chomp $row2;
					#print "$row2\n";
						foreach my $sequence (@cluster){
						#print "$sequence\n";
							if ($row2 =~m/^>(gi\|$sequence\|.*)/gi){
							system <<EOF;
							grep -A 999999 ">gi\|$sequence\|" $file2
| awk 'NR>1 && /^>/{exit} 1' >> MB_cluster$z
EOF
							}
						}
				}
		}
close IN2;
exit;
```

9.1.11 Identificação de sobreposições entre os resultados do hmmsearch

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira

#  perl script [.tbl 1] [.tbl 2]
#Program to find intersections between hmmsearch results of two models
#Input    files .tbl 1 e .tbl 2, outputs of hmmsearch
#Output print the intersection between two models, repeted file, and all unic sequences, uniq file

#Open the first file .tbl, output from hmmsearch
my $arquivo= $ARGV[0];
open (IN, "<", $arquivo) || die "File 1 .tbl not open\n";
my @ids;
#Read each line of the result file .tbl 1
while (my $row = <IN>) {
	#remove \n of the line end
	chomp $row;
	#separate by space to get only the sequences indentifiers that showed match in hmmsearch result
	my ($target) = split /\s/, $row;
	#not consider lines beging with hash
	if ($target ne "\#") {
```

```
		push (@ids, $target);
		}
}
#Open the second file .tbl, output from hmmsearch
my $arquivo2= $ARGV[1];
open (IN2, "<", $arquivo2) || die "File 2 .tbl not open\n";
#open a file where the unics IDS will be stored
my $arquivo3 = "unics_$ARGV[0]_$ARGV[1]";
open (OUT, ">> $arquivo3") or die "NÃ£o foi possÃ-vel abrir o arquivo '$arquivo3' \n";
#open a file where the repeated IDS will be stored
my $arquivo4 = "repeted_$ARGV[0]_$ARGV[1]";
open (OUT2, ">> $arquivo4") or die "NÃ£o foi possÃ-vel abrir o arquivo '$arquivo4' \n";
#Reads each line of the result file 2 .tbl
while (my $row2 = <IN2>) {
		#remove \n of the line end
		chomp $row2;
		#separate by space to get only the sequences indentifiers that showed match in hmmsearch result
		my ($target2) = split /\s/, $row2;
		#not consider lines beging with hash
		if ($target2 ne "\#") {
			#check if the sequence have alread showed match with the fisrt model
			if (grep {$_ eq $target2} @ids) {
				print OUT2 "$target2\n";
				} else{
				print OUT "$target2\n";
				}
		}
}
close (IN);
close (IN2);
close (OUT);
exit;
```

9.1.12 Identificação do nó dos clados correspondentes às classes de BLs

Autor: Fábio Mota

9.1.13 Construção de arquivos multi-FASTA com as sequências em cada clado

Autor: Fábio Mota

9.1.14 Separação das sequências resultado do hmmsearch que pertencem ao clado da classe de BL e as que não pertencem

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# perl script [hmmsearch result] [phylogeny result]
#Program to separe sequences identified by hmmsearch that belong to the clade from that that not belong
#Input    hmmsearch result and phylogeny result
```

```
#Output tab file

#Open the file with the sequences identified by hmmsearch
my $arquivo= $ARGV[0];
open (IN, "<", $arquivo) || die "File cluster.tbl not open\n";
my %score = () ;
while (my $row = <IN>) {
	# retira \n do final da linha
	chomp $row;
	if ($row=~m/^(?!#)(^.*\d)\s+-\s{10}.*\s+.*\s+(\d+\.*\d*e-\d+)\s*(\d+\.\d+).*$/gi){ #find the line with
sequences
		$score{$1} .= $3; #store sequence and score into a hash
	}
}
#Open the file where are the sequences that were allocated in class D according to the phylogeny
my $arquivo2 = $ARGV[1];
open (IN2, "< $arquivo2") or die "File phylogene not open \n";
my @subjects;
while (<IN2>) {
	chomp $_;
	if ($_=~m/(cath.*\|.*\/\d+-\d+)||(cath.*\|.*\/\d+\d+)_.*$/gi){
	#print "$1\n";
	push (@subjects, $1); #an array to store the sequences from the clade
	}
}
my @keys = keys %score;
my @values= values %score;
my $sizeKeys= (scalar @keys) -1;
for (my $i=0; $i <= $sizeKeys; $i++) {
	if ( grep {$_ eq $keys[$i]} @subjects ){
		print "$keys[$i]\t$values[$i]\n";#print in the first and second collums the sequences (and
score) identidied by hmmsearch and the phylogeny
	}else{
		print "\t\t$keys[$i]\t$values[$i]\n";#print in the third and fourth collums the sequences (and
score) identidied only by hmmsearch
	}
}
exit;
```

9.1.15 Criar arquivos multi-FASTA do resultado da busca usando hmmsearch

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [all_sequences.fasta]
#Program to make multifasta file corresponding to hmmsearch results
#Input    multifasta file with all seqs; hmmsearch output .tbl
#Output multifastafile corresponding to the groups

my $arquivo= $ARGV[0];
#open de multifasta file with all sequences
open (IN, "<", $arquivo) || die "File $arquivo not open\n";
#put the database in a array
my @database;
while (<IN>){
	chomp $_;
	if ($_ =~ m/^>(.*)/){
		my @headers = split (/\s/,$1);
		push @database, $headers[0];
```

```
	} else {push @database, $_;}

}
#print "@database\n";
#open the output of hmmsearch
my @tbl = `cd /home/melise/projeto_doutorado/genomas_completos_bac/correcao_plasmidios/hmmsearch &&
ls SA_c.tbl`; #to extract from a specific .tbl file, you can use “ls [A-Z][A-Z].tbl`”

foreach my $tbl_outs (@tbl) {
	chomp $tbl_outs;
	open (IN2, "<", $tbl_outs) || die "File $tbl_outs not open\n";
	my @extractname1 = split (/\./, $tbl_outs);
	my $arquivo2 = "./$extractname1[0]\.fasta"; #creat output path
	open (OUT, ">", $arquivo2) || die "File $arquivo2 not open\n";

	#read the cluster file
	#my $count = 0; #create a gi
	while (my $row = <IN2>) {
		# remove \n of line end
		chomp $row;
		#read the row saying which cluster sequences belong
		if (!($row=~m/#/gi)){
			my @ids = split (/\s-\s/,$row);
			#print "$ids[0]\n";
			my @ids2 = split (/\s/,$ids[0]);
			my $print = $ids2[0];
			#print "$print\n";
				if (grep {$_ eq $print} @database){
				#print ">$print\n$database[$print]\n";
				my $search_for = "$print";
				my( $index )= grep { $database[$_] eq $search_for } 0..$#database;
				my $seq = $index + 1;
				#$count++;
				print OUT ">$print\n$database[$seq]\n";
				}
		}
	}
	close IN2;
	#exit;
}
close IN;
exit;
```

9.1.16 Extrair a anotação das sequências nos clusters resultantes do BLASTClust após BLASTP

```
#!/usr/bin/perl -w
use strict;
 # Author: Melise Chaves Silveira
# > perl script [BLATp output]
#Program to get the sequence’s annotation
#Input BLASTp output
#Output beta-lactamase enzyme names

#open the blastp out file
my $input_fasta=$ARGV[0];
open(IN,"<$input_fasta") || die ("Error opening $input_fasta $!");
#put the annotation
my %annotation;
```

```
my $match_mumber;
while (<IN>){
	chomp $_;
	my @outs = split (/\t/,$_);
	$match_mumber++;
	%annotation->{$outs[4]} = $outs[3];
	#push @annotation, $outs[4];
}
my $size = 0;
my $count = 0;
my @enzymes;

for my $key ( keys %annotation ) {
	$size = $size + $annotation{$key};
	$count++;
	if (($key=~ /.*(\s([a-zA-Z]{3,4}-\d{1,3})\s).*/) | ($key=~ /.*(\s([A-Z]{1}[a-z]{2}[A-Z]{1})\s).*/)) {
	push @enzymes, $1;
	}
}
my %hashTemp = map { $_ => 1 } @enzymes;
my @sorted_enzymes = sort keys %hashTemp;
#my @unique_enzymes = do { my %seen; grep { !$seen{$_}++ } @enzymes };
print @sorted_enzymes;
my $sizeMedio= ($size/$count);
#print "\n$size\n";
#print "\n$count\n";
print "\n$sizeMedio\n";
print "$match_mumber\n";
exit;
```

9.1.17 Identificação de patterns específicos em sequências de BLs

```
#!/usr/bin/perl -w
use strict;
# Author: Melise Chaves Silveira
# > perl script [multi-FASTA file]
#Program to identify motifs
#Input    multi-FASTA file with sequences
#Output number of residues before the motif

#open the multifasta file
my $input_fasta=$ARGV[0];
open(IN,"<$input_fasta") || die ("Error opening $input_fasta $!");
my $header;
my $pre;
my $pre_length;
while (<IN>) {
	chomp $_;
	if ($_ =~ /^$/){
	next;
	}
	if ($_ =~ /^>/g) {
	$header = $_;
	#print "$header\n";
		} elsif ($_ =~ /(^\w*)(K\w{2}S)\w*/gi) { # motif, ex
[PA]\wS[ST]FK[LIV][PALV]\w[STA][LI]
				$pre = $1;
				$pre_length = length ($pre);
				print "$pre_length\n";
				print "$header\n$2\n";
```

```
                        } #else {
                                #print "$header\n$_\n";
                        #}
}
exit;
```

9.1.18 Criação de um arquivo com as sequências cromossômicas e outro com as sequências plasmidiais

```
#!/usr/bin/env perl
use warnings;
use strict;
use File::chdir;
use Cwd;
use LWP::Simple;
# Author: Melise Chaves Silveira
# > perl script [directory with the organism files]

#Program to make FASTA files with chromosome and plasmidial sequences
#Input    directory with the organism files
#Output two FASTA files, one with chromosome sequences and other with the plasmidial ones

#open the directory of the organism
my $dir2= $ARGV[0];
opendir my $dh2, $dir2 or die "Could not open '$dir2' 2 for reading: $!\n";
chdir($ARGV[0]) or die "$!"; #go to the organism directory to open the files
my $content;
my $url;
        while (my $file = readdir $dh2) { #read the files in the directory
                next if $file =~ /^\./ or $file =~ /config_file/;
                if ($file =~ /(.*)\.faa/){ #get NCBI code
                #print "\n$1\n";
                my $url = "https://www.ncbi.nlm.nih.gov/nuccore/$1";#acess NCBI webpage correspoding o
the code
                $content = get $url;
                die "Could not get $url" unless defined $content;
                #print "\n$content\n";
                        if (($content =~ m/<title>.*complete\sgenome.*<\/title>/i) || ($content =~
m/<title>.*chromosome.*<\/title>/i) ||($content =~ m/<title>.*genome.*<\/title>/i)) { #check if its a
chromosome or plasmid file
                        #print "chromosome\n";
                        system "cat $file >>
/home/melise/projeto_doutorado/genomas_completos_bac/correcao_plasmidios/chromosome.fasta"
                        } elsif ($content =~ m/<title>.*plasmid.*<\/title>/i) {
                          #print "plasmid\n";
                       system "cat $file >>
/home/melise/projeto_doutorado/genomas_completos_bac/correcao_plasmidios/plasmid.fasta"
                        } else {
                                print "not classified $file\n";
                                }
                }
        }
exit;
```

9.1.19 Extrair o gênero do isolado de origem da sequência

```
 #!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
#> perl script [headers]
#Program to extract the genero from GenBank header
#Input    hearder
#Output list with genero

#open the file with the headers
my $header= $ARGV[0];
open (IN, "<", $header) || die "File '$header' not open\n"; #open the files
while (<IN>){
        if ($_=~m/>gi.+\[(\w+)\s*.*\]/gi){
        #print the genero
        print "$1\n";
        }
}
close (IN);
exit;
```

9.1.20 Anotar o filo correspondente de cada gênero a partir das informações do Genome Online Database (GOLD)

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [genus] [Genome Online Database (GOLD) table]
#Program to phyla annotation
#Input      Genus
#Output      Phyla

#open the file with genus
my $genus= $ARGV[0];
open (IN, "<", $genus) || die "File '$genus' not open\n"; #open the files
my %genus;
while (<IN>){
        if ($_=~m/\s(\d+)\s+(\w+)/gi){
        #get the genus
        $genus{$2} = $1;
        }
}
close (IN);
#open the a file with phyla and classes as model
my $row= $ARGV[1];
open (IN2, "<", $row) || die "File '$row' not open\n"; #open the files
my %genus_phylo;
while (<IN2>){
        if ($_=~m/(\w+)+\t(\w+)\t(\w+)\t(\w+)\t(\w+)\t(\w+)/gi){
                $genus_phylo{$6} = $2;
        }
}
close (IN2);
foreach my $key (keys %genus) {
        if (exists $genus_phylo{$key}) {
                print "$genus{$key}\t$key\t$genus_phylo{$key}\n";
        } else {
                print "$genus{$key}\t$key\n";
        }
```

```
}
exit;
```

9.1.21 Somar os filos

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [phyla]
#Program to add the phyla
#Input      phyla
#Output    phyla sum

#open file with phyla after “sort” command
my $filos= $ARGV[0];
open (IN, "<", $filos) || die "File '$filos' not open\n"; #open the files
my $filo = "Acidobacteria";
my $number = 0;
my %final;
my @check_uniqs;
while (<IN>){
        chomp $_;
        if ($_=~m/(\w+)\t(\d+)/gi){
        push @check_uniqs, "$1";
                if ($1 eq $filo){
                        $number = $number + $2;
                } else {
                        $final{$filo}= $number;
                    $filo = $1;
                        $number = $2;
                        }
        }
}
$final{$filo}= $number;
my %arr_counts;
for (@check_uniqs) {
        $arr_counts{$_}++
        };
foreach my $key (keys %final) {
        print "$key\t$final{$key}\n";
}
close (IN);
exit;
```

9.1.22 Verificar se existe mais de uma sequência de BL de cada subclasse por cromossomo

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [headers] |
#Program to check for more than one sequence per genome
#Input    hearder
#Output list with genero
```

```
#open the file with the headers
my $header= $ARGV[0];
open (IN, "<", $header) || die "File '$header' not open\n"; #open the files
my @genoma;
while (<IN>){
        if ($_=~m/>gi.+\[(.*)\]/gi){
        #print the genero
        push @genoma, "$1";
        }
}
my %genome_repeat;
foreach my $element( @genoma ) {
  ++$genome_repeat{$element};
}
foreach my $element( keys %genome_repeat ) {
  print "$element = $genome_repeat{$element}\n";
}
close (IN);
exit;
```

9.1.23 Escolher o melhor hit de proteína "Rep" para os plasmídios após BLASTp

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [directory with the organism files]
#Program to choose the best protein hit "Rep" for all plasmids after BLASTp

#Input    BLASTp output
#Output best hit

#Open the file with blastp output
my $arquivo= $ARGV[0];
open (IN, "<", $arquivo) || die "File $arquivo not open\n";

#check if the file is empty
if ( -z $arquivo ) {
        exit;
}
# Declara o Hash
my %hash;
#Read the lines of the result
while (my $row = <IN>) {
        #remove \n of the line end
        chomp $row;
        #split to get the evalue
        my @result = split /\t/, $row;
        my $evalue = $result[10];
        #get only evalue = 0.0
        if ($evalue == "0.0") {
        $hash{$result[3]}{$result[1]}=$arquivo;
        }
}
close (IN);
#Open the file with blastp output 2
```

```
my $arquivo2= $ARGV[0];
open (IN2, "<", $arquivo2) || die "File $arquivo2 not open\n";
#in case there is no evalue = 0.0
my %hash2;
unless (!!%hash) {
	my $best = 0;
	while (my $row2 = <IN2>) {
	#remove \n of the line end
	chomp $row2;
	#split to get the evalue
	my @result2 = split /\t/, $row2;
	my $evalue = $result2[10];
	my @exponent = split /\-/, $evalue;
	if ($exponent[1] >= 5) {
		$hash2{$exponent[1]}{$result2[1]}=$arquivo;
		}
	}
my $length2= 0 ;
my $inc2;
foreach my $key3 ( keys %hash2 ) {
	foreach my $key4 (keys %{$hash2{$key3}}) {
	if ($length2 < $key3){
		$length2 = $key3;
		$inc2= $key4;
		#print "$key $key2\n";
		}
	}
	}
#print the inc name with evalue >=1E-05 with higher exponent
print "$arquivo2\t$inc2\n";
} else {
	my $length= 0 ;
	my $inc;
	foreach my $key ( keys %hash ) {
	#print "$key\n";
		foreach my $key2 (keys %{$hash{$key}}) {
		if ($length < $key){
			$length = $key;
			$inc= $key2;
			#print "$key $key2\n";
		}
		}
	}
#print the inc name with evalue 0.0 with higher length
print "$arquivo2\t$inc\n";
}
exit;
```

9.1.24 A partir do identificador da proteína "Rep", atribui o grupo de incompatibilidade

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [best hit Rep]
#Program to from the protein identifier "Rep", assigns the incompatibility group
#Input blastp output
#Output plasmid followed by the incompatibility group
```

```
#Open the file with blastp output
my $arquivo= $ARGV[0];
open (IN, "<", $arquivo) || die "File $arquivo not open\n";
#Read each line of the result
while (my $row = <IN>) {
	#remove \n of the line end
	chomp $row;
	#split to get inc file
	my @result = split /\t/, $row;
	my $inc = $result[1];
	if ($inc eq "BAA78894.1") {
	print "$result[0]\tIncFII\n";
	} elsif ($inc eq "BAA78895.1"){
		print "$result[0]\tIncFII\n";
		} elsif ($inc eq "BAA97903.1"){
			print "$result[0]\tIncFI\n";
			}elsif ($inc eq "BAA97915.1"){
				print "$result[0]\tIncFI\n";
				}elsif ($inc eq "AAF69874.1"){
					print "$result[0]\tIncH\n";
					}elsif ($inc eq "NP_863360.1"){
						print "$result[0]\tIncI\n";
						}elsif ($inc eq "AAL13416.1"){

							print "$result[0]\tIncN\n";
							}elsif ($inc eq "CAK02642.1"){
								print "$result[0]\tIncP\n";
								}elsif ($inc eq "YP_009182140.1"){
									print "$result[0]\tIncW\n";
								}

}
exit;
```

9.1.25 Identificar a quais plasmídios pertencem as sequências de BLs identificadas

```
#!/usr/bin/env perl
use warnings;
use strict;
# Author: Melise Chaves Silveira
# > perl script [headers] [plasmid file]
#Program to identify the plasmids with beta-lactamases
#Input the headers of the files output from BLs search
#Output list of plasmid's name per subclass

#Open the file with BL
my $arquivo= $ARGV[0];
open (IN, "<", $arquivo) || die "File $arquivo not open\n";
my @ptns;
while (my $row = <IN>) {
	#remove \n of the line end
	chomp $row;
	my @line = split /\|/, $row;
	#print "$line[0]\n";
	push @ptns, $line[1];
}
close (IN);
```

```
#Open the file plasmids
my $arquivo2= $ARGV[1];
open (IN2, "<", $arquivo2) || die "File $arquivo2 not open\n";
while (my $row2 = <IN2>) {
	chomp $row2;
	#print "$row2\n";
	foreach my $x (@ptns){
		#print "$x\n";
		if ($row2=~m/.*$x.*/gi){
		#print "$x\n";
		print "$arquivo2\n";
		}
	}
}
exit;
```

9.1.26 Relacionar a lista de plasmídios com BLs ao arquivo com os grupos de incompatibilidade

```
#!/usr/bin/env perl

use warnings;
use strict;

# Author: Melise Chaves Silveira

# > perl script  [plasmids with beta-lactamase file] [incompatibility groups file]

#Program to relates the output of the plasmids with beta-lactamase to that of the incompatibility groups
#Input list of plasmid's name per subclass
#Output incompatibility groups

#Open the file with BL
my $arquivo= $ARGV[0];
open (IN, "<", $arquivo) || die "File $arquivo not open\n";
my @file;
#guardar os arquivos de plasmídeo que tem BL
while (my $row = <IN>) {
	#remove \n of the line end
	chomp $row;
	my @line = split /\|/, $row;
	#print "$line[0]\n";
	push @file, $line[0];
}
close (IN);
#Open the file plasmids
my $arquivo2= $ARGV[1];
open (IN2, "<", $arquivo2) || die "File $arquivo2 not open\n";

while (my $row2 = <IN2>) {
	chomp $row2;
	#print "$row2\n";
	my @line2 = split /\t/, $row2;
	foreach my $x (@file){
		if ($line2[0]=~m/.*$x.*/gi){
		print "$line2[1]\n";
		}
	}
}
```

```
exit;
```

9.1.27 Determinar o tamanho das sequências nos arquivos FASTA

```
#!/usr/bin/perl -w
use strict;

# Author: Melise Chaves Silveira
# > perl script [sequences file]
#Program to determine the size of the sequences in the FASTA files
#Input sequences multifasta file
#Output size list

#open the multifasta file
my $input_fasta=$ARGV[0];
open(IN,"<$input_fasta") || die ("Error opening $input_fasta $!");
my $x= 0;
while (my $line = <IN>){
        chomp $line;
        if (!($line=~/^$/)){
        if ($line=~m/(^>)(.*)/) {
                #print $1,$2,"\n";
        }
                else {
                        my $size = length($line);
                        print "$size\n";
                }
        }
}
exit;
```

9.1.28 Selecionar sequências com um tamanho específico

```
#!/usr/bin/perl -w
use strict;
# Author: Melise Chaves Silveira
# > perl script [multifasta that you want to get sequence greater than X]
#Program to select sequences with specific size
#Input    sequences multifasta file
#Output new sequences multifasta file

#open the multifasta file
my $input_fasta=$ARGV[0];
open(IN,"<$input_fasta") || die ("Error opening $input_fasta $!");
#put the database in a array
my @database;
while (<IN>){
        chomp $_;
        if ($_ =~ m/^>(.*)/){
                my @headers = $_; #split (/\s/,$1) o que era
                push @database, $headers[0];
        } else {push @database, $_;}
}
close IN;
open(IN2,"<$input_fasta") || die ("Error opening $input_fasta $!");
my $x= -1; #to make correspondence to array elements
while (<IN2>){
```

```
		chomp $_;
		$x++;
		if ($_ =~ m/^(\w+)$/gi){
			my $header = ($x - 1);
			my $size = length($1);
			#print "$1\n";
			if ($size > 99  && $size < 301){ #$size > 100 && $size < 1000
				#print "$size\n";
				print "$database[$header]\n$1\n"; # tirei o sinal de maior
			}
		}
}
close IN2;
exit;
```

9.2 Tabelas

9.2.1 PDB IDs correspondentes às sequências em cada cluster da classificação hierárquica

Clusters	PDB IDs
SA1	1ALQ;1AXB;1BLC;1BLH;1BSG;1BT5;1BTL;1BUE;1BUL;1BZA;1DY6;1ERM;1ERO;1ERQ;1FQG;1G68;1HTZ;1HZO;1I2S;1I2W;1IYO;1IYP;1IYQ;1IYS; 1JTD;1JTG;1JWZ;1LHY;1LI0;1LI9;1M40;1N4O;1N9B;1O7E;1ONG;1PIO;1PZO;1PZP;1Q2P;1SHV;1TDG;1TDL;1TEM;1VM1;1W7F;1XPB;1XXM; 1YLJ;1YLP;1YLT;1YLW;1YLY;1YLZ;1YM1;1YMS;1YMX;1YT4;1ZG4;2A3U;2A49;2B5R;2BLM;2CC1;2G2U;2G2W;2GDN;2H5S;2OV5;2P74;2V1Z; 2V20;2WK0;2X71;2Y91;2ZD8;3B3X;3BFC;3BFE;3BFF;3BFG;3BLM;3BYD;3C4O;3C4P;3C7U;3C7V;3CG5;3D4F;3DTM;3DWZ;3E2K;3E2L;3G2Y;3G2Z; 3G30;3G31;3G32;3G34;3G35;3HRE;3HUO;3HVF;3IQA;3LEZ;3LY3;3LY4;3M2J;3M6B;3M6H;3MKE;3MKF;3MXR;3MXS;3N4I;3N6I;3N7W;3N8L; 3N8R;3N8S;3NBL;3NC8;3NCK;3NDE;3NDG;3NY4;3P09;3Q07;3Q1F;3QHY;3RXW;3RXX;3SH7;3SH8;3SH9;3SOI;3TOI;3TSG;3V3R;3V3S;3VFF; 3VFH;3W4O;3W4P;3W4Q;3ZDJ;3ZHH;4A5R;4B88;4BLM;4C6Y;4C75;4DDS;4DDY;4DE0;4DE1;4DE2;4DE3;4DF6;4EBL;4EBN;4EBP;4EQI;4EUZ; 4FCF;4FD8;4FH2;4FH4;4GD6;4GD8;4GDB;4GKU;4HBT;4HBU;4IBR;4IBX;4ID4;4JLF;4JPM;4LEN;4M3K;4MBF;4MBH;4MBK;4MXH;4N92;4N9K; 4N9L;4OQG;4Q8I;4QB8;4QHC;4QY5;4QY6;4R3B;4R4R;4R4S;4S2I;4X69;4X6T;4XUZ;4XXR;4ZAM;4ZBE;5A90;5A91;5A92;5A93;5E2E;5E43;5EE8; 5EEC;5FA7;5FAP;5HW3;5HX9;5IHV
SA2	1E25;4D2O
TEM+MBP	4DXB;4DXC
PBP5	3MZF;3MZE;3MZD;3BEC;3BEB
SC	1BLS;1C3B;1FCO;1FR1;1FR6;1FSW;1FSY;1GA0;1GA9;1GCE;1IEL;1IEM;1KDS;1KDW;1KE0;1KE3;1KE4;1KVM;1L2S;1LL5;1LL9;1LLB;1MXO; 1MY8;1ONH;1Q2Q;1RGY;1RGZ;1S6R;1XGI;1XGJ;1XX2;1Y54;1ZC2;1ZKJ;2BLS;2HDQ;2HDR;2HDS;2HDU;2I72;2P9V;2PU2;2PU4;2Q9M;2Q9N; 2QZ6;2R9W;2R9X;2RCX;2WZX;2WZZ;2ZC7;2ZJ9;3BLS;3BM6;3GQZ;3GR2;3GRJ;3GSG;3GTC;3GV9;3GVB;3O86;3O87;3O88;3S1Y;3S22;3W8K; 3WRT;4E3I;4E3J;4E3K;4E3L;4E3M;4E3N;4E3O;4GZB;4HEF;4JXG;4JXS;4JXV;4JXW;4KG2;4KZ3;4KZ4;4KZ5;4KZ6;4KZ7;4KZ8;4KZ9;4KZA;4KZB; 4LV0;4LV1;4LV2;4LV3;4NET;4NK3;4OKP;4OLD;4OLG;4OOY;4U0T;4U0X;4WBG;4WYY;4WZ4;4X68;4XUX;5CGS;5E2G;5E2H;5EVI;5EVL;
ClbP	3O3V;4E6W;4E6X
Pab87	2QMI
SD1	1E3U;1E4D;1EWZ;1FOF;1H8Y;1H8Z;1K38;1K4E;1K4F;1K54;1K55;1K56;1K57;1K6R;1K6S;2JC7;2WGI;2WKH;2X02;3FV7;3FYZ;3FZC;3G4P;3HBR; 3LCE;3MBZ;3QNB;3QNC;3ZNT;4F94;4IED;4JF4;4JF5;4JF6;4K0W;4K0X;4S2J;4S2K;4S2L;4S2M;4S2N;4S2O;4S2P;4WMC;4WZ5;4Y0O;4Y0T;4Y0U; 4YIN;5BOH;5CTM;5CTN;5E2F;5FAQ;5FAS;5FAT;
SD2	1M6K;3ISG;4GN2;4MLL
MB1.1	1A7T;1A8T;1BC2;1BMC;1BVT;1DD6;1HLK;1JJE;1JJT;1KO2;1KO3;1KR3;1M2X;1MQO;1VGN;1ZNB;2BC2;2BMI;2DOO;2WHG;2WRS;2YZ3;2ZNB; 3BC2;3FCZ;3I0V;3I11;3I13;3I14;3I15;3L6N;3PG4;3Q6X;3RKJ;3RKK;3SBL;3SFP;3SPU;3SRX;3WXC;3ZNB;3ZR9;4BZ3;4C09;4C1C;4C1D;4C1E; 4C1F;4C1G;4C1H;4EXS;4EXY;4EY2;4EYB;4EYF;4EYL;4H0D;4HKY;4HL1;4HL2;4NQ2;4NQ4;4NQ5;4NQ6;4NQ7;4RBS;4RL0;4RL2;4RM5;4TYT; 4U4L;5A5Z;5A87;5ACU;5ACV;5ACW;5AC
MB1.2	2FHX;4BP0
ME	1JT1;1K07;1SML;2AIO;2FM6;2FU6;2FU7;2FU8;2FU9;2GFJ;2GFK;2GMN;2H6A;2HB9;2QDT;3M8T;3VPE;3VQZ;5AEB

9.2.2 Códigos das sequências da superfamília DD-peptidase/beta-lactamase do CATH identificados pelo perfil da classe SA e seus respectivos valores HMM bit score

Profile class SA									
seqs	**score**	**seqs**	**score**	**seqs**	**score**	**seqs**	**score**	**seqs**	**score**
1ylzA	401.2	2jbfA	117.0	1axbA	412.8	3jyiE	409.9	1mblA	386.1
1ymsA	398.0	3gmwA	412.0	1ck3A	412.8	3c7uC	412.8	2v20A	401.4
3g32B	397.5	1blpA	313.3	3toiA	392.9	2ov5B	383.8	2j8yB	112.6
1ylyB	396.7	1i2wA	393.8	2j8yC	112.6	1li9A	412.5	3bfdC	396.2
1es5A	23.5	2gdnA	354.7	2ov5C	383.8	3sh7B	384.5	3mxrA	390.9
3bfgB	395.9	1zg6A	409.6	1ermA	406.5	1m40A	415.4	3niaA	323.7
3sh8A	368.2	3bffB	395.9	2jbfB	117.0	3ny4A	351.3	2odsA	383.8
3e2kA	382.4	1n9bA	389.2	3huoA	394.4	1omeA	292.0	1nymA	415.4
2qpnA	319.0	3tsgB	322.1	2j7vB	112.6	3p98A	413.1	1pzoA	415.4
3bffC	395.9	2g2wA	391.5	1xxmA	410.2	1q2pA	390.9	3p98B	413.1
3g30A	397.5	3v3sA	319.4	3soiB	377.6	2h10A	386.6	3bfeD	396.5
3b3xA	394.1	2x71B	394.1	1tdlA	387.7	3nc8A	351.9	2j8yA	112.6
1temA	412.8	2j8yD	112.6	1e25A	260.9	3bfeC	396.5	2j7vC	112.6
3jyiD	409.9	3e2lB	369.5	3tsgA	322.1	1jvjA	410.1	1kgeA	328.8
1ylyA	398.0	1o7eB	337.7	1ny0A	415.4	1g6aA	334.0	3vffC	323.3
1s0wA	414.0	1lhyA	412.1	3m2jA	385.5	3c5aA	384.9	1vm1A	390.9
2wuqB	41.4	3ndgA	351.9	3e2kB	382.4	3mxsA	390.9	1w7fB	394.9
3mkfA	390.9	3rxxA	384.7	3kgnA	371.7	3g2yB	397.2	3sh9A	371.7
3sh9B	385.5	3v50A	387.7	2j9oB	117.0	3g2yA	397.5	3bfdA	396.2
1jtgC	402.1	1i2sA	393.8	1shvA	390.9	1ylpA	400.7	3q1fB	393.7
3hreA	394.9	3kgmB	385.5	3bfdD	396.2	2j9oA	117.0	3jyiF	409.9
1jtgA	405.3	3qhyA	375.3	1s0wB	414.0	3ly4A	386.5	1blhA	333.2
2j7vA	112.6	3c7vC	412.8	3huoB	393.7	3dwzA	354.7	3sh7A	386.5
3sh8B	367.2	3vffB	323.3	1jtdA	412.5	3vfhA	351.9	3hvfA	394.9
2a49A	388.1	1bsgA	341.2	3hvfB	394.4	1pioA	336.7	3g2zB	397.2
3lezA	359.2	3kgoA	368.2	1esuA	412.8	1htzB	413.5	1omeB	291.0
1djcA	330.3	3dtmA	408.4	3hreB	393.7	3cjmA	47.0	2odsB	383.3
1xpbA	414.0	3cmzA	414.7	3vfhB	351.9	1we4A	398.1	1zg4A	412.8
3m6hA	354.7	1rcjA	388.1	1jwzA	411.9	3bffA	395.9	3ly3A	387.1
1jwpA	415.4	3n8sA	351.9	1erqA	414.0	2wyxA	393.4	1g56A	390.9
1ylzB	400.0	3bfcB	395.9	2j9oC	117.0	2h0yA	386.6	2zqcA	396.8
1i2wB	394.8	1djbA	330.3	2h5sA	390.9	2y91A	386.0	1bzaA	393.0
2zq9A	395.2	2wuqA	44.0	1o7eA	337.8	2a3uA	388.1	2zqaA	396.8
3blmA	333.2	3n8rA	351.9	1htzD	413.5	3oppA	361.5	3dw0A	383.8
3m6bA	354.7	1xxmB	410.2	1dy6B	377.3	3cg5A	354.7	1hzoA	389.5
3oprA	358.1	2x71A	394.1	1iyqA	395.7	3bfcC	395.9	3vffD	351.9
3iqaA	354.7	1i2sB	393.8	3v5mA	389.5	2p74A	397.5	1tdgA	387.7
3ni9A	323.7	1ymxB	397.2	3n6iA	351.9	1mfoA	315.7	3p09B	269.8
3kgnB	385.5	1es2A	21.8	3g31B	397.2	1nxyA	415.4	1htzE	413.5
1mblB	386.1	1w7fA	394.9	1ongA	390.9	1btlA	412.8	2ov5A	383.8
3bfgA	395.9	3bydA	389.8	2wk0A	394.9	2j7vD	112.6	3n8lA	351.9
3nckA	323.3	1ghmA	327.6	3bfgC	395.9	3n4iA	391.6	3jyiB	409.9
3bfcA	395.9	1bulA	364.5	1bueA	364.5	1jwvA	412.1	3mkeA	390.9
2xqzA	399.3	3g32A	397.5	2qpnB	319.6	2cc1A	315.7	1ymsB	397.2
3b3xB	394.1	1iypA	395.3	3v3rA	321.2	3hlwA	394.9	3toiB	391.4
2jbfD	117.0	3q1fA	394.4	3rxwA	384.9	1kggA	327.7	3dw0B	383.3
1iysA	398.7	2v1zA	401.6	3ophA	387.5	1l7uA	415.4	2jbfC	117.0
3vfhD	351.9	1nyyA	415.4	1yltA	401.2	3bfgD	395.9	2zd8A	390.9
3bfdB	396.2	1esiA	24.3	3bfeA	396.5	3c7vA	412.8	3g35A	390.0
1yt4A	410.7	1g68A	332.0	3bffD	395.9	3bfeB	396.5	3ndeA	351.9
1eroA	414.0	3q07B	393.7	3gmwC	412.0	1skfA	24.2	3p09A	268.2
1dy6A	377.3	3d4fA	390.9	3kgoB	367.2	2b5rA	409.1	2b5rB	409.1
3m2jB	386.5	3v3rB	321.1	3m2kB	362.4	1kgfA	330.1	1fqgA	411.5
2p74B	396.7	3vfhC	351.9	3soiA	377.6	1iyoA	395.3	1ym1A	398.0
1htzA	413.5	2j9oD	116.9	3g34A	397.5	1n4oA	337.8	1alqA	318.8
3c4oA	390.9	3jyiA	409.9	3nblA	333.2	3bfcD	395.9	1blcA	333.2
3g2zA	397.5	2y91B	386.0	3v3sB	319.5	1ghiA	327.6	3g34B	397.2
1htzC	413.5	2h0tA	386.6	2xr0A	399.3	1pioB	336.7	1ylwA	397.0
3g35B	397.2	1pzpA	413.2	1ymxA	397.5	1ghpA	327.6	3hlwB	394.4
3q07A	394.4	3m2kA	349.3	1htzF	413.5	3jyiC	409.9	3n7wA	351.9
3oplA	387.7	2zq7A	396.8	2g2uA	390.9	1ym1B	396.7	1djaA	330.3
3c4pA	390.9	1n4oB	337.7	3g31A	397.5	1bt5A	412.8	1yljA	398.0
2zq8A	399.8	3ni9B	323.7	2wk0B	394.9	2zqdA	396.8	3kgmA	386.5
3vffA	351.9	3e2lA	374.1	1li0A	413.8	1hzoB	389.5	3c7uA	412.8

9.2.3 Códigos das sequências da superfamília DD-peptidase/beta-lactamase do CATH identificados pelo perfil da classe SC e seus respectivos valores HMM bit score

Profile class SC													
seqs	**score**	**seqs**	**score**	**seqs**	**score**	**seqs**	**score**	**seqs**	**score**	**seqs**	**score**	**seqs**	**score**
1sdeA	106.9	1l0eB	556.3	1pwgA	106.9	3blsB	558.3	3s1yA	539.5	1pwcA	106.9	1l2sB	559.8
3ixbB	558.4	3gv9A	559.8	1ga9B	538.5	1ci8B	87.9	2hdqB	559.8	1pw8A	106.9	1kdsA	559.8
3pteA	106.9	2pu2A	559.8	3o3vC	313.9	2zc7B	561.5	1q2qA	555.9	2ffyB	559.2	1pw1A	106.9
3gr2A	559.8	1ke3A	559.8	1ikgA	106.9	1pi5B	559.2	1xx2A	567.7	2zc7D	561.5	1iemB	559.8
3ixbA	539.2	3gv9B	559.8	3iwqB	558.4	2ffyA	559.2	1fcnA	530.2	1l0eA	556.3	2qmiF	413.2
3hldA	37.8	1zkjA	536.4	3tg9A	104.0	3gtcB	559.8	2zj9B	541.3	2blsB	558.3	2qmiG	413.2
1ei5A	72.4	3o86B	559.8	1xgiB	559.8	3hlbA	38.7	1ci8A	87.9	1gceA	562.9	3ixgB	555.4
2r9wA	559.8	1l0fA	539.8	1ll5A	539.7	2hduA	559.8	3tg9B	102.5	1y54A	569.7	1fr1B	562.6
1fcnB	549.4	1ielB	559.8	1ll9B	559.8	1ga9A	559.8	1o07B	549.4	2q9nA	567.7	1fsyA	559.2
1l0gB	555.8	3ixdA	527.9	1kvlB	555.8	3hlgA	34.6	1kdsB	559.8	3gvbA	559.8	1l0dB	555.6
3ixgA	455.5	1fr6A	562.6	2bltB	567.7	2i72A	559.8	1llbB	559.8	2hdrA	540.6	1ikiA	106.9
3fkvA	548.1	3ixhB	554.9	3o3vB	313.9	2wzzA	538.6	3bm6B	559.8	2qmiE	413.2	2q9mA	567.7
2qmiD	413.2	1kvmA	543.7	3fkwB	556.8	2hdqA	559.8	1ll5B	559.8	2rcxB	559.8	1l0gA	543.7
1scwA	106.9	3hlcA	40.1	2wzxA	539.5	3fkwA	540.3	3ixdB	527.9	1fr6B	562.6	3ixhA	554.9
3gsgB	559.8	1my8B	559.8	1blsB	567.7	2hduB	559.8	1xx2B	567.7	3hl9D	38.8	2d83E	90.4
3fkvB	556.8	1kvlA	555.8	3hlbD	38.8	1kvmB	559.8	1blsA	567.7	1yqsA	106.9	3iwqA	541.8
2qmiH	413.2	1kdwB	559.8	3bm6A	559.8	3gqzA	559.8	1pi5A	559.2	2pu4A	559.8	3gr2B	559.8
1ielA	544.8	1ke0A	559.8	1fcmA	532.2	1fswA	559.2	3iwoA	535.7	3s22A	539.5	3blsA	558.3
1o07A	537.0	3hl9C	38.8	2blsA	558.3	2hdrB	559.8	1kdwA	559.8	1iemA	559.8	1rgyA	558.2
1cefA	106.9	1ke4A	540.6	1c3bA	559.8	3grjA	546.7	1ci9B	87.9	1mxoB	559.8	3grjB	559.8
1l0dA	535.5	2p9vA	538.7	3s4xA	566.1	3i7jA	28.8	1c3bB	559.8	1xgjB	559.8	3gsgA	559.8
2zc7C	561.5	2hdsB	559.8	3iwoB	554.9	1ke4B	559.8	1zc2A	559.1	3iwiA	532.4	1rgzA	566.3
1fcoA	542.9	2bltA	567.7	1llbA	559.8	3o88A	559.8	1ga0A	566.3	3gqzB	557.8	2hdsA	559.8
1pi4B	559.2	2d83F	90.9	1i5qA	544.5	1l0fB	556.0	2qmiA	413.2	3o3vA	313.9	3i7jB	28.8
1mplA	106.9	2p9vB	559.8	3gtcA	559.8	3o88B	559.8	1onhA	566.3	2qmiB	413.2	1cegA	106.9
1s6rA	569.0	1ke3B	559.8	1ll9A	559.8	3gvbB	559.8	3o87A	539.7	3hl9A	38.8	1fswB	559.2
1mxoA	559.8	1xgiA	521.7	2pu2B	559.8	2r9xB	559.8	3hleA	34.6	1fsyB	559.2	3hlfA	34.6
2pu4B	559.8	2qmiC	413.2	1fcmB	549.4	3hlbC	38.7	3o87B	559.8	1l2sA	546.7	1pwdA	106.9
3o86A	539.7	2zc7A	561.5	3hl9B	38.8	1ci9A	87.9	2qz6A	532.5	1xgjA	539.7	1zc2B	557.5
2rcxA	559.8	2r9xA	559.8	3hlbB	38.7	1fcoB	559.8	2i72B	530.1	1pi4A	559.2	3iwiB	552.5
2zj9A	541.3	1ke0B	559.8	1fr1A	562.6	1my8A	559.8	2r9wB	559.8	1hvbA	106.9	1i5qB	556.2

9.2.4 Códigos das sequências da superfamília DD-peptidase/beta-lactamase do CATH identificados pelo perfil da classe SD e seus respectivos valores HMM bit score

Profile class SD							
seqs	**score**	**seqs**	**score**	**seqs**	**score**	**seqs**	**score**
1xa1A	182.3	1xa1D	173.4	1k54D	362.9	1k6rB	317.0
1k4fB	365.3	2wgwA	370.9	1h8yB	367.5	2hp5D	329.8
1e4dB	363.9	3lceA	364.6	1mwuA	38.8	3qnbB	370.4
1xa7A	129.2	1k57C	359.0	1k4eB	326.4	3qncA	367.1
1fofA	375.2	1k38A	299.5	1k57A	363.3	1xa7B	128.6
3hbrA	345.6	3isgA	278.5	2x01A	370.8	3q7zB	184.5
1mwtA	33.9	2hpbA	367.8	2wkiA	371.6	2hp5A	367.5
3qnbD	369.8	1xa1B	187.2	2wgiA	331.5	2rl3B	369.3
1k56C	337.9	1xkzC	187.2	1e3uC	358.9	3lceC	364.6
1k56D	355.2	1ewzB	373.5	1k55A	364.6	1e4dA	363.3
2hp6A	367.8	1k4eA	326.8	3if6C	258.2	2iwcA	195.9
1mwtB	31.4	1e3uA	373.5	1h8zB	333.5	1ewzD	364.9
1xa1C	177.0	3q81A	184.5	1xkzD	182.5	2jc7A	331.7
1k54C	355.6	2wkhB	372.3	3q82B	184.5	3fyzA	322.1
1k54B	364.6	2hp9B	369.1	3lceB	364.6	2hp5B	336.2
2wgiB	369.1	3uy6A	190.0	1mwsB	33.7	3if6A	258.4
1k55D	374.7	1k57D	350.5	3qnbA	370.4	3qncB	367.8
1vqqA	40.7	2wkhA	371.6	2rl3A	368.7	1e4dD	364.6
3q7vB	174.9	1mwuB	31.2	1ewzC	373.5	1k56A	363.3
2wgwB	371.5	3hbrB	354.7	3q81B	184.5	1ewzA	373.5
2x01B	371.5	1xkzA	182.5	1m6kA	278.1	1k57B	364.6
3q7zA	184.5	3fzcA	322.1	3if6B	300.1	2iwbA	206.5
1k6sA	363.9	1k6rA	374.8	2x02A	364.6	1fofB	375.2
3isgB	278.5	2wkiB	372.9	1e3uD	357.1	1k4fA	363.9
3uy6B	190.0	3pagA	325.8	3mbzA	322.1	1h5xB	324.7
1k55C	374.7	3fv7A	322.1	1k6sB	364.3	1m6kB	278.1
2hp6B	369.1	2wgvA	370.9	3q7vA	184.5	1mwxB	41.1
2hp5C	324.0	3hbrD	348.5	3lceD	364.6	3hbrC	347.7
3paeB	330.1	1h5xA	332.1	3q82A	184.5	1h8yA	367.5
1vqqB	41.0	1nrfA	227.6	3g4pA	322.1	3paeA	330.1
3qnbC	370.4	1h8zA	333.5	1xkzB	186.9	1k38B	286.3
1k54A	364.0	2x02B	365.3	1e3uB	373.5	1mwsA	31.2
2wgvB	371.5	2hp9A	367.8	1mwxA	40.8		
1k55B	364.6	1e4dC	364.6	2hpbB	369.1		
3pagB	325.8	2iwdA	196.1	1k56B	363.9		

10 ARTIGOS PUBLICADOS REFERENTES À TESE

- Silveira, M.C., Catanho, M., de Miranda, A.B., 2018. Genomic analysis of bifunctional Class C-Class D β-lactamases in environmental bactéria. Mem Inst Oswaldo Cruz, 113(8): e180098**.**

- Silveira, M.C., Silva, R.A., Mota, F.F., Catanho, M., Jardim, R., Guimarães, A.C.R., de Miranda, A.B., 2018. Systematic Identification and Classification of β-Lactamases Based on Sequence Similarity Criteria. Evolutionary Bioinformatics, 14: 1–11.

Genomic analysis of bifunctional Class C-Class D β-lactamases in environmental bacteria

Melise Chaves Silveira[1]/+, Marcos Catanho[2], Antônio Basílio de Miranda[1]

[1]Fundação Oswaldo Cruz-Fiocruz, Instituto Oswaldo Cruz, Laboratório de Biologia Computacional e Sistemas, Rio de Janeiro, RJ, Brasil
[2]Fundação Oswaldo Cruz-Fiocruz, Instituto Oswaldo Cruz, Laboratório de Genômica Funcional e Bioinformática, Rio de Janeiro, RJ, Brasil

β-lactamases, which are found in several bacterial species and environments, are the main cause of resistance to β-lactams in Gram-negative bacteria. In 2009, a protein (LRA-13) with two β-lactamase domains (one class C domain and one class D domain) was experimentally characterised, and an extended action spectrum against β-lactams consistent with two functional domains was found. Here, we present the results of searches in the non-redundant NCBI protein database that revealed the existence of a group of homologous bifunctional β-lactamases in the genomes of environmental bacteria. These findings suggest that bifunctional β-lactamases are widespread in nature; these findings also raise concern that bifunctional β-lactamases may be transferred to bacteria of clinical importance through lateral gene transfer mechanisms.

Key words: bifunctional β-lactamase - antibiotic resistance - health surveillance

β-lactamases are part of a large group of diverse and widely distributed enzymes, encoded by genes located on both the chromosome and on mobile genetic elements (Bush 2001, Srivastava et al. 2014). Bacteria containing β-lactamases have been found in a wide range of envi- ronmental conditions, including soil, water and in hu- man and animal microbiota (Allen et al. 2009, Gibson et al. 2015, Fróes et al. 2016). The production of bacterial β-lactamases is the main cause of β-lactam resistance in Gram-negative bacteria (Bush 2001), and it is essential to know their spectrum of action and distribution (Bush 2001, Gibson et al. 2015).

A few years ago, a novel β-lactamase, LRA-13, was identified in the metagenome of uncultured bacteria isolated from Alaskan soil (Allen et al. 2009). LRA-13 contains two serine-β-lactamase domains - one belong- ing to class C and one to class D. The fusion of these domains expands the hydrolytic capacity of the protein beyond what either could display alone, thereby causing resistance to amoxicillin, ampicillin, cephalexin (class C and carbenicillin (class D), as demonstrated experi- mentally (Allen et al. 2009). The identification of bi- functional β-lactamases in other bacterial species may indicate that more attention should be given to genes encoding this class of enzyme, particularly since lateral gene transfer events are common among prokaryotes (Soucy et al. 2015), and these genes could theoretically transfer to human bacterial pathogens.

To determine whether bifunctional β-lactamases are present in other bacterial species, we searched the non- redundant NCBI protein database (July 2017) utilising the BLAST programme (Altschul et al. 1997) to identify potential homologs of the LRA-13 enzyme. We identi- fied nine putative homologs encoded in the genomes of nine different bacterial species or isolates (Table). The sequence of these nine proteins is highly conserved be- tween the nine species (≥ 94%) and align closely with the reference sequence of LRA-13 (≥ 65%). All nine pro- teins have two complete characteristic domains of class C (COG1680, PRK11289) and one of class D (COG2602) according to the Conserved Domain Database (CDD, Batch CD-search tool) (Marchler-Bauer et al. 2017). In addition, these proteins display characteristic ac- tive site patterns of both class C (PS00336) and class D (PS00337) domains according to PROSITE (Sigrist et al. 2012), including the serine (S) catalytic residue.

We then examined whether these bifunctional β-lactamases are encoded in genomic islands or near prophage sequences using IslandViewer 4 (Bertelli et al. 2017), which integrates four different genomic island pre- diction methods, and the PHAge Search Tool (PHAST) (Zhou et al. 2011), which identifies prophage sequences in bacterial genomes. Bifunctional β-lactamase was not encoded in genomic islands, and only the strain *Massilia* sp. Root351 showed the presence of an incomplete 8.4Kb prophage located approximately 3.4Kb downstream from the gene encoding a bifunctional β-lactamase.

The β-lactamases with fused domains found in this work were identified in the genomes of bacterial strains belonging to three distinct Gram-negative genera (Bal- dani et al. 2014): *Duganella* spp., *Janthinobacterium* sp. and *Massilia* sp. The original annotation of the gene products encoding these enzymes is either "class C β-lactamase" or "class D β-lactamase" (Table). In all cases, the upstream protein-coding gene is originally annotated as "class D β-lactamase", and their prod- ucts display a complete characteristic domain of class D (COG2602). Additionally, these products display complete or incomplete domains of methicillin resistance regulatory

doi: 10.1590/0074-02760180098
MCS was supported by CAPES.
+Corresponding author: melisechaves@gmail.com
Received 23 February 2018 Accepted 20 April 2018

TABLE

Genomes, original annotation and genomic context of genes enconding

		Bifunctional β-				Upstream				Downstream			
		/	A		/		A		/	A			/
Uncul. Bacterium		EU408352.1		L	ACH58991.1	resp. reg.		ACH58892.1		glycoside	ACN58887.1		
BLR13 Janthinobacterium		NZ_AMWD01000002.		RA-13	WP_008451281	class	D	WP_008451277		hydrolase	WP_008451283		
sp. HH01 Massilia	sp.	1 LMEC01000020.1			.1 KQW93884.1	BL	class	.1 KQW93885.1		transaldolase	.1 KQW93883.1		
Root418 Massilia	sp.	NZ_LMDJ01000033.1		cla	WP_082552146	D	BL	WP_057157847		diguanylate	WP_057157849.		
Root351 Massilia	sp.	FQWU01000002.1		ss C BL	.1 SHH20105.1	class	D	.1 SHH20059.1		cyclase	1 SHH20125.1		
CF038 Duganella	sp.	LRHV01000029.1		hypot.	OEZ55387.1	BL	class	OEZ55388.1		diguanylate	OEZ55386.1		
HH105 Duganella	sp.	FOOF01000012.1		protein class	SFG43659.1	D	BL	SFG43677.1		cyclase	SFG43637.1		
CF458 Duganella	sp.	NZ_LMIC01000034.1		C BL class	WP_082591432	class	D	WP_082591444		hypothetical	WP_082507115		
Root198D2 Duganella	sp.	NZ_LMDB01000002.1		C BL class	.1	BL	class	.1		protein	.1		
Root336D2 Duganella	sp.	NZ LMFZ01000003.1		C BL class	WP 082507116	D	BL	WP 082507139		transaldolase	WP 082507115		

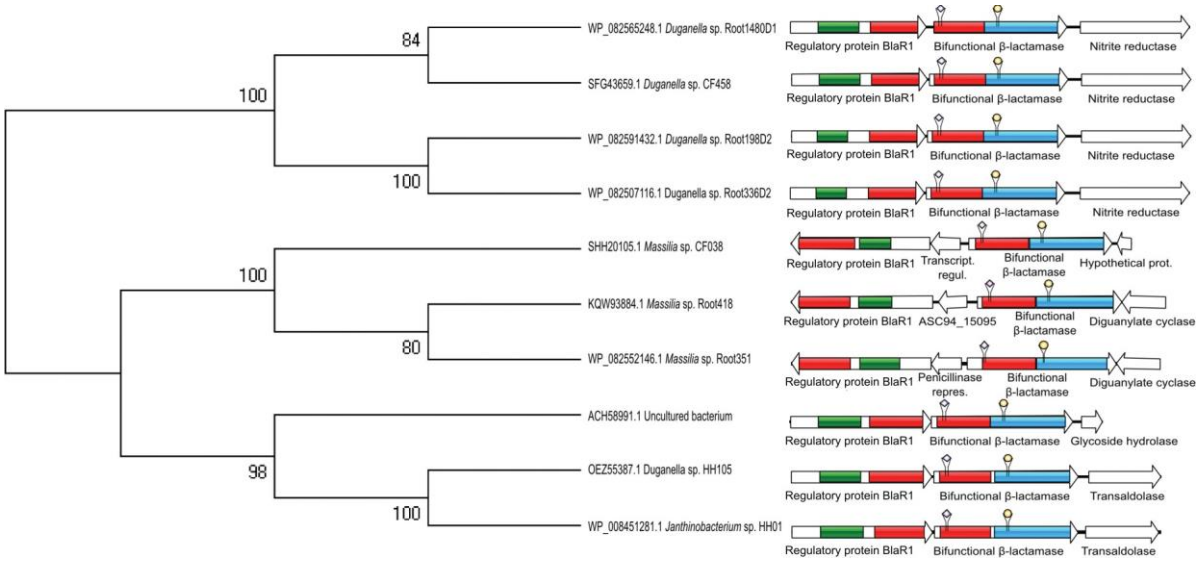
Uncul.: uncultured; BL: β-lactamase; hypot: hypothetical; resp. reg.: putative

cd07341), corresponding to the structure of the signal-transducing integral membrane protein that regulates the β-lactam resistance in the Gram-positive species *Staphylococcus aureus* (Wilke et al. 2004). Genes lo- cated downstream of the bifunctional β-lactamase vary among distinct bacterial genera and fall into four func- tional categories: "diguanylate cyclase", "transaldolase", "glycoside hydrolase" or "hypothetical" (Table). All *Massilia* strains harbor a gene encoding a transcrip- tional regulator which is upstream and inverted relative to the gene encoding the bifunctional β-lactamase. This transcriptional regulator mediates the expression of the regulatory protein BlaR1, which is also inverted in rela- tion to the bifunctional β-lactamase (Figure).

The strains *Janthinobacterium* sp. HH01 and *Du- ganella* sp. HH105 were isolated from an aquatic envi- ronment and exhibit an ampicillin resistance phenotype (Hornung et al. 2013, Haack et al. 2016). The strains *Duganella* sp. CF458 (Gp0136797) and *Massilia* sp. CF038 (Gp0136806) were isolated in 2016 from the root of a Populus tree in Tennessee, USA (NCBI BioProject PRJEB18228), while the other strains of *Duganella* sp. and *Massilia* sp. were isolated from the Arabidopsis root microbiota (Bai et al. 2015). These three genera belong to the family *Oxalobacteraceae* (*Betaproteobacteria* group); they are (supposedly) non-pathogenic to humans, animals and plants and are known for their antifungal ef- fect (Yin et al. 2013, Haack et al. 2016). Bacteria from this family have few phenotypic differences, and their classification in distinct genera is mainly based on 16S rRNA gene sequencing (Kämpfer et al. 2007). Function-al metallo-β-lactamases (class B) have already been de-scribed in *Janthinobacterium lividum* and *Massilia oculi* (Docquier et al. 2004, Gudeta et al. 2016). These genes are phylogenetically related and share common ances- tors with acquired β-lactamases produced by clinical pathogens, which could have been acquired from mem- bers of *Oxalobacteraceae* (Gudeta et al. 2016).

According to Allen et al. (2009), the LRA-13 β-lactamase appears to be the result of an ancient natu- ral fusion of genes encoding complete enzymes, not due to modern selective pressure caused by the extensive use of antibiotics. In two cases, the bifunctional β-lactamase sequences are virtually identical (*Duganella* sp. Root 198D2 vs. *Duganella* sp. Root 336D2, and *Duganella* sp. HH105 vs. *Janthinobacterium* sp. HH01) with 99% and 96% amino acid identity over their entire sequences, respectively. However, LRA-13 is not the only example of a bifunctional enzyme implicated in antibiotic resis- tance. Some aminoglycoside transferases are capable of conferring resistance to practically all antibiotics of this class via modifications to the antibiotic molecule at two different sites. However, unlike LRA-13, their origin ap- pears to be recent and caused by the clinical (mis)use of aminoglycosides (Kim et al. 2007, Zhang et al. 2009).

The absence of genes encoding bifunctional β-lactamases in genomic islands or near prophage se-quences, and the fact that these genes are shared among all currently sampled representatives of three genera be- longing to the same family (*Oxalobacteraceae*), suggest



Phyletic pattern of genes encoding bifunctional β-lactamases, and their genomic organisation including surrounding genes. Left: a dendrogram representing the relationships between the bifunctional β-lactamases in this study. Right: a panel displaying the order and orientation of the genes encoding the bifunctional β-lactamases and surrounding genes. The boxes represent distinct domains: red, class D; green, MecR1/BlaR1; blue, class C. Arrows indicate gene orientation. Diamonds and circles above the bifunctional β-lactamases indicate the location of class D and class C active sites, respectively. Sequences were globally aligned using MAFFT version 7 (Katoh et al. 2017). The dendrogram was constructed with MEGA version 7 (Kumar et al. 2016), applying the NJ algorithm and 500 bootstrap replicates. The panel containing genes, domains and active sites was drawn using the IBS (Liu et al. 2015).

curred naturally and long ago. Indeed, several benefits of bearing a bifunctional enzyme can be assumed, such as the concomitant mobilisation of two different func- tions, the potential for complementary and extended resistance, and the simultaneous selection of two enzy- matic activities by the selective pressure exerted by a single antibiotic (Zhang et al. 2009).

The evidence presented here suggests that bifunctional β-lactamases are part of a new class of enzyme with poten- tially broad spectrums of action. The first reported enzyme within this class (LRA-13) was found in a non-cultivable bacterium from a remote soil sample, but proteins with the same characteristics can be found in different bacterial genera present in water, soil, and even sharing the same niche. To date, there is no evidence of a clinically signifi- cant role for bifunctional β-lactamases, but this possibility cannot be ignored. Chromosomal location and degree of sequence conservation suggest that these enzymes might be characteristic of the family *Oxalobacteraceae*. Since our knowledge of the environmental microbiota is far from complete, it is necessary to examine the eventual dissemi- nation of these bifunctional β-lactamases to bacteria that are pathogenic to humans and other animals.

AUTHORS' CONTRIBUTION

MCS and ABM conceived of the manuscript outline; MCS, MC and ABM wrote, edited and revised the manuscript.

REFERENCES

Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J. Func- tional metagenomics reveals diverse β-lactamases in a remote Alaskan soil. ISME J. 2009; 3(2): 243-51.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17): 3389-3402.

Bai Y, Müller D, Srinivas G, Garrido-Oter R, Potthoff E, Rott M, et al. Functional overlap of the Arabidopsis leaf and root micro- biota. Nature. 2015; 528(7582): 364-9.

Baldani JI, Rouws L, Cruz LM, Olivares FL, Schmid M, Hartmann A. The family *Oxalobacteraceae*. In: Rosenverg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, eds. The prokaryotes - alphaproteobacteria and betaproteobacteria. Berlin/Heidelberg: Springer-Verlag; 2014. p. 919-74.

Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Simon Fraser University Research Computing Group, et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res. 2017; 45(W1): W30-5.

Bush K. New β-lactamases in gram-negative bacteria: diversity and impact on the selection of antimicrobial therapy. Clin Infect Dis. 2001; 32: 1085-9.

Docquier JD, Lopizzo T, Liberatori S, Prenna M, Thaller MC, Frère JM, et al. Biochemical characterization of the THIN-B metal- lo-beta-lactamase of *Janthinobacterium lividum*. Antimicrob Agents Chemother. 2004; 48(12): 4778-83.

Fróes AM, da Mota FF, Cuadrat RRC, D'avila AMR. Distribution and classification of serine β-lactamases in Brazilian hospital sewage and other environmental metagenomes deposited in public databases. Front Microbiol. 2016; 7: 1-15.

Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2015; 9(1): 1-10.

Haack F, Poehlein A, Kröger C, Voigt C, Piepenbring M, Bode H, et al. Molecular keys to the *Janthinobacterium* and *Duganella* spp. Interaction with the plant pathogen *Fusarium graminearum*. Front Microbiol. 2016; 7: 1668.

Hornung C, Poehlein A, Haack F, Schmidt M, Dierking K, Pohlen A, et al. The *Janthinobacterium* sp. HH01 genome encodes a homo- logue of the *V. cholerae* CqsA and *L. pneumophila* LqsA autoin- ducer synthases. PLoS ONE. 2013; 8(2): e55045.

Kämpfer P, Rosselló-Mora R, Hermansson M, Persson F, Huber B, Falsen E, et al. *Undibacterium pigrum* gen. nov., sp. nov., isolated from drinking water. Int J Syst Evol Microbiol. 2007; 57(Pt 7): 1510-5.

Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualiza- tion. Brief Bioinform. 2017; doi: 10.1093/bib/bbx108.

Kim C, Villegas-Estrada A, Hesek D, Mobashery S. Mechanistic characterization of the bifunctional aminoglycoside-modifying enzyme AAC(3)-Ib/AAC(6')-Ib' from *Pseudomonas aeruginosa*. Biochemistry. 2007; 46(17): 5270-82.

Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33(7): 1870-4.

Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, et al. IBS: an illustrator for the presentation and visualization of biological sequences. Bioin- formatics. 2015; 31(20): 3359-61.

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/ SPARCLE: functional classification of proteins via subfamily do- main architectures. Nucleic Acids Res. 2017; 45(D1): D200-3.

Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Ac- ids Res. 2012; 41: D344-7.

Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015; 16(8): 472-82.

Srivastava A, Singhal N, Goel M, Virdi JS, Kumar M. CBMAR: a comprehensive β-lactamase molecular annotation resource. Da- tabase (Oxford). 2014; 2014: bau111.

Wilke MS, Hills TL, Zhang HZ, Chambers HF, Strymadka NCJ. Crystal structures of the Apo and penicillin-acylated forms of the BlaR1 β-lactam sensor of *Staphylococcus aureus*. J Biol Chem. 2004; 279(45): 47278-87.

Yin C, Hulbert S, Schroeder K, Mavrodi O, Mavrodi D, Dhingra A, et al. Role of bacterial communities in the natural suppression of *Rhizoctonia solani* bare patch disease of wheat (*Triticum aesti- vum L.*). Appl Environ Microbiol. 2013; 79(23): 7428-38.

Zhang W, Fisher J, Mobashery S. The bifunctional enzymes of antibi- otic resistance. Curr Opin Microbiol. 2009; 12(5): 505-11.

Zhou Y, Liang Y, Lynch K, Dennis JJ, David S, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011; 39(2): W347-352.

Systematic Identification and Classification of β-Lactamases Based on Sequence Similarity Criteria: β-Lactamase Annotation

Evolutionary Bioinformatics Volume 14: 1–11

DOI: 10.1177/1176934318797351

Melise Chaves Silveira[1], Rangeline Azevedo da Silva[1], Fábio Faria da Mota[1], Marcos Catanho[2], Rodrigo Jardim[1], Ana Carolina R Guimarães[2] and Antonio B de Miranda[1]

[1]Laboratório de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil. [2]Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil.

ABSTRACT: β-lactamases, the enzymes responsible for resistance to β-lactam antibiotics, are widespread among prokaryotic genera. However, current β-lactamase classification schemes do not represent their present diversity. Here, we propose a workflow to identify and classify β- lactamases. Initially, a set of curated sequences was used as a model for the construction of profiles Hidden Markov Models (HMM), specific for each β-lactamase class. An extensive, nonredundant set of β-lactamase sequences was constructed from 7 different resistance proteins databases to test the methodology. The profiles HMM were improved for their specificity and sensitivity and then applied to fully assembled genomes. Five hierarchical classification levels are described, and a new class of β-lactamases with fused domains is proposed. Our profiles HMM provide a better annotation of β-lactamases, with classes and subclasses defined by objective criteria such as sequence similarity. This classification offers a solid base to the elaboration of studies on the diversity, dispersion, prevalence, and evolution of the different classes and subclasses of this critical enzymatic activity.

KeywoRdS: β-lactamase, class, subclass, identification, sequence similarity

ReCeIVed: March 15, 2018. **ACCePTed:** August 8, 2018.

TyPe: Original Research

FundIng: The author(s) received no financial support for the research, authorship, and/or publication of this article.

deCLARATIon oF ConFLICTIng InTeReSTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CoRReSPondIng AuTHoR: Melise Chaves Silveira, Laboratório de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz, Fiocruz, Av. Brasil 4365, Manguinhos, Rio de Janeiro RJ 21040-900, Brazil. Email: melisechaves@gmail.com

Introduction

The increasing amounts of genomic data produced by next- generation sequencing technologies made β-lactamases (BLs), the enzymes responsible for the irreversible inactivation of β- lactam antibiotics, one of the most numerous families of pro- teins studied to date.[1] First described by Abraham and Chain,[2] BLs can be found in pathogenic or commensal bacteria, iso- lated from humans or varied environments.[3] Due to its great medical importance and clinical impact, several groups directed efforts in the development of a proper BL classification scheme, usually based on functional or structural criteria.[4–6]

Function-based classification is achieved using experimental data to link the enzyme to its clinical role.[6,7] The determination of these parameters for a large number of BLs, however, may be rela- tively costly and time-consuming. Also, they do not generate sequence data, essential for studies involving molecular evolution.[8] Currently, the most widely used classification scheme for BLs is the Ambler structural classification, which is based on sequence similarity, and separates BLs into 4 classes: the classes A, C, and D of serine-β-lactamases (SBLs) and the class B of metallo-β- lactamases (MBLs).[4,9,10] Class B is further divided into sub- classes B1, B2, and B3, using sequence conservation data.[11]

Despite SBLs and MBLs are able to break amide and ester bonds (EC 3.5.2.6), they belong to 2 distinct protein super- families that do not share a common ancestor.[12] Considering SBL's tertiary structures, they are similar enough among themselves to be considered homologous,[5] whereas the differ- ences between their primary structures and catalytic mecha- nisms justify their division into classes A, C, and D.[13]

Experimental data indicate that the 3 subclasses of MBL should not be treated as equally separated groups. Subclasses B1 and B2 have detectable sequence similarity between them but not with B3,[14] and structural evidence strongly suggest dif- ferent Most Recent Common Ancestor between the group formed by subclasses B1 and B2 and the subclass B3.[15]

Based on this structural information, a reorganization of BL at 4 hierarchical levels was proposed by Hall and Barlow.[5] In this scheme, the former subclasses B1 and B2 were merged and renamed as class MB, whereas subclass B3 was renamed as class ME. Thus, the 5 BL classes (third classification level) are SA, SC, SD, MB, and ME, and subclasses MB1 and MB2 rep- resent the

reflects the distinct evolutionary origins of SBLs and MBLs.[5] Application of this scheme to 2774 assembled bacterial genomes provided improvements in BL annotation and in the knowledge about BLs distribution among bacteria phyla, con- firming previous studies suggesting new BL subclasses.[1,16] Finally, our results propose the existence of a new BL class with fused domains and extended action spectrum.

Methodology

Data collection and preparation

On March 22, 2016, an online search for BL structures using the EC number (3.5.2.6) in the National Center for Biotechnology Information (NCBI) Structure Database[17] returned 516 entries. Records with the word "mutant"in the description were excluded. The PDB (Protein Data Bank) IDs were retrieved and used for downloading PDB files and their corresponding FASTA files from RCSB PDB.[18] The data were filtered using the following criteria: duplicate atoms positions were removed, the resolution of the structure should be less than 3Å, and only monomers or the chain A from homomultimers were used.

A Non-Redundant Beta-Lactamase Dataset (NRBLD) was constructed with protein sequences retrieved from 7 dif- ferent antibiotic resistance databases. Four databases are spe- cific for BLs: (1) Comprehensive Beta-lactamase Molecular Annotation Resource (CBMAR, downloaded on August 2015),[19] (2) The Institute Pasteur Database (downloaded on October 2015),[20] (3) DLact Antimicrobial Resistance Gene Database (downloaded on October 2015),[21] and (4) Lactamase Engineering Database (LacED, downloaded on August 2015).[22] The Metallo-Beta-Lactamase Engineering Database (MBLED, release 1.0) is the only specific for MBLs.[23] The remaining 2 databases, Comprehensive Antibiotic Resistance Database v1.0.0 (CARD)[24] and Resfams v1.2,[3] have protein sequences related to antibiotic resistance in general. To recover only BL sequences, we search for those with the term "bla" and "beta" in the headers. Identical sequences have been removed using CD-HIT[25] at a 100% identity threshold. The following methodology steps are summarized in Figure 1.

Clustering curated BL structures and sequences

The clustering assays were made using BL structures and their corresponding amino acid sequences from PDB, in an attempt to reproduce the BL classes and subclasses proposed by previous works. The MaxCluster program[26] was used to cluster the structures based on root mean-squared deviation and hierar- chical clustering, applying 3 tests: single, average, and maximum linkage. The BLASTClust v2.2.26 program[27] was used for the hierarchical clustering of the amino acid sequences, which performs a single linkage type clustering based on pair- wise matches found by BLAST. Different threshold values of "BLAST score density" (BLAST score divided by the align- ment length) and "minimum length coverage" were tested. An E value of 1E−05 was used.

Building a profile HMM for each BL class

The profiles HMM were constructed using the clusters of sequences corresponding to the 5 BL classes.[5] Because these sequences came from proteins with a resolved structure, they were considered reliable to be used to construct the profiles. Sequences from each cluster were saved in separate multi- FASTA files. Identical sequences were removed using CD-HIT[25] at a 100% identity threshold. Sequences in each file were aligned using MUSCLE.[28] Profiles HMM were built from each alignment using the program *hmmbuild* from the HMMER package v3.1b2.[29]

Calibration and validation of the profiles HMM

The profiles HMM generated were used against the protein sequences at NRBLD and superfamilies 3.40.710.10 (DD-peptidase/BL like) and 3.60.15.10 (Metallo-BL like) from Protein Structure Classification Database (CATH) (down- loaded on March 17, 2016).[30] *Hmmsearch* from the HMMER package v3.1b2 with an E value of 1E−05 was employed.[29]

The Class profile Specificity index (CpSp) evaluates whether each profile identifies a unique group of sequences. CpSp was calculated by dividing the number of NRBLD sequences identified exclusively with a given profile (Ne) by the total number of NRBLD sequences recovered by it (T), includ- ing intersections with others profiles results [CpSp= (Ne/T)*100]. Profiles with CpSp below 100% were calibrated. Sequences from other classes that should not be identified were used as "negative training sequences," following the HMM-ModE protocol.[31] After this, the *hmmsearch* Gathering Threshold (GA) parameter was used, substituting the E value.

A total of 851 amino acid sequences of DD-peptidase/ BL-like superfamily downloaded from CATH[30] were used to construct an unrooted Maximum Likelihood phylogenetic tree in MEGA-CC v7.0.18,[32] using the Jones-Taylor-Thornton model, partial deletion for gaps/missing data treatment (95% site coverage cutoff) and 500 bootstrap replicates. Using in- house Perl scripts, the BL sequences were manually labeled with their respective class, and the clade node of each BL class was identified. The sequences in these clades were allocated in corresponding multi-FASTA files.

Graphics were created with all the sequences retrieved by each profile against the DD-peptidase/BL superfamily with their respective HMM bit score. These results were used to calibrate the profiles. A new *hmmsearch* parameter of HMM bit score threshold was established to separate true BL from other sequences in the superfamily. The Function Specificity (FSp)

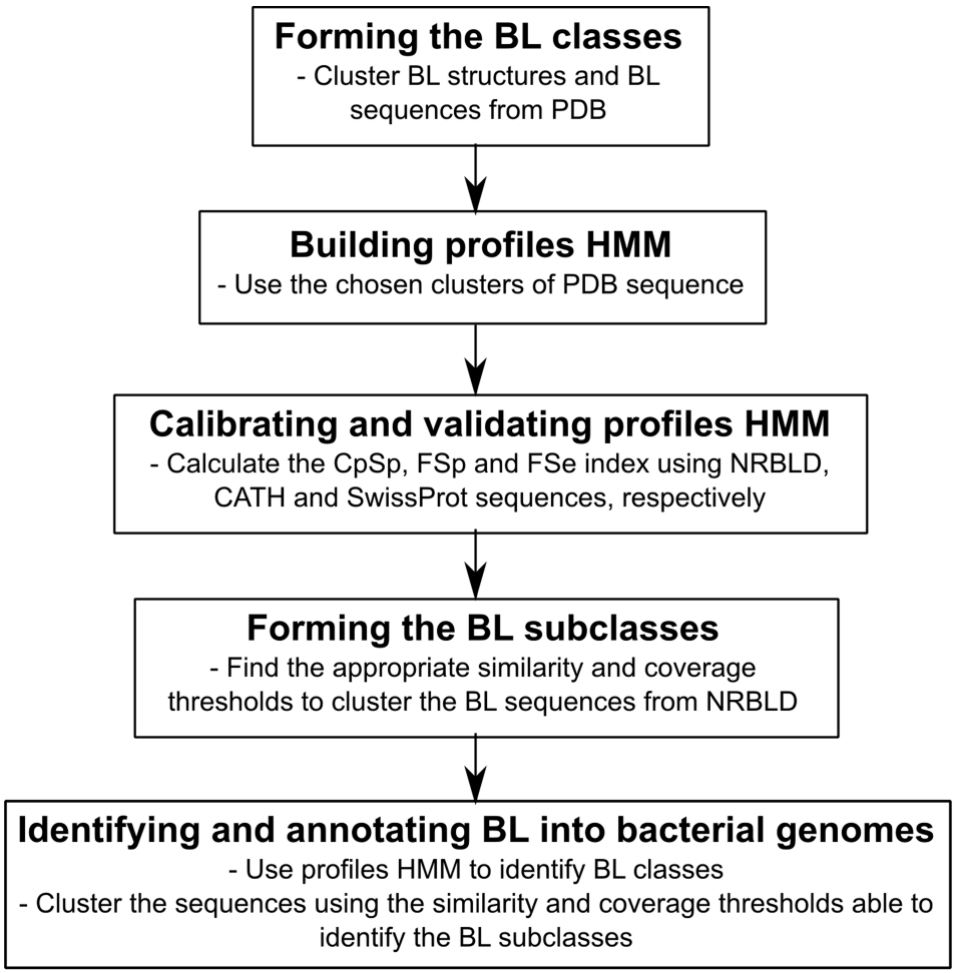


Figure 1. Main conceptual steps of the workflow.
BL indicates β-lactamase; CATH, Protein Structure Classification Database; PDB: Protein Data Bank; CpSp, Class profile Specificity; FSe, Function Sensibility; FSp, Function Specificity; NRBLD, Non-Redundant Beta-Lactamase Dataset.

A total of 144 sequences are attributed to the Gene Ontology (GO) molecular function "BL activity" (GO:0008800) in the SwissProt database (downloaded on March 2, 2016).[33] The validation index Function Sensitivity (FSe) evaluates the ability of all profiles together to identify the sequences annotated with the BL function in SwissProt. To calculate this index, the number of identified SwissProt sequences attributed to the BL GO term (Ngo) was divided by 144 [FSe = (Ngo/144)*100].

Validation of thresholds used to form BL subclasses

The results obtained with the curated PDB data set were used as a basis to group the NRBLD sequences into subclasses. The BLs from PDB were added to NRBLD, and the profiles HMM were used against them.

In order to reproduce BL subclasses, different thresholds of "minimum length coverage"were tested to cluster the sequences in each class, together with the "BLAST score density" thresholds previously established using the PDB sequences.The sequences in each cluster were annotated using the BLASTP[34] best hit (v2.2.28) against the nonredundant NCBI protein database.[17]

The size of specific domains of BLs was used to stipulate a minimum size necessary for the enzyme to be functional.[35] The Pfam[36] models most common in MBL (PF00753), class SA (PF13354), and class SC (PF00144) have a length of 197, 202, and 330 residues, respectively. A total of 214 residues domain have been attributed to class SD enzymes.[37]

The clusters formed were categorized as follows: (1) clusters corresponding to those obtained in the hierarchical classifica- tion of functionally characterized BLs (PDB sequences) and (2) clusters containing sequences that do not fit into the previ- ous established similarity and coverage thresholds.

Comparison of the improved profiles HMM with Pfam profiles and BL motifs

To test the efficiency of our profiles HMM in identifying and classifying BL sequences, these were compared with profiles available in the Pfam database[36] and with motifs specific to BL classes from different sources.

The profiles HMM from Pfam[36] were used to search for BL sequences from PDB, and the CpSp indexes were calculated accordingly. Seven Pfam profiles (PF00144.22, PF13354.4,

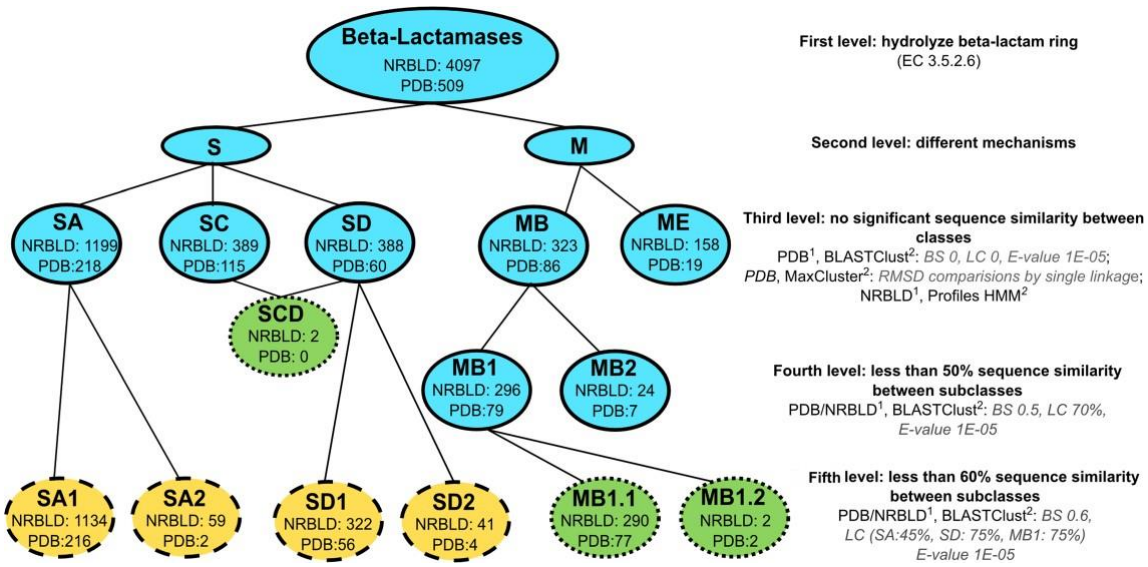


Figure 2. Hierarchical classification of β-lactamases.
The number of Protein Data Bank (PDB) and Non-Redundant Beta-Lactamase Dataset (NRBLD) sequences in each cluster after clustering is shown (PDB identifiers can be obtained in Table S2). 1 indicates the sequence dataset used for clustering. 2 indicates the program used for clustering, followed by the parameters used. Blue: Ambler's classification with real relationships as shown by Hall and Barlow[5]; green and dotted: new groups proposed for the first time in this study; yellow and dashed: groups described recently and confirmed in this work[1,16]; BS: BLAST score density; LC: minimum length coverage; BL: Beta-lactamase; S: Serine-BL; M: Metallo-BL; SA: Serine-BL class SA; SC: Serine-BL class SC; SD: Serine-BL class SD; MB: Metallo-BL class MB (former subclasses B1 and B2); ME: Metallo-BL class ME (former subclass B3); SCD: Serine BL class SC-class SD; MB1: Metallo-BL subclass MB1; MB2: Metallo-BL subclass MB2; MB1.1: Metallo-BL subclass MB1.1; MB1.2000000000: Metallo-BL subclass MB1.2; SA1: Serine-BL subclass SA1; SA2: Serine-BL subclass SA2; SD1: Serine-BL subclass SD1; SD2: Serine-BL subclass SD2.

PF00753.25, PF12706.5, PF13483.4, PF14597.4, and PF16661.3) were downloaded on December 2016.

Motifs specific for BL classes obtained from the literature were evaluated for their ability to distinguish between classes and subclasses. An in-house Perl script was developed to iden- tify the motifs in the BL sequences from PDB.

Identification and classification of BLs in fully assembled genomes

The workflow proposed here was applied to identify and clas- sify BL sequences present in 2774 bacterial strains with fully assembled genomes that were deposited in NCBI.[17] A genome was defined as the complete set of chromosome and plasmids of each strain. The protein sequences present in each genome were downloaded (June 2016). Using the profiles HMM, the BL classes were formed, which were then separated into their respective subclasses applying the BLASTClust program[27] with the "BLAST score density" and "minimum length cover- age" thresholds set in the previous step. The taxonomic infor- mation of each sequence was determined using the Genome Online Database (GOLD)[38] and in-house Perl scripts.

The scripts, profiles HMM, and instructions required to apply the workflow presented here, in addition to the data used for the searches, are available at https://github.com/melisesil-veira/betaLactamase-classification.git.

Results

Clustering curated BLs structures and sequences

In total, 509 PDB structures and their respective sequences were used in the clustering and in the construction of profiles HMM. Of these, 208 were monomers and 301 were homomul- timeric proteins.

Clustering of the curated set of structures using single link- age was consistent with the BL classification as proposed by Hall and Barlow, producing the same 5 classes (Figure 2).[5] Average and maximum linkage resulted in 7 and 18 clusters, respectively, and were not further used.

The PDB sequences clustering applying "BLAST score density" thresholds of 0 and 0.5 were also consistent with the Hall and Barlow's classification scheme,[5] producing the 5 BL classes and 2 MB subclasses, respectively. Using a "BLAST score density" threshold equal to 0.6, the subclasses proposed in previous works for classes SA and SD were reproduced,[1,16] as well as the division of MB1 into 2 groups (Figure 2). The length coverage did not influence these results, and therefore a "minimum length coverage" of zero was chosen.

Separate clusters, one composed by 2 BL TEM-1 fused to a maltose-binding protein and another with 5 *Escherichia coli* penicillin-binding protein 5 (PBP5), were created in all clus- tering assays and were excluded because they are not true BLs, being wrongly associated with EC 3.5.2.6 in PDB. All sequence and structural clustering results are presented in Tables S1 to S3.

Building, calibration, and validation of the profiles HMM

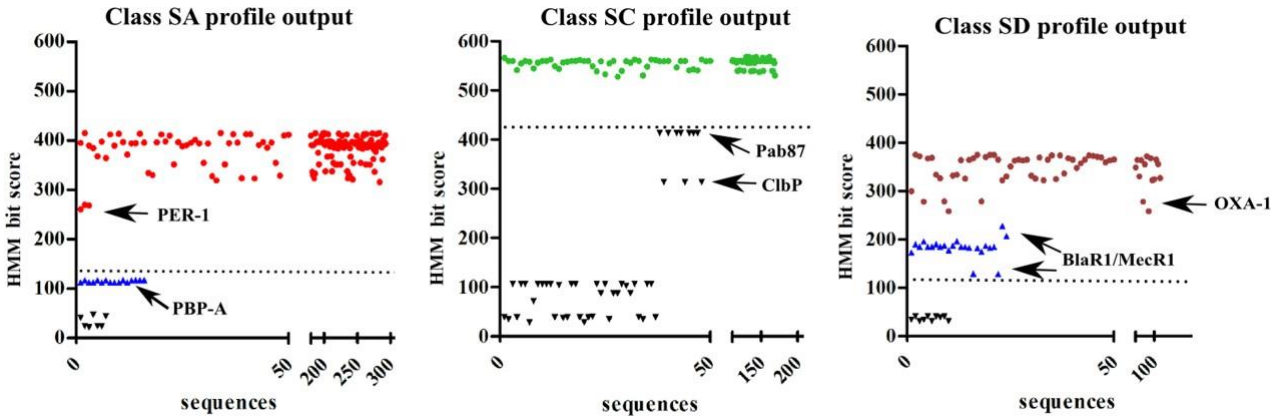


Figure 3. Plot of the 851 DD-peptidase/β-lactamase–like superfamily (CATH 3.40.710.10) sequences.
Each dot represents a sequence identified by the profile with its HMM bit score result. The HMM bit score reflects whether the sequence is a better match to the profile model (positive score) or the null model of nonhomologous sequences (negative score). The dashed lines in the graphs indicate the established HMM bit score threshold for each profile. X: numbered sequence order, Y: HMM bit score. Red circle: 295 sequences from BL class SA, green circle: 177 sequences from BL class SC, brown circle: 103 sequences from BL class SD, blue triangle: proteins inserted within the BL clades that are not functionally characterized as BLs, and black inverted triangle: other non-BL sequences from DD-peptidase/β-lactamase–like superfamily. Arrows indicate particular proteins quoted in the text; BL: Beta-lactamase.

DD-peptidase/BL- and MBL-like superfamilies, with 851 and 399 protein sequences, respectively; and (3) SwissProt/ UniProt[33] database, with 550 552 sequences. The profiles for the SBL and MBL classes were analyzed separately as these 2 groups belong to different CATH superfamilies. The data and profiles HMM are available at https://github.com/melisesil-veira/betaLactamase-classification.git.

The profiles for classes SA, SC, and SD were searched against NRBLD recovering 1292, 1121, and 396 sequences, respectively. The profile of the class SA was 100% specific for the class (CpSp), whereas the profiles of the classes SC and SD were 99% specific. Two sequences were identified by both pro- files, BL LRA-13 (ACH58991.1) and an enzyme annotated as "class C BL" of *Janthinobacterium* sp. (WP_008451281.1). Both have domains of SC and SD classes, which led us to pro- pose a new class, SCD, composed of BLs with fused domains (Figure 2).

The profiles for MBL classes recovered 527 (MB, CpSp= 84%) and 612 (ME, CpSp= 86%) NRBLD protein sequences, 84 of which were identified by both profiles. Therefore, the profiles were optimized, and after the calibra- tion they recovered 323 (MB) and 159 (ME) sequences, with- out intersections, reaching 100% CpSp.

To compare the calibrated and noncalibrated MBL profiles, they were tested against the 598 MBL protein sequences from the MBLED database.[23] Initial profiles identified 440 (MB, 85% CpSp) and 222 (ME, 70% CpSp) sequences, which repre- sents 99.8% of the database. However, calibrated profiles iden- tified 424 (MB, 100% CpSp) and 164 (ME, 100% CpSp) sequences, representing 98.3% of the database. The remaining 1.7% was composed of protein

fragments ranging from 75 to 131 amino acids, meaning that the calibrated profiles were able to identify all the complete BL sequences.

The phylogenetic relationships of the SBL superfamily pro- teins showed a few sequences with no BL activity inserted into BL clades (Figure S1). The penicillin-binding proteins A (PBP-A) are in the inner branches of class SA and have a

structure very similar to the BL PER-1 (subclass SA2) but do not have BL activity.[39] The regulatory proteins BlaR1 (and its cognate MecR1) are in inner branches of the class SD. Their extracellular domains are phosphorylated by β-lactams and, consequently, these proteins regulate resistance to these antibi- otics in *Staphylococcus aureus*.[40]

Initially, the profiles HMM for SBLs identified these sequences and others outside the BLs clades (all non-BLs). However, based on their HMM bit score values, a total of 75 non-BL sequences could be excluded (Figure 3). The separa- tion of true SD2 BLs from BlaR1 proteins in this step was not possible because the HMM bit score of the single structure available of this subclass present in the CATH database[30] (OXA-1) was very close to some BlaR1 proteins (Figure 3). Additional tests have shown that when other variants of this subclass are included, their scores are smaller than those of some BlaR1 sequences. This can be explained by the structural homology of the extracellular sensor domain of BlaR1 to BLs from subclass SD2.[41]

The enzymes ClbP and Pab87 are associated with the BL EC number in the PDB and share a significant similarity between their active sites with the BLs from class SC.[42] However, they can be separated from the class SC BLs by both phylogeny (Figure S1) and HMM bit score (Figure 3). Sequences in each BL clade and their respective HMM bit score are shown in Table S4.

HMM bit score thresholds of 120, 430, and 120 were defined for classes SA, SC, and SD, respectively (Figure 3). After the utilization of these bit scores, the FSp was equal to 100%, 100%, and 81%, for classes SA, SC, and SD, respectively. These thresholds retrieved 1199, 389, and 388 sequences from NRBD, respectively. After the calibrations, all BL profiles together retrieved 132 proteins from SwissProt[33]: 82, 10, 24, 15, and 1 sequences with

The remaining 19 sequences not identified as BLs presented one of the following annotations: BL fragments, BL-like pro- tein, DacA carboxypeptidase, hydroxyacylglutathione hydro- lases, and ribonuclease. Also unidentified were Hcp family proteins and a sequence described as a class SA BL PenA.[43] Hcp proteins, a family of cysteine-rich PBP, do not have a sig- nificant sequence or structural similarity to known BLs.[44] A BLASTP[34] search in the NCBI protein database[17] with the putative PenA returned as best-hit a class SC protein with only 13% coverage, meaning that it is probably not a BL and cer- tainly not a PenA. Seven BlaR1/MecR1 proteins were also identified by the profiles but are regulatory proteins not associ- ated with the GO term for BL.

Validation of BLASTClust thresholds to form BLs subclasses

The "BLAST score density" values applied to the curated BL sequences to form the subclasses were validated in the PDB sequence set plus NRBLD. A significant length variation was observed between the BL sequences in PDB and those in NRBLD. The sequences in PDB range from 219 to 447 amino acids, whereas in NRBLD, they range from 96 to 619 amino acids. Therefore, in addition to the previous values of "BLAST score density," different "minimum length coverage" thresholds were chosen to cluster NRBLD sequences. Clustering results are available in Tables S5 and S6 and the thresholds in Figure 2. Application of the similarity and coverage thresholds stip- ulated for clustering resulted in the separation of true BLs from other sequences such as partial domains and the regula- tory proteins BlaR1/MecR1. For instance, in the case of class SA, most non-BL sequences have a larger average size (345- 637 amino acids) than BLs in subclasses SA1 and SA2 (285 and 300 amino acids, respectively). One cluster contains a functionally characterized BL (LRA-5, non-BL9, Table S5). No non-BL sequence was observed for the class SC. All clus- ters containing non-BL sequences from class SD have one sequence, which is similar in size to BlaR1/MecR1 proteins (~585 amino acids) or partial domains (<214 amino acids). Only one of them (YP_612206, non-BL13, Table S5) is simi- lar in size to class SD BL (274 amino acids) and its best hit in the NCBI nonredundant protein database[17] shows 43% iden- tity with a sequence annotated as "class D BL" from *Oceanicaulis alexandrii* (E value of 1E−67). All non-BL sequences from class MB are partial domains (<197 amino acids). Among the 2 clusters containing non-BL sequences formed from MB1, one possesses partial domain sequences and the other has a 340-amino acid sequence from *Stigmatella aurantiaca*, considerably larger than BL sequences (~250 amino acids). The ME2 subclasses were not maintained after clustering of NRBLD sequences, even with higher minimum coverage thresholds (90%), not corroborating what was observed when only sequences from PDB were clustered.

Comparison of the improved profiles HMM with Pfam profiles and BL motifs

Seven Pfam[36] profiles were tested against the curated data set of BLs from PDB, showing low specificity for BL classes. The 2 Pfam profiles for SBL (PF00144.22 and PF13354.4) identi- fied 339 and 227 enzymes out of a total of 399 SBLs, with an intersection of 220 sequences (35% and 3% CpSp, respectively). Three profiles for MBL (PF00753.25, PF12706.5, and PF16661.3) identified 98, 35, and 12 enzymes out of a total of 105 MBLs available (64%, 0%, and 0% CpSp, respectively). The other 2 MBL profiles did not identify any enzyme (PF13483.4 and PF14597.4).

Different sources were used to select 13 motifs related to the various BL classes, which were used to search among the 509 PDB sequences. About 11 specific motifs for the third classifi- cation level (SA, SC, and SD classes) and 2 motifs specific only for the second level (MBL) were used (Tables 1 and 2). The efficiency of these motifs was tested to separate BL sequences into subclasses (fourth and fifth levels).

No motifs developed for MBL nor motifs for class SA are present in all the sequences allocated to their respective groups.The KxxS motif[47] was found in all class SC sequences, whereas the motif SxV[6] is present in all sequences allocated in the class SD. None of the motifs analyzed was specific to BL subclasses (MB1 or MB2, SA1 or SA2, SD1 or SD2; Tables 1 and 2).

Identification and classification of BLs in fully assembled genomes

A total of 1476 BL sequences were identified in 2774 prokary- otic genomes. SA, SC, SD, MB, and ME profiles recovered 616, 280, 366, 103, and 111 sequences, respectively. No SCD class members were found in the genomes surveyed.

After the clustering and annotation process, 123 (8.3%) sequences were considered non-BLs (Table S7). The remain- ing 1352 sequences (91.7%) were distributed among 12 phyla and classified according to the BL subclasses to which they belong (Table 3). All subclasses were found in the *Proteobacteria* phylum, excepted for MB1.2. SD1 is the most disseminated subclass among the phyla analyzed, whereas SC, SD2, and MB2 were mostly restricted to *Proteobacteria*. A clear differ- ence can be observed between the phyla where BL sequences of the subclasses SA1 (*Proteobacteria, Firmicutes, and Actinobacteria*) and SA2 (*Bacteroidetes, Cyanobacteria, and Spirochaetes*) were identified. It should be noted that 49% of all analyzed strains belong to the phyla *Proteobacteria*. Regarding MBL subclasses, 51% (46) of MB1.1 sequences are in

Table 1. Efficiency of serine-β-lactamase motifs.

MOTIF	TARGET	CLASS	SuBCLASS 1*	SuBCLASS 2**
ExxLN[a]	SA	174/218 (80%)	174/216 (81%)	0/2 (0%)
SDN[a]	SA	183/218 (84%)	181/216 (83%)	2/2 (100%)
KTG[a]	SA	169/218 (78%)	167/216 (77%)	2/2 (100%)
S-[DG]-N-x(1,2)-A-[ACGNST]-x(2)-[ILMV]-x(4)-[AGSTV][b]	SA	107/218 (49%)	105/216 (48%)	2/2 (100%)
[FY]-x-[LIVMFY]-{E}-S-[TV]-x-K-x(3)-{T}-[AGLM]-{D}-{KA}-[LC][c]	SA	192/218 (88%)	190/216 (87%)	2/2 (100%)
YxN[a]	SC	116/119 (97%)	—	—
KxxS[d]	SC	119/119 (100%)	—	—
[FY]-E-[LIVM]-G-S-[LIVMG]-[SA]-K[c]	SC	118/119 (99%)	—	—
SxV[a]	SD	60/60 (100%)	56/56 (100%)	4/4 (100%)
SxxxxS[d]	SD	50/60 (83%)	46/56 (82%)	4/4 (100%)
[PA]-x-S-[ST]-F-K-[LIV]-[PALV]-x-[STA]-[LI][c]	SD	43/60 (72%)	41/56 (73%)	2/4 (50%)

[a]Bush.[6]
[b]Singh et al.[45]
[c]PROSITE.[46]
[d]MACiE.[47] Target: BL class for which the motif was developed; BL: Beta-lactamase; SA: Serine-BL class SA; SC: Serine-BL class SC; SD: Serine-BL class SD.
*Subclasses SA1 and SD1; **subclasses SA2 and SD2; — represents classes that have no subclass.

Table 2. Efficiency of metallo-β-lactamase (MBL) motifs.

MOTIF	TARGET	MBL	ME	MB	MB1	MB2
[LI]-x-[STN]-[HN]-x-H-[GSTAD]-D-x(2)-G-[GP]-x(7,8)-[GS][a]	MBL	55/105 (52%)	12/19 (63%)	43/86 (51%)	36/79 (46%)	7/7 (100%)
P-x(3)-[LIVM](2)-x-G-x-C-[LIVMF](2)-K[a]	MBL	46/105 (44%)	0	46/86 (54%)	39/79 (50%)	7/7 (100%)

[a]PROSITE.[46] Target: BL class for which the motif was developed; BL: Beta-lactamase; MBL: Metallo-BL; MB: Metallo-BL class MB; ME: Metallo-BL class ME; MB1: Metallo-BL subclass 1; MB2: Metallo-BL subclass MB2.

A higher absolute number of BL sequences were observed in *Proteobacteria*. The distribution of BLs was also analyzed at the taxonomic class level for this phylum. *Gammaproteobacteria* class has 60% (545) of the BL sequences, divided among all BL subclasses. SD1 is the only subclass found in all classes, and the unique BL found in *Epsilonproteobacteria* (Table4).

A total of 100% of the genomic sequences retrieved by the profiles (after the exclusion of non-BL sequences) were allo- cated to some subclass of BL. About 70% of their original annotations were designated only as BL (first level), 24% had information on the second or third level of classification or the gene name, and there were still 6% with erroneous or imprecise annotations, such as "hypothetical protein" (Figure 4).

Discussion

Classification schemes for BLs are of utmost importance due to the diversity of these enzymes and their importance in the scenario of bacterial resistance to antibiotics.[1,4,5,16] In gen- eral, the identification of new sequences is most often done by sequence comparison methods,[48] such as the BLASTP program.[34] Profiles HMM and other profile-sequence comparison methods led to a significant improvement in sensi- tivity over sequence comparison approaches and are already used in the identification of antibiotic resistance proteins.[3,49]

The workflow developed here systematizes the annotation of BLs based mainly on 2 steps: searches using profiles HMM, followed by clustering the resulting sequences. The calibrated profiles HMM can assign a sequence to a specific class. They also discriminate functionally characterized BLs from proteins with other biochemical functions that belong to the same superfamily and therefore share fold signals that make this separation difficult.[31] In addition, the calibrated profiles allowed the recognition of sequences erroneously attributed to the BL GO term ("BL activity") in SwissProt[33] and also enable the identification of sequences imprecisely described as BL in different antibiotic resistance databases. In the clustering step, the established thresholds of similarity and coverage allowed the clearing of non-BL sequences, providing coherent BL subclasses.

Table 3. β-lactamase sequences identified in bacterial genomes.

PHYLA	GENOMES	SA		SC	SD		MB			ME	TOTAL
		SA1	SA2		SD1	SD2	MB1.1	MB1.2	MB2		
Proteobacteria	1176	302	5	276	132	68	24	—	5	91	903
Firmicutes	583	148	—	—	30	—	46	—	—	3	227
Actinobacteria	283	107	—	4	4	—	—	—	—	1	116
Bacteroidetes	88	—	16	—	14	—	18	—	—	3	51
Cyanobacteria	73	—	2	—	17	—	—	—	—	—	19
Spirochaetes	60	—	1	—	3	—	2	1	—	7	14
Other	135	2	4	—	9	1	—	—	—	6	22
Total	2398	559	28	280	209	69	90	1	5	111	1352

Genomes: number of strains analyzed by phylum. — represents phyla where no sequences were found. Other: Verrucomicrobia, Acidobacteria, Fusobacteria, Gemmatimonadetes, Chlorobi, and Chlamydiae; BL: Beta-lactamase; SA: Serine-BL class SA; SC: Serine-BL class SC; SD: Serine-BL class SD; MB: Metallo-BL class MB (former subclasses B1 and B2); ME: Metallo-BL class ME (former subclass B3); MB1: Metallo-BL subclass MB1; MB2: Metallo-BL subclass MB2; SA1: Serine-BL subclass SA1; SA2: Serine-BL subclass SA2; SD1: Serine-BL subclass SD1; SD2: Serine-BL subclass SD2; MB1.1: Metallo-BL subclass MB1.1; MB1.2: Metallo-BL subclass MB1.2.

Table 4. β-lactamase sequences identified in bacterial genomes from *Proteobacteria* phylum.

CLASS	GENOMES	SA		SC	SD		MB		ME	TOTAL
		SA1	SA2		SD1	SD2	MB1.1	MB2		
Gammaproteobacteria	546	148	3	216	77	20	13	4	64	545
Alphaproteobacteria	321	75	2	29	15	15	2	—	22	160
Betaproteobacteria	146	76	—	31	13	29	—	1	4	154
Epsilonproteobacteria	104	—	—	—	21	—	—	—	—	21
Deltaproteobacteria	59	3	—	—	6	4	9	—	1	23
Total	1176	302	5	276	132	68	24	5	91	903

Genomes: number of strains analyzed by phylum. — represents phyla where no sequences were found BL: Beta-lactamase; SA: Serine-BL class SA; SC: Serine-BL class SC; SD: Serine-BL class SD; MB: Metallo-BL class MB (former subclasses B1 and B2); ME: Metallo-BL class ME (former subclass B3); MB1: Metallo-BL subclass MB1; MB2: Metallo-BL subclass MB2; SA1: Serine-BL subclass SA1; SA2: Serine-BL subclass SA2; SD1: Serine-BL subclass SD1; SD2: Serine-BL subclass SD2; MB1.1: Metallo-BL subclass MB1.1.

Among the sequences from BL databases used to con- struct NRBLD, some non-BL sequences presented partial domains, suggesting assembly/sequencing artefacts or anno- tation errors (eg, non-BL3, Table S5).[35] Others had much larger sizes than BLs, such as the BlaR1 protein, homologous to the BLs of the class SD, but with different functions. However, 2 sequences classified as non-BL did not cluster within any BL subclass despite being similar in size to them. The LRA-5 protein, described and functionally characterized as a class SA BL, has low similarity and is a distant relative to functionally characterized BLs and their ancestors.[50] As the experimental and HMM profile data indicate that LRA-5 is a BL of class SA, we speculate that it may belong to a third subclass (SA3) considering its low similarity to other sequences. Availability of new sequences in the future will help confirm the existence of this new. The second exception (YP_612206) is 43% identical to a sequence annotated as "class D BL" of the dimorphic rods *O alexandrii*, although the activity of this protein has not been demonstrated.[51]

It has been shown that the use of Pfam[36] profiles to identify sequences from the "Ser-BL-like superfamily" may capture unrelated sequences.[16] In our comparisons, the improved pro- files HMM displayed higher specificity when compared with Pfam profiles and BL motifs from literature. Recently, indi- vidualized subgroups of the class SA have been demonstrated, such as LSBL or TEM/SHV and CARB clusters.Characteristic residues have already been attributed to each of them, but in this work we have chosen to test motifs attributed to BL classes.[1] However, subclass-specific motifs, such for SA1 and SA2, should also be tested and compared in future studies.

Some BL subclasses that were previously described based on phylogenetic criteria were identified here using sequence similarity criteria. The sequences in subclasses MB1 and MB2 correspond to the subclasses "B1" and "B2" in the work of

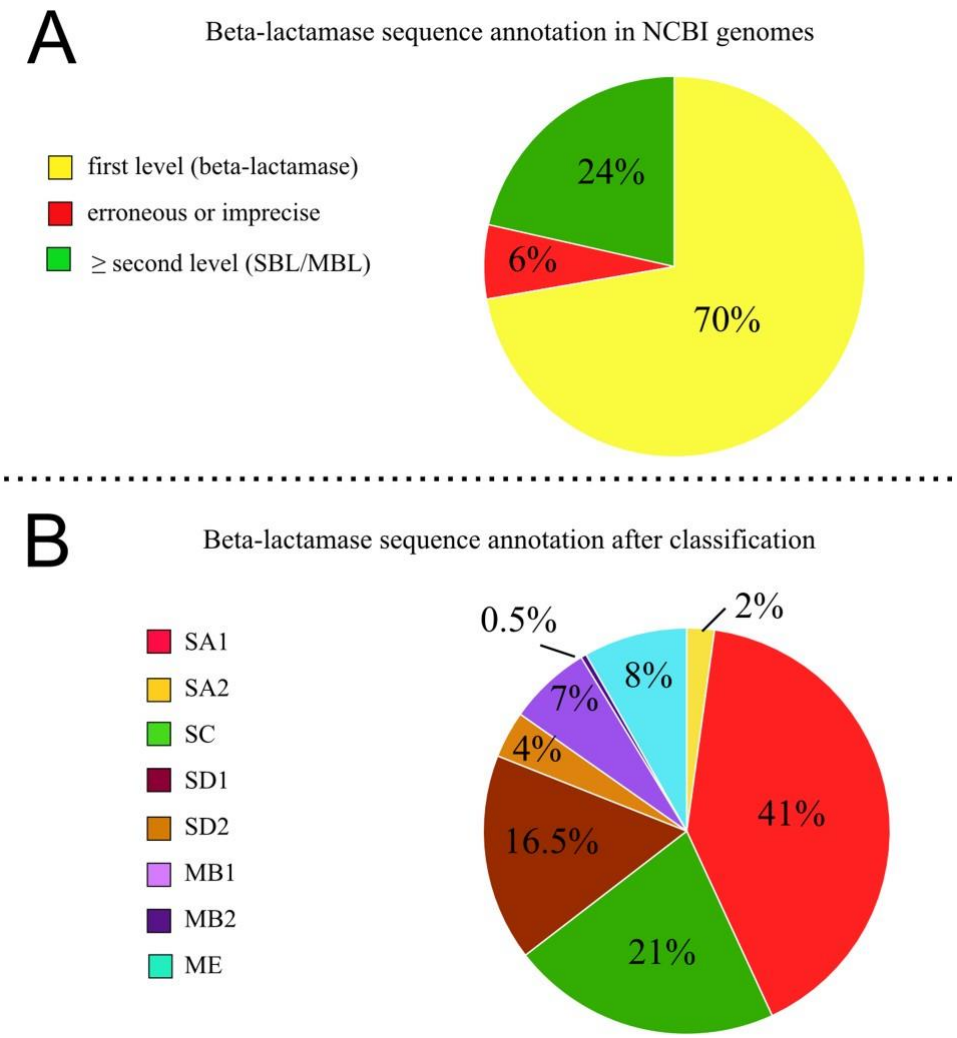


Figure 4. Original (A) annotation and (B) reannotation of the 1352 sequences according to the hierarchical levels of BL classification. BL indicates
β-lactamase; MBL: metallo-β-lactamase; SBL: serine-β-lactamase; MB1: Metallo-BL subclass MB1; MB2: Metallo-BL subclass MB2; SA1: Serine-BL subclass SA1; SA2: Serine-BL subclass SA2; SD1: Serine-BL subclass SD1; SD2: Serine-BL subclass SD2; ME: Metallo-BL class ME.

Galleni et al.[52] The sequences in the subclass SA2 correspond to BLs isolated mainly from the group Cytophagales- Flavobacteriales- Bacteroidales.[1,16,53] The OXA alleles in subclass SD2 are the same as those found in the class called "D2" by Brandt et al.[16]

A new class of BLs with fused domains, the class SCD, is proposed. Two NRBLD proteins were captured by both SC and SD profiles without using the HMM bit score threshold. One of them, LRA-13, isolated from a noncultivated soil bac- terium in Alaska, was confirmed experimentally as a BL, dis- playing a broad hydrolytic profile.[50]

Subclass MB1.2 was first presented here, formed by mem- bers of the BL SPM family. It has been suggested that SPM-1 may be a structural hybrid between MB1 and MB2 sub- classes.[54] SPM-1 genes were found only in isolates of *Pseudomonas aeruginosa*, a *Proteobacteria*, and its chromosomal location may have contributed to its isolation to other BL fam- ilies of subclass MB1.[55] Curiously, we found that *S smaragdi- nae*, a spirochaete, carries a protein with the SPM-1 domain (WP_013255389.1). Further studies are needed to establish the evolutionary relationships between these proteins.

The distribution of BL sequences obtained from fully assembled genomes in different subclasses confirmed and com- plemented previous observations. For instance, the observed enrichment of BLs in *Actinobacteria* relative to other phyla[3] is caused mainly by members of subclass SA1. The BLs of sub- class SA2, related to the *Bacteroidetes* phylum,[1,16] were also found in other phyla. The majority presence of class SC in the *Proteobacteria* phylum is a consensus, and the few sequences found in *Actinobacteria* confirm the occasional isolation of this class.[16] The wide distribution of the subclass SD1 between the phyla analyzed confirms that class SD, more precisely the sub- class SD1, has been underestimated.[16] The recently described presence of this subclass in Gram-positive bacteria has also been confirmed here.[56] The association of BL sequences from subclass SD2 to *Proteobacteria* may be related to the fact that they are naturally occurring intrinsic genes, most likely chro- mosomally located.[16] The occurrence of MB1 in *Bacteroidetes*

and *Firmicutes* phyla has been associated with chromosome, whereas the *Proteobacteria* MB1 enzymes are mostly mobile.[57] The association of class ME with soil bacteria may be related to its distribution among different phyla, which can share the same environment.[3] New BLs of class ME have already been described in metagenomes, and the majority diverge deeply from other known enzymes of this class,[50] which reinforces the idea that class ME is widely distributed and diverse.

Acidobacteria are abundant mainly in the soil, but their culti- vation is difficult.[58] Although few genomes of this phylum were available in 2016, the number of BLs identified was significant, suggesting that this is an important reservoir of BLs in nature.

The class *Gammaproteobacteria* includes common human pathogens such as *Enterobacteriaceae* and *Pseudomonadaceae*.[16] These pathogens are exposed continuously to evolutionary pressure exerted by antibiotics, favoring the acquisition of resistance genes, which may be related to the presence of differ- ent BL subclasses in this group. *Epsilonproteobacteria*, which are widely distributed bacteria including genera with pathogenic species for humans such as *Helicobacter* and *Campylobacter*, pre- sented BLs of only the SD1 subclass. Indeed, the production of 2 major BLs (OXA-61 and OXA-184) was detected in 85% of the *Campylobacter* strains.[59]

Conclusions

The workflow developed in this study presents better specific- ity when compared with available BL motifs and Pfam profiles. Application of the improved profiles HMM and sequence similarity clustering parameters resulted in a 5-level hierarchi- cal classification, consistent with previous BL classification scheme and recent proposed subclasses based on phylogeny. We also emphasize that the number of amino acids may be an important criterion for characterizing BLs, although there may be exceptions, such as the new class of fused domain enzymes (class SCD), proposed here. The workflow presented here, which can be further improved by the addition of functional and phylogenetic data, will be of great help in studies on the prevalence, distribution, and evolution of this critical enzy- matic activity.

Acknowledgements

The authors would like to thank Dr Reema Singh and Dr Harpreet Singh for providing the DLact database sequences and Dr Alex Herbert for giving us the script to remove alterna- tive positions of the atoms. MCS recognizes CAPES (Coordination for the Improvement of Higher Education Personnel, Brazil) for supporting her with a scholarship during her DSc program.

Author Contributions

MCS and ABM conceived of the manuscript outline; MCS, FFM, MC, RJ, ACRG, and ABM jointly developed the meth- odology; MCS, RAS, FFM, MC, and ABM wrote, edited and revised the manuscript.

RefeRenCes

1. Philippon A, Slama P, Dény P, Labia R. A structure-based classification of class A β-lactamases, a broadly diverse family of enzymes. *Clin Microbiol Rev*. 2016;29:29–57.
2. Abraham EP, Chain E. An enzyme from bacteria able to destroy penicillin. *Nature*. 1940;146:837.
3. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*. 2014;9:207–216.
4. Ambler RP. The structure of beta-lactamases. *Philos Trans R Soc Lond B Biol Sci*. 1980;289:321–331.
5. Hall BG, Barlow M. Revised Ambler classification of β-lactamases. *J Antimicrob Chemother*. 2005;55:1050–1051.
6. Bush K. The ABCD's of β-lactamase nomenclature. *J Infect Chemother*. 2013;19:549–559.
7. Bush K, Jacoby GA, Medeiros AA. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrob Agents Chemother*. 1995;39:1211–1233.
8. Ouzounis CA, Coulson RMR, Enright AJ, Kunin V, Pereira-Leal JB. Classification schemes for protein structure and function. *Nat Rev Genet*. 2003;4:508–519.
9. Jaurin B, Grundström T. AmpC cephalosporinase of *Escherichia coli* K-12 has a different evolutionary origin from that of beta-lactamases of the penicillinase type. *Proc Natl Acad Sci U S A*. 1981;78:4897–4901.
10. Ouellette M, Bissonnette L, Roy PH. Precise insertion of antibiotic resistance determinants into Tn21-like transposons: nucleotide sequence of the OXA-1 beta-lactamase gene. *Proc Natl Acad Sci U S A*. 1987;84:7378–7382.
11. Rasmussen BA, Bush K. Carbapenem-hydrolyzing beta-lactamases. *Antimicrob Agents Chemother*. 1997;41:223–232.
12. Gherardini PF, Wass MN, Helmer-Citterich M, Sternberg MJE. Convergent evolution of enzyme active sites is not a rare phenomenon. *J Mol Biol*. 2007;372:817–845.
13. Frère JM, Galleni M, Bush K, Dideberg O. Is it necessary to change the classification of β-lactamases? *J Antimicrob Chemother*. 2005;55:1051–1053.
14. Hall BG, Salipante SJ, Barlow M. The metallo-β-lactamases fall into two distinct phylogenetic groups. *J Mol Evol*. 2003;57:249–254.
15. Alderson RG, Barker D, Mitchell JBO. One origin for metallo-β-lactamase activity, or two? An investigation assessing a diverse set of reconstructed ancestral sequences based on a sample of phylogenetic trees. *J Mol Evol*. 2014;79:117–129.
16. Brandt C, Braun SD, Stein C, et al. In silico serine β-lactamases analysis reveals a huge potential resistome in environmental and pathogenic species. *Sci Rep*. 2017;7:43232. doi:10.1038/srep43232.
17. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2015;43:D6–D17.
18. Burley SK, Berman HM, Christie C, et al. RCSB Protein Data Bank: sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci*. 2017;27:316–330.
19. Srivastava A, Singhal N, Goel M, Virdi JS, Kumar M. CBMAR: a comprehensive β-lactamase molecular annotation resource. *Database*. 2014;2014:bau111.
20. Bialek-Davenet S, Criscuolo A, Ailloud F, et al. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. *Emerg Infect Dis*. 2014;20:1812–1820.
21. Singh R, Singh H. DLact: an antimicrobial resistance gene database. *J Comput Intell Bioinform*. 2008;1:93–108.
22. Thai QK, Bös F, Pleiss J. The lactamase engineering database: a critical survey of TEM sequences in public databases. *BMC Genomics*. 2009;10:390.
23. Widmann M, Pleiss J, Oelschlaeger P. Systematic analysis of metallo-β-lactamases using an automated database. *Antimicrob Agents Chemother*. 2012;56:3481–3491.
24. McArthur AG, Waglechner N, Nizam F, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother*. 2013;57:3348–3357.
25. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26:680–682.
26. MaxCluster: a tool for protein structure comparison and clustering. http://www.sbg.bio.ic.ac.uk/maxcluster/index.html. Accessed March, 2016.
27. Wei D, Jiang Q , Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics*. 2012;13:174.
28. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–1797.
29. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*.

30. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*. 2012;28:2685–2686.
31. Bateman A, Martin MJ, O'Donovan C, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45:D158–D169.
32. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–3402.
33. Triant DA, Pearson WR. Most partial domains in proteins are alignment and annotation artifacts. *Genome Biol*. 2015;16:99.
34. Finn RD, Coggill P, Eberhardt RY, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016;44:D279–D285.
35. Pratap S, Katiki M, Gill P, Kumar P, Golemi-Kotra D. Active-site plasticity is essential to carbapenem hydrolysis by OXA-58 class D β-lactamase of *Acineto- bacter baumannii*. *Antimicrob Agents Chemother*. 2016;60:75–86.
36. Mukherjee S, Stamatis D, Bertsch J, et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res*. 2017;45:D446–D456.
37. Urbach C, Evrard C, Pudzaitis V, Fastrez J, Soumillion P, Declercq JP. Struc- ture of PBP-A from *Thermosynechococcus elongatus*, a penicillin-binding protein closely related to class A beta-lactamases. *J Mol Biol*. 2009;386:109–120.
38. Boudreau MA, Fishovitz J, Llarrull LI, Xiao Q , Mobashery S. Phosphorylation of BlaR1 in manifestation of antibiotic resistance in methicillin-resistant Staphylococ- cus aureus and its abrogation by small molecules. *ACS Infect Dis*. 2016;1:454–459.
39. Wilke MS, Hills TL, Zhang HZ, Chambers HF, Strymadka NCJ. Crystal structures of the Apo and penicillin-acylated forms of the BlaR1 beta-lactam sensor of *Staphylococcus aureus*. *J Biol Chem*. 2004;279:47278–47287.
40. Dubois D, Baron O, Cougnoux A, et al. ClbP is a prototype of a peptidase sub- group involved in biosynthesis of nonribosomal peptides. *J Biol Chem*. 2011;286:35562–35570.
41. Proenca R, Niu WW, Cacalano G, Prince A. The *Pseudomonas cepacia* 249 chro- mosomal penicillinase is a member of the AmpC family of chromosomal beta- lactamases. *Antimicrob Agents Chemother*. 1993;37:667–674.
42. Lüthy L, Grütter MG, Mittl PRE. The crystal structure of *Helicobacter pylori* cysteine-rich protein B reveals a novel fold for a penicillin-binding protein. *J Biol Chem*. 2002;277:10187–10193.
43. Singh R, Saxena A, Singh H. Identification of group specific motifs in beta-lac- tamase family of proteins. *J Biomed Sci*. 2009;16:109.
44. Sigrist CJA, De Castro E, Cerutti L, et al. New and continuing developments at PROSITE. *Nucleic Acids Res*. 2013;41:344–347.
45. Holliday GL, Andreini C, Fischer JD, et al. MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res*. 2012;40:783–789.
46. Feuermann M, Gaudet P, Mi H, Lewis SE, Thomas PD. Large-scale inference of gene function through phylogenetic annotation of gene ontology terms: case study of the apoptosis and autophagy cellular processes. *Database*. 2016; 2016:baw155.
47. Fróes AM, da Mota FF, Cuadrat RRC, D'avila AMR. Distribution and classifi- cation of serine β-lactamases in Brazilian hospital sewage and other environ- mental metagenomes deposited in public databases. *Front Microbiol*. 2016;7:1790.
48. Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J*. 2009;3:243–251.
49. Oh HM, Kang I, Vergin KL, Lee K, Giovannoni SJ, Cho JC. Genome sequence of *Oceanicaulis* sp. strain HTCC2633, isolated from the western Sargasso Sea. *J Bacteriol*. 2011;193:317–318.
50. Galleni M, Lamotte-brasseur J, Rossolini GM. Standard numbering scheme for class B beta-lactamases. *Society*. 2001;45:660–663.
51. Hall BG, Barlow M. Evolution of the serine beta-lactamases: past, present and future. *Drug Resist Updat*. 2004;7:111–123.
52. Bebrone C. Metallo-beta-lactamases (classification, activity, genetic organiza- tion, structure, zinc coordination) and their superfamily. *Biochem Pharmacol*. 2007;74:1686–1701.
53. Silveira MC, Albano RM, Asensi MD, Carvalho-Assef AP. Description of genomic islands associated to the multidrug-resistant *Pseudomonas aeruginosa* clone ST277. *Infect Genet Evol*. 2016;42:60–65.
54. Toth M, Antunes NT, Stewart NK, et al. Class D β-lactamases do exist in Gram-positive bacteria. *Nat Chem Biol*. 2016;12:9–14.
55. Berglund F, Marathe NP, Österlund T, et al. Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metage- nomic data. *Microbiome*. 2017;5:134.
56. Kielak AM, Barreto CC, Kowalchuk GA, Van Veen JA, Karumae EE. The ecology of Acidobacteria: moving beyond genes and genomes. *Front Microbiol*. 2016;7:744.
57. Weis AM, Storey DB, Taff CC, et al. Genomic comparison of *Campylobacter* spp. and their potential for zoonotic transmission between birds, primates, and livestock. *Appl Environ Microbiol*. 2016;82:7165–7175.