UNIVERSIDADE DE SÃO PAULO
Programa Interunidades de Pós-Graduação em Bioinformática

**Uma abordagem computacional integrada para a construção e
utilização de HMMs de perfil para análises genômicas e
metagenômicas**

Liliane Santana Oliveira Kashiwabara

Orientador: Prof. Dr. Arthur Gruber

Co-orientador: Prof. Dr. Alan Mitchell Durham

SÃO PAULO

2019

Uma abordagem integrada para a construção e utilização de HMMs de perfil para análises genômicas e metagenômicas

Liliane Santana Oliveira Kashiwabara

Tese apresentada à Universidade de São Paulo,
como parte das exigências do Programa de Pós-
graduação Interunidades em Bioinformática,
para obtenção do título de Doutor em Ciências.

Área de Concentração: Bioinformática

Orientador: Prof. Dr. Arthur Gruber

Co-orientador: Prof. Dr. Alan Mitchell Durham

SÃO PAULO

2019

FICHA CATALOGRÁFICA

K19 Kashiwabara, Liliane Santana Oliveira
Uma abordagem integrada para a construção e utilização de HMMs de perfil para análise genômicas e metagenômicas. / Liliane Santana Oliveira Kashiwabara, Arthur Gruber, [orient.] e Alan Mitchell Durham [coorient.]. São Paulo : 2019. 100 p.

Tese (Doutorado) - Universidade de São Paulo
Orientador: Prof. Dr. Arthur Gruber
Coorientador: Prof. Dr. Alan Mitchell Durham

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Bioinformática. 2. Modelos para Processos Estocásticos. 3. Desenvolvimento de Software. 4. Vírus. 5. Marcador Molecular. I. Gruber, Arthur, orientador. II. Durham, Alan Mitchell, coorientador. III. Universidade de São Paulo. IV. Título.

CDD – 572.8

UNIVERSIDADE DE SÃO PAULO
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Candidato(a): Liliane Santana Oliveira Kashiwabara

Título da Tese: Uma abordagem integrada para a construção e utilização
de HMMs de perfil para análises genômicas metagenômicas

Orientador(a): Arthur Gruber

Co-orientador(a): Alan Michell Durham

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a/...../.....,

☐ Aprovado(a) ☐ Reprovado(a)

Examinador(a): Assinatura:

Nome:

Instituição:

Examinador(a): Assinatura:

Nome:

Instituição:

Examinador(a): Assinatura:

Nome:

Instituição:

Agradecimentos

Agradeço primeiramente aos meus pais, Maria José e Geraldo, pelo apoio em todas as áreas da minha vida. Eu não seria a pessoa que sou sem vocês e não teria chegado até aqui. Aos meus irmãos, Lidiane e João Geraldo, pelo carinho e apoio durante toda a vida. Ao meu esposo, André Kashiwabara, pelo amor e grande ajuda dispendido durante esses quatros. Agradeço também pela compreensão durante as ausências, sem o seu apoio teria sido impossível a obtenção desse título. Amo muito todos vocês!

Agradeço também ao querido orientador Arthur Gruber, pelos conhecimentos passados, paciência, carinho e pelo grande esforço para a realização de todos os nossos projetos. Sem essa ajuda o desenvolvimento desse lindo trabalho não teria sido possível. Agradeço também ao meu co-orientador Alan Mitchell por toda a atenção dada a mim durante esses anos, e ao meu supervisor do período sanduíche, Alejandro Reyes, por me receber em seu laboratório. Agradeço também ao Prof. Bas Dutilh (Utrecht University, Holanda) pela parceria durante o desenvolvimento desse projeto, e pela imensa ajuda nos testes e validações das ferramentas. Ao Prof. François Enault (Université Clermont Auvergne, França) pelas sequências virais gentilmente cedidas para a validação das ferramentas desenvolvidas durante esse trabalho.

Não posso deixar de agradecer a pessoas especiais da minha família: à minha madrinha Rita de Cássia, por toda a torcida e apoio durante todos os momentos da minha vida. Às minhas tias Ana e Lourdes, que sempre estiveram presentes em todos os momentos, mesmo que à distância. Aos meus tios Gervásio, Maria Aparecida e Amintas, pela eterna preocupação com o meu bem-estar e torcida pelo meu sucesso! Vocês estão no meu coração e espero que possamos comemorar muito nossas pequenas e grandes conquistas!

Agradeço a todos os meus amigos, tanto os feitos durante esse período (Almir Ferreira, Wendel Hime e Alda Backx), como os antigos (Ulysses Oliveira, Kátia Ribeiro, Gabrielle Eler, Henrique Shishido, Alexandre Paschoal e Flávia Cal). Muito obrigada por todo o apoio, pelas risadas e pelos momentos de distração e desabafo durante esse tempo! Elas foram crucias para me manterem de pé!

Enfim, muito obrigada a todos que direta ou indiretamente contribuíram para que eu chegasse aqui.

Apoio financeiro

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

Kashiwabara, L.S.O. **Uma abordagem integrada para a construção e utilização de HMMs de perfil para análises genômicas e metagenômicas.** [tese (Doutorado em Ciências)] - Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo, 2019.

Resumo

HMMs de perfil são um método poderoso para modelar a diversidade de sequências biológicas e constituem uma abordagem muito sensível para a detecção de ortólogos remotos. Uma potencial aplicação de tais modelos é a detecção de vírus emergentes e novos elementos genéticos móveis. Nosso grupo desenvolveu recentemente o GenSeed-HMM, um programa que emprega HMMs de perfil como sementes para montagem progressiva de genes-alvo, utilizando tanto dados genômicos como metagenômicos. No presente trabalho foi desenvolvido o TABAJARA, um programa para o desenho racional de HMMs de perfil. Partindo de um alinhamento de múltiplas sequências, o TABAJARA é capaz de encontrar blocos que são (1) conservados ou (2) discriminativos para dois ou mais grupos de sequências. O programa utiliza diferentes métricas para atribuir pontuações posição-específicas ao longo de todo o alinhamento e utiliza então uma janela deslizante para encontrar as regiões com maiores pontuações. Blocos de alinhamento selecionados são então extraídos e utilizados para construir HMMs de perfil. Para validar o método, o programa TABAJARA foi empregado para a construção de modelos para vírus do gênero *Flavivirus* e para fagos da família *Microviridae*. Em ambos os grupos virais foi possível se obter modelos de ampla abrangência, capazes de detectar todos os membros de um respectivo grupo taxonômico, e modelos de abrangência mais restrita, específicos para espécies distintas de *Flavivirus* (ex. DENV, ZIKV ou YFV) ou subfamílias de *Microviridae* (ex. *Alpavirinae*, *Gokushovirinae* e *Pichovirinae*). Em outra validação, foram utilizadas sequências da endonuclease Cas1 para se obter modelos capazes de diferenciar CRISPRs de casposons, esses últimos representando uma superfamília de transposons de DNA autossintetizantes, os quais originaram o sistema de imunidade CRISPR-Cas de procariotos. O TABAJARA conseguiu gerar modelos específicos de Cas1 derivada de casposons, permitindo sua diferenciação em relação aos seus ortólogos de CRISPRs. No presente trabalho foi desenvolvido ainda o HMM-Prospector, uma ferramenta que utiliza um conjunto de HMMs de perfil para a triagem de dados de sequenciamento genômico ou metagenômico. O programa informa quais são os modelos mais reconhecidos pelas leituras, sob valores de corte de pontuação definidos pelo usuário, assim como quantas leituras são detectadas por cada modelo. Com esta informação, os modelos mais relevantes podem ser utilizados como sementes em montagens progressivas com o programa GenSeed-HMM, dentro de uma abordagem integrada para a construção de modelos e sua aplicação. Finalmente, foi desenvolvido o e-Finder, um aplicativo genérico para a detecção e extração de elementos multigênicos a partir de genomas ou metagenomas montados utilizando HMMs de perfil. O e-Finder executa buscas de similaridade entre os HMMs de perfil e as sequências traduzidas dos dados montados e checa, em seguida, se os critérios de sintenia pré-definidos foram atendidos, incluindo o número mínimo de genes, a ordem dos genes e as distâncias intergênicas. As sequências dos elementos são então extraídas, as regiões codificantes (ORFs) identificadas e traduzidas conceitualmente em sequências completas de proteínas. Para validar esta ferramenta, foram empregados dois estudos de caso, profagos da família *Microviridae* e casposons, utilizando-se HMMs de perfil específicos, construídos com o programa TABAJARA. Em ambos os casos, o e-Finder foi executado usando-se a base de dados PATRIC, um repositório com mais de 135.000 genomas de bactérias e arqueias.

Foram identificados um total de 91 contigs positivos para casposons a partir de 79 genomas distintos. No caso dos *Microviridae*, foram encontrados 104 profagos candidatos, estendendo o conhecimento da gama de hospedeiros bacterianos. Em ambos os casos, análises filogenéticas confirmaram a correta atribuição taxonômica das sequências positivas. Os programas desenvolvidos neste trabalho podem ser utilizados isoladamente ou em combinação para detectar e discriminar sequências conhecidas ou remotamente relacionadas. Juntamente com o GenSeed-HMM, estes programas constituem um conjunto integrado de ferramentas com potencial aplicação na busca de novos vírus e elementos genéticos móveis, bem como em qualquer outra tarefa relacionada à detecção e/ou discriminação de subgrupos de famílias de sequências nucleotídicas ou proteicas.

Palavras-chave: Modelos ocultos de Markov, HMMs de perfil, Detecção de vírus, Metagenômica, Sintenia, Famílias proteicas.

Kashiwabara, L.S.O. **An integrated approach for the construction and application of profile HMMs for genomic and metagenomic analyses.** [thesis (Doctorate degree in Science)] - Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo, 2019.

Abstract

Profile HMMs are a powerful way of modeling sequence diversity and constitute a very sensitive approach to detect remote orthologs. A potential application of such models is the detection of emerging viruses and novel mobile genetic elements. Our group has recently developed GenSeed-HMM, a tool that employs profile HMMs as seeds for gene-targeted progressive assembly using either genomic or metagenomic data. In this work we developed TABAJARA, a program for the rational design of profile HMMs. Starting from a multiple sequence alignment, TABAJARA is able to find blocks that are either (1) conserved across all sequences or (2) discriminative for two or more specific groups of sequences. The program uses different metrics to ascribe position-specific scores along the whole alignment and then uses a sliding-window to find top-scoring regions. Selected alignment blocks are then extracted and used to build profile HMMs. To validate the method, we employed TABAJARA to construct models for viruses of the *Flavivirus* genus and phages of the *Microviridae* family. In both viral groups we were able to obtain wide-range models, able to detect all members of the respective taxonomic group, and models that are specific to particular *Flavivirus* species (e.g. DENV, ZIKV or YFV) or *Microviridae* subfamilies (e.g. *Alpavirinae*, *Gokushovirinae* and *Pichovirinae*). In another validation, we used sequences of the endonuclease Cas1 to obtain models capable of differentiating CRISPRs from casposons, the latter elements representing a superfamily of self-synthesizing DNA transposons that originated the prokaryotic CRISPR-Cas immunity. TABAJARA succeeded to generate models specific to casposon-derived Cas1, enabling their differentiation from CRISPR orthologs. We also developed HMM-Prospector, a tool that can use a batch of profile HMMs to screen genomic or metagenomic sequencing data, reporting which profile HMMs are mostly recognized under user-defined score cutoff values, and how many reads are detected by each model. With this information, the most relevant models can be used as seeds in progressive assemblies with GenSeed-HMM program, providing an integrated approach for model construction and application. Finally, we developed e-Finder, a generic application for detecting and extracting multigene elements from assembled genomes or metagenomes using profile HMMs. e-Finder runs similarity searches of profile HMMs against translated sequences of the assembled data and then checks if pre-defined syntenic criteria have been fulfilled, including minimum number of genes, gene order and intergenic distances. Element sequences are then extracted, their ORFs identified and conceptually translated into full-length protein sequences. To validate the tool, we employed two distinct case studies, prophages of the *Microviridae* family and casposons, using specific profile HMMs constructed by TABAJARA. In both cases, we executed e-Finder using the PATRIC database, a repository with over 135,000 bacterial and archaeal genomes. We identified in total 91 casposon-positive contigs from 79 distinct genomes. In the case of *Microviridae*, we found a total of 104 provirus candidates, extending the known range of bacterial hosts. In both cases, phylogenetic analyses confirmed the correct taxonomic assignment of the positive sequences. The programs developed in this work can be used alone or in combination to detect and discriminate known or distantly related sequences. Together with GenSeed-HMM, these programs provide an integrated toolbox with

potential application in the search of novel viruses and mobile genetic elements, as well as in any other task related to the detection and/or discrimination of subgroups of DNA or protein sequences.

Keywords: Hidden Markov models, Profile HMMs, Virus detection, Metagenomics, Synteny, Protein families.

Lista de Abreviaturas e Siglas

BLAST	Basic Local Alignment Search Tool
HMM	Hidden Markov Model
PSSM	Position-specific scoring matrix
MSA	Multiple sequence alignment
ICTV	International Committee on Taxonomy of Viruses
IPG	Identical Protein Groups
MUSCLE	Multiple Sequence Comparison by Log
NCBI	National Center for Biotechnology Information
PATRIC	Pathosystems Resource Integration Center
TABAJARA	Tool for Alignment Block Analysis Joining Appropriate Rational Approaches
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
PFAM	Protein Families database
REFSEQ	Reference sequence database
CDD	Conserved Domains Database
CDS	Coding Sequence
ORF	Open Reading Frame
POGs	Phage Orthologous Groups
pVOGs	Prokaryotic Virus Orthologous Groups
GOM	Genomic Organization Model
PPHMM	Protein Profile Hidden Markov Model

Lista de Figuras

Figura 1– Exemplo de um HMM de perfil.....	13
Figura 2 – Gráficos de dispersão com <i>jittering</i> representando os resultados de buscas de similaridade de HMMs de perfil construídos a partir de sequências inteiras de poliproteínas de <i>Flavivirus</i>	21
Figura 3 – Gráficos de dispersão com <i>jittering</i> representando os resultados de buscas de similaridade de HMMs de perfil construídos a partir de sequências de proteínas inteiras de VP1 de <i>Microviridae</i>	23
Figura 4. Avaliação de um repertório completo de blocos de alinhamento de tamanho variável como candidatos para a construção de HMMs de perfil discriminativos para o gênero <i>Flavivirus</i>	26
Figura 5 – Avaliação de um repertório completo de blocos de alinhamento de tamanho variável como candidatos para a construção de HMMs de perfil discriminativos para a família <i>Microviridae</i>	28
Figura 6 – Fluxo de trabalho de do TABAJARA.	32
Figura 7- Efeito da variação do valor mínimo do score por posição do alinhamento (parâmetro $-\tau$) no desempenho dos HMMs de perfil construídos com o TABAJARA.	38
Figura 8 - Efeito da variação do valor de porcentagem mínima de posições significativas em uma janela deslizante (parâmetro $-\rho$) no desempenho dos HMMs de perfil construídos com o TABAJARA.	39
Figura 9. Efeito da variação do tamanho dos blocos de alinhamento selecionados no desempenho dos HMMs de perfil construídos com o TABAJARA.	40
Figura 10. O efeito da combinação de múltiplos HMMs de perfil na capacidade de detectar e discriminar sequências de <i>Microviridae</i>	49
Figura 11. Efeito do tamanho da amostragem de sequências no desempenho de HMMs de perfil construídos com TABAJARA.	55
Figura 12 – Fluxo de trabalho do HMM-Prospector.....	65
Figura 13 – Execução do e-Finder para a busca de um elemento multigênico composto por quatro genes.	68
Figura 14 – Fluxo de trabalho do e-Finder.....	69
Figura 15 – Árvore filogenética construída com sequências da proteína Cas1 (endonuclease Cas) dos casposons identificados pelo e-Finder nos genomas de bactérias e arqueias da base <i>Patric</i>	73

Figura 16 - Árvore filogenética construída com sequências da proteína VP1 da família *Microviridae*.78

Figura 17 – Fluxo de trabalho da abordagem integrada de produção e uso de HMMs de perfil.84

Lista de Tabelas

Tabela 1 – Classificação de Baltimore para os genomas virais.	4
Tabela 2 – Sensibilidade e especificidade de HMMs de perfil construídos com os alinhamentos de proteínas inteiras de VP1 de <i>Microviridae</i>	24
Tabela 3. Resultado das buscas de similaridade realizadas com modelos conservados construídos para Flavivirus com o TABAJARA.	42
Tabela 4. Resultado das buscas de similaridade realizadas com modelos conservados construídos para a família <i>Microviridae</i> com o TABAJARA.	43
Tabela 5. Resultado das buscas de similaridade realizadas com modelos discriminativos construídos para as espécies <i>Dengue virus</i> , <i>Zika virus</i> e <i>Yellow Fever virus</i> do gênero <i>Flavivirus</i>	46
Tabela 6. Resultado das buscas de similaridade realizadas com modelos discriminativos construídos para as subfamílias <i>Alpavirinae</i> , <i>Gokushovirinae</i> e <i>Pichovirinae</i> da família <i>Microviridae</i>	47
Tabela 7 – Validação cruzada k-fold de HMMs de perfil específicos do vírus da Dengue construídos com TABAJARA.	51
Tabela 8 - Validação cruzada k-fold de HMMs de perfil específicos de vírus da subfamília Gokushovirinae construídos com TABAJARA.	52
Tabela 9 – Quantidade de HMMs de perfil curtos conservados gerados para cada uma das subfamílias de <i>Microviridae</i> com o TABAJARA.	74
Tabela 10 – Lista dos profagos de <i>Microviridae</i> encontrados pelo e-Finder para as subfamílias <i>Alpavirinae</i> , <i>Gokushovirinae</i> e <i>Pichovirinae</i>	75
Tabela 11 - Lista dos profagos de <i>Microviridae</i> encontrados pelo e-Finder para a nova subfamília <i>Katanvirinae</i>	76
Tabela 12 – Comparação entre as características das ferramentas e-Finder, MultigeneBlast e MacsyFinder.....	82

Sumário

1	Introdução	1
2	Conceitos biológicos e ferramentas para a prospecção viral.....	3
2.1	Introdução aos vírus	3
2.1.1	Bacteriófagos da família <i>Microviridae</i>	5
2.1.2	Vírus do gênero <i>Flavivirus</i>	6
2.2	Casposons.....	7
2.3	Ferramentas existentes para a prospecção e classificação viral	7
3	HMMs de perfil e prospecção viral.....	11
3.1	HMMs de perfil.....	11
3.2	Bases virais de HMMs de perfil.....	13
3.3	Aplicação de HMMs de perfil na detecção e reconstrução de sequências virais 14	
3.4	Metodologias para a classificação viral que utilizam HMMs de perfil.....	15
4	Objetivos	18
4.1	Geral.....	18
4.2	Específicos	18
5	TABAJARA.....	19
5.1	Uso de proteínas inteiras para a construção dos HMMs de perfil.....	20
5.2	Uso de blocos de alinhamento curtos para a construção de HMMs de perfil ..	24
5.3	Desenvolvimento da metodologia.....	28
5.3.1	Conceitos de teoria da informação	29
5.3.2	Implementação do TABAJARA	31
5.4	Validação da ferramenta.....	36
5.4.1	Otimização dos parâmetros do TABAJARA.....	36
5.4.2	Validação de modelos construídos para <i>Flavivirus</i> e <i>Microviridae</i>	41
5.4.3	Uso cumulativo de modelos curtos para a detecção de sequências.....	47
5.4.4	Capacidade de generalização dos HMMs de perfil.....	49
5.4.5	Influência da amostragem no desempenho dos HMMs de perfil.....	53
5.5	Discussão.....	57
6	Viral MinionDB.....	62
7	HMM-Prospector	63
7.1	Implementação da ferramenta	64
8	e-Finder.....	66
8.1	Implementação da ferramenta	66

8.2	Validação da ferramenta.....	70
8.2.1	Busca de Casposons em genomas de bactérias e arqueias	70
8.2.2	Busca de profagos de Microviridae em genomas da base PATRIC	72
8.3	Discussão.....	77
9	Uma abordagem integrada para a prospecção e classificação de sequências virais	83
10	Conclusões e trabalhos futuros.....	85
11	Referências bibliográficas	87

1 Introdução

Os vírus são as entidades mais abundantes e diversas do planeta, estão presentes em diversos ambientes e são considerados os principais protagonistas dos ecossistemas naturais (Roux *et al.*, 2014; Zhang *et al.*, 2018). Apesar dessa abundância, a diversidade viral ainda é pouco conhecida, uma vez que os métodos clássicos de detecção de vírus se baseiam na cultura e isolamento dos mesmos em laboratório, o que exige o conhecimento seus hospedeiros e as condições necessárias para o cultivo *in vivo* (Fancello *et al.*, 2012). Como a grande maioria dos vírus não é facilmente cultivável por técnicas tradicionais (Mokili *et al.*, 2012), estima-se que apenas uma pequena parcela das sequências virais seja conhecida. Por exemplo, Rose *et al.* (2016) estimam que pelo menos 320.000 vírus que infectam mamíferos ainda não tenham sido detectados.

A metagenômica, por ser uma técnica independente de cultivo e que utiliza tecnologias de sequenciamento de nova geração, tem permitido aumentar enormemente a detecção de novos vírus (Dutilh *et al.*, 2017). Vários estudos têm utilizado essas técnicas para identificar vírus em amostras de diferentes tipos de ambiente, como solo (Segobola *et al.*, 2018; Fancello *et al.*, 2013), águas marinhas (Coutinho *et al.*, 2017), ambientes de água doce (Roux *et al.*, 2012A), o corpo humano (Moore *et al.*, 2015), entre outros. Em geral, a identificação e classificação dos vírus nesses estudos é feita através de buscas de similaridade por meio de um alinhamento pareado das sequências montadas contra bancos de dados de genomas e/ou proteínas virais conhecidas. Embora o alinhamento pareado tenha sido o método que permitiu identificar quase todos os vírus atualmente conhecidos, essa metodologia não é suficientemente sensível para detectar vírus muito divergentes em relação às sequências disponíveis nos bancos de dados (Krishnamurthy & Wang, 2017). De fato, Mokili *et al.* (2012), fizeram um levantamento da literatura cobrindo metagenomas de diversos ambientes, como águas marinhas, fezes humanas e sedimentos marinhos, e observaram que todos os trabalhos usaram buscas de similaridade com o programa BLAST, tendo falhado em identificar entre 44 a 99% das sequências. De acordo com Brenner *et al.* (1998), métodos de busca de similaridade pareada são capazes de detectar sequências protéicas com até 30% de identidade. Quando esse percentual de identidade fica entre 20 e 30%, esses métodos deixam de detectar pelo menos metade das sequências. Considerando-se que vírus apresentam altas taxas de mutação, especialmente aqueles com genoma de RNA (Drake, 1993; Peck & Luring, 2018), esses métodos podem deixar de identificar sequências virais mais remotas.

Métodos de busca de similaridade baseados em perfis, os quais podem agregar e quantificar as variações existentes em um conjunto de sequências, podem representar uma alternativa mais sensível do que buscas com alinhamento pareado. De fato, Park *et al.* (1998) demonstraram para sequências com até 30% de similaridade que métodos de comparação de sequências que utilizam perfis (ISS, PSI-BLAST e SAM-T98) são capazes de detectar até três vezes mais sequências do que métodos baseados em alinhamento pareado. HMMs de perfil (Krogh *et al.*, 1994) são modelos probabilísticos lineares construídos a partir de alinhamentos de múltiplas sequências, os quais funcionam como conjunto de treinamento. Esses modelos podem, portanto, incorporar toda a diversidade de resíduos de cada posição do alinhamento, incluindo a ocorrência de inserções e deleções (*indels*). Existem atualmente alguns bancos de HMMs de perfil de sequências virais publicamente disponíveis. Entretanto, essas bases possuem uma série de limitações, incluindo a pequena representação de sequências em cada modelo e a falta de relação direta dos modelos com grupos taxonômicos específicos (Reyes *et al.*, 2017).

Os vírus não compartilham genes que podem ser utilizados como marcadores universais, ao contrário de organismos procariotos e eucariotos, nos quais se pode empregar marcadores como, por exemplo, o RNA ribossômico 16S e 18S, respectivamente. Contudo, utilizando-se um conjunto de sequências de um determinado grupo viral, é possível se obter um alinhamento múltiplo de suas sequências para a construção de modelos específicos para esse grupo. Isso pode ser feito utilizando-se todo o comprimento das sequências do alinhamento ou identificando regiões que detectem de forma específica os organismos desse grupo. Os modelos gerados podem então ser utilizados para a detecção e classificação de sequências virais.

Esse documento está estruturado da seguinte maneira: no capítulo 2 é feita uma descrição geral a respeito de vírus, bem como uma revisão da literatura sobre as ferramentas disponíveis publicamente para a prospecção viral em conjuntos de sequências. No capítulo 3 é introduzido o conceito de HMMs de perfil, os bancos de HMMs de perfil virais disponíveis publicamente e os métodos existentes para a desenvolvimento de HMMs de perfil virais. No capítulo 4 são descritos os objetivos do projeto desenvolvido e os capítulos 5, 6, 7, 8 e 9 descrevem as metodologias e ferramentas desenvolvidas durante esse trabalho. Por fim, no capítulo 10 são feitas algumas conclusões a respeito do projeto, bem como sugeridos alguns trabalhos futuros, e no capítulo 11 são listadas as referências utilizadas.

2 Conceitos biológicos e ferramentas para a prospecção viral

Os vírus são entidades biológicas encontradas em diferentes ambientes, como solo, águas marinhas e diversas partes do corpo humano (sangue, trato respiratório, gastrointestinal, boca, entre outros). Nesse capítulo faremos uma breve explanação sobre os vírus, bem como sua caracterização e classificação taxonômica.

Para a validação as metodologias e ferramentas desenvolvidas durante esse projeto, foram utilizados dois grupos virais, bacteriófagos da família *Microviridae* e vírus do gênero *Flavivirus*, e uma superfamília de elementos genéticos móveis, os casposons. Esses três grupos biológicos também serão brevemente descritos durante esse capítulo.

Por fim, nesse capítulo serão ainda apresentadas algumas ferramentas disponíveis publicamente para a identificação de sequências virais em amostras genômicas e metagenômicas.

2.1 Introdução aos vírus

Apesar de não serem considerados seres vivos, uma vez que são parasitas intracelulares obrigatórios, os vírus são as entidades biológicas mais abundantes do planeta e compreendem uma grande proporção do bioma do planeta (Flint *et al.*, 2015). Alguns desses são conhecidos, como o *HIV* e o *Herpesvirus*, enquanto outros são novos ou emergentes, como os vírus da *Zika* e *Chikungunya*. Desde que foram descobertos no final do século 19, os vírus são frequentemente estudados como causadores de doenças em seres humanos, animais e plantas economicamente importantes (Zhang *et al.*, 2018; Güemes *et al.*, 2016). Contudo, os vírus sabidamente patogênicos representam apenas uma pequena parcela da virosfera. Muitos vírus têm papéis ecológicos ainda indefinidos, mas claramente têm uma grande importância, especialmente em procariotos. De fato, mais de 94% de todas as partículas contendo ácido nucléico presentes nos oceanos pertencem a esses organismos e eles são 15 vezes mais abundantes que bactérias e arqueias (Flint *et al.*, 2015).

Bacteriófagos, ou fagos, são vírus que infectam bactérias e estão presentes de forma ubíqua nas bactérias de diversos ambientes. Estudos mostram que esses fagos desempenham papéis-chave nos microbiomas, os quais afetam os ambientes em geral (Clokier *et al.*, 2011). As bactérias presentes no intestino humano de indivíduos saudáveis hospedam uma comunidade viral diversa de bacteriófagos que potencialmente ajudam a

moldar a estrutura e a função do microbioma intestinal (Manriquea *et al.*, 2016). Esses vírus podem se reproduzir utilizando dois ciclos reprodutivos: o lítico e o lisogênico. No ciclo lítico, após a infecção da célula bacteriana, os fagos se reproduzem e rompem a membrana da célula, causando a sua morte. Já no ciclo lisogênico, após a invasão da célula hospedeira, esses fagos integram seus genomas ao genoma da célula e se mantêm num estado de dormência. Nesse caso, a expressão gênica viral é represada e o vírus passa a ser replicado pelo hospedeiro (Flint *et al.*, 2015). Quando estão integrados ao genoma do hospedeiro, esses fagos são chamados de profagos. Os fagos que são capazes de trocar seus ciclos de reprodução entre o lítico e o lisogênico são chamados de fagos temperados.

Os vírus apresentam genoma constituído de DNA ou RNA de fita simples ou dupla, linear ou circular, que pode ser segmentado ou não. O material genético de todos os vírus é envolto por um capsídeo, e alguns vírus possuem uma membrana adicional denominada envelope. O tamanho dos genomas virais é bastante variável, podendo conter menos de 2.000 a 2.5 milhões de pares de bases. A quantidade de genes codificados varia igualmente: de dois a 2.500 genes (Aiewsakun & Simmonds, 2018). Baseado nas características dos genomas virais, foi elaborada uma classificação clássica que ficou conhecida como classificação de Baltimore (Baltimore, 1971). Essa classificação divide os vírus em sete classes, de acordo com o tipo de material genético e do método de transmissão da informação genética de uma geração para outra. A Tabela 1 relaciona as classes de Baltimore com os respectivos genomas.

Tabela 1 – Classificação de Baltimore para os genomas virais.

Classe	Genoma
I	DNA dupla fita
II	DNA fita simples positiva
III	RNA dupla fita
IV	RNA fita simples positiva
V	RNA fita simples negativa
VI	RNA fita simples positiva RT
VII	DNA dupla fita RT

Os vírus são formalmente classificados em ordens, famílias, gêneros e espécies pelo *International Committee on Taxonomy of Viruses* (ICTV) (Lefkowitz *et al.*, 2018). A versão atual do ICTV, atualizada em março de 2019, possui 5.560 espécies de vírus

distribuídas entre 1.056 gêneros, 97 subfamílias, 152 famílias e 14 ordens, além de numerosos vírus que ainda não classificados e deverão no futuro compor novos gêneros ou famílias.

Embora os vírus tenham em geral uma pequena quantidade de hospedeiros, eles podem sofrer mutações adaptativas e passar a infectar indivíduos de outras espécies (Flint *et al.*, 2015). Podemos citar o *HIV*, o vírus *Ebola*, o vírus da síndrome respiratória aguda grave (SARS) e o vírus da Zika como exemplos de vírus que foram capazes de cruzar a barreira de espécies e passaram a infectar humanos, causando doenças graves. Além da infecção de outros hospedeiros, as mutações em vírus podem torná-los resistentes a drogas e a anticorpos dos hospedeiros (Peck & Luring, 2018). As taxas de mutação são diferentes entre os tipos de vírus, variando entre 10^{-8} e 10^{-4} substituições por nucleotídeo por infecção celular. Os vírus de RNA sofrem mais mutação que os vírus de DNA e os vírus de fita simples mais mutação que o vírus de fita dupla (Peck & Luring, 2018). A taxa de mutação viral parece ainda se relacionar com o tamanho do genoma: quanto maior o genoma, menor a taxa de mutação (Sanjuán & Domingo-Calap 2016; Peck & Luring, 2018). As mutações virais não são causadas apenas por erros da polimerase, mas também pela capacidade de um vírus corrigir erros de correspondência de DNA por revisão ou reparo pós-replicativo. Outras formas de mutação incluem enzimas do hospedeiro, dano espontâneo de ácido nucléico e elementos genéticos especiais localizados dentro de alguns genomas virais, cuja função específica é produzir novas mutações (Sanjuán & Domingo-Calap 2016).

2.1.1 Bacteriófagos da família *Microviridae*

Os vírus da família *Microviridae* são bacteriófagos de fita simples de DNA (ssDNA) circular, não envelopados, com formato icosaédrico de diâmetro aproximado de 30 nanômetros. Os vírus dessa família são encontrados em diversos ambientes, como diversos ecossistemas aquáticos e no intestino humano (Cherwa & Fane, 2011; Doore & Fane, 2016). Com base em diferenças de estrutura genômica e de estudos filogenéticos, a família *Microviridae* é atualmente dividida em quatro grupos: gênero *Microvirus*, com vírus que infectam enterobactérias, subfamília *Gokushovirinae*, composta por vírus que infectam *Chlamydia*, *Bdellovibrio* e *Spiroplasma*; subfamília *Alpavirinae*, com vírus que infectam bactérias do filo Bacteroidetes (gêneros *Prevotella* e *Bacteroides*), e a

subfamília *Pichovirinae*, com vírus encontrados em ecossistemas aquáticos (Roux *et al.*, 2012).

O genoma dos *Microviridae* apresenta um pequeno número de genes, sendo que três estão presentes em todos os vírus da família e que codificam as seguintes proteínas: a proteína principal de capsídeo (VP1), a proteína piloto de DNA/*minor spike protein* (VP2) e a proteína de iniciação de replicação (VP4) (Roux *et al.*, 2012). Por ser mais conservada entre os indivíduos da família, a proteína VP1 é geralmente utilizada na realização de análises filogenéticas.

Com exceção da subfamília *Alpavirinae*, que possui fagos que se integram no genoma de algumas bactérias (profagos) (Krupovic & Forterre, 2011), não foram evidenciados até o momento profagos de outros vírus da família *Microviridae*, o que sugere que seus ciclos biológicos envolvam partículas virais dentro dos organismos infectados.

2.1.2 Vírus do gênero *Flavivirus*

O gênero *Flavivirus* (família *Flaviviridae*) é composto de vírus esféricos, envelopados, com cerca de 50 nm de diâmetro, contendo um genoma monosegmentado de fita simples positiva de RNA (ssRNA+) linear (Carstens, 2012; Daep *et al.*, 2014; Burrell *et al.*, 2017). O genoma desses vírus possui uma única fase de leitura aberta (*ORF* – *Open Reading Frame*) de aproximadamente 11 kb, a qual é traduzida em um poliproteína. Essa poliproteína, após clivagem, gera 10 proteínas, sendo três delas estruturais (proteínas de capsídeo, de membrana e de envelope), e sete não estruturais (proteína 1, 2A, 2B, 3, 4A, 4B, e 5) (Harris *et al.*, 2006; Lindenbach & Rice, 2003).

O gênero *Flavivirus* compreende mais de 75 vírus distintos, que infectam uma grande variedade de espécies de vertebrados e artrópodes (Burrell *et al.*, 2017). Alguns vírus têm uma quantidade limitada de hospedeiros vertebrados, como primatas, enquanto outros podem infectar e se replicar em uma grande quantidade de espécies. A maioria dos *Flavivirus* utiliza-se de hospedeiros artrópodes hematófagos, os quais funcionam como vetores na sua transmissão para hospedeiros vertebrados. Cerca de 50% dos *Flavivirus* conhecidos são transmitidos por mosquitos, 28% por carrapatos e o restante por roedores ou morcegos sem vetores artrópodes conhecidos (Carstens, 2012).

Mais de 50% dos desses vírus estão associados a doenças humanas, algumas delas de enorme importância no Brasil, como os vírus da Febre Amarela, Dengue e Zika (Burrell

et al., 2017). As doenças causadas por *Flavivirus* podem estar associadas a sintomas do sistema nervoso central (como meningite e encefalite), febre, erupção cutânea e febre hemorrágica (Carstens, 2012). Outros vírus do gênero *Flavivirus* são patogênicos para animais domésticos ou selvagens e causam doenças economicamente importantes.

2.2 Casposons

Os casposons são uma superfamília de elementos genéticos móveis que codificam a enzima endonuclease Cas1, que é a enzima chave do sistema imune-adaptativo CRISPR-Cas (Kuprovic *et al.*, 2014; Kuprovic *et al.*, 2016; Kuprovic *et al.*, 2017). Nos casposons, o gene de *Cas1* está associado a genes que codificam DNA polimerases da família B (*PolB*) e outras proteínas conservadas (Kuprovic *et al.*, 2017), bem como repetições invertidas terminais (*TIR – terminal inverted repeats*) e sítios de duplicação (*TSD – target site duplications*).

O tamanho dos casposons varia de 8 a 20 kb e esses elementos são encontrados em bactérias e arqueias. Com base em genômica comparativa, análise filogenética e distribuição taxonômica, os casposons são classificados em quatro famílias (Kuprovic *et al.*, 2016). Os elementos da família 1 estão presentes em arqueias do filo Thaumarchaeota, enquanto que os elementos das famílias 2 e 3 são encontrados respectivamente em organismos do filo Euryarchaeota (em especial em metanogênicas) e em bactérias. Por fim, a família 4 é a que contém menos elementos e está restrita a certas espécies de *Methanosarcina mazei*.

2.3 Ferramentas existentes para a prospecção e classificação viral

A busca e identificação de vírus conhecidos e emergentes tem se tornado de grande importância. Devido a isso, várias ferramentas para a prospecção de sequências virais em conjuntos de dados genômicos e metagenômicos têm sido desenvolvidas e disponibilizadas publicamente. Em geral, essas ferramentas aceitam como entrada leituras brutas resultantes de processo de sequenciamento ou *contigs* montados e realizam buscas de similaridade contra bases de dados de sequências conhecidas, utilizando ferramentas que executam buscas de similaridade pareada, como o BLAST (Altschul *et al.*, 1990). Em alguns casos, essas ferramentas também executam buscas de similaridade utilizando HMMs de perfil de famílias como o Pfam (El-Gebali *et al.*, 2019) e/ou

domínios proteicos disponíveis como PSSMs (*position-specific substitution matrices*) em bases com o CDD (Marchler-Bauer *et al.* 2009).

Dentre essas ferramentas, podemos citar a plataforma MG-RAST (Meyer *et. al.*, 2008), que é uma das ferramentas mais conhecidas e utilizadas para análise de metagenomas. O MG-RAST é um servidor *web* público para a anotação e análise de metagenomas em geral, que utiliza como entrada leituras brutas ou *contigs* montados e realiza uma série de buscas de similaridades contra bancos de dados de sequências com as ferramentas BLAST e BLAT (Kent, 2002) para a anotação das sequências. A partir dos resultados, são construídos gráficos da diversidade das amostras e árvores filogenéticas. A ferramenta permite ainda a comparação dos resultados encontrados com os resultados de outros metagenomas disponíveis publicamente.

A ferramenta VIROME (Wommack *et al.*, 2012), também um servidor *web*, consiste de um *pipeline* desenvolvido para análise e caracterização de comunidades virais. Essa ferramenta aceita como entrada leituras brutas e *contigs* montados, realiza a análise das sequências de entrada e segue com a predição e anotação das regiões codificantes (ORFs - *Open Reading Frames*). O processo de anotação é realizado por meio de buscas de similaridade utilizando o BLAST contra os bancos de dados de proteínas. Da mesma forma que nas ferramentas anteriores, os resultados são exibidos por meio de gráficos e relatórios.

O Metavir (Roux *et al.*, 2014) é um servidor *web* público desenvolvido para a análise de metagenomas virais. Da mesma forma que o MG-RAST, o Metavir, permite ao usuário informar como entrada tanto leituras brutas como *contigs* montados. Quando os dados de entrada são leituras brutas, a ferramenta identifica a composição taxonômica do viroma através de buscas de similaridade realizadas com o BLAST contra as sequências virais da base RefSeq (O'Leary *et. al.*, 2016). Quando *contigs* montados são informados, a ferramenta primeiro realiza a predição das ORFs e depois executa buscas de similaridades com as ferramentas BLASTp e *hmmsearch* (pacote HMMER) contra as bases RefSeq e Pfam, respectivamente. Em ambos os casos são gerados gráficos da composição taxonômica dos metagenomas e árvores filogenéticas utilizando um conjunto de marcadores virais. O Metavir também permite que o usuário realize a comparação entre metagenomas.

A ferramenta VirFind (Ho & Tzanetakis, 2014) é um programa *online* para a identificação e descoberta de sequências virais em amostras não montadas. Como

primeira etapa, o VirFind processa as leituras de forma a eliminar *primers* e adaptadores, e utiliza o Bowtie2 (Langmead & Salzberg, 2012) para remover leituras pertencentes a hospedeiros virais. As leituras remanescentes são montadas com o Velvet (Zerbino & Birney, 2008) e os contigs resultantes com tamanho mínimo de 90 bp são utilizados em buscas de similaridade com o BLAST contra as bases nt e nr do Genbank. As sequências sem similaridade com essas bases de dados são então traduzidas e submetidas a buscas contra o banco CDD. Ao final, são gerados relatórios com os resultados encontrados e as sequências montadas são salvas em arquivos fasta.

Outra ferramenta disponível, denominada ViromeScan (Rampelli *et. al.*, 2016), foi desenvolvida para se determinar a composição taxonômica da comunidade viral eucariótica dentro de um microbioma, e recebe como entrada apenas leituras brutas. Essa ferramenta utiliza o Bowtie2 para mapeamento das leituras contra genomas virais de eucariotos conhecidos. Ao contrário das demais ferramentas, o ViromeScan está disponível apenas para instalação local.

O programa VIP (Li *et. al.*, 2016) consiste de um *pipeline* para a identificação e descoberta de vírus em amostras metagenômicas, e recebe como entrada apenas leituras brutas. Essas leituras são inicialmente processadas para a remoção de adaptadores, leituras de baixa qualidade e complexidade e leituras oriundas de genomas dos hospedeiros (caso os genomas estejam disponíveis). O VIP pode ser executado de dois modos: rápido e normal. No modo rápido, a ferramenta simplesmente mapeia as leituras contra os bancos de dados de vírus patogênicos ViPR (Pickett *et. al.*, 2011) e IRD (Squires *et. al.*, 2012) com o Bowtie2. No outro modo, as leituras são mapeadas contra as proteínas do RefSeq com a ferramenta RAPSearch (Ye *et. al.*, 2011).

Por fim, o FastViromeExplorer (Tithi *et al.*, 2018) é também um *pipeline*, que foi desenvolvido para a detecção e quantificação de abundância de vírus e fagos em grandes conjuntos de dados metagenômicos. Essa ferramenta executa suas análises em duas etapas: a primeira usa o programa kallisto para mapear leituras curtas contra as sequências do RefSeq. A segunda etapa filtra os resultados encontrados utilizando em critérios mínimos de cobertura. Como resultado são disponibilizadas informações sobre os tipos de vírus e suas abundâncias nas amostras.

As ferramentas descritas acima foram desenvolvidas para a identificação geral de vírus em amostras metagenômicas. Além disso, existe um outro conjunto de ferramentas desenvolvidas especificamente para a identificação de profagos inseridos em genomas.

Entre elas, podemos citar a ferramenta *Phosphinder* (Lima-Mendez *et al.* 2008), *PHAST* (Zhou *et al.*, 2011; Arndt *et al.*, 2016) e *VirSorter* (Roux *et al.*, 2015). O *Phosphinder* é um programa que combina buscas de similaridade, detecção de regiões enriquecidas por genes de fagos e identificação do contexto sintênico dos genes para predição de profagos. Como entrada, essa ferramenta recebe o genoma procarioto anotado em formato GenBank. As CDS preditas do genoma são então utilizadas em buscas de similaridade com o BLAST contra o banco de proteínas de fagos ACLAME (Leplae *et al.*, 2004) para a identificação dos prováveis genes de profagos. Os resultados encontrados são analisados utilizando um conjunto de regras de forma a verificar se existe um profago integrado ao genoma.

O *PHAST* é um servidor web capaz de identificar, anotar e exibir graficamente sequências de profagos dentro de genomas bacterianos. O funcionamento da ferramenta é bastante similar com a anterior, porém essa ferramenta aceita como entrada sequências de DNA anotadas em formato Genbank ou sem anotação. As sequências não anotadas têm seus genes preditos pelo *Glimmer3* (Salzberg *et al.* 1998), que são anotados através de buscas de similaridade utilizando o BLAST contra a base nr. Esses genes são então submetidos a buscas de similaridade também com o BLAST contra o banco de proteínas de fagos criado a partir da integração de dados de vários bancos. Utilizando também um conjunto de regras, a ferramenta define quais sequências de entrada podem conter profagos inseridos em seus genomas. Essa ferramenta foi atualizada em 2016 e passou a se chamar *PHASTER*.

Por fim, a ferramenta *VirSorter* utilizada uma abordagem diferente das anteriores: ao invés de executar buscas de similaridades pareadas, um banco de HMMs de perfil foi construído a partir de sequências proteicas de fagos. Como primeira análise, a ferramenta verifica quais das sequências de entrada são circulares. Com essas sequências, a ferramenta executa a predição dos genes e sequências que possuem mais de dois genes são submetidas a buscas de similaridade utilizando o banco de HMMs de perfil e os modelos e do Pfam. Com os resultados, um conjunto de análises similar ao *Phosphinder* é realizado com o objetivo de verificar se as sequências de entrada possuem profagos.

3 HMMs de perfil e prospecção viral

Como vimos na seção anterior, a maioria das ferramentas desenvolvidas para identificação de sequências virais em conjuntos de dados genômicos e metagenômicos realizam buscas de similaridades pareadas para a identificação da diversidade taxonômica das amostras. Embora muitas das sequências virais conhecidas tenham sido descobertas utilizando essa técnica, essas ferramentas podem não ser capazes de identificar sequências mais divergentes, uma vez que esses métodos falham em identificar sequências com valores de identidade inferiores a 30% (Brenner *et al.*, 1998). O uso de métodos de busca de similaridade baseados em perfis pode ser mais eficiente para a detecção de sequências divergentes, uma vez que esses modelos são capazes de agregar e quantificar as variações existentes em um conjunto de sequências (Krishnamurthy & Wang, 2017; Park *et al.*, 1998). Algumas das ferramentas descritas no capítulo anterior realizam buscas de similaridade contra bases de perfis de famílias/domínios proteínas como o Pfam e CDD. Contudo, estima-se que menos de 20% das famílias de proteínas virais conhecidas estejam representadas no Pfam (Skewes-Cox *et al.*, 2014). Dessa forma, seria ideal o desenvolvimento de metodologias para a criação de modelos capazes de representar a diversidade das famílias virais existentes. A ferramenta o VirSorter utiliza conjuntos de sequências virais para a construção de HMMs de perfil. Porém, essa ferramenta foi somente desenvolvida para a análise de profagos inseridos em genomas bacterianos.

Nesse capítulo faremos uma breve definição sobre HMMs de perfil. Em seguida, descreveremos um exemplo de uso de HMMs de perfil virais para a detecção e reconstrução de sequências virais utilizando a ferramenta GenSeed-HMM (Alves *et al.*, 2016), bem como duas metodologias desenvolvidas para a criação de HMMs de perfil virais.

3.1 HMMs de perfil

HMMs de perfil (Krogh *et al.*, 1994) são modelos probabilísticos lineares capazes de modelar perfis de sequências. A partir de um alinhamento de múltiplas sequências (MSA), um HMM de perfil representa a diversidade de resíduos em cada posição do alinhamento, permitindo inclusive a adição de informação a respeito de inserções e deleções. Para essa modelagem, são utilizados um conjunto de estados, que podem ser de três tipos: estados de correspondência, de inserção e de deleção.

Os estados de correspondência representam a ocorrência de um símbolo em cada posição do alinhamento, bem como modelam as frequências de cada símbolo existente em cada posição. Já os estados de inserção e de deleção servem respectivamente para modelar as inserções e exclusões de símbolos em cada posição. Os estados de inserção, da mesma forma que os estados de correspondência, modelam as frequências dos símbolos inseridos em cada coluna do alinhamento, enquanto que os estados de deleção são silenciosos, pois apenas representam a deleção de um símbolo. A Figura 1 (Reyes *et al.*, 2017) ilustra um exemplo de HMM de perfil. Os retângulos vermelhos na parte inferior da figura representam os estados de correspondência para cada posição do alinhamento com os respectivos valores de probabilidade de cada um dos símbolos possíveis. Os losangos azuis representam os estados de inserção, bem como a probabilidade dos símbolos que podem ser inseridos. Por fim, os círculos verdes modelam os estados de deleção de símbolos em cada coluna do alinhamento. Os estados são conectados uns aos outros por setas que representam as probabilidades de transição de um estado a outro. Como exemplo, começando pelo estado inicial, a probabilidade da primeira coluna do alinhamento de conter uma inserção é de 0,1, uma exclusão é de 0,2 e um símbolo é de 0,7. Sendo a primeira coluna um símbolo, as probabilidades desses símbolos serem os aminoácidos A ou L são, respectivamente, 0,05 e 0,95. Todas as colunas do alinhamento são modeladas dessa forma até o estado final.

Figura 1– Exemplo de um HMM de perfil. Os retângulos em vermelho ilustram os estados de início e fim, que representam o início e final do alinhamento de múltiplas sequências, e os estados de correspondência, que modelam os símbolos existentes em cada coluna do alinhamento com os seus valores de probabilidade. Os losangos azuis e os círculos verdes representam respectivamente as inserções e deleções em cada coluna. As setas representam as probabilidades de transição de um estado para outro (figura extraída de Reyes et al., 2017).

3.2 Bases virais de HMMs de perfil

Existem atualmente algumas bases de HMMs de perfil derivados de sequências virais publicamente disponíveis. Dentre elas, podemos citar as bases vFam (Skewes-Cox *et al.* 2014), viralOGs (Huerta-Cepas *et al.*, 2015) e pVOGs (Grazziotin *et al.*, 2017). O vFam é uma base de HMMs de perfil construída a partir de proteínas virais inteiras disponíveis na base RefSeq. Para a construção dos modelos foi desenvolvido um *pipeline* que obtém as sequências do RefSeq, realiza uma série de processamentos para a remoção de sequências redundantes, agrupamento das mesmas de acordo com seus níveis de similaridade e construção dos modelos. Como resultado, foram gerados 5.585 HMMs de perfil, construídos a partir de 29.655 proteínas. Juntamente com os modelos, a base vFam disponibiliza os arquivos das sequências protéicas utilizadas na construção de cada modelo, bem como os arquivos com a anotação. Vale ressaltar que esse banco não possui modelos para fagos.

Os *viralOGs* são uma parte do *eggNOG* (Huerta-Cepas *et al.*, 2015), uma base de grupos ortólogos que tem sido utilizada há vários anos para a classificação de novas sequências protéicas em grupos ortólogos existentes e anotação transitiva de sua função. A nova versão do *eggNOG* (versão 4.5) incorporou 352 genomas virais, dos quais foram derivadas 5.228 proteínas, que resultaram na construção de 2.605 HMMs de perfil. Além de fornecer os modelos, as listas de grupos ortólogos, alinhamentos múltiplos, árvores filogenéticas e anotações funcionais das proteínas são igualmente disponibilizadas para os usuários.

Os *pVOGs*, por sua vez, são HMMs de perfil construídos a partir de grupos ortólogos de proteínas de fagos que infectam bactérias e arqueias. Esse banco é uma atualização da antiga base *POGs* (*Phage Orthologous Groups*) (Kristensen *et al.*, 2011; Kristensen *et al.*, 2013), e é composto de 9.518 grupos ortólogos, construídos a partir de uma metodologia desenvolvida para a criação de grupos ortólogos microbianos utilizando 296.595 proteínas. Esses grupos foram então utilizados para a construção dos HMMs de perfil. Além dos modelos, a base de dados fornece outros arquivos como as sequências de proteínas utilizadas para a construção dos grupos ortólogos, seus alinhamentos e anotações funcionais.

Apesar dessas bases terem sido construídas com sequências de proteínas virais, elas possuem uma série de limitações, incluindo a pequena representação de sequências na maioria dos modelos, alto viés devido à grande representação de alguns poucos grupos virais e à falta de representação de outros, e a falta de relação direta dos modelos com grupos taxonômicos específicos (Reyes *et al.*, 2017).

3.3 Aplicação de HMMs de perfil na detecção e reconstrução de sequências virais

HMMs de perfil podem potencialmente ser utilizados para a detecção e descoberta de vírus a partir de dados genômicos e metagenômicos. Um exemplo desse tipo de aplicação é o programa *GenSeed-HMM* (Alves *et al.*, 2016), uma ferramenta desenvolvida pelo nosso grupo que se baseia em dois princípios: montagem progressiva como alternativa para a reconstrução de sequências e (2) o uso de HMMs de perfil como sementes para a reconstrução alvo-dirigida. Como validação da ferramenta, o *GenSeed-HMM* foi utilizado para a reconstrução de sequências virais da subfamília *Alpavirinae* (família *Microviridae*) em conjuntos de dados metagenômicos de amostras fecais

humanas e de esgoto bruto. Para tanto, foram construídos HMMs de perfil de *Alpavirinae* a partir da seleção manual de regiões do alinhamento de múltiplas sequências de 83 sequências da proteína principal de capsídeo (VP1) de vírus da família *Microviridae* (Roux *et al.*, 2012) pertencentes às subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae*. A partir da inspeção manual desse alinhamento, foram selecionadas algumas regiões discriminativas para *Alpavirinae* e, utilizando-se a ferramenta `hmmbuilder` do pacote HMMER (Eddy, 2011), HMMs de perfil para essas regiões foram construídos. Esses modelos foram então utilizados como semente nas montagens com o GenSeed-HMM nos metagenomas. Como resultado, dezenas de genomas completos e algumas centenas de incompletos, todos inéditos, foram montados. Esses resultados demonstraram a viabilidade de utilização de HMMs de perfil para a detecção de sequências virais conhecidas e inéditas, bem como para a reconstrução dessas sequências utilizando esses modelos como sementes em montagem progressiva.

3.4 Metodologias para a classificação viral que utilizam HMMs de perfil

Aiweesakun e Simmonds (2018) desenvolveram um *pipeline* com o objetivo inicial de avaliar a relação entre conteúdo genômico das sequências virais e a classificação taxonômica atribuída pelo ICTV, que recebeu o nome de GRAViTy (*Genome Relationships Applied to Virus Taxonomy*). O GRAViTy é composto de seis sub-*pipelines*, um para cada classe de Baltimore (as classes VI e VII foram consideradas como um único grupo), que utiliza HMMs de perfil construídos com proteínas virais de sequências da classe e modelos de organização genômica, definidos como GOMs (*Genomic Organization Models*), para a caracterização das sequências do respectivo grupo. Para a construção dos modelos, foram utilizados 3.854 genomas virais de eucariotos pertencentes às sete classes de Baltimore, compreendendo 12 ordens, 103 famílias e 472 gêneros. Os HMMs de perfil de cada uma das classes foram construídos com as proteínas extraídas dos genomas selecionadas de acordo com um conjunto de regras e posteriormente agrupadas com o algoritmo MCL (Enright *et al.*, 2002). Os grupos gerados foram então alinhados com a ferramenta MUSCLE (Edgar, 2004) e os alinhamentos utilizados para a criação dos HMMs de perfil com a programa `hmmbuild` (pacote HMMER). Esses modelos foram definidos como PPHMMs (*Protein Profile Hidden Markov Models*). Para o banco de GOMs, foram construídas tabelas com a localização

das regiões codificantes de proteínas de cada genoma da classe. Os níveis de relação entre as sequências virais foram definidos mediante o uso de análises de distância de Jaccard calculadas de acordo com os resultados obtidos pelos PPHMMs e GOM. Foi observada no trabalho uma forte relação entre a taxonomia e a composição genômica dos vírus para a maioria das famílias descritas no ICTV, embora ainda tenham sido encontrados alguns problemas de classificação. Apesar do objetivo do primeiro trabalho ter sido apenas o de mostrar a relação entre a classificação taxonômica do ICTV e o conteúdo genômico das sequências, os autores tornaram a ferramenta pública.

Recentemente foi publicada por Chibani *et al.* (2019A) uma metodologia para a classificação de profagos baseada em HMMs de perfil, denominado *ClassiPhage*. Os autores desenvolveram um *pipeline* que constrói HMMs de perfil específicos para famílias de fagos utilizando genomas anotados e emprega protocolos de refinamento para garantir a especificidade dos modelos. O primeiro passo do *pipeline* consiste na obtenção dos genomas de fagos de cada uma das famílias selecionadas do NCBI em formato Genbank. As proteínas de cada família são extraídas dos respectivos genomas e posteriormente agrupadas com o algoritmo MCL (Enright *et al.*, 2002). As sequências redundantes de cada grupo são então removidas com o programa CD-HIT (Li & Godzik, 2006) e as remanescentes alinhadas e utilizadas para a construção dos HMMs de perfil com programa *hmmbuild*. Por fim, os modelos construídos passam por um protocolo de refinamento com o objetivo de verificar se novas sequências podem ser adicionadas aos modelos e se sequências divergentes precisam ser removidas dos mesmos.

Para a validação dessa metodologia, foram construídos modelos para quatro famílias de fagos que infectam Vibriófagos: *Inoviridae*, *Myoviridae*, *Podoviridae* e *Siphoviridae*. Como resultado, foram gerados nove HMMs de perfil para a família *Inoviridae*, 242 para *Myoviridae*, 96 para *Podoviridae* e 54 para *Siphoviridae*. Os autores fizeram vários testes de validação e observaram que múltiplos modelos eram necessários para se obter uma classificação taxonômica confiável dos diferentes grupos taxonômicos. A segunda versão dessa ferramenta, denominada *ClassiPhages2.0* (Chibani *et al.* 2019B), descreve a utilização de uma rede neural para a classificação de fagos. Nesse trabalho foram construídos 5.920 HMMs de perfil derivados de 7.342 genomas de fagos pertencentes a 12 famílias (*Cystoviridae*, *Fuselloviridae*, *Haloviruses*, *Inoviridae*, *Leviviridae*, *Ligamenvirales*, *Microviridae*, *Myoviridae*, *Pleolipoviridae*, *Podoviridae*, *Siphoviridae* e *Tectiviridae*) aplicando-se a mesma metodologia do trabalho anterior. Os

modelos foram utilizados para analisar os 7.342 genomas de fagos e os resultados obtidos aplicados para treinar a rede neural, que foi então utilizada para classificar as sequências. Embora a acurácia estimada tenha sido alta, segundo os autores esse resultado foi fortemente influenciado pelos altos valores de especificidade, uma vez que a sensibilidade da predição foi inferior a 50% em nove das 12 famílias. Os autores destacaram ainda a presença de falsos positivos nas famílias *Microviridae*, *Myoviridae*, *Podoviridae* e *Siphoviridae*, e sugeriram que esse fato pode indicar que tenha havido trocas de genes entre espécies de fagos não relacionados. A ferramenta até o presente momento ainda não está publicamente disponível.

4 Objetivos

Em função do atual estado do conhecimento, descrito nos itens anteriores, o objetivo do presente trabalho foi o desenvolvimento de metodologias para o desenho racional de HMMs de perfil a partir de alinhamentos de múltiplas sequências e sua aplicação na prospecção de dados de sequenciamento de genomas e metagenomas. Embora a motivação inicial do trabalho tenha sido o desenvolvimento de abordagens metodológicas para a representação, classificação e prospecção de sequências virais, as ferramentas foram desenvolvidas para uso genérico, permitindo sua utilização em sequências de quaisquer organismos.

4.1 Geral

Desenvolver metodologias para o desenho racional de HMMs de perfil que sejam altamente sensíveis e específicos para grupos de sequências de interesse e sua aplicação na prospecção de dados de sequenciamento de genomas e metagenomas.

4.2 Específicos

- Desenvolver e implementar uma metodologia para o desenho de HMMs de perfil capazes de detectar sequências de um grupo com alta sensibilidade ou discriminar subgrupos de sequências com alta sensibilidade e especificidade;
- Desenvolver e implementar um método para a prospecção de dados de sequenciamento com HMMs de perfil;
- Desenvolver e implementar uma abordagem metodológica para o uso de HMMs de perfil na detecção de elementos contendo múltiplos genes em contexto sintênico específico;
- Estabelecer protocolos para a integração das ferramentas desenvolvidas nesse trabalho.

5 TABAJARA

Conforme discutido anteriormente, apesar dos HMMs de perfil virais disponíveis publicamente terem sido construídos especificamente para vírus, eles possuem um conjunto de limitações, principalmente no que se relaciona à falta de relação com grupos taxonômicos específicos. Os métodos desenvolvidos para a construção de HMMs de perfil GRAVITY e ClassiPhage se preocupam com a manutenção da relação taxonômica dos modelos, porém o simples agrupamento das sequências de uma mesma família com a remoção de sequências redundantes não garante que os modelos construídos sejam altamente sensíveis e específicos, mesmo que alguns protocolos de refinamento sejam aplicados.

Embora HMMs de perfil possam potencialmente ser excelentes modelos para a representação da diversidade de um conjunto de sequências, é preciso que sejam desenvolvidas abordagens apropriadas para que esses modelos sejam suficientemente sensíveis e específicos para a detecção e classificação de um grupo de sequências. Em função disso, nosso grupo decidiu desenvolver uma metodologia própria para o desenho racional de HMMs de perfil a partir de um conjunto de sequências. Essa metodologia foi implementada na ferramenta TABAJARA (*Tool for Alignment Block Analysis Joining Appropriate Rational Approaches*). Nesse capítulo serão descritos os experimentos executados para o desenvolvimento da metodologia, bem como a implementação e validação do TABAJARA utilizando diferentes conjuntos virais.

Tanto nos experimentos realizados durante o desenvolvimento da metodologia como na validação do TABAJARA, foram utilizadas sequências de dois grupos virais: gênero *Flavivirus* e família *Microviridae*. Esses dois grupos foram escolhidos por representarem vírus de eucariotos e procariotos, respectivamente, e apresentarem graus de divergência muito diferentes entre as suas sequências. Enquanto as sequências do gênero *Flavivirus* são bastante similares entre si e permitem uma clara distinção entre as diferentes espécies de cada grupo (similaridade média entre as suas sequências de 60,5%¹), as sequências de *Microviridae* são bastante divergentes entre si, inclusive entre membros de uma mesma subfamília (similaridade média entre as sequências de

¹ Similaridade entre cada par de sequências calculada com programa Needle do pacote EMBOSS (Rice *et al.* 2000).

Alpavirinae, *Gokushovirinae* e *Pichovirinae* de 37,0%, 59,6% e 54%, respectivamente), o que torna a construção de modelos discriminativos mais desafiadora.

Para a criação dos modelos de *Flavivirus*, foi utilizado um conjunto de treinamento composto de 127 sequências de poliproteínas do gênero *Flavivirus*, contendo 20 sequências de *Dengue virus* pertencentes aos quatro sorotipos, 10 de *Zika virus*, 10 sequências de Febre Amarela e 87 pertencentes à 63 espécies desse gênero. Já o conjunto de teste desse gênero era composto de 6.364 sequências de poliproteínas de *Flavivirus* de diferentes espécies, sendo 3.919 sequências de Dengue, 327 de Zika e 63 de Febre Amarela. Tanto as sequências do conjunto de treinamento como as do conjunto de teste foram obtidas do NCBI.

Para a construção dos modelos de *Microviridae*, foi utilizado um conjunto de treinamento composto de 113 sequências da proteína principal de capsídeo (VP1) de *Microviridae*, compreendendo as subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae* (33, 44 e 36 sequências, respectivamente), que une 83 sequências de VP1 de *Microviridae* descritas em Roux *et al.* (2012) e 30 sequências de *Pichovirinae* gentilmente cedidas pelo Prof. François Enault (Université Clermont Auvergne, França). O conjunto de teste dos *Microviridae* foi composto de 1.836 sequências de VP1 de *Microviridae*, contendo 501 sequências da subfamília *Alpavirinae*, 1.040 de *Gokushovirinae* e 295 sequências de *Pichovirinae*, todas gentilmente cedidas pelo Prof. François Enault (Université Clermont Auvergne, França).

5.1 Uso de proteínas inteiras para a construção dos HMMs de perfil

Como discutido anteriormente, a maioria dos HMMs de perfil virais disponíveis nas bases públicas (vFAMs, viralOGs e pVOGs) foram construídos a partir de sequências protéicas inteiras, utilizando-se grupos ortólogos, sem a preocupação das relações taxonômicas das sequências. Assim, decidimos em uma primeira abordagem avaliar a utilização de proteínas virais inteiras para a detecção e classificação de grupos taxonômicos virais. Para isso, foi inicialmente construído um HMM de perfil específico para a espécie Dengue vírus (gênero *Flavivirus*), a partir da seleção das 20 sequências de poliproteínas inteiras de Dengue presentes no conjunto de treinamento dos *Flavivirus*. Essas sequências foram alinhadas com o programa MUSCLE (Edgar, 2004) e o alinhamento múltiplo resultante utilizado para a construção de um HMM de perfil com a ferramenta `hmmbuild` do pacote HMMER (Eddy, 2011).

A fim de testar o potencial de detecção e discriminação do modelo construído, foi realizada uma busca de similaridade com o programa *hmmsearch*, também do pacote HMMER, do modelo construído para Dengue contra o conjunto de teste dos *Flavivirus* (composto de 6.364 poliproteínas desse gênero), utilizando os parâmetros padrão da ferramenta (*e-value* 10). Como resultado, todas as sequências do conjunto de teste foram detectadas pelo modelo (Figura 2A). Esse resultado mostra que, embora esse modelo tenha sido construído com poucas sequências de Dengue, ele foi capaz de detectar toda a diversidade de sequências conhecidas para essa espécie (100% de sensibilidade ou *recall*). Contudo, sequências de todas as outras espécies de *Flavivirus* presentes no conjunto de teste também foram detectadas, mostrando que o modelo construído não possui capacidade discriminativa entre diferentes espécies de vírus do gênero. Esse resultado pode ser atribuído a regiões da poliproteína que são conservadas em todos os representantes do gênero *Flavivirus* e, portanto, levam o modelo a detectar sequências não relacionadas a *Dengue virus*.

Figura 2 – Gráficos de dispersão com *jittering* representando os resultados de buscas de similaridade de HMMs de perfil construídos a partir de sequências inteiras de poliproteínas de *Flavivirus*. Modelos construídos para os vírus da dengue (A), Zika (B) e Febre Amarela (C) foram utilizados em buscas de similaridade contra o conjunto de teste dos *Flavivirus* e os pontos representam os valores de escores obtidos pelas sequências detectadas por cada um dos modelos. As linhas horizontais tracejadas referem-se aos valores de corte de escore correspondentes a 70, 80 e 90% do valor de escore mais baixo observado para o respectivo grupo no conjunto de treinamento.

Como próximo passo, resolvemos então inspecionar os valores de escore de alinhamento obtidos pelo *hmmsearch* nas sequências detectadas pelo modelo. Como podemos ver na Figura 2A, as sequências de Dengue detectadas pelo modelo apresentaram valores de escore bastante superiores aos obtidos pelas demais sequências. Assim, uma maneira relativamente simples de discriminar as sequências dessa espécie das demais seria o estabelecimento de um valor de corte de escore adequado ao modelo. A fim de avaliarmos a atribuição de diferentes valores de corte ao modelo, com o objetivo

de torna-lo discriminativo, foi realizada então uma busca de similaridade do modelo construído para Dengue contra o conjunto de treinamento dos *Flavivirus*, composto de 127 sequências e que contém as 20 sequências de Dengue utilizadas na construção do modelo. Com base nos resultados, foram calculados três valores de corte de escore, equivalentes a 70, 80 e 90% do valor de escore mais baixo observado para *Dengue virus* no conjunto. Aplicando esses valores de corte aos resultados das buscas de similaridade obtidos contra o conjunto das 6.364 sequências de *Flavivirus*, podemos observar que o uso dos valores de corte de 80% e 90% (respectivamente linhas verde e azul no gráfico) tornam o modelo discriminativo para *Dengue virus* (Figura 2A). Esses experimentos foram repetidos com os vírus da Zika e Febre Amarela e os resultados obtidos foram bastante similares (Figuras 2B e 2C). Esse experimento demonstrou que é possível se obter um modelo discriminativo para algumas espécies do gênero *Flavivirus* utilizando-se como conjunto de treinamento um alinhamento múltiplo de proteínas inteiras, desde que um valor de corte de escore adequado seja utilizado.

Como as sequências de poliproteína do gênero *Flavivirus* são bastante similares entre si, sem regiões de lacunas (*gaps*), resolvemos verificar se essa metodologia poderia ser também aplicada a conjuntos de sequências mais divergentes para a construção de modelos discriminativos. Dessa forma, esse experimento foi repetido para a família *Microviridae*, utilizando-se sequências da proteína principal de capsídeo (VP1), que são bastante divergentes entre si. Nesse caso, foi construído um HMM de perfil para a subfamília *Alpavirinae*, utilizando-se 33 sequências de VP1 dessa subfamília contidas no conjunto de treinamento com o mesmo protocolo anterior. O modelo construído foi utilizado em uma busca de similaridade contra o conjunto de teste dos *Microviridae* (composto por 1.836 sequências de VP1 de *Microviridae*). De forma similar aos resultados obtidos com as espécies do gênero *Flavivirus*, o modelo construído foi capaz de detectar todas as sequências do conjunto de treinamento (Figura 3A). Entretanto, a partir da análise dos escores das sequências detectadas pelo modelo, foi possível se observar que não há uma clara separação entre as sequências de *Alpavirinae* das demais sequências de *Microviridae*, diferentemente do que ocorreu com os *Flavivirus*. Como pode ser visto na Figura 3A, a faixa de valores de escores das sequências de *Alpavirinae* detectadas pelo modelo é bastante ampla e se sobrepõe aos valores de escores obtidos por sequências das subfamílias *Gokushovirinae* e *Pichovirinae*.

Apesar dessa sobreposição, foram calculados valores de corte de escore para o modelo com base no conjunto de treinamento, utilizando-se o protocolo anterior. Dessa

forma, o modelo construído foi utilizado em uma busca de similaridade contra o conjunto de treinamento de *Microviridae* e valores de corte de escore correspondentes a 70, 80 e 90% do valor de escore mais baixo observado para *Alpavirinae* foram calculados. Como pode ser visto na Figura 3A, nenhum valor de corte definido foi capaz de separar completamente as sequências de *Alpavirinae* das demais. No entanto, um valor de escore de 90% poderia ser utilizado para tornar o modelo completamente discriminativo para *Alpavirinae*, com capacidade de detecção de 67% das sequências de *Alpavirinae* do conjunto de teste (Tabela 2). Esse mesmo experimento foi realizado utilizando sequências da subfamília *Gokushovirinae* e *Pichovirinae* (Figuras 3B e 3C, respectivamente). No caso da subfamília *Gokushovirinae*, observou-se uma porcentagem de detecção maior que aquela apresentada pelo modelo de *Alpavirinae*, uma vez que a divergência entre as sequências de *Gokushovirinae* é menor que nas sequências de *Alpavirinae* (Figura 3B). Em relação à subfamília *Pichovirinae*, nenhum dos três valores de corte tornou o modelo completamente discriminativo para essa subfamília. O uso do valor de escore de 90% fez com que o modelo apresentasse 90% de sensibilidade e 57% de especificidade.

Figura 3 – Gráficos de dispersão com *jittering* representando os resultados de buscas de similaridade de HMMs de perfil construídos a partir de sequências de proteínas inteiras de VP1 de *Microviridae*. Modelos construídos para *Alpavirinae* (A), *Gokushovirinae* (B) e *Pichovirinae* (C) foram utilizados em buscas de similaridade contra o conjunto de teste dos *Microviridae* e os pontos representam os valores de escores obtidos pelas sequências detectadas por cada um dos modelos. As linhas horizontais tracejadas ilustram os valores de corte de escore correspondentes a 70, 80 e 90% do valor de escore mais baixo observado para o respectivo grupo no conjunto de treinamento.

Com base nos experimentos realizados, podemos concluir que proteínas inteiras geram HMMs de perfil que não têm, em princípio, capacidade discriminativa. Contudo, utilizando-se valores de corte adequados, é possível obter um determinado potencial de discriminação. Um ponto importante que deve ser observado é que, quando as sequências do grupo em análise são altamente divergentes entre si, dentro do próprio grupo, como as sequências das subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae*, a aplicação do

escore de corte pode reduzir a sensibilidade do modelo. A depender do grau de divergência, o modelo pode ser altamente específico (Figura 3A) ou não (Figura 3C).

Tabela 2 – Sensibilidade e especificidade de HMMs de perfil construídos com os alinhamentos de proteínas inteiras de VP1 de *Microviridae*. Os modelos foram construídos para *Alpavirinae*, *Gokushovirinae* e *Pichovirinae* a partir do alinhamento de proteínas. Os três modelos foram utilizados em buscas de similaridade contra o conjunto de teste dos *Microviridae* sem e com a aplicação dos valores de corte calculados a partir do conjunto de treinamento.

Valores de corte	<i>Alpavirinae</i>		<i>Gokushovirinae</i>		<i>Pichovirinae</i>	
	Sensib.	Especif.	Sensib.	Especif.	Sensib.	Especif.
Sem corte	1,00	0,00	1,00	0,00	1,00	0,00
70%	0,81	0,95	0,98	0,93	1,00	0,34
80%	0,71	0,99	0,97	1,00	0,99	0,38
90%	0,67	1,00	0,92	1,00	0,98	0,57

5.2 Uso de blocos de alinhamento curtos para a construção de HMMs de perfil

No experimento anterior se observou que proteínas inteiras podem ser utilizadas para a construção de HMMs de perfil discriminativos para grupos virais específicos, desde que valores de corte de escore adequados sejam atribuídos aos modelos. No entanto, quando desejamos utilizar esses modelos para a identificação de sequências em conjuntos de dados compostos por sequências curtas, como, por exemplo, dados brutos (não montados) de metagenomas, esse tipo de abordagem não pode ser utilizada. Como as leituras da grande maioria das plataformas são relativamente curtas, os blocos de alinhamentos obtidos também seriam curtos e, portanto, os valores de corte estabelecidos para sequências inteiras da proteína seriam excessivamente estridentes para a sua detecção.

Uma alternativa ao uso de sequências proteicas inteiras seria o desenvolvimento de uma metodologia capaz de identificar blocos de alinhamento curtos e discriminativos a partir de um alinhamento de múltiplas sequências. Com base nessa premissa, decidiu-se realizar um segundo experimento com o objetivo de se avaliar se HMMs de perfil construídos a partir de blocos de alinhamento curtos de um alinhamento múltiplo sem nenhum critério de seleção, cobrindo todo o comprimento do alinhamento, teriam a capacidade de detectar de forma sensível e específica um determinado subgrupo de

sequências. Em caso positivo, objetivou-se ainda determinar qual seria a faixa de tamanho ideal para a utilização desses blocos.

Assim como no experimento anterior, foram utilizadas como conjunto de treinamento as 20 sequências de poliproteínas de *Dengue virus* alinhadas com o programa MUSCLE. Para isso, foi construído um *script* que utiliza uma janela deslizante de tamanho n para selecionar regiões contíguas do alinhamento múltiplo de entrada, caminhando uma posição (base) por vez. Para o teste, foram utilizadas janelas de diferentes tamanhos, variando entre 10 a 110 posições, com intervalos de 10. Esses valores foram escolhidos de forma arbitrária. Cada região selecionada pela janela deslizante foi então extraída e usada para a construção de um HMM de perfil.

Todos os modelos gerados foram usados em buscas de similaridade com o programa *hmmsearch* contra as 6.364 sequências de *Flavivirus* do conjunto de teste, e valores de sensibilidade e especificidade foram determinados para cada um. Esses valores foram então agrupados de acordo com o tamanho da janela usada para a seleção dos respectivos blocos de alinhamento, e seus valores médios foram calculados. Como pode ser visto na Figura 4A (linhas verde e lilás), modelos de *Dengue virus* construídos com blocos de tamanho curto apresentaram os melhores valores sensibilidade e especificidade. Com um comprimento de 10 bases, os blocos apresentaram valores médios de 0,9 e 0,84, respectivamente. A partir de um comprimento de 20, os valores de especificidade vão se tornando muito baixos. Esse experimento foi repetido para os vírus da Zika (Figura 4B) e Febre Amarela (Figura 4C), utilizando-se também sequências dessas espécies presentes no conjunto de treinamento dos *Flavivirus*. Os resultados obtidos foram muito similares, sugerindo que, pelo menos dentro do modelo de *Flavivirus*, o uso de regiões curtas de até dez resíduos de aminoácidos permite gerar modelos que têm um desempenho relativamente razoável de detecção e discriminação, enquanto regiões mais longas implicam numa redução extremamente acentuada da especificidade.

Em função desse resultado, decidimos testar se a aplicação de valores de corte de escore (*cutoff scores*) aos modelos poderia melhorar a especificidade dos mesmos. Para isso, foi implementada no *script* anterior uma abordagem heurística para a atribuição de valores de corte personalizados para cada HMM de perfil. Para cada um dos modelos gerados, o *script* passou a executar uma busca de similaridade contra as 127 sequências do conjunto de treinamento dos *Flavivirus* e as sequências detectadas foram listados em ordem decrescente do escore de alinhamento. Com base nos resultados, foi implementada uma rotina que atribui um valor de corte ao modelo, arbitrariamente calculado em 80%

do valor do menor escore observado pelas sequências do grupo de interesse (nesse caso, pelas sequências de *Dengue virus*). O experimento foi então repetido com essa nova versão do *script*. Conforme podemos ver na Figura 4A (linha azul), o uso de valores de corte melhorou dramaticamente os valores de especificidades para os modelos construídos, independente do tamanho da janela utilizado.

Figura 4. Avaliação de um repertório completo de blocos de alinhamento de tamanho variável como candidatos para a construção de HMMs de perfil discriminativos para o gênero *Flavivirus*. Janelas deslizantes de diferentes tamanhos, começando com um comprimento mínimo de 10 aa e variando em incrementos de 10 aa cada, foram utilizadas em alinhamentos de múltiplas sequências de poliproteínas inteiras específicas para *Dengue virus* (A), *Zika virus* (B) e *Yellow Fever virus* (C). Blocos de alinhamento representando todo o repertório obtido para cada tamanho de janela foram usados para construir os HMMs de perfil. Todos os modelos foram testados com o programa *hmmsearch* contra um conjunto de dados de 6.364 sequências de *Flavivirus* compreendendo uma grande variedade de espécies virais. As linhas de sensibilidade e especificidade, usando (+ CS) ou não (-CS) pontos de corte, representam médias aritméticas calculadas a partir da combinação dos múltiplos modelos gerados para cada tamanho de bloco de alinhamento. As áreas sombreadas indicam valores de desvio padrão. Os escores de corte correspondem a 80% do valor de escore mais baixo observado por cada modelo individual em relação ao conjunto de treinamento correspondente.

O uso de valores de corte nos modelos foi, portanto, capaz de torná-los mais específicos para a identificação de sequências de *Dengue virus*, sem a alteração da sensibilidade dos mesmos. Esses experimentos foram realizados novamente com as sequências de *Zika virus* (Figura 4B) e *Yellow Fever virus* (Figura 4C) e os resultados foram bastante similares.

Com o objetivo de verificarmos se modelos construídos a partir de regiões curtas do alinhamento de grupos de sequências mais divergentes entre si apresentavam resultados similares aos encontrados para as três espécies de *Flavivirus*, esse experimento foi repetido utilizando sequências da família *Microviridae*. Nesse caso, o *script* foi novamente executado, utilizando agora como entrada o alinhamento múltiplo das 33

sequências de VP1 de *Alpavirinae* (descritas por Roux *et al.*, 2012) contidas no conjunto de treinamento dos *Microviridae*. Para o cálculo dos valores de corte dos modelos, foram utilizadas as 113 sequências de VP1 de *Microviridae* do conjunto de treinamento.

Todos os modelos gerados foram submetidos a buscas de similaridade contra o conjunto de teste dos *Microviridae*, composto de 1.836 sequências de VP1, sem e com a aplicação dos valores de corte. Da mesma forma que no experimento com os *Flavivirus*, valores de sensibilidade e especificidade dos modelos foram calculados para cada modelo e os valores médios dessas métricas calculados para cada tamanho de janela. Como podemos observar na Figura 5A, a aplicação de valores de corte de escore aos modelos gerados para *Alpavirinae* também foi capaz de torná-los mais específicos para essa subfamília. Contudo, como esse grupo viral é bem mais diverso que os anteriores, os valores médios de sensibilidade foram baixos (inferiores a 0,4 para qualquer tamanho de janela), e afetados pelo uso dos valores de corte de escore.

Esses experimentos foram repetidos para a subfamília *Gokushovirinae* e *Pichovirinae* e resultados similares foram obtidos (Figura 5B e 5C, respectivamente). Como podemos observar nos gráficos, os valores de sensibilidade médios para as subfamílias *Gokushovirinae* e *Pichovirinae* foram maiores em relação à subfamília *Alpavirinae*, porém esses valores foram igualmente afetados pelo uso dos valores de corte dos modelos. Assim, conclui-se que a aplicação de valores de corte de escore para uso dos HMMs de perfis tornou-os de modo geral muito mais específicos, tanto no modelo de *Flavivirus* como em *Microviridae*. Entretanto, o uso de um repertório de modelos, derivados de todas as possibilidades de blocos de alinhamento de diferentes tamanhos, demonstrou que somente uma fração desses modelos é discriminativa, especialmente aqueles modelos construídos a partir de sequências curtas. Portanto, o uso de modelos curtos, sem processo de escolha criterioso, não é suficiente para permitir seu uso geral.

Do ponto de vista conceitual isso se explica porque nem todas as regiões do alinhamento são boas para a construção de modelos altamente específicos e relativamente sensíveis. De fato, algumas regiões podem ser conservadas em outros grupos virais e não serem, portanto, taxonomicamente discriminativas. Outras regiões podem conter grande quantidade de lacunas no alinhamento ou, ainda, não serem representarem toda a diversidade do grupo viral em questão. O conjunto desses resultados nos estimulou a desenvolver uma metodologia racional, capaz de identificar em um alinhamento de múltiplas sequências, as regiões mais conservadas entre todas as sequências ou, alternativamente, aqueles mais discriminativas para um subconjunto de sequências.

Figura 5 – Avaliação de um repertório completo de blocos de alinhamento de tamanho variável como candidatos para a construção de HMMs de perfil discriminativos para a família *Microviridae*. Janelas deslizantes de diferentes tamanhos, começando com um comprimento mínimo de 10 aa e variando em incrementos de 10 aa cada, foram utilizadas em alinhamentos de múltiplas sequências da proteína principal de capsídeo (VP1) de *Alpavirinae* (A), *Gokushovirinae* (B) e *Pichovirinae* (C). Blocos de alinhamento representando todo o repertório obtido para cada tamanho de janela foram usados para construir os HMMs de perfil. Todos os modelos foram testados com o programa *hmmsearch* contra um conjunto de dados de 1.836 sequências da proteína VP1 de *Microviridae*. As linhas de sensibilidade e especificidade, usando (+CS) ou não (-CS) pontos de corte, representam médias aritméticas calculadas a partir da combinação dos múltiplos modelos gerados para cada tamanho de bloco de alinhamento. As áreas sombreadas indicam valores de desvio padrão. Os escores de corte correspondem a 80% do valor de escore mais baixo observado por cada modelo individual em relação ao conjunto de treinamento correspondente.

5.3 Desenvolvimento da metodologia

A metodologia desenvolvida consiste na utilização de um conjunto de algoritmos implementados em uma ferramenta denominada TABAJARA (*Tool for Alignment Block Analysis Joining Appropriate Rational Approaches*). Começando de um alinhamento de múltiplas sequências, o TABAJARA pode utilizar todo o comprimento das sequências presentes no alinhamento ou selecionar sub-regiões específicas para a construção dos modelos. Em ambos os casos, a ferramenta pode ser executada em dois modos: (1) Conservação, para a construção de modelos capazes de representar todas as sequências do alinhamento, ou (2) Discriminação, que constrói modelos apropriados para a distinção de um grupo específico de sequências das demais presentes no alinhamento. Nas próximas seções serão descritos os conceitos computacionais aplicados no desenvolvimento da ferramenta, bem como a sua implementação.

5.3.1 Conceitos de teoria da informação

Para a identificação de sub-regiões específicas do alinhamento de entrada, a primeira ação do programa consiste na atribuição de valores de escore posição-específicas, variando entre 0 a 1. Esses valores podem se referir aos graus mínimos ou máximos, respectivamente, de conservação ou de discriminação, dependendo do modo de execução do programa. Quando o programa TABAJARA é executado em modo Conservação e o conjunto de sequências de entrada é composto por sequências protéicas, o método utilizado para calcular os escores é a divergência de Jensen-Shannon (JSD), conforme implementado por Capra & Singh (2007). Esse método quantifica a similaridade entre duas distribuições de probabilidade. No caso da implementação de Capra & Singh (2007), uma das distribuições consiste da distribuição de fundo dos aminoácidos estimada para a construção da matriz BLOSUM62, que é utilizada para aproximar a distribuição de sítios de aminoácidos sujeitos a nenhuma pressão evolutiva. Então, posições que possuem distribuições de aminoácidos muito diferentes da distribuição de fundo são, provavelmente, funcionalmente importantes ou restringidas pela evolução. Assim, para cada coluna c do alinhamento, a divergência de Jensen-Shannon D_c^{JS} é definida da seguinte forma:

$$D_c^{JS} = \lambda RE_{p_{c,r}} + (1 - \lambda) RE_{(q,r)} \quad (1)$$

onde RE é a entropia relativa ou distância de Kullback–Leibler (Kullback & Leibler, 1951), p_c consiste da distribuição de aminoácidos na coluna c , $r = \lambda p_c + (1 - \lambda)q$, q é a distribuição de fundo dos aminoácidos e λ é um peso a priori (com valor de $\frac{1}{2}$, definido por Capra & Sing, 2007 como sendo o valor de λ que obteve os melhores resultados).

Quando o TABAJARA é executado no modo Conservação utilizando sequências nucleotídicas, o método da entropia de Shannon (Shannon, 1948) é utilizado para calcular as pontuações de cada coluna do alinhamento. Thomas & Cover (2006) descrevem a entropia de Shannon como sendo a medida da incerteza de uma variável aleatória. Em outras palavras, a entropia de Shannon reflete a quantidade de informação necessária para descrever uma variável. Aplicando-se ao nosso problema, quanto mais variável é uma coluna do alinhamento, maior é a sua entropia, uma vez que maior é a incerteza em relação

a sua conservação. Em termos formais, a entropia de Shannon E_c para cada coluna c de um MSA, é calculada da seguinte forma:

$$E_c = - \sum_{m=\{A,C,G,T\}} p_{c_m} \log_b p_{c_m} \quad (2)$$

onde p_{c_m} é a probabilidade de ocorrência de uma base de DNA m e b é a base do logaritmo (nesse caso, o valor $b = 2$ foi utilizado). Como nesse caso o MSA é composto de sequências nucleotídicas, 5 símbolos (ou estados) podem estar presentes em cada coluna: A, C, G, T, '-'. Assim, os valores de entropia mínimo e máximo são, respectivamente zero, quando apenas um tipo de nucleotídeo está presente na coluna (não há variação de símbolos), e aproximadamente 2,32 (entropia máxima), quando os 5 símbolos aparecem com a mesma frequência. Como os valores de escore de cada coluna devem variar entre 0 e 1, os escores obtidos são normalizados a partir da divisão desses valores pelo valor de entropia máxima. Como conceituamos anteriormente, os valores de escore variam entre 0 e 1, onde 0 significa nenhuma conservação, e 1 máxima conservação. Porém, após a normalização, os valores obtidos indicam o contrário: 0 indica máxima conservação e 1 nenhuma conservação. Dessa forma, para que os valores variassem da forma que queríamos, os valores normalizados de entropia são então subtraídos de 1. Assim, podemos dizer que o escore de conservação SC_c de cada coluna c do MSA composto de sequências nucleotídicas é calculado da seguinte maneira:

$$SC_c = 1 - E_c/E_{max} \quad (3)$$

Para medir a habilidade discriminativa das colunas do alinhamento múltiplo de entrada entre dois grupos de sequências, o TABAJARA utiliza duas métricas: Informação mútua (MI – *Mutual Information*) (Thomas & Cover, 2006) e Desarmonia de Sequências (SD – *Sequence Disharmony*), tanto para sequências nucleotídicas como para sequências protéicas. De acordo com Thomas & Cover (2006), a informação mútua é a medida da dependência entre duas variáveis aleatórias. Em outras palavras, essa métrica reflete a associação estatística existente entre duas variáveis. No caso do TABAJARA, a informação mútua é utilizada para verificar, em cada coluna do alinhamento múltiplo, como um determinado conjunto de símbolos (composto por nucleotídeos ou aminoácidos) se relaciona com um grupo específico de sequências. Ou seja, essa métrica informa se o conjunto de nucleotídeos ou aminoácidos de um determinado subgrupo de sequências é

capaz de diferenciá-lo das demais sequências presentes no alinhamento. Conceitualmente, o valor de MI_c para cada coluna c do alinhamento é definido como:

$$MI_c = \sum_{i=1}^2 \sum_{\alpha=1}^n p_c(\alpha, i) \log \frac{p_c(\alpha, i)}{p_c(\alpha)p_c(i)} \quad (4)$$

onde n representa o número de símbolos possíveis em cada coluna incluindo o gap ($n = 5$ para sequências nucleotídicas e 21 para protéicas), $p_c(\alpha, i)$ é a função probabilidade conjunta de um símbolo α e do grupo i e $p_c(\alpha)$ e $p_c(i)$ são as funções de probabilidades marginais do caractere α e do grupo i , respectivamente.

Já a Desarmonia de Sequências (DS) é uma medida desenvolvida pelo nosso grupo, baseada no método de Harmonia de Sequências (SH) implementado por Pirovani *et al.* (2006). Essa medida verifica o grau de semelhanças entre dois grupos de sequência presentes em um alinhamento múltiplo. Dados dois grupos de sequências A e B presentes em um MSA, o valor de DS_c para cada coluna c do alinhamento é calculado da seguinte forma:

$$SD_c = 1 + \frac{1}{2} (E_c^{A+B} - E_c^A - E_c^B) \quad (5)$$

e

$$E_c^{A+B} = - \sum (p_A + p_B) \log(p_A + p_B) \quad (6)$$

onde as entropias individuais de cada grupo (E_c - equação 2) e combinadas (E_c^{A+B} - equação 6) são calculadas.

5.3.2 Implementação do TABAJARA

O programa TABAJARA foi implementado na linguagem PERL e pode ser usado em qualquer sistema operacional compatível com POSIX, como distribuições UNIX e Linux com um interpretador Perl instalado (<http://www.perl.org>). Como programas auxiliares, TABAJARA usa o MUSCLE (Edgar, 2004) e hmmsearch, hmmbuild e nhmmer do pacote HMMER (Eddy, 2011). A Figura 6 ilustra o fluxo de trabalho do TABAJARA.

Figura 6 – Fluxo de trabalho de do TABAJARA. Um alinhamento de múltiplas sequências (A) é utilizado como conjunto de treinamento para a construção de HMMs de perfil a partir de seleção de blocos de alinhamento curtos (B) ou empregando todo o comprimento das sequências presentes no alinhamento (C). Para modelos curtos, o programa calcula os escores por posição do alinhamento usando métricas diferentes, de acordo com o tipo das sequências (nucleotídeo ou proteína) e do modo de execução. As regiões de maior pontuação são selecionadas usando uma abordagem de janela deslizante e posteriormente extraídas. Modelos construídos a partir de proteínas inteiras tem fluxos diferentes para a construção de modelos conservados ou discriminativos. No modo Discriminação, sequências específicas da categoria escolhida são selecionadas e realinhadas com MUSCLE. No caso do modo Conservação, o alinhamento original é usado para a construção do modelo. Na etapa final (D), os HMMs de perfil são construídos com a ferramenta `hmmbuild` e todos os modelos são submetidos a um procedimento de validação envolvendo testes de sensibilidade e especificidade, de acordo com os critérios definidos pelo usuário. Valores de corte de escore podem ser opcionalmente atribuídos.

O TABAJARA utiliza como entrada um conjunto de sequências, nucleotídicas ou protéicas, representadas por meio de um alinhamento de múltiplas sequências nos formatos FASTA ou CLUSTAL. A partir desse alinhamento, a ferramenta pode utilizar todo o comprimento das sequências ou sub-regiões específicas para a construção dos modelos. Nos dois casos, o TABAJARA pode ser executado nos modos (1) Conservação, que constrói modelos capazes de detectar todas as sequências do alinhamento de entrada, ou (2) Discriminação, para a construção de modelos capazes de detectar e discriminar sequências de um grupo específico em relação às demais presentes no alinhamento. Quando o TABAJARA utiliza todo o comprimento das sequências, apenas um modelo é gerado. No modo Conservação, o programa simplesmente utiliza o `hmmbuilder` para construir o HMM de perfil a partir do alinhamento original. No modo Discriminação, por outro lado, as sequências pertencentes ao grupo de interesse são inicialmente extraídas do alinhamento original, realinhadas com a ferramenta `MUSCLE` e o alinhamento resultante é utilizado para a construção de um HMM de perfil específico com o programa `hmmbuilder`.

Quando o objetivo é a construção de modelos curtos, a ferramenta seleciona sub-regiões específicas do alinhamento para a construção de modelos. Nesse caso, o TABAJARA primeiramente atribui valores de escore a cada coluna do alinhamento múltiplo de entrada, que varia entre 0 e 1 e é calculado utilizando diferentes métricas, a depender do tipo das sequências de entrada e do modo de execução (seção 5.3.1). Colunas que apresentam uma proporção de lacunas maior que valor permitido (padrão = 30%), têm o valor zero atribuído, uma vez que é improvável que colunas com muitos *gaps* em um alinhamento sejam funcionalmente importantes (Capra & Singh, 2007). Com os escores calculados para todas as posições, o TABAJARA utiliza uma janela deslizante e, com base em um conjunto de parâmetros (como tamanho da janela e porcentagem mínima de posições da mesma com escore superior a um valor mínimo), decide se a janela é válida ou não. As janelas válidas são associadas compondo blocos de alinhamento, os quais são então extraídos e processados de forma a eliminar colunas compostas somente por lacunas (*gaps*) e sequências redundantes.

É possível ainda selecionar os tamanhos mínimo e máximo desses blocos. Caso alguns blocos tenham um tamanho maior do que o máximo desejado, o programa usa uma segunda janela deslizante para selecionar a sub-região com a maior somatória de escores. Quando o modo Discriminação é utilizado, o TABAJARA executa o processo de seleção

dos blocos por meio de três etapas. Na primeira o programa utiliza o método de MI para pontuar cada coluna do alinhamento de entrada e seleciona blocos discriminativos de acordo com o protocolo descrito acima. Na segunda, o TABAJARA repete esse processo utilizando o método SD para pontuar cada coluna do MSA. Por fim, os dois conjuntos de blocos discriminativos encontrados nas etapas anteriores são unidos e processados com o objetivo de remover blocos redundantes e unir blocos sobrepostos, de forma a gerar um conjunto final de blocos discriminativos. Com os blocos finais, o TABAJARA utiliza então o programa `hmmbuilder` para construir os HMMs de perfil.

Todos os modelos gerados são submetidos a um processo de validação para que somente modelos que atendam a um conjunto de critérios de qualidade sejam selecionados. Nessa validação, os modelos são submetidos a buscas de similaridade contra as sequências do conjunto de treinamento (MSA original) utilizando as ferramentas `nhmmer` (sequências nucleotídicas) ou `hmmsearch` (sequências protéicas). Após as buscas, o TABAJARA inspeciona os resultados de forma a verificar se os modelos atendem aos critérios de controle de qualidade definidos pelo usuário. No modo Conservação, cada modelo é avaliado em relação a sua capacidade de detecção: os modelos devem ser capazes de detectar uma porcentagem mínima de sequências do conjunto de treinamento (padrão = 80%).

No modo Discriminação, como o objetivo é a construção de modelos capazes de discriminar as sequências de uma determinada categoria (grupo A) em relação às demais sequências presentes no conjunto de treinamento (grupo B), o conjunto de regras é mais complexo. Para tal, o TABAJARA inspeciona o arquivo tabular dos resultados em ordem decrescente de escore. A primeira condição é que a sequência com resultado de escore de alinhamento mais alto pertença à categoria de interesse, isto é, seja membro do grupo A. Em caso negativo, o modelo é descartado. Em caso positivo, se sequências do grupo B também tenham sido detectadas pelo modelo, o TABAJARA irá calcular a proporção entre o valor de escore da primeira sequência positiva do grupo B e o escore da sequência imediatamente anterior de grupo A. Esse número deve ser igual ou menor que um valor definido pelo usuário (por padrão, 80%). Como um último passo, todos os modelos considerados válidos são testados quanto à capacidade de detectar sequências da categoria escolhida. O percentual mínimo de detecção é também definido pelo usuário, com valor padrão de 80%. Caso seja de interesse do usuário, os modelos descartados pelo TABAJARA podem ser armazenados em um diretório para análises posteriores.

Em geral, os HMMs de perfil são utilizados para a prospecção de sequências, como, por exemplo, vírus em bases de dados de sequências de DNA ou proteína. O uso desses modelos nos permite identificar sequências mais divergentes que as sequências conhecidas, uma vez que eles modelam a variabilidade de um grupo de sequências, posição a posição. Como pudemos observar nos experimentos realizados com HMMs de perfil construídos com conjuntos de proteínas inteiras, a aplicação de valores de corte de escore aos modelos foi bastante útil para tornar os modelos bastante específicos. Na maioria dos nossos experimentos, o uso de um valor de corte de escore correspondente a 80% do valor de escore observado para o último hit do conjunto de treinamento apresentou uma boa relação entre sensibilidade e especificidade. Com base nisso, nós decidimos implementar uma abordagem heurística no TABAJARA que atribui automaticamente valores de corte de escore personalizados para cada modelo, utilizando os resultados das buscas de similaridade realizadas na etapa de validação.

Para a atribuição de valores de corte de escore aos modelos conservados, o TABAJARA calcula 80% do valor de escore apresentado pelo o último hit do conjunto de treinamento e atribui esse valor ao modelo. No caso dos modelos discriminativos, se todas as sequências de conjunto de treinamento detectadas pelo modelo durante o procedimento de validação pertencem à categoria escolhida (por exemplo, grupo A), o cálculo do valor de corte é similar ao anterior, porém o valor da porcentagem é definido pelo usuário (padrão = 80%). Quando sequências de outras categorias (por exemplo, grupo B) são também detectadas pelo modelo, o programa inspeciona os resultados da busca de similaridade em ordem decrescente de escore e calcula a diferença absoluta entre os escores da primeira sequência positiva do grupo B e a sequência imediatamente anterior do grupo A. Um valor correspondente a 80% desta diferença é adicionado ao escore da sequência de melhor desempenho do grupo B, e esse valor é atribuído como valor de corte ao modelo. Por exemplo, digamos que o escore da sequência com maior escore do grupo B seja 100 e que o valor de escore da sequência imediatamente precedente do grupo A seja 180. O valor de corte de escore sugerido para esse modelo será: $100 + (0,8 * (180 - 100)) = 164$. A ideia dessa heurística é permitir que sequências mais divergentes pertencentes à categoria desejada sejam identificadas, uma vez que existe uma margem de tolerância de valores de escore, e garantir que sequências que não pertencem à categoria sejam excluídas. Os valores de corte de escore são adicionados aos HMMs de perfil através de uma etiqueta juntamente com o valor de corte calculado (por exemplo, `CUTOFF SCORE 164`). Essa etiqueta é neutra para programas do pacote HMMER

(`hmmsearch` e `nhmmer`), mas esses valores podem ser utilizados para filtrar os resultados das buscas de similaridade por meio da atribuição desse valor no parâmetro `-T` desses programas (por exemplo, `-T 164`). Algumas ferramentas desenvolvidas pelo nosso grupo, como GenSeed-HMM (Alves *et al.*, 2016) e e-Finder (ver item 8), são capazes de ler automaticamente os valores de corte de escore de cada HMM de perfil e utilizá-los nas buscas de similaridade.

5.4 Validação da ferramenta

A fim de validarmos os modelos gerados pelo TABAJARA e, consequentemente, a metodologia desenvolvida, foi executado um conjunto de experimentos. Nessa seção são descritos esses experimentos, bem como os resultados obtidos tanto pelos modelos conservados como discriminativos construídos pela ferramenta. Nesse processo foram utilizados os mesmos conjuntos de treinamento e teste de sequências de *Flavivirus* e *Microviridae* utilizadas durante o desenvolvimento da metodologia e descritos no início desse capítulo.

5.4.1 Otimização dos parâmetros do TABAJARA

O TABAJARA é uma ferramenta flexível e configurável, que permite ao usuário definir uma grande quantidade de parâmetros, sendo alguns deles obrigatórios e outros opcionais. Quando o TABAJARA é executado com o objetivo de gerar modelos curtos, é necessário que alguns parâmetros para a seleção dos blocos, como tamanho mínimo do bloco de alinhamento (parâmetro `-b`), o valor mínimo de escore por posição (parâmetro `-t`) e porcentagem mínima de posições significativas dentro da janela deslizante (parâmetro `-p`), sejam definidos.

A fim de avaliar o impacto da variação desses parâmetros no desempenho de modelos discriminativos construídos pelo TABAJARA, foi executada uma série de experimentos, utilizando-se o conjunto de treinamento com 127 sequências de diferentes espécies de *Flavivirus*. Nesses experimentos, foram construídos modelos discriminativos para *Dengue virus*, *Zika virus* e *Yellow Fever virus* com o TABAJARA sendo executado diversas vezes, com a alteração de apenas um parâmetro por vez. Todos os modelos gerados foram submetidos a buscas de similaridade com a ferramenta `hmmsearch`

contra o conjunto de teste de *Flavivirus*, composto de 6.364 sequências da poliproteína, com e sem a aplicação dos valores de corte sugeridos pelo TABAJARA. Para se avaliar o desempenho dos modelos, valores de sensibilidade e especificidade para os mesmos foram calculados, bem como os valores médios dessas métricas obtidos a partir da combinação dos múltiplos modelos gerados para cada valor de parâmetro.

No primeiro experimento, o valor mínimo de escore por posição do alinhamento (parâmetro $-\tau$) foi modificado a cada execução do TABAJARA, variando de 0 a 1, com incrementos de 0,1 a cada execução. Os demais parâmetros foram mantidos inalterados e definidos da seguinte forma: janela deslizante de tamanho 15, com pelo menos 50% de suas posições ($-p \geq 50$) apresentando escore mínimo definido pelo parâmetro $-\tau$. O tamanho do bloco de alinhamento selecionado para a criação dos modelos variou entre 15 e 60 aminoácidos. A Figura 7A apresenta os resultados obtidos para *Dengue virus* utilizando-se diferentes valores de τ .

Como podemos observar, os HMMs de perfil gerados com valores de τ acima de 0,6 apresentaram valores médios de sensibilidade e especificidades muito próximos a 1,0. Por outro lado, valores inferiores a 0,6 resultaram em modelos com valores menores de especificidade. Por fim, o TABAJARA não foi capaz de gerar modelos com valores mínimos de escore superiores a 0,7, uma vez que essa condição se tornou muito rigorosa e nenhum bloco foi selecionado. Quando os valores de corte de escore sugeridos pelo TABAJARA foram utilizados nas buscas de similaridade, independentemente do valor mínimo de escore por posição utilizado para se construir os modelos, os valores médios de sensibilidade e especificidade foram muito próximos de 1,0. Esses resultados mostram que, independentemente dos valores de τ utilizados para construir os modelos, quando os valores de corte sugeridos pelo TABAJARA são utilizados nas buscas de similaridade, todos os modelos se mostram altamente específicos. Resultados semelhantes foram observados com os modelos gerados para os vírus da Zika (Figura 7B) e Febre Amarela (Figura 7C).

Figura 7- Efeito da variação do valor mínimo do escore por posição do alinhamento (parâmetro $-\tau$) no desempenho dos HMMs de perfil construídos com o TABAJARA. O TABAJARA foi executado em modo discriminativo para a construção de modelos curtos utilizando como conjunto de treinamento o alinhamento múltiplo de 127 sequências de poliproteína de *Flavivirus*. Regiões específicas de *Dengue virus* (A), *Zika virus* (B) e *Yellow Fever virus* (C), com tamanho variando entre 15 e 60 aminoácidos, foram selecionadas utilizando uma janela deslizante de tamanho 15, onde pelo menos 50% de suas posições deveriam ter um valor mínimo de escore, que variou de 0 a 1,0. Todos os modelos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Flavivirus*, sem e com a aplicação dos valores de corte de escore sugeridos para cada modelo. As linhas representam as médias aritméticas dos valores de sensibilidade e especificidade dos modelos gerados com cada valor de τ . As áreas sombreadas indicam os valores de desvio padrão.

No segundo experimento, o parâmetro alterado ($-p$) a cada iteração foi a porcentagem mínima de posições significativas dentro da janela deslizante. Nesse caso, uma porcentagem mínima de posições dentro de uma janela deslizante de tamanho 15 deveria ter um escore de pelo menos 0,5 ($-\tau = 0,5$) para que a janela fosse considerada válida. O valor de porcentagem mínima nos testes variou de 0 a 100, sendo incrementado em 10 a cada iteração. Da mesma forma que no experimento anterior, os blocos selecionados para a construção dos modelos tiveram tamanhos entre 15 e 60 aminoácidos.

Como podemos observar na Figura 8A, quando os modelos construídos para *Dengue virus* foram submetidos a buscas de similaridade sem a aplicação dos escores de corte sugeridos pelo TABAJARA, modelos gerados com valores de p maiores que 60 apresentaram valores de sensibilidade e especificidade muito próximos a 1,0, enquanto que os modelos gerados com valores menores que 60 apresentaram valores de especificidade mais baixos. Por outro lado, quando os escores de corte foram empregados, o desempenho dos modelos melhorou significativamente, com os valores de especificidade muitos próximos a 1,0 para todos os valores de p testados. Com valores de p acima de 85, o TABAJARA não foi capaz de gerar modelos. Esses testes foram também

realizados para a geração de modelos discriminativos para *Zika virus* (Figura 8B) e *Yellow Fever virus* (Figura 8C) e resultados similares foram encontrados.

Figura 8 - Efeito da variação do valor de porcentagem mínima de posições significativas em uma janela deslizante (parâmetro $-p$) no desempenho dos HMMs de perfil construídos com o TABAJARA. O TABAJARA foi executado em modo discriminativo para a construção de modelos curtos utilizando como conjunto de treinamento o alinhamento múltiplo de 127 sequências de poliproteína de *Flavivirus*. Regiões específicas de *Dengue virus* (A), *Zika virus* (B) e *Yellow Fever virus* (C), com tamanho variando entre 15 e 60 aminoácidos, foram selecionadas utilizando uma janela deslizante de tamanho 15, onde uma porcentagem variável de suas posições (de 0 a 100) deveriam ter escore de pelo menos 0,5. Todos os modelos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Flavivirus*, sem e com a aplicação dos valores de corte de escore sugeridos para cada modelo. As linhas representam as médias aritméticas dos valores de sensibilidade e especificidade dos modelos gerados com cada valor de t . As áreas sombreadas indicam os valores de desvio padrão.

Nos experimentos relatados na seção 5.2, o tamanho dos blocos de alinhamento curtos mostrou influência direta na capacidade discriminativa nos HMMs de perfil gerados a partir desses blocos. Assim, resolvemos avaliar a influência do tamanho dos blocos de alinhamento selecionados pelo TABAJARA no desempenho dos modelos discriminativos gerados. Para isso, foram utilizados tamanhos mínimo e máximo dos blocos (parâmetros $-b$ e $-mb$, respectivamente) de valores iguais, de forma a restringir precisamente todos os modelos para o comprimento desejado. Os demais parâmetros definidos para o experimento foram os seguintes: janela deslizante de tamanho 10 ($-w$ 10), que foi considerada válida se pelo menos 40% ($-p$ 40) de suas posições da janela tivessem um escore de pelo menos 0,4 ($-t$ 0.4). As Figuras 9A, 9B e 9C ilustram os resultados obtidos respectivamente para *Dengue virus*, *Zika virus* e *Yellow Fever virus*.

Como podemos observar em todos os experimentos, para qualquer tamanho do bloco de alinhamento utilizado, os valores médios de sensibilidade foram muito altos, independente do uso dos valores de corte de escore sugeridos para os modelos. No caso

do experimento com *Zika virus*, esses valores foram levemente afetados para alguns tamanhos quando os valores de corte foram utilizados, porém os valores médios continuaram bastante altos. Por outro lado, sem o uso de escores de corte, os valores de especificidade foram altos para modelos curtos (10 aa), mas apresentaram uma queda drástica para tamanhos de blocos selecionados superiores a 20 aa.

Esse resultado pode ser explicado pelo fato de as sequências de *Flavivirus* serem muito similares e, assim, os trechos específicos de uma determinada espécie são muito curtos, enquanto trechos longos do alinhamento podem conter posições que são conservadas em múltiplas espécies. Quando os valores de corte de escore sugeridos pelo TABAJARA foram utilizados, todos os modelos se tornaram altamente específicos para a detecção de sequências da espécie para a qual foram desenvolvidos. Devido aos diferentes graus de variação intra-específica entre as sequências de *Dengue virus*, *Zika virus* e *Yellow Fever virus*, o tamanho máximo do bloco selecionado em cada uma das espécies foi diferente (respectivamente 290, 280 e 420 aminoácidos). Com o objetivo de deixar os gráficos consistentes, modelos construídos com blocos de alinhamento maiores que 200 aminoácidos e que apresentaram valores de especificidade médios abaixo de 0,025 foram omitidos dos gráficos.

Figura 9. Efeito da variação do tamanho dos blocos de alinhamento selecionados no desempenho dos HMMs de perfil construídos com o TABAJARA. O TABAJARA foi executado em modo discriminativo para a construção de modelos curtos utilizando como conjunto de treinamento o alinhamento múltiplo de 127 sequências de poliproteína de *Flavivirus*. Regiões específicas de *Dengue virus* (A), *Zika virus* (B) e *Yellow Fever virus* (C), foram selecionadas utilizando uma janela deslizante de tamanho 10, onde pelo menos 40% de suas posições deveriam ter escore mínimo de 0,4, com variação do tamanho das regiões selecionadas a cada iteração. Todos os modelos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Flavivirus*, sem e com a aplicação dos valores de corte de escore sugeridos para cada modelo. As linhas representam as médias aritméticas dos valores de sensibilidade e especificidade dos modelos gerados com cada valor de τ . As áreas sombreadas indicam os valores de desvio padrão.

Como pode ser concluído a partir dos experimentos apresentados nas Figuras 7 a 9, os modelos discriminativos construídos pelo TABAJARA podem ser altamente específicos para diferentes grupos virais, sob diferentes conjuntos de parâmetros, desde que as pontuações de corte atribuídas pelo programa sejam aplicadas. Por outro lado, caso os valores de corte de pontuação sugeridos pelo TABAJARA não sejam utilizados, somente uma pequena fração de blocos de alinhamento curtos e altamente discriminativos pode ser obtida e, ainda assim, sob combinações específicas e otimizadas de parâmetros.

5.4.2 Validação de modelos construídos para *Flavivirus* e *Microviridae*

Tendo testado a influência dos diferentes conjuntos de parâmetros do programa TABAJARA no desempenho dos modelos gerados, resolvemos então construir modelos conservados e discriminativos utilizando um conjunto padronizado de parâmetros. Nesse teste foram construídos modelos para sequências de vírus do gênero *Flavivirus* e da família *Microviridae*. Foram utilizados como conjuntos de treinamentos alinhamentos múltiplos com 127 sequências de poliproteína de *Flavivirus* e o com 113 sequências da proteína VP1 de *Microviridae*. Todos os modelos gerados foram testados em buscas de similaridade contra seus respectivos conjuntos de teste utilizando o `hmmsearch`. No caso de modelos conservados, foram feitos testes com e sem a utilização de valores de corte de pontuação sugeridos pelo TABAJARA. No caso dos modelos discriminativos, os testes foram feitos sempre com os valores de corte, visto que eles são essenciais para a obtenção de altos valores de especificidade (Figuras 7 a 9).

O primeiro experimento teve como objetivo estimar a sensibilidade (*recall*) dos modelos conservados construídos com o TABAJARA. Foram obtidos dez modelos curtos e um modelo de proteínas inteiras para *Flavivirus* e, no caso de *Microviridae*, sete modelos curtos e um modelo de proteínas inteiras. A Tabela 3 mostra que todos os modelos conservados construídos para o gênero *Flavivirus* foram capazes de detectar praticamente todas as sequências do conjunto de teste, independentemente do uso dos valores de corte de pontuação (porcentagem de detecção superior a 99% em todos os casos).

Tabela 3. Resultado das buscas de similaridade realizadas com modelos conservados construídos para *Flavivirus* com o TABAJARA. O TABAJARA foi executado em modo Conservação para a construção de modelos curtos utilizando como conjunto de treinamento o alinhamento múltiplo de 127 sequências de poliproteína de *Flavivirus*. Todos os modelos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Flavivirus*, sem e com a aplicação dos valores de corte de escore sugeridos para cada modelo.

HMM de perfil	Com valores de corte		Sem valores de corte	
	Nº de positivos	% de positivos	Nº de positivos	% de positivos
Proteína inteira	6.340	99,62	6.364	100,00
Modelo curto 2039-2058	6.356	99,87	6.360	99,94
Modelo curto 2969-2987	6.358	99,91	6.361	99,95
Modelo curto 3102-3120	6.358	99,91	6.360	99,94
Modelo curto 3242-3261	6.354	99,84	6.360	99,94
Modelo curto 3369-3388	6.360	99,94	6.361	99,95
Modelo curto 3349-3367	6.354	99,84	6.361	99,95
Modelo curto 3432-3446	6.353	99,83	6.362	99,97
Modelo curto 3498-3517	6.349	99,76	6.359	99,92
Modelo curto 3663-3682	6.358	99,91	6.359	99,92
Modelo curto 3711-3730	6.359	99,92	6.359	99,92
Média dos modelos curtos	6.356	99,87	6.360	99,94
Modelos curtos combinados	6.360	99,94	6.362	99,97

Com base nesses dados, podemos concluir que, para a identificação de sequências de *Flavivirus* em bases de dados compostas por proteínas inteiras, o modelo construído com proteínas inteiras e os modelos curtos geram resultados similares. Contudo, observou-se que o tempo gasto para a execução das buscas com o modelo construído com proteínas inteiras é muito maior (cerca de 100 vezes) do que com os modelos curtos contra o conjunto de teste. Outra vantagem do uso dos modelos curtos se relaciona com a prospecção de dados metagenômicos não montados. Como o tamanho médio das leituras é relativamente curto, o uso de modelos de sequências inteiras não poderia ser feito com valores de corte de pontuação, já que eles seriam restritivos demais para permitirem a detecção de blocos de alinhamento curtos. Com isso, a sensibilidade de detecção poderia ser alta, mas a especificidade poderia ser inaceitavelmente baixa.

O mesmo tipo de experimento foi realizado para a família *Microviridae* e, conforme pode ser visto na Tabela 4, o uso de valores de corte causou uma forte redução

da sensibilidade, tanto nos modelos de proteínas inteiras como nos modelos curtos. O modelo construído a partir de proteínas inteiras sem a aplicação do valor de corte sugerido pelo TABAJARA nas buscas de similaridade foi capaz de identificar 100% das sequências do conjunto de teste. Isso pode ser explicado pelo fato de tal modelo incorporar todos os sítios conservados do alinhamento múltiplo. Contudo, quando o valor de corte de pontuação sugerido pelo TABAJARA foi utilizado nas buscas de similaridade, a sensibilidade do modelo caiu para 70%.

Tabela 4. Resultado das buscas de similaridade realizadas com modelos conservados construídos para a família *Microviridae* com o TABAJARA. O TABAJARA foi executado em modo Conservado para a construção de modelos curtos utilizando um conjunto de treinamento composto por 113 sequências VP1 de *Microviridae*. Todos os modelos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Microviridae*, sem e com a aplicação dos valores de corte de escore sugeridos para cada modelo.

HMM de perfil	Com valores de corte		Sem valores de corte	
	# de positivos	% de positivos	# de positivos	% de positivos
Proteína inteira	1.294	70,48	1.836	100,00
Modelo 45-64	1.328	72,33	1.608	87,58
Modelo 91-110	1.417	77,18	1.693	92,21
Modelo 553-572	1.397	76,09	1.679	91,45
Modelo 900-919	1.369	74,56	1.656	90,20
Modelo 996-1015	1.498	81,59	1.697	92,43
Modelo 1038-1055	1.557	84,80	1.706	92,92
Modelo 1128-1147	1.428	77,76	1.749	95,26
Média dos modelos curtos	1.294	70,48	1.697	92,41
Modelos curtos combinados	1.765	96,13	1.818	99,01

Esse resultado pode ser atribuído ao fato das sequências de cada uma das subfamílias da família *Microviridae* serem bastante divergentes entre si. O valor de corte sugerido pelo modelo é calculado de acordo com as sequências do conjunto de treinamento, as quais apresentam um certo grau de divergência entre si. Se, contudo, o conjunto de teste contiver sequências mais divergentes, o valor de corte de pontuação, atribuído pela amostragem do conjunto de treinamento, pode acabar sendo alto demais para permitir a detecção dessas sequências, limitando assim a sensibilidade do modelo. Dessa forma, grupos taxonômicos que apresentam maior divergência interna tendem a

tornar a tarefa de obtenção de modelos conservados bem mais complexa. Esse foi justamente o caso observado nas diferentes subfamílias da família *Microviridae*, quando comparadas às diferentes espécies de vírus do gênero *Flavivirus*.

A divergência relativamente alta das sequências virais de *Microviridae* também foi desafiadora para o uso modelos curtos, os quais contém as regiões mais conservadas do alinhamento de entrada. Quando utilizados de forma individual para a detecção de sequências, nenhum dos modelos curtos conservados para a família *Microviridae* foi capaz de detectar 100% das sequências do conjunto de teste (Tabela 4). Porém, quando esses modelos foram utilizados de forma combinada, o percentual de sequências detectadas ficou acima de 96%, independentemente do uso dos valores de corte. Isso mostra que, no caso dos *Microviridae*, o uso do alinhamento composto de proteínas inteiras para a geração de um modelo conservado foi capaz de detectar a diversidade das sequências de *Microviridae*, uma vez que todas as regiões informativas do alinhamento foram adicionadas ao modelo. Contudo, como as sequências de *Microviridae* são muito divergentes entre si, o uso do valor de corte sugerido foi alto e terminou por filtrar algumas sequências que deveriam ser detectadas no conjunto de teste.

Por outro lado, quando o TABAJARA seleciona as regiões mais conservadas para a construção de modelos curtos, as regiões não informativas (não conservadas) não são incluídas aos modelos. Isso garante que essas regiões selecionadas pelo TABAJARA representem de fato assinaturas do grupo viral e permitam a detecção dos membros desse grupo. O uso combinado de todos os modelos curtos fez com que todos sítios conservados do alinhamento múltiplo fossem utilizados de forma conjunta, sem a adição de ruído. Como consequência, quase 100% das sequências foram identificadas com essa abordagem. Em relação ao tempo gasto na execução das buscas de similaridade, da mesma que forma que observado nos *Flavivirus*, o tempo de execução do modelo construído a partir das proteínas inteiras foi também bastante superior ao tempo gasto por todos os modelos curtos (aproximadamente 15 segundos contra 3 segundos, respectivamente). Com base nesses resultados, podemos concluir que o uso combinado de modelos curtos oferece vantagens em relação ao uso de HMMs de perfil construídos a partir de sequências inteiras de proteínas.

Um outro teste consistiu na construção de modelos discriminativos para três espécies de *Flavivirus* (*Dengue virus*, *Zika virus* e *Yellow Fever virus*) e três subfamílias de *Microviridae* (*Alpavirinae*, *Gokushovirinae* e *Pichovirinae*). Nesse experimento, os conjuntos de treinamento e teste utilizados foram os mesmos do experimento anterior.

Foram construídos modelos utilizando proteínas inteiras, assim como sub-regiões grupo-específicas definidas pelo TABAJARA a partir do alinhamento. Para o gênero *Flavivirus*, foram gerados oito modelos curtos para *Dengue virus*, 12 para *Zika virus* e 12 para *Yellow Fever virus*, além de modelos de proteínas inteiras para cada espécie. Para as subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae* (família *Microviridae*), foram gerados quatro, sete e três modelos curtos, respectivamente, além dos modelos de sequências inteiras de cada uma. Todos os modelos foram submetidos a buscas de similaridade contra os respectivos conjuntos de teste utilizando os valores de corte de pontuação sugeridos para cada modelo. A Tabela 5 mostra os resultados obtidos pelos modelos construídos para as espécies de *Flavivirus*. Como muitos modelos curtos foram gerados, nessa tabela são apenas exibidos os valores médios de sensibilidade e especificidade desses modelos, bem como os valores dessas métricas obtidos pela combinação de todos os modelos curtos.

Como podemos perceber, tanto os modelos construídos com proteínas inteiras como modelos curtos apresentaram valores de sensibilidade e especificidade muito próximos do valor um (100%). Isso aconteceu porque, embora as sequências de *Flavivirus* sejam bastante similares entre si, o nível de similaridade entre sequências de uma mesma espécie é bem maior que entre sequências de espécies diferentes, o que permite a clara separação entre as sequências de espécies distintas. Assim, de forma similar aos modelos conservados de *Flavivirus*, os modelos curtos e aqueles construídos com proteínas inteiras geraram resultados bastante similares entre si. A diferença observada entre os dois tipos de modelo foi novamente em relação ao tempo de execução. Enquanto os modelos construídos com as proteínas inteiras levaram em média 11 minutos para realizarem as buscas, a execução de todos os modelos curtos de cada espécie levou aproximadamente seis segundos.

Tabela 5. Resultado das buscas de similaridade realizadas com modelos discriminativos construídos para as espécies *Dengue virus*, *Zika virus* e *Yellow Fever virus* do gênero *Flavivirus*. O TABAJARA foi executado em modo Discriminativo para a construção de modelos curtos utilizando como conjunto de treinamento o alinhamento múltiplo de 127 sequências de poliproteína de *Flavivirus*. Todos os modelos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Flavivirus*, com a aplicação dos valores de corte de escore sugeridos para cada modelo.

HMM de perfil	Sensibilidade	Especificidade
<i>Dengue virus</i>		
Proteína inteira	1,00	1,00
Média dos modelos curtos	1,00	1,00
Modelos curtos combinados	1,00	1,00
<i>Zika virus</i>		
Proteína inteira	1,00	1,00
Média dos modelos curtos	0,98	1,00
Modelos curtos combinados	1,00	1,00
<i>Yellow Fever virus</i>		
Proteína inteira	1,00	1,00
Média dos modelos curtos	0,98	1,00
Modelos curtos combinados	1,00	1,00

Em relação aos HMMs de perfil discriminativos construídos para as subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae* (Tabela 6), embora todos os modelos tenham sido altamente específicos (especificidade próxima ou igual a 1), a sensibilidade dos modelos utilizados de forma individual não foi muito alta. No caso dos modelos de proteína inteira, a sensibilidade observada foi de 0,63 para *Alpavirinae*, 0,87 para *Gokushovirinae* e 0,75 para *Pichovirinae*. Para os modelos curtos, a sensibilidade média dos modelos individuais variou de 0,61, 0,66 e 0,77 para as três subfamílias, respectivamente. O uso combinado dos diferentes modelos curtos aumentou consideravelmente a sensibilidade, com valores de 0,82, 0,94 e 0,92, respectivamente.

Portanto, pode-se concluir que modelos curtos utilizados de forma combinada apresentam valores de sensibilidade bastante superiores aos observados com modelos individuais, sejam eles curtos ou de proteínas inteiras. Assim, podemos novamente concluir que no caso de grupos de sequências divergentes, é mais vantajoso construir modelos curtos, sejam eles conservados ou discriminativos, e utilizar esses modelos de forma combinada para a identificação de sequências. Por fim, em relação ao tempo de execução das buscas de similaridade, o tempo gasto pelo modelo construído com

proteínas inteiras foi bastante superior ao gasto por todos os modelos curtos, da mesma forma que em todos os casos anteriores. Para *Alpavirinae*, o modelo de proteína inteira levou aproximadamente 25 segundos, enquanto que todos os modelos curtos levaram apenas 1 segundo para realizarem as buscas. Com os modelos de *Gokushovirinae*, o tempo do modelo construído com proteínas inteiras foi de 21 segundos e o dos modelos curtos foi de 3 segundos. Por fim, para *Pichovirinae*, o modelo gerado com proteínas inteiras levou 15 segundos, enquanto que todos os modelos curtos gastaram apenas 1 segundo.

Tabela 6. Resultado das buscas de similaridade realizadas com modelos discriminativos construídos para as subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae* da família *Microviridae*. O TABAJARA foi executado em modo Discriminativo para a construção de modelos curtos utilizando um conjunto de treinamento composto por 113 sequências VP1 de *Microviridae*. Todos os modelos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Microviridae*, com a aplicação dos valores de corte de escore sugeridos para cada modelo.

HMM de perfil	Sensibilidade	Especificidade
<i>Alpavirinae</i>		
Proteína inteira	0,63	1,00
Média dos modelos curtos	0,61	1,00
Modelos curtos combinados	0,82	1,00
<i>Gokushovirinae</i>		
Proteína inteira	0,87	1,00
Média dos modelos curtos	0,66	1,00
Modelos curtos combinados	0,94	1,00
<i>Pichovirinae</i>		
Proteína inteira	0,75	1,00
Média dos modelos curtos	0,77	0,99
Modelos curtos combinados	0,92	0,97

5.4.3 Uso cumulativo de modelos curtos para a detecção de sequências

Como observado no experimento anterior, o uso de modelos curtos combinados para a detecção de sequências virais em conjuntos de dados mais divergentes pode ser mais sensível que o uso individual de modelos, sejam eles curtos ou de proteínas inteiras. Como no experimento anterior foi somente verificado o potencial de detecção da

combinação de todos os modelos, resolvemos investigar o efeito do uso cumulativo de modelos curtos com todas as combinações possíveis de uso. Assim, foram construídos modelos curtos específicos para *Alpavirinae*, *Gokushovirinae* e *Pichovirinae* com o programa TABAJARA, utilizando um conjunto de treinamento composto de 113 sequências da proteína VP1 de *Microviridae*. Todos os HMMs de perfil construídos foram submetidos a buscas de similaridade contra o conjunto de teste dos *Microviridae*, composto de 1.836 sequências de VP1, utilizando os valores de corte de pontuação sugeridos pelo TABAJARA.

Para cada subfamília, os modelos discriminativos curtos gerados foram testados utilizando todas as combinações possíveis de um número n de modelos, onde n variou de um até o número total de modelos. Para cada modelo, foram calculados os valores de sensibilidade e especificidade e, para cada valor de n , os valores médios dessas métricas foram obtidos. Os resultados de cada subfamília foram plotados em gráficos, como mostrado na Figura 10. Como podemos observar nas três subfamílias, a menor média de sensibilidade foi observada com o uso de modelos individuais. À medida que a quantidade de modelos combinados aumentou, o valor médio de sensibilidade também cresceu, chegando ao valor máximo possível quando todos os modelos disponíveis foram utilizados em conjunto. É possível perceber que o uso de todos os modelos gera uma curva próxima da saturação, mas, mesmo assim, não se observa 100% de sensibilidade.

Esse resultado é mais acentuado no caso da subfamília *Alpavirinae* (Figura 10A) do que nas subfamílias *Gokushovirinae* e *Pichovirinae* e pode ser explicado pelo fato da subfamília *Alpavirinae* apresentar uma similaridade média entre sequências de seus membros de 37%, enquanto *Gokushovirinae* e *Pichovirinae* apresentam valores médios de 59,6% e 54%, respectivamente (dados não mostrados). Essa característica própria de cada subfamília pode ser claramente evidenciada na árvore filogenética obtida por (Roux *et al.* 2012), na qual claramente se observa ramos muito mais longos entre os membros da subfamília *Alpavirinae* do que nas demais. Essa maior divergência dificulta a tarefa de se encontrar regiões que sejam ao mesmo tempo discriminativa entre membros de subfamílias distintas de um lado e, por outro lado, que essas regiões sejam suficientemente conservadas para detectar a grande maioria dos membros dessa subfamília.

Em relação à especificidade, o valor médio obtido para as subfamílias *Alpavirinae* (Figura 10A) e *Gokushovirinae* (Figura 10B) foi muito próximo de 1,0. Para a subfamília *Pichovirinae* (Figura 10C), a especificidade apresentou uma pequena queda para 97%

com o aumento da quantidade de modelos combinados. Isso ocorreu devido a alguns modelos dessa subfamília detectarem algumas sequências falso-positivas de *Gokushovirinae* (dados não mostrados), que é filogeneticamente mais próxima da subfamília *Pichovirinae*.

Figura 10. O efeito da combinação de múltiplos HMMs de perfil na capacidade de detectar e discriminar sequências de *Microviridae*. O TABAJARA foi executado em modo Discriminativo utilizando um conjunto de treinamento composto por 113 sequências VP1 de *Microviridae*. Modelos específicos para as subfamílias *Alpavirinae* (A), *Gokushovirinae* (B) e *Pichovirinae* (C) foram testados em buscas de similaridade utilizando o programa *hmmsearch* contra um conjunto de dados de 1.836 sequências de VP1 de *Microviridae* compreendendo membros das três subfamílias da família utilizando os valores de corte de pontuação sugeridos pelo TABAJARA. Os testes foram realizados usando múltiplas combinações de um número n de modelos selecionados aleatoriamente, variando de um até o número total de modelos construídos para cada subfamília. As linhas de sensibilidade e especificidade, representam médias aritméticas calculadas a partir dos resultados observados para a combinação múltipla de modelos gerados para cada valor de n . Áreas sombreadas indicam valores de desvio padrão

5.4.4 Capacidade de generalização dos HMMs de perfil

Em todos os experimentos relatados anteriormente para *Flavivirus*, foi utilizado como conjunto de treinamento o alinhamento múltiplo de 127 sequências de poliproteína de diversas espécies. O próximo passo no desenvolvimento da abordagem metodológica de construção dos modelos consistiu na avaliação do desempenho dos mesmos frente a conjuntos de treinamento compostos de diferentes combinações de sequências. Para essa finalidade, foi executado um protocolo de validação cruzada para se gerar diferentes conjuntos de treinamento destinados à construção de modelos pelo TABAJARA.

O método de validação cruzada escolhido foi o k -fold (Arlot & Celisse, 2010), uma técnica utilizada para se avaliar a capacidade de generalização de uma metodologia, a qual consiste na divisão do conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho. Em cada uma das k iterações, um subconjunto é utilizado como conjunto de teste e os $(k - 1)$ subconjuntos remanescentes são utilizados de forma

combinada como conjunto de treinamento. Em nossa validação, o valor de k foi definido como 10, utilizando-se o conjunto de 6.364 sequências de poliproteínas de *Flavivirus*, composto de 3.919 sequências de *Dengue virus*, 327 de vírus da *Zika* e 63 de vírus da Febre Amarela.

A primeira validação realizada teve como objetivo a construção de modelos discriminativos para o vírus da dengue. Dessa forma, o total de sequências foi dividido em um conjunto positivo, composto de 3.919 sequências de *Dengue virus*, e um conjunto negativo, composto de 2.445 sequências pertencentes as demais espécies de *Flavivirus*. O conjunto de teste foi então repartido em 10 subconjuntos, mantendo-se a proporção de sequências positivas e negativas em cada conjunto. Assim, cada subconjunto foi composto de 319 sequências de *Dengue virus* e 244 sequências das demais espécies de *Flavivirus*. A distribuição das sequências em cada subconjunto foi feita de forma arbitrária. A cada iteração (considerando-se 10 iterações), um subconjunto foi definido como sendo de teste e os nove restantes unidos e alinhados com a ferramenta MUSCLE para comporem o conjunto de treinamento. As sequências alinhadas foram então utilizadas para a construção de modelos discriminativos curtos de *Dengue virus* com o TABAJARA. Todos os modelos gerados foram submetidos a buscas de similaridade com a ferramenta *hmmsearch* contra o conjunto de teste, utilizando ou não valores de corte de pontuação sugeridos pelo TABAJARA.

Para a avaliação do desempenho dos modelos, foram calculados os valores de sensibilidade e especificidade de cada um. Para cada iteração, foram calculadas as médias aritméticas desses valores. Como podemos ver na Tabela 7, quando os valores de corte de pontuação foram utilizados, todos os modelos foram altamente sensíveis e específicos. Porém, quando esses valores não foram empregados, a especificidade média dos modelos em todas as iterações foi de aproximadamente 40%. Esse resultado corrobora mais uma vez a grande importância de se utilizar os valores de corte de pontuação definidos pelo TABAJARA. Esse experimento foi repetido para a construção de modelos de *Zika virus* e *Yellow Fever virus*, mantendo, a cada iteração, a proporção de sequências positivas e negativas. Os resultados obtidos foram bastante similares aos observados para *Dengue virus* (dados não mostrados).

Tabela 7 – Validação cruzada *k-fold* de HMMs de perfil específicos do vírus da Dengue construídos com TABAJARA. Um conjunto composto por 6.364 sequências de poliproteínas de *Flavivirus*, compreendendo 3.919 sequências de *Dengue virus* e 2.445 de outras espécies de *Flavivirus*, foi usado em um procedimento de validação cruzada *k-fold*, com valor de *k* igual a 10. Cada grupo consistiu em sequências selecionadas aleatoriamente, preservando a proporção original de sequências de *Dengue virus* e não-*Dengue virus*. Conjuntos de sequências compostos de $(k - 1)$ grupos foram alinhados o programa MUSCLE e o alinhamento resultante utilizado como entrada para a execução do programa TABAJARA para a construção de HMMs de perfil curtos discriminativos para *Dengue virus*. Todos os modelos foram testados em buscas de similaridade usando o programa hmmsearch, com ou sem os valores de corte de pontuações atribuídos pelo TABAJARA, contra um conjunto de dados de validação composto por um único grupo. As médias aritméticas de sensibilidade e especificidade foram calculadas a partir dos resultados combinados dos múltiplos modelos obtidos para cada grupo.

Iteração	Nº de modelos	Com valores de corte		Sem valores de corte	
		Média da sensibilidade	Média da especificidade	Média da sensibilidade	Média da especificidade
Iteração 1	30	1,00	1,00	1,00	0,34
Iteração 2	31	1,00	1,00	1,00	0,37
Iteração 3	33	1,00	1,00	1,00	0,40
Iteração 4	32	1,00	1,00	1,00	0,38
Iteração 5	33	1,00	1,00	1,00	0,44
Iteração 6	33	1,00	1,00	1,00	0,43
Iteração 7	32	1,00	1,00	1,00	0,38
Iteração 8	31	1,00	1,00	1,00	0,38
Iteração 9	31	1,00	1,00	1,00	0,39
Iteração 10	32	1,00	1,00	1,00	0,40

O experimento de validação cruzada foi também realizado com sequências da família *Microviridae*, que apresenta um maior grau de divergência entre seus membros. Dentro dessa família, escolhemos a subfamília *Gokushovirinae*, que apresenta uma maior similaridade entre as suas sequências do que as demais subfamílias. O experimento foi realizado com um conjunto de sequências composto de 1.866 sequências da proteína VP1 de *Microviridae*, sendo 501 sequências pertencentes à subfamília *Alpavirinae*, 1.040 à *Gokushovirinae* e 325 à *Pichovirinae*. Esse conjunto foi dividido em dois subconjuntos, sendo um positivo, composto de 1.040 sequências de *Gokushovirinae*, e um outro negativo, composto de 826 sequências pertencentes as demais subfamílias de *Microviridae*. O conjunto de teste foi particionado em 10 subconjuntos, mantendo-se a proporção de sequências positivas e negativas em cada subconjunto.

Como é possível observar na Tabela 8, a quantidade de modelos encontrados em cada uma das iterações foi bastante inferior do que o obtido para *Dengue virus* (Tabela 7). Dada a maior divergência entre as sequências de *Gokushovirinae*, foram gerados apenas seis modelos (Figura 10), visto que o programa TABAJARA possivelmente teve maior dificuldade para identificar regiões do alinhamento que são ao mesmo tempo conservadas para *Gokushovirinae* e discriminativas em relação às demais sequências. Em relação à sensibilidade e especificidade dos modelos (Tabela 8), quando os valores de corte de pontuação foram utilizados nas buscas de similaridade, todos os modelos foram altamente específicos, porém a sensibilidade desses modelos foi notadamente reduzida, mas ainda dentro de valores relativamente altos (85-92%). Quando os valores de corte não foram utilizados, a sensibilidade se manteve próxima de 1,0, mas a especificidade dos modelos caiu consideravelmente.

Tabela 8 - Validação cruzada *k-fold* de HMMs de perfil específicos de vírus da subfamília *Gokushovirinae* construídos com TABAJARA. Um conjunto composto por 1.866 sequências da proteína VP1 de *Microviridae*, compreendendo 1.040 sequências de *Gokushovirinae* e 826 de outras subfamílias de *Microviridae*, foi usado em um procedimento de validação cruzada *k-fold*, com valor de *k* igual a 10. Cada grupo consistiu em sequências selecionadas aleatoriamente, preservando a proporção original sequências de *Gokushovirinae* e não-*Gokushovirinae*. Conjuntos de sequências compostos de (*k* - 1) grupos foram alinhados o programa MUSCLE e o alinhamento resultante utilizado como entrada para a execução do programa TABAJARA para a construção de HMMs de perfil curtos discriminativos para *Gokushovirinae*. Todos os modelos foram testados em buscas de similaridade usando o programa *hmmsearch*, com ou sem os valores de corte de pontuações atribuídos pelo TABAJARA, contra um conjunto de dados de validação composto por um único grupo. As médias aritméticas de sensibilidade e especificidade foram calculadas a partir dos resultados combinados dos múltiplos modelos obtidos para cada grupo.

Iteração	Nº de modelos	Com valores de corte		Sem valores de corte	
		Média da sensibilidade	Média da especificidade	Média da sensibilidade	Média da especificidade
Iteração 1	5	0,88	1,00	0,99	0,60
Iteração 2	5	0,89	1,00	0,99	0,61
Iteração 3	5	0,85	1,00	0,98	0,59
Iteração 4	4	0,90	1,00	0,99	0,60
Iteração 5	6	0,88	1,00	0,99	0,52
Iteração 6	5	0,90	1,00	0,99	0,57
Iteração 7	5	0,85	1,00	0,99	0,52
Iteração 8	5	0,90	1,00	0,99	0,50
Iteração 9	5	0,87	1,00	1,00	0,60
Iteração 10	5	0,92	1,00	1,00	0,63

Esse conjunto de resultados parece sugerir que a detecção específica de diferentes espécies de *Flavivirus* não é significativamente influenciada pelos diferentes conjuntos de treinamento testados. Contudo, tais resultados têm que ser interpretados com extrema cautela, uma vez que não necessariamente se aplicam a outros grupos virais, especialmente em casos nos quais haja taxa de divergências maiores entre seus membros. De fato, quando tentamos repetir o mesmo experimento para as sequências das demais subfamílias da família *Microviridae*, o programa TABAJARA sequer conseguiu gerar modelos. Além disso, os experimentos de validação cruzada relatados acima foram feitos com conjuntos de sequências bem definidos. Em aplicações metagenômicas mais amplas, é muito provável que a diversidade real de sequências de qualquer grupo viral seja maior do que aquela observada em conjuntos definidos como aqueles testados neste trabalho. Com isso, sequências ainda desconhecidas, e mais remotamente relacionadas às aquelas dos conjuntos de treinamento, podem vir a não ser detectadas. A vantagem de se utilizar os HMMs de perfil em relação a métodos baseados em alinhamento pareado é a de reduzir essa limitação, mas ela certamente pode persistir em graus variáveis.

5.4.5 Influência da amostragem no desempenho dos HMMs de perfil

Na seção 5.4.1 foram realizados alguns experimentos com o objetivo de verificarmos como a alteração de alguns parâmetros poderia afetar o potencial de detecção e discriminação dos modelos discriminativos curtos gerados pelo TABAJARA. Os resultados obtidos mostraram que os diferentes parâmetros do programa TABAJARA têm influência no desempenho dos modelos construídos. Contudo, o fator mais importante para se obter altos valores de sensibilidade e especificidade foi, sem dúvida, a utilização dos valores de corte de pontuação atribuídos pelo TABAJARA. Dentro do desenvolvimento e otimização da metodologia de construção de HMMs de perfil, decidimos também investigar a influência da amostragem do conjunto de treinamento no desempenho dos modelos gerados.

Um conjunto de treinamento ideal deve ser composto de sequências capazes de representar a diversidade do grupo de interesse. Utilizando novamente os grupos virais *Flavivirus* e *Microviridae* como modelos de estudo, o primeiro experimento consistiu na determinação do número ideal de sequências que um conjunto de treinamento deve ter para gerar modelos com altas taxas de sensibilidade e especificidade. Dentro do gênero

Flavivirus, escolhemos a espécie *Dengue virus* para fazer esta avaliação. Assim, o conjunto de 6.364 sequências de *Flavivirus* foi alinhado com o programa MUSCLE e o alinhamento resultante foi utilizado como conjunto de treinamento para a execução do TABAJARA, para a seleção de blocos de alinhamento discriminativos para *Dengue virus*. A partir dos pares de coordenadas de cada bloco, foi utilizado um *script* auxiliar implementado por nós para selecionar aleatoriamente um número n de sequências de *Dengue virus* cobrindo as regiões respectivas de cada um desses blocos e, em seguida, construir um HMM de perfil com o programa `hmmbuild` a partir de cada um deles. O valor de n variou de 2 a 40 em incrementos de 2. Para cada valor de n , esse experimento foi repetido 1.000 vezes. Todos os modelos construídos receberam valores de corte de pontuação, calculados utilizando-se a mesma heurística implementada no TABAJARA. Para esse cálculo, as sequências de *Dengue virus* utilizadas na construção do modelo foram definidas como controle positivo e o conjunto das 2.445 sequências das outras espécies do gênero *Flavivirus* como controle negativo.

Os modelos construídos foram submetidos a buscas de similaridade com a ferramenta `hmmsearch` contra o conjunto das 6.364 sequências de *Flavivirus* subtraído das sequências de *Dengue virus* utilizadas para a construção respectiva de cada modelo, aplicando-se os valores de corte de pontuação calculados para cada um. Com base nos resultados dessas buscas, valores de sensibilidade e especificidade foram calculados para cada modelo e, para cada valor de n , foram calculados os valores médios dessas métricas. A Figura 11A mostra os resultados de uma das regiões selecionadas como discriminativa pelo TABAJARA (coordenadas 10-117 do MSA).

Como podemos observar, todos os modelos foram altamente específicos, independentemente do número de sequências utilizadas para a construção dos mesmos, o que confirma que o TABAJARA é altamente eficaz em encontrar blocos de alinhamento que geram HMMs de perfil grupo-específicos. Quanto à sensibilidade, observamos uma relação entre o número de sequências utilizadas para a construção dos modelos e os valores médios de sensibilidade dos mesmos. Modelos construídos com apenas duas sequências apresentaram aproximadamente 60% de sensibilidade, enquanto que os modelos construídos com seis sequências apresentaram sensibilidade próxima de 90%. A partir de 20 sequências, a sensibilidade atingiu um valor de cerca de 98% e números superiores de sequências resultaram na saturação da curva de sensibilidade próximo a 100%. Concluindo, dentro desse modelo de estudo, 20 sequências parece ser um número

mínimo de sequências de *Dengue virus* necessárias para a construção de modelos discriminativos altamente sensíveis e específicos para essa espécie. As demais regiões definidas como discriminativas pelo TABAJARA apresentaram resultados bastante similares a esse (dados não mostrados), comprovando que esse número de sequências é bastante adequado para a construção de HMMs de perfil de *Dengue virus*. As áreas sombreadas do gráfico da Figura 11A representam os desvios padrão e refletem a variação de desempenho dos modelos causada pelas diferentes composições de sequências de cada iteração. Conforme esperado, quanto maior o número de sequências de cada amostra, menor é o efeito das diferenças de composição dessa amostra, com a resultante redução dos valores de desvios-padrão.

Figura 11. Efeito do tamanho da amostragem de sequências no desempenho de HMMs de perfil construídos com TABAJARA. Uma região de alinhamento específica de *Dengue virus* (coordenadas 10-117), derivada de um MSA composto por 6.364 sequências de poliproteína de *Flavivirus* foi selecionada. Um segundo teste foi realizado com uma região de alinhamento específica de *Gokushovirinae* (coordenadas 2034-2063), derivada de uma MSA composta por 1.866 sequências VP1 de *Microviridae* compreendendo representantes das subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae*. Blocos de alinhamento cobrindo as regiões selecionadas, compostas de um número n de sequências específicas de cada táxon aleatoriamente amostradas, com 1000 pseudoreplicatas para cada valor n , foram usados para construir os HMMs de perfil. Todos os modelos construídos para *Dengue virus* foram testados com o programa *hmmsearch* contra um conjunto de teste composto de 6.364 sequências de poliproteínas de *Flavivirus* (A), compreendendo uma grande diversidade de espécies virais. Para *Microviridae* (B), os HMMs de perfil foram validados contra um conjunto de teste composto por 1.866 sequências de VP1 de *Microviridae*, compreendendo membros das três subfamílias. Em todos os testes, os conjuntos de dados foram especificamente depletados das sequências de *Dengue virus* ou *Gokushovirinae* utilizadas para construir os respectivos modelos, e os valores de corte de pontuação sugeridos pelo TABAJARA aos modelos foram utilizados. As linhas de sensibilidade e especificidade representam médias aritméticas calculadas a partir dos resultados combinados de modelos gerados com as pseudoreplicatas de cada tamanho de amostragem. Áreas sombreadas indicam valores de desvio padrão.

Utilizando o mesmo protocolo anterior, o experimento foi repetido para a construção de HMMs de perfil curtos discriminativos para a subfamília *Gokushovirinae* (família *Microviridae*). Para isso, foi construído um MSA utilizando-se o conjunto de 1,866 sequências de VP1 de *Microviridae* com o programa MUSCLE. Com esse

alinhamento, o TABAJARA foi executado em modo Discriminativo para a subfamília *Gokushovirinae*. Da mesma forma que no teste anterior, as regiões selecionadas pelo TABAJARA foram processadas pelo *script* de forma a gerar modelos com diferentes quantidades e composições de sequências de *Gokushovirinae*. A quantidade de sequências de *Gokushovirinae* utilizadas para a construção dos modelos variou entre 2 a 100, com incrementos de 2 sequências e um total de 1.000 pseudoreplicatas.

Como podemos observar na Figura 11B (coordenadas 2034-2063 do MSA), semelhantemente ao observado para os modelos de *Dengue virus*, todos os HMMs de perfil gerados foram altamente específicos. Em relação à sensibilidade, também foi possível notar que existe uma relação entre a quantidade de sequências utilizadas para a construção dos modelos e os valores de sensibilidade observados. Quando os modelos foram construídos com 20 sequências de *Gokushovirinae*, a média da sensibilidade foi de aproximadamente 66%. Com a utilização de 44 sequências, a sensibilidade média dos modelos aumentou para aproximadamente 80%. A curva de sensibilidade apresentou uma saturação em torno desse valor, sem aumento significativo com o uso de um número maior de sequências.

Comparando-se os experimentos realizados com *Dengue virus* e *Gokushovirinae*, pôde-se observar duas diferenças importantes: (1) foi necessário um número maior de modelos em *Gokushovirinae* comparado a *Dengue virus* para se atingir o valor máximo de sensibilidade e (2) esse valor máximo de sensibilidade foi nitidamente menor em *Gokushovirinae*. Esses dois resultados podem ser atribuídos ao maior grau de divergência existente entre as sequências de *Gokushovirinae* (similaridade média de 59,6%), quando comparada à *Dengue virus* (similaridade média de 88%). Uma menor similaridade entre sequências implicaria na necessidade de se utilizar um número maior de sequências para representar de forma adequada essa diversidade. Por outro, o fato da curva de sensibilidade dos modelos de *Gokushovirinae* não atingir valores próximos a 100% sugere que a região utilizada para a construção dos modelos testados não está representada em cerca de 20% das sequências do conjunto de teste. Isso também é possivelmente consequência da maior divergência entre as sequências de *Gokushovirinae*. Em outras palavras, a detecção desse conjunto de 20% de sequências provavelmente só seria possível com modelos derivados de outras regiões da proteína VP1, as quais não necessariamente detectariam todas as demais 80% de sequências. Essa hipótese é corroborada pelos resultados apresentados na Figura 11B, que mostram que o uso

combinado de múltiplos modelos, de forma cumulativa, tem relação direta com o aumento da sensibilidade de detecção das sequências.

5.5 Discussão

Neste capítulo foi descrito o desenvolvimento de uma metodologia para o desenho racional de HMMs de perfil e sua implementação em uma ferramenta computacional, denominada TABAJARA. Essa ferramenta recebe como entrada um alinhamento de múltiplas sequências e permite a construção de modelos conservados entre todas as sequências do alinhamento e discriminativos para um determinado subconjunto, utilizando todo o comprimento das sequências ou sub-regiões específicas do alinhamento.

Como demonstrado nos experimentos descritos na seção 5.1 (Figuras 2 e 3), é possível se construir modelos discriminativos a partir de todo o comprimento das sequências presentes no MSA, mas, nesse caso, é absolutamente necessária a utilização de valores de corte de pontuação adequados. Para a atribuição desse valor de corte de pontuação, foi desenvolvida uma heurística que, baseada no conjunto de treinamento, calcula e atribui um valor de corte de pontuação personalizado aos modelos construídos. Como mostrado nos resultados, os modelos discriminativos gerados com proteínas inteiras foram altamente específicos para a classe desenvolvida quando esses valores de corte foram aplicados. Em alguns casos, contudo, a aplicação dos valores de corte sugeridos pode levar à diminuição da sensibilidade do modelo, especialmente quando os conjuntos de dados envolvem grupos de sequências mais divergentes, como é o caso de fagos da família *Microviridae*.

Modelos construídos a partir de proteínas inteiras devem ser utilizados para a prospecção de sequências em conjuntos de dados de sequências inteiras, uma vez que a aplicação desses modelos como ferramentas de detecção e classificação taxonômica requer o uso de valores de corte de pontuação, sob pena de se obter uma especificidade muito baixa. Como esses valores de corte de pontuação são calculados pelo alinhamento global da sequência, o uso de tais modelos em dados de leituras curtas resultaria em blocos de alinhamentos curtos cuja pontuação seria sempre muito inferior aos valores de corte e, portanto, seriam considerados negativos. Em função disso, o uso de HMMs de perfil com especificidade taxonômica em dados genômicos e metagenômicos de leituras curtas somente é viável com a utilização de modelos construídos a partir de blocos de

alinhamento relativamente curtos. Nossos resultados indicam que regiões entre 20 a 60 resíduos de aminoácidos são ideais para tal propósito.

Com base nisso, foi também desenvolvida uma metodologia utilizando diferentes métricas, seja para a estimativa de conservação de resíduos em todas as sequências ou, alternativamente, a medição do grau de discriminação de cada posição em um subconjunto de sequências em relação às demais sequências do alinhamento. Finalmente, elaborou-se uma metodologia para a seleção e extração das regiões mais significativas do alinhamento e a construção de HMMs de perfil curtos. Implementou-se ainda a heurística para o cálculo e atribuição de valores de corte de pontuação aos HMMs de perfil, bem como testes automáticos de validação dos modelos frente ao conjunto de treinamento.

Os resultados obtidos em diferentes testes, usando conjuntos de dados distintos, mostraram que os modelos curtos gerados pelo TABAJARA, sejam eles conservados ou discriminativos, foram capazes de detectar sequências nos conjuntos de teste com uma sensibilidade relativamente alta na maioria dos casos. Ainda, naqueles casos em que a sensibilidade de modelos individuais era baixa, foi demonstrado que o uso combinado de múltiplos modelos permitiu aumentar a sensibilidade para níveis altos. Conforme apresentado nos experimentos descritos na seção 5.4.1, independentemente do conjunto de parâmetros utilizados para a construção dos HMMs de perfil, a aplicação dos valores de corte de pontuação nas buscas de similaridade tornou os modelos altamente discriminativos para a identificação e classificação taxonômica de sequências. Isso comprova que os valores de corte de pontuação podem ser aplicados tanto em modelos longos como curtos de forma eficiente, garantindo uma alta especificidade.

Comparando-se o desempenho de modelos curtos com modelos derivados de proteínas inteiras contra um conjunto de teste de sequências longas, o desempenho dos dois tipos é similar desde que o grau de divergência entre as sequências seja baixo, como foi o caso dos modelos construídos para o gênero *Flavivirus* (Tabela 3). Quando o nível de divergência entre as sequências é grande, como é o caso dos fagos da família *Microviridae*, é mais vantajoso que modelos curtos sejam utilizados nas buscas de forma combinada (Tabelas 4 e 6), tanto no caso de modelos conservados como discriminativos.

Em relação ao tempo gasto nas buscas de similaridade em bancos de dados de sequências longas, podemos concluir que o uso de múltiplos modelos curtos para a detecção de sequências é mais vantajoso que a utilização de modelos derivados de proteínas inteiras, uma vez que o tempo gasto por esses modelos longos é sempre muito superior ao tempo total gasto pelo conjunto dos modelos curtos.

Além dos conjuntos de sequências virais utilizados nos testes descritos nesse trabalho, o TABAJARA tem sido utilizado em nosso laboratório para a construção de HMMs de perfil conservados e discriminativos para diversos conjuntos de sequências. Por exemplo, a ferramenta foi utilizada para a construção de modelos curtos conservados e discriminativos para famílias virais da ordem *Bunyavirales*, fagos da subfamília *Spounavirinae*, entre outros. Todos os modelos apresentaram excelentes resultados para a detecção e discriminação de sequências de interesse em conjuntos de dados, inclusive em metagenomas.

Outra aplicação da ferramenta foi para a construção de modelos curtos para a proteína endonuclease Cas1, que permitiram discriminar dois grupos de elementos genéticos móveis, os casposons e os CRISPRs e foram utilizados com sucesso para a identificação de elementos casposons com o *e-Finder*, outra ferramenta desenvolvida no presente trabalho (ver item 9). O TABAJARA também permitiu construir um HMM de perfil capaz de discriminar duas proteínas muito semelhantes, relacionadas à síntese de canabinoides: a CBDA (*cannabidiolic acid synthase*) e a THCA (*THCA synthase*). Esse modelo foi empregado com sucesso pela *Phylos Bioscience*, empresa norte-americana de biotecnologia de *Cannabis*, para a prospecção da diversidade dos respectivos genes em diferentes linhagens da planta.

Não foi possível realizar uma comparação dos modelos gerados pelo TABAJARA com aqueles publicamente disponíveis nas bases *vFam*, *viralOGs* e *pVOGs*, uma vez que os modelos dessas bases foram construídos com base em relação de ortologia de sequências proteicas e sem uma validação em termos de relações com grupos taxonômicos. De fato, a classificação de vírus ainda é um assunto ainda bastante controverso. A classificação oficial oferecida pelo ICTV é majoritariamente baseada em dados morfológicos e na composição de genomas virais. Apenas em anos recentes o ICTV tem incorporado dados moleculares em seus critérios de classificação.

Nesse sentido, foi recentemente desenvolvida a plataforma *Gravity* (Aiewsakun *et al.*, 2018B) para a classificação em larga de escala de grupos virais, baseando-se no relacionamento de dados genômicos. Nesse método, a classificação é obtida mediante o uso de análises de distância de Jaccard baseada em similaridade de genes e de dados de estrutura genômica que incluem a ordem e orientação dos genes. Os resultados obtidos pelos autores foram bastante promissores no sentido de se desenvolver uma taxonomia viral baseada em dados genômicas, como alternativa às classificações

tradicionais baseadas em morfologia. Em outro trabalho (Aiewsakun & Simmonds, 2018), os autores demonstraram que a abordagem do Gravity oferece um conjunto de regras que permitir fazer atribuições taxonômicas em níveis de famílias e ordens, ao mesmo tempo que preserva os princípios gerais e organizacionais da taxonomia do ICTV.

Conforme será apresentado no capítulo 6, já estamos utilizando o TABAJARA para criar HMMs de perfil para todos os grupos virais baseando-nos na taxonomia do ICTV. No futuro, uma vez estabelecida uma classificação taxonômica viral baseada em dados genômicos, como a proposta no Gravity, o TABAJARA poderá se tornar uma ferramenta muito útil e complementar para a produção em larga escala de modelos cobrindo todos os grupos virais taxonomicamente definidos. Com esses modelos, pode-se tornar muito mais simples o processo de prospecção de dados metagenômicos, bem como a detecção de vírus emergentes, relacionados aos grupos taxonômicos já conhecidos.

Chibani *et al.* (2019) reportaram o desenvolvimento do ClassiPhages 2.0, um sistema de classificação de fagos que partiu de um total de 8.721 genomas de fagos e, através de múltiplas etapas de processamento, resultou numa classificação nas diferentes famílias de fagos. Primeiramente, os autores geraram um conjunto de 5.920 HMMs de perfil, os quais foram utilizados em múltiplos conjuntos de dados de sequências de fagos para refinar os modelos e gerar uma matriz de pontuação que por sua vez foi utilizada para treinar uma rede neural artificial para classificar as sequências dentro das 12 famílias de fagos. Os testes de validação cruzada demonstraram uma especificidade muito alta, na maioria das vezes próxima de 100%. Contudo, a sensibilidade observada foi muito baixa, com somente três das doze famílias mostrando valores acima de 50% e seis famílias com sensibilidade inferior a 20%.

No caso da família *Microviridae*, os autores obtiveram uma sensibilidade de 25,4% em um conjunto de 734 sequências. Em nosso caso, utilizando um HMM de perfil construído com a sequência inteira da VP1, obtivemos uma sensibilidade de 70% com o uso de valores de corte de pontuação e 100% sem o uso (Tabela 4). Em termos de especificidade, nossos resultados não podem ser comparados com os de Chibani *et al.* (2019) porque não testamos o modelo contra sequências de outras famílias de fagos. Contudo, dada a grande divergência existente entre fagos, é muito provável que a especificidade seja próxima a 100% com o uso de valores de corte. Com o uso combinado de múltiplos HMMs de perfil curtos, derivados da VP1, obtivemos uma sensibilidade de

96,13% (Tabela 4). Esses resultados mostram mais uma vez que o TABAJARA é uma ferramenta robusta, capaz de construir modelos altamente sensíveis e específicos para um determinado conjunto de sequências, associados a grupos taxonômicos definidos.

6 Viral MinionDB

Como descrito rapidamente no capítulo anterior, o nosso grupo está utilizando o TABAJARA para gerar um banco de HMMs de perfil de todos as famílias virais disponíveis publicamente no NCBI, denominado Viral MinionDB. Para tanto, o estudante Wendel Hime Lino Castro, ex-estagiário de IC do nosso grupo, desenvolveu um programa na linguagem de programação Python que realiza as seguintes etapas: (1) *download* das sequências do grupo taxonômico de interesse; (2) alinhamento de cada grupo de múltiplas sequências; e (3) execução do TABAJARA para a construção de HMMs de perfil e (4) geração de relatórios.

A ferramenta gera ainda um banco de dados relacional acoplado a uma página *web*. Essa interface possibilita ao usuário realizar buscas por modelos específicos para cada grupo de vírus, permitindo especificar diferentes características dos modelos nas consultas, como, por exemplo, número mínimo de sequências e tamanho do bloco de alinhamento usado para sua construção. Esses modelos têm diferentes graus de especificidade taxonômica, para famílias, subfamílias e gêneros. Atualmente estamos depurando os programas relacionados ao banco, incrementando seus recursos e adicionando modelos de proteínas inteiras. Esse banco deverá ser liberado em breve à comunidade científica e se tornar um repositório público de HMMs de perfil de origem viral com especificidade taxonômica.

7 HMM-Prospector

A prospecção de sequências virais em conjuntos metagenômicos é geralmente realizada pela separação das sequências em grupos (*binning*) e/ou montagem das leituras, seguida de buscas de similaridade contra bancos de sequências conhecidas (Wommack *et al.*, 2012; Roux *et al.*, 2014). Como vimos nos capítulos anteriores, o uso de alinhamentos pareado apresenta limitações para a identificação de sequências virais mais divergentes. Em função disso, métodos baseados em perfis como os PSSMs (*Position-Specific Scoring Matrices*) e HMMs de perfil virais passaram a ser também utilizados para a prospecção de sequências.

Em geral, as buscas de similaridade utilizando HMMs de perfil são executadas com as ferramentas `hmmsearch` e `hmmScan`, ambas do pacote HMMER. Esses programas retornam os resultados em dois tipos de arquivos texto, em formato tabular e contendo os alinhamentos completos. No primeiro, as sequências identificadas pelos modelos são listadas juntamente com os valores de pontuação e `e-value` obtidos. Essas ferramentas permitem estabelecer somente um único valor de corte de pontuação em cada execução (usando-se o parâmetro `-T`), mesmo que seja utilizado um arquivo com múltiplos HMMs de perfil concatenados.

Com o intuito de facilitar as buscas de similaridade contra conjuntos de sequências utilizando HMMs de perfil, bem como permitir que os valores de corte inseridos nos modelos sejam utilizados de forma personalizada nas buscas, foi desenvolvida a ferramenta HMM-prospector. Esse programa lê a `tag CUTOFF SCORE` introduzida pelo TABAJARA no cabeçalho de cada modelo e executa o `hmmsearch` com o parâmetro `-T` específico para cada HMM de perfil, otimizando assim as buscas com arquivos de entrada contendo múltiplos modelos. O programa extrai as principais informações dos resultados da busca e permite filtrá-los de acordo com diferentes parâmetros definidos pelo usuário. Por fim, essa ferramenta também permite que resultados de execuções anteriores do `hmmsearch` sejam utilizadas como entrada. Esse tipo de execução é útil quando desejamos apenas filtrar resultados de buscas de similaridade previamente executadas, utilizando valores de `score` ou `e-value` específicos, sem ter que executar novamente as buscas com o `hmmsearch`, etapa mais lenta da análise. Nesse capítulo será descrita a implementação da ferramenta, bem como alguns experimentos executados para a sua validação.

7.1 Implementação da ferramenta

O HMM-prospector é uma ferramenta desenvolvida tanto para a análise de leituras brutas resultantes de um processo de sequenciamento ou de sequências montadas (nucleotídicas ou protéicas) utilizando HMMs de perfil, como para a filtragem de resultados obtidos por buscas executadas previamente com o `hmmsearch`. O programa foi implementado na linguagem Perl e utiliza como programas auxiliares o `transeq` do pacote EMBOSS (Rice *et al.* 2000), o `hmmsearch` do pacote HMMER (Eddy, 2011) e o `FASTQ-to-FASTA` do pacote FASTX-Toolkit. A Figura 12 apresenta um diagrama de fluxo de trabalho do programa.

A entrada de dados pode ser feita com um arquivo de sequências em formato FASTQ ou FASTA ou, ainda, no formato tabular gerado pelo `hmmsearch`. Os HMMs de perfil que serão utilizados na busca devem ser concatenados em um único arquivo, o qual também usado como entrada do programa. A critério do usuário, o programa pode utilizar os valores de corte sugeridos pelo TABAJARA nos modelos (*tag CUTOFF SCORE*) ou um valor de corte único de pontuação ou de *e-value* para todos os modelos. Caso o arquivo de entrada esteja em formato FASTQ, a ferramenta primeiramente converte esse arquivo para o formato FASTA utilizando o programa `FASTQ-to-FASTA` do pacote FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) e, em seguida, utiliza o programa `transeq` (do pacote EMBOSS) para a tradução conceitual em todas as fases de leitura. Se o arquivo de entrada estiver no formato FASTA e for composto de sequências nucleotídicas, o HMM-Prospector executa diretamente a tradução conceitual das sequências. O programa também pode utilizar como entrada um arquivo FASTA de sequências protéicas. Os HMMs de perfil são então submetidos a buscas de similaridade contra as sequências proteicas com o programa `hmmsearch`, com a opção de utilizar os valores de corte de pontuação de cada modelo. Por fim, se o arquivo de entrada estiver no formato tabular de saída do `hmmsearch`, a ferramenta somente utiliza valores de corte de pontuação (atribuídos de forma geral para todos os HMMs de perfil ou individuais de cada modelo sugeridos pelo TABAJARA) ou *e-value* para filtrar os resultados.

Ao final, o HMM-Prospector gera dois arquivos tabulares de saída: um com a listagem de todas as sequências identificadas como positivas juntamente com o modelo que a detectou e seus respectivos valores de *score* e *e-value*, e outro com a quantidade

de sequências detectadas por cada HMM de perfil utilizado na busca. No caso específico de HMMs de perfil da base vFam (Skewes-Cox *et al.*, 2014), o programa pode utilizar os dados de anotação dos modelos (disponível no *sítio web* da base vFam) um terceiro arquivo de saída é criado, classificando e quantificando os resultados de acordo com dados taxonômicos.

Figura 12 – Fluxo de trabalho do HMM-Prospector. O processo inicial da ferramenta depende do tipo de arquivo de entrada (A). Se o arquivo for do tipo FASTQ, a ferramenta converte o arquivo de entrada em FASTA com a ferramenta FASTQ-to-FASTA. Com o arquivo convertido ou se o arquivo de entrada for do tipo FASTA e composto por sequências nucleotídicas, o programa somente faz a tradução das sequências com o programa transeq. Já se o arquivo for do tipo FASTA composto por sequências protéicas, nenhum processamento inicial é necessário (B). Com o arquivo das sequências protéicas, o programa realiza buscas de similaridade com o hmmsearch utilizando os HMMs de perfil como entrada (B). Caso o arquivo de entrada seja do formato tabular resultante de um processamento anterior do hmmsearch, a etapa anterior é ignorada. Com os resultados das buscas, se um valor de corte de escore ou e-value for definido, o programa filtra os resultados e gera ao final duas tabelas com os resultados encontrados: uma com todos os resultados positivos e outra com a quantidade de sequências que tiveram similaridade com cada HMM de perfil utilizado na busca (C).

8 e-Finder

Alguns elementos, como profagos, operons e transposons, são formados por agrupamento específicos de genes que seguem uma determinada arquitetura. Essa arquitetura define um contexto sintênico, podendo ser bem conservada entre os indivíduos ou evoluir rapidamente por meio de rearranjos, inserções, deleções e duplicações (Medema *et al.*, 2013). A detecção de elementos multigênicos pode ser feita mediante a busca de um ou mais genes, mas o método torna-se muito mais específico se o contexto sintênico da localização desses genes puder ser definido. Para implementar tal abordagem, foi desenvolvida o e-Finder, uma ferramenta genérica para a detecção e extração de elementos multigênicos a partir de genomas montados. Nessa ferramenta, a identificação dos genes é feita com o uso de perfis HMM de sequências proteicas, o que permite representar a diversidade das proteínas codificadas por cada gene buscado. Além disso, a identificação do elemento no genoma considera critérios de composição e ordem gênicas definidas pelo usuário.

8.1 Implementação da ferramenta

O e-Finder foi também implementado na linguagem Perl e utiliza como programas auxiliares o `transeq` do pacote EMBOSS (Rice *et al.* 2000), o `hmmsearch` do pacote HMMER (Eddy, 2011) e os programas `blast` e `tblastn` do pacote BLAST+ (Camacho *et al.*, 2009). O programa aceita como entrada um conjunto de sequências nucleotídicas em formato FASTA. Como a busca é feita por elementos com múltiplos genes, é recomendável utilizar sequências montadas e não dados fragmentados na forma de leituras curtas. Para cada um dos genes a serem buscados, o usuário deve prover um arquivo composto por um ou vários HMMs de perfil concatenados. O programa permite ao usuário definir a quantidade mínima de genes presentes no elemento (por padrão, dois), um valor de distância intergênica máxima permitida entre pares de genes, de forma genérica (a mesma distância para todos) ou específica (uma distância distinta para cada par), e o tamanho mínimo do elemento que engloba todos os genes. É possível ainda definir uma ordem e orientação específicas para os genes.

A Figura 13 apresenta um diagrama que ilustra o processo de busca de elementos multigênicos com o programa e-Finder e a Figura 14 descreve o fluxo de trabalho da

ferramenta. O primeiro passo da execução da ferramenta consiste na tradução conceitual das sequências nucleotídicas de entrada em todas as fases de leitura com o programa `transeq`. Com as sequências traduzidas, o `e-Finder` executa buscas de similaridade contra os HMMs de perfil com o programa `hmmsearch`. Após essa etapa, cada sequência positiva passa a ser verificada individualmente, com o objetivo de identificar possíveis elementos. Assim, o programa verifica se a quantidade mínima de genes foi encontrada na sequência. Em seguida, caso uma ordem gênica obrigatória tenha sido definida pelo usuário, o programa verifica se a ordem dos genes encontrados na sequência está de acordo com esta ordem. Em caso positivo, as coordenadas das ORFs são identificadas, definidas pelas sequências delimitadas por códons de terminação. Posteriormente, o programa busca o primeiro códon de iniciação na região 5' do gene (permitindo o uso de códons de iniciação alternativos) e define a região codificadora (CDS) de cada gene. Com base nessas coordenadas, o programa verifica se as distâncias intergênicas obedecem aos parâmetros definidos e se a região que engloba todos os genes apresenta o tamanho mínimo definido. Os elementos válidos podem ser adicionados de regiões flangeadoras 5' e 3' terminais com comprimento também definido pelo usuário. Todas as sequências nucleotídicas dos elementos são então extraídas e salvas em arquivos para análises posteriores. O `e-Finder` também armazena as proteínas encontradas nos elementos, sendo um arquivo salvo para cada tipo de proteína. Por fim, o programa gera um arquivo tabular com as informações de todos os elementos encontrados, incluindo os nomes dos genes, coordenadas de início e fim, orientação, etc.

Figura 13 – Execução do e-Finder para a busca de um elemento multigênico composto por quatro genes. Inicialmente o e-Finder executa buscas de similaridade dos HMMs perfil de cada gene contra as sequências de entrada traduzidas com o programa *hmmsearch*. A primeira verificação realizada é se a quantidade mínima de genes foi encontrada no genoma. Caso o usuário tenha definido uma ordem obrigatória para os genes, o programa verifica se a ordem dos genes encontrada pelas buscas é compatível com o padrão definido. Logo após, o programa identifica as ORFs e as CDS de cada um dos genes, e então verifica se a distância entre os genes está de acordo com a distância máxima definida. Por fim, o programa verifica se o tamanho do elemento obedece ao tamanho mínimo estabelecido. Em caso positivo, o elemento é extraído e as regiões flangeadoras adicionadas ao elemento.

Figura 14 – Fluxo de trabalho do e-Finder. O programa aceita como entrada sequências montadas em formato FASTA e um HMMs de perfil para de cada um dos genes a serem buscados (A). Como primeiro passo da execução, o programa verifica se as sequências de entrada estão traduzidas (B). Em caso negativo, o programa transeq (pacote EMBOSS) é utilizado para traduzir as sequências em todas as fases de leitura. O e-Finder então executa buscas de similaridade dos HMMs perfil contra as sequências traduzidas com programa hmsearch. A partir dos resultados obtidos, uma série de verificações é realizada individualmente para cada sequência (C). Inicialmente, o programa confere se a quantidade mínima de genes definida pelo usuário foi encontrada. Em caso positivo, a análise segue. Se uma ordem obrigatória para os genes tiver sido definida, o programa irá verificar se a ordem dos genes encontrada pelas buscas é compatível com o padrão definido. Logo após, o programa identifica as ORFs e as CDS de cada um dos genes, e então checka se a distância entre os genes está de acordo com a distância máxima definida. Continuando, o programa verifica se o tamanho do elemento obedece ao tamanho mínimo estabelecido. Em caso positivo, o elemento é extraído e as regiões flanqueadoras adicionadas a ele. As sequências de cada uma das proteínas são também extraídas dos elementos e salvas em arquivos individuais. Por fim (D), os arquivos gerados arquivos em formato FASTA com os elementos encontrados nas buscas e para cada tipo de proteínas proteína encontrados. Um arquivo tabular com informações dos elementos encontrados é igualmente gerado.

8.2 Validação da ferramenta

Com o objetivo de validar a metodologia implementada no e-Finder, utilizamos dois modelos de estudo compreendendo elementos multigênicos: (1) elementos genéticos móveis da família dos casposons (Krupovic *et al.*, 2014; 2016; 2017) e (2) profagos da família *Microviridae*. Para ambos os elementos, a prospecção foi realizada na base *Patric* (Wattam *et al.*, 2014 - <https://www.patricbrc.org/>), um banco de genomas montados de bactérias e arqueias publicamente disponível. A versão dessa base utilizada para a busca desses elementos era composta de 135.979 genomas. Nessa seção serão descritos os dois estudos e os resultados obtidos.

8.2.1 Busca de Casposons em genomas de bactérias e arqueias

Como descrito anteriormente, casposons são elementos transponíveis encontrados em genomas de bactérias e arqueias e são constituídos de um número variável de genes, incluindo sempre os que codificam as proteínas endonuclease Cas1 e DNA Polimerase B (PolB) (Krupovic *et al.*, 2014; 2016; 2017). Visando realizar um levantamento de casposons no banco de sequências genômicas *Patric*, inicialmente foram construídos HMMs de perfil para Cas1 e PolB.

Para essa construção, foram utilizados conjuntos de sequências de referência das proteínas endonuclease Cas1 e RNA polimerase B (PolB), enviadas pelo Dr. Mart Krupovic (Instituto Pasteur de Paris, França). Como a proteína Cas1 apresenta ortólogos em todos os elementos CRISPR, o nosso objetivo foi construir modelos capazes de diferenciar casposons de CRISPRs. Dessa forma, as sequências de Cas1 dos casposons foram alinhadas com o programa *MUSCLE* e juntamente com sequências de referência de Cas1 dos CRISPRs e esse alinhamento foi utilizado para a construção de modelos curtos discriminativos de Cas1 de casposons com o *TABAJARA*. Os modelos gerados foram validados por buscas de similaridade com o programa *hmmsearch* contra o conjunto de treinamento e demonstraram ser capazes de detectar de forma específica apenas sequências de Cas1 de casposons.

Em relação à proteína PolB, devido ao fato dessa proteína estar presente de forma ubíqua em todas as bactérias e arqueias, utilizou-se uma abordagem diferente. Primeiramente as sequências de referência de PolB de casposons foram alinhadas com o

MUSCLE e usadas para a construção de modelos curtos com o TABAJARA no modo Conservação. Esses modelos foram então testados contra um banco de cerca de 2.000 sequências de PolB de bactérias e arqueias, que correspondem às sequências utilizadas originalmente para a construção do modelo PF03175 (*DNA polymerase type B, organellar and viral*) do banco Pfam. Para isso, os modelos foram submetidos a buscas de similaridade com *hmmsearch* e as sequências positivas foram então alinhadas com as sequências de referência de PolB de casposons e algumas sabidamente não pertencentes a estes elementos. O alinhamento foi usado para uma análise filogenética com o programa FastTree e somente as sequências pertencentes ao clado das proteínas de casposons foram consideradas verdadeiro-positivas. Com isso, obteve-se um modelo capaz de identificar especificamente as proteínas PolB de casposons, discriminando-as dos ortólogos de bactérias e arqueias.

Para a prospecção de casposons nos genomas da base PATRIC com o e-Finder, foi definido um contexto sintênico com a presença obrigatória dos genes *CasI* e *Polb*, sem ordem gênica específica e com uma distância intergênica máxima de 40kb. Para a extração dos elementos, foi definido que regiões flangeadoras de 10 kb seriam adicionados aos mesmos. Para a busca, foram usados cinco HMMs de perfil de CasI e um modelo de PolB.

Como resultado, foram obtidos no total 91 elementos válidos pertencentes a 79 genomas procariotos distintos. Esse valor é muito superior ao número de casposons descritos na literatura (Krupovic *et al.*, 2014; 2016; 2017) e corresponde a mais do que o dobro do número total conhecido (Mart Krupovic, comunicação pessoal). As sequências protéicas de CasI de todos os elementos foram utilizadas em uma reconstrução filogenética. Para isso, essas sequências foram alinhadas utilizando-se o programa MUSCLE. O alinhamento gerado foi então passado ao programa *fasttree* (Price *et al.*, 2010) para a reconstrução filogenética utilizando-se o modelo de evolução Whelan-And-Goldman (WAG) e uma estimativa γ discreta com 20 categorias para as taxas de evolução de aminoácidos nos sítios. A árvore gerada confirmou a existência de quatro famílias de casposons (Figura 15).

Na família 1, o e-Finder encontrou 18 casposons, sendo todos pertencentes a arqueias do filo Thaumarchaeota, enquanto Krupovic *et al.* (2016) descreveram apenas cinco elementos dessa família. Na família 2, foram encontrados oito casposons além dos definidos por Krupovic *et al.* (2016), todos derivados de arqueias da

ordem Methanosarcinales. Na família 3 também foram encontrados casposons em genomas de bactérias que não estão descritas por Krupovic *et al.* (2016). Por fim, na família 4, foi encontrado um elemento a mais em *Methanosarcina* sp. E03.2. Em relação ao arranjo dos dois genes, foram encontrados arranjos gênicos alternativos nas famílias 1, 2 e 3. Nas famílias 2, 3 e 4 há um arranjo predominante, sendo <cas pol> o mais frequente nas famílias 2 e 3, enquanto que o arranjo pol>cas> é o mais observado na família 4. Nos elementos da família 1 foram encontrados dois arranjos: cas>pol> e pol><cas, sendo o primeiro detectado pela primeira vez.

8.2.2 Busca de profagos de *Microviridae* em genomas da base PATRIC

O segundo estudo de caso envolveu o levantamento de profagos da família *Microviridae* nos genomas do banco Patric. Para esse levantamento, foram definidos como marcadores as proteínas VP1 (proteína principal de capsídeo), VP2 (a proteína piloto de DNA / *minor spike protein*) e VP4 (proteína de iniciação de replicação) dessa família, por serem ubíquas em todas as subfamílias de *Microviridae* (Roux *et al.*, 2012). Para a construção dos HMMs de perfil, foi utilizado um conjunto de 1.893 sequências da proteína VP1, 1.877 sequências da proteína VP2 e 1.893 sequências de VP4, todos gentilmente cedidos pelo Prof. François Enault (Université Clermont Auvergne, França). Os três conjuntos continham sequências pertencentes às subfamílias *Alpavirinae*, *Gokushovirinae*, *Pichovirinae* e *Bullavirinae*.

Para a construção dos modelos de cada uma das proteínas, os conjuntos foram individualmente alinhados com a ferramenta MUSCLE e os alinhamentos gerados utilizados para a construção de modelos curtos conservados com o TABAJARA. Como os vírus da família *Microviridae* são bastante divergentes, os alinhamentos múltiplos das sequências das três proteínas foram altamente fragmentados, o que não permitiu ao TABAJARA identificar regiões conservadas para a construção de modelos. Em razão disso, resolvemos então construir modelos curtos conservados para cada subfamília de forma individual. No caso da subfamília *Alpavirinae*, que é a mais divergente de todas, foram construídos modelos para diferentes grupos de proteínas definidos por filogenia molecular pelo Prof. François Enault. Ao final, foi possível construir modelos conservados para todas as subfamílias das três proteínas. A Tabela 9 lista a quantidade de

Figura 15 – Árvore filogenética construída com sequências da proteína Cas1 (endonuclease Cas) dos casposons identificados pelo e-Finder nos genomas de bactérias e arqueias da base Patric. As famílias de casposons conforme a classificação de Krupovic *et al.* (2016) estão destacadas em diferentes cores, bem como o arranjo dos genes cas e pol.

modelos construídos para cada proteína. No total, foram construídos 25 HMMs de perfil para a proteína VP1, 16 para a proteína VP2 e 24 para a proteína VP4.

Tabela 9 – Quantidade de HMMs de perfil curtos conservados gerados para cada uma das subfamílias de *Microviridae* com o TABAJARA.

Subfamílias	Número de modelos		
	VP1	VP2	VP4
<i>Alpavirinae</i>	18	11	18
<i>Bullavirinae</i>	2	3	2
<i>Gokushovirinae</i>	2	1	2
<i>Pichovirinae</i>	3	1	2
Total	25	16	24

Os HMMs de perfil construídos foram então utilizados para a busca de profagos da família *Microviridae* com o programa e-*Finder* sem a definição de uma ordem gênica obrigatória, apenas especificando que pelo menos dois genes deveriam estar presentes no elemento e que a distância intergênica máxima entre os genes deveria ser de até 3 kb. Todos os elementos encontrados foram extraídos com sequências flangeadoras de 20 kb. Como o objetivo desse estudo era a detecção de profagos inseridos em genomas de procariotos, os elementos encontrados foram posteriormente filtrados por tamanho. Como os genomas de *Microviridae* possuem um tamanho variando entre 4 e 6,5 kb (Roux *et al.*, 2012), decidiu-se descartar todos os elementos menores do que 9 kb, de forma a garantir que os elementos encontrados estivessem inseridos no genoma de bactérias e/ou arqueias.

Com resultado, foram obtidos 104 elementos candidatos a profagos, que foram submetidos a um *pipeline* de anotação funcional com a plataforma de anotação EGene2 (Durham *et al.*, 2005). Como a proteína VP1 de *Microviridae* é a mais conservada dentre todas as demais e é comumente utilizada como marcador filogenético, as sequências das proteínas VP1 dos elementos encontrados pelo e-*Finder* foram utilizadas em conjunto com as sequências de VP1 descritas por Roux *et al.* (2012) e com algumas sequências do conjunto de dados do Prof. François Enault para uma análise filogenética. A partir da curadoria manual dos resultados obtidos pelo processo de anotação, com o objetivo de analisar a sintenia dos genes encontrados, juntamente com a análise da árvore filogenética gerada, foram selecionados 80 candidatos a profagos, sendo 28 deles pertencentes a

subfamília *Alpavirinae*, 37 a *Gokushovirinae*, 15 a *Pichovirinae*. A Tabela 10 lista os hospedeiros e as quantidade de profagos encontrados em cada subfamília.

Tabela 10 – Lista dos *profagos* de *Microviridae* encontrados pelo e-Finder para as subfamílias *Alpavirinae*, *Gokushovirinae* e *Pichovirinae*. Para cada subfamília são listados os hospedeiros e a respectiva quantidade de profagos encontrados.

Subfamília	Gênero do hospedeiro	Nº de profagos
<i>Alpavirinae</i>	<i>Bacteroides</i>	7
	<i>Prevotella</i>	12
	<i>Bacteroides</i>	1
	<i>Bacteroides</i>	4
	<i>Alistipes</i>	3
	<i>Dysgonomonas</i>	1
<i>Gokushovirinae</i>	<i>Escherichia</i>	29
	<i>Kosakonia</i>	1
	<i>Enterobacter</i>	1
	<i>Salmonella</i>	1
	<i>Acetobacter</i>	1
	<i>Erysipelato</i>	1
	<i>Clostridium</i>	1
	<i>Flavonifractor</i>	1
	<i>Treponema</i>	1
<i>Pichovirinae</i>	Não definido (ordem Bacteroidales)	2
	<i>Parabacteroides</i>	6
	<i>Elizabethkingia</i>	6
	Não definido (classe Elusimicrobia)	1

Como dois elementos encontrados pelo e-Finder não apresentaram similaridade significativa contra os genes de VP1 de *Microviridae*, foi realizada uma investigação mais detalhada, que revelou similaridade desses genes com proteínas não anotadas de algumas bactérias das ordens Rhodobacterales e Rhizobiales. Essas sequências foram então alinhadas e alguns HMMs de perfil foram construídos. O mesmo procedimento foi feito para as proteínas VP2 e VP4. Os conjuntos de modelos dessas três proteínas, derivadas desse novo grupo, de fagos foram então usados com o e-Finder na base PATRIC e permitiram detectar um número maior de profagos desse grupo.

Os novos elementos encontrados foram também anotados com o EGene2 e as proteínas VP1 dos mesmos, em conjunto com as proteínas de VP1 utilizadas

anteriormente, foram utilizadas em uma nova reconstrução filogenética. Após uma etapa de curadoria manual, foram encontrados 17 elementos em bactérias das ordens Rhodobacterales e Rhizobiales que, de acordo com os resultados de anotação funcional e filogenia, sugerem que possam constituir uma nova subfamília. Estamos propondo que esta nova subfamília seja nomeada *Katanvirinae* (do Hebraico: katan = pequeno). A árvore filogenética construída com a proteína VP1 dos elementos encontrados é ilustrada na Figura 16. Como podemos observar, todas as subfamílias formaram clados individuais, e as sequências de *Katanvirinae* formam um clado irmão da subfamília *Bullavirinae*, sugerindo que esses elementos são diferentes dos já conhecidos e, muito provavelmente, descrevem uma nova subfamília. A Tabela 11 lista os hospedeiros e as respectivas quantidades de profagos de *Katanvirinae* encontrados pelo e-Finder.

Tabela 11 - Lista dos profagos de *Microviridae* encontrados pelo e-Finder para a nova subfamília *Katanvirinae*. Nela são listados os hospedeiros e a respectiva quantidade de profagos encontrados.

Ordem	Gênero do hospedeiro	Nº de profagos
Rhodobacterales	<i>Oceanicola</i>	1
	<i>Pseudooceanicola</i>	1
	<i>Rhodovulum</i>	3
	<i>Ruegeria</i>	1
	<i>Rhodobacter</i>	3
	<i>Phaeobacter</i>	1
	<i>Paracoccus</i>	1
	<i>Nioella</i>	1
Rhizobiales	<i>Devosia</i>	2
	<i>Rhizobium</i>	1
	<i>Agrobacterium</i>	1
	<i>Labrys</i>	1

Como descrito anteriormente, na literatura há a descrição de alguns profagos de *Microviridae* apenas na subfamília *Alpavirinae* (Krupovic & Forterre, 2011). Os resultados da anotação funcional e curadoria manual dos elementos encontrados pelo e-Finder sugerem fortemente que tenhamos detectado pela primeira vez profagos das subfamílias *Gokushovirinae* e *Pichovirinae*, bem como de uma nova subfamília, a

Katanvirinae. Além da identificação dos profagos em si, tais resultados revelaram novos hospedeiros bacterianos para esses fagos, contribuindo para um melhor conhecimento das relações fago-hospedeiro em diferentes ambientes.

8.3 Discussão

O objetivo do desenvolvimento do programa *e-finder* foi a prospecção de elementos multigênicos em conjuntos de sequências montadas, de conjuntos de dados genomas ou metagenomas. O *e-finder* foi criado como um programa genérico, de forma a permitir que qualquer elemento multigênico possa ser detectado utilizando-se HMMs de perfil e aplicando-se um conjunto de parâmetros para descrever a arquitetura dos genes nas sequências-alvo. Como os HMMs de perfil são mais sensíveis do que métodos baseados em alinhamento pareado (Park *et al.*, 1998), eles permitem a detecção de elementos contendo genes mais divergentes.

Para testar o potencial da ferramenta, foram selecionados dois tipos de elementos multigênicos: casposons e profagos da família *Microviridae*. Como vimos nos resultados de dois experimentos (itens 8.2.1 e 8.2.2, respectivamente), a ferramenta permitiu de maneira muito simples e direta a descoberta de um grande número de novos elementos em ambos os casos, inclusive em genomas hospedeiros nunca descritos. Esse conjunto de resultados demonstra inequivocamente o grande potencial do programa *e-Finder* para a descoberta de elementos compostos por múltiplos genes em diferentes contextos sintênicos.

Existem duas outras ferramentas publicamente disponíveis que fazem a busca de elementos formados por agrupamento de genes em conjuntos de sequências (Tabela 12). A primeira delas, denominada *MultiGeneBlast* (Medema *et al.*, 2013), é uma ferramenta que realiza buscas de similaridade com a ferramenta BLAST para a detecção de grupos de genes e operons em genomas. O programa recebe como entrada genomas anotados em formato Genbank e o conjunto de genes a ser buscado. Esses genes podem ser passados para a ferramenta em arquivos de sequências proteicas em formato FASTA ou, ainda, como coordenadas da região que engloba o grupo de genes de interesse no arquivo de anotação do genoma (formatos Genbank ou EMBL). O *MultiGeneBlast* executa então buscas de similaridade com o programa BLAST

Figura 16 - Árvore filogenética construída com seqüências da proteína VP1 da família *Microviridae*. Seqüências da proteína VP1 dos profagos identificados pelo e-Finder foram unidas a proteínas VP1 conhecidas e utilizadas em uma análise filogenética. As seqüências destacadas em vermelho pertencem à subfamília *Alpavirinae*, em azul *Gokushovirinae*, em verde *Pichovirinae*, em cinza *Bullavirinae* e em amarelo para a nova subfamília *Katanvirinae*.

(blastp) entre as sequências proteicas de entrada e aquelas identificadas pela anotação dos genomas. Para o agrupamento dos genes, ou seja, para a identificação dos elementos, a distância a cada par de genes é verificada. Para que os genes sejam considerados como membros do mesmo grupo, essa distância deve ser menor que uma distância máxima permitida. Para cada elemento encontrado é calculado um valor de pontuação que leva em conta o número de genes encontrado e sua sintenia. Todos os elementos encontrados são disponibilizados para o usuário em arquivos nos formatos XHTML (para análise interativa dos elementos encontrados a partir de um navegador internet) e TXT, em ordem decrescente de pontuação. O MultiGeneBlast pode ser executado tanto em linha de comando como utilizando uma interface gráfica.

Em comparação ao e-Finder, podemos identificar algumas limitações no programa MultiGeneBlast. Primeiramente, o MultiGeneBlast aceita somente sequências anotadas, enquanto o e-Finder recebe como entrada sequências nucleotídicas de genomas/metagenomas montados, sem necessidade de anotação funcional. Caso seja de interesse do usuário, a anotação dos elementos encontrados pode ser posteriormente realizada, uma vez que o e-Finder disponibiliza um arquivo FASTA com as sequências nucleotídicas desses elementos (conforme foi feito com os casposons e profagos de *Microviridae* encontrados nas análises descritas na seção 8.2). Outra limitação se relaciona ao processo de busca dos genes de interesse nos genomas. Enquanto o MultiGeneBlast utiliza buscas de similaridade por alinhamento pareado (via programa BLAST), o e-Finder utiliza HMMs de perfil, permitindo que sequências mais distantemente relacionadas sejam identificadas. Diferentemente do e-Finder, o MultiGeneBlast não permite que a orientação dos genes e o tamanho mínimo dos elementos sejam usados como critérios de busca. Essa ferramenta também não considera a topologia dos genomas e não disponibiliza as sequências nucleotídicas dos elementos, nem as sequências das proteínas encontradas.

O MacSyFinder (Abby *et al.*, 2014) é outro programa desenvolvido para a identificação de elementos compostos por grupamentos de genes (Tabela 12). Esse programa compreende um *framework* desenvolvido para a modelagem de sistemas moleculares e que, da mesma forma que o e-Finder, utiliza HMMs de perfil para detectar os componentes dos sistemas. Essa ferramenta utiliza como entrada um conjunto composto de sequências proteicas em formato FASTA, que pode ser de três tipos: sequências proteicas sem informação de ordem gênica e origem genômica (ex. proteínas de metagenomas); sequências de proteínas de um único genoma sem a informação de

ordem gênica; e proteínas de um único replicon², que estão ordenadas de acordo com a posição de seus genes no genoma. O usuário tem ainda a opção de informar um arquivo com múltiplos replicons únicos pertencentes a diferentes genomas.

A definição da sintenia (ou arquitetura) dos genes no MacSyFinder é feita a partir de um arquivo XML, no qual são descritas informações de quais genes serão buscados (representados pelos HMMs de perfil) com seus respectivos tipos (obrigatórios, opcionais ou proibidos), a quantidade mínima de genes e a distância máxima permitida entre eles. O MacSyFinder realiza buscas de similaridade de todos os HMMs de perfil definidos nos arquivos XMLs contra as sequências de entrada.

Quando a ferramenta recebe como entrada sequências proteicas sem informação de ordem gênica e origem genômica, o programa simplesmente retorna as sequências que apresentaram similaridade com os modelos de cada arquitetura utilizada nas buscas. Quando sequências de proteínas de um único genoma sem a informação de ordem gênica são utilizadas, os resultados encontrados para cada modelo de sintenia consideram a quantidade mínima de genes definida pelo modelo. Por fim, quando as sequências são proteínas de replicons ordenados, os genes são inicialmente agrupados de acordo com uma distância intergênica máxima permitida. Nesse caso, diferentemente do e-Finder e do MultigeneneBlast, que medem essa distância em números de bases, o MacsyFinder considera essa distância como sendo a quantidade de genes que não apresentaram similaridade com nenhum HMM de perfil informado. Para cada grupo de genes encontrado, é verificado se a arquitetura de seus genes é compatível com algum dos modelos definidos no arquivo XML de entrada. Os resultados encontrados são disponibilizados em arquivos de texto e em formato JSON, o qual pode ser visualizado na ferramenta gráfica disponibilizada pelo MacsyFinder.

Comparando-se o MacsyFinder com o e-Finder (Tabela 12), a primeira diferença se relaciona com os dados de entrada. Enquanto o e-Finder aceita sequências nucleotídicas dos genomas de interesse, as sequências de entrada no MacsyFinder têm que ser proteicas. Ou seja, da mesma forma que no MultigeneBlast, é preciso que algum processamento inicial seja realizado no conjunto de entrada antes que as sequências sejam passadas para a ferramenta. Apesar das duas ferramentas utilizarem HMMs de perfil para a detecção dos genes, a análise da sintenia dos elementos encontrados no MacsyFinder considera apenas os resultados das buscas de

² Unidade de DNA na qual ocorre um ato de replicação (Lewin, 2008), ou seja, é uma região do DNA que se duplica.

similaridade. No *e-Finder*, após a validação da quantidade mínima de genes e da ordem, as CDS de cada uma das regiões de similaridade são definidas, e as análises prosseguem considerando esses elementos. Isso evita que regiões falso-positivas sejam utilizadas, pois nesse caso os genes reais são utilizados no processo de validação dos elementos.

Outra diferença se relaciona com a utilização de valores de corte de pontuação nas buscas de similaridade com os HMMs de perfil. Enquanto o *e-Finder* permite que os valores de corte de pontuação inseridos nos modelos sejam utilizados de forma personalizada, o *MacsyFinder* apenas permite que valores de corte de *e-value* e cobertura mínima sejam aplicados nas buscas de forma geral. Diferentemente do *MacsyFinder*, que utiliza apenas um HMM de perfil para cada gene a ser buscado, o *e-Finder* permite utilizar múltiplos modelos para representar um gene (ex.: modelos curtos construídos com diferentes regiões dos alinhamentos múltiplos das proteínas). Ao contrário do *e-Finder* e semelhantemente ao *MultiGeneBlast*, informações a respeito de orientação dos genes e tamanho mínimo do elemento não são usados como critérios de busca pelo *MacsyFinder*. Por fim, as sequências nucleotídicas dos elementos e das proteínas encontradas não são disponibilizadas por essa ferramenta. A Tabela 12 ilustra de forma comparativa as diferentes características do *e-Finder* e das ferramentas *MultiGeneBlast* e *MacsyFinder*.

Tabela 12 – Comparação entre as características das ferramentas e-Finder, MultiGeneBlast e MacsyFinder.

Característica	e-Finder	MultiGeneBlast	MacSyFinder
Arquivo de entrada	FASTA – sequências nucleotídicas de genomas/ metagenomas montados.	Genbank – genomas anotados.	FASTA – conjunto de sequências proteicas.
Ordem dos genes	Pode ser especificada pelo usuário (opcional)	Definida pela anotação ou ordem dos genes no arquivo FASTA	Definida no XML
Orientação dos genes	Pode ser especificada pelo usuário (opcional)	Não	Não
Número mínimo de genes	Sim	Não	Sim
Uso de HMMs de perfil	Sim	Não	Sim
Diferencia entre topologia linear e circular	Sim	Não	Sim
Extração do elemento	Sim	Não	Não
Arquivos de saída	Arquivos FASTA com sequências nucleotídicas e proteicas dos elementos encontrados. Arquivo CSV com informações sobre os elementos e seus genes.	Arquivos em formato de texto e XHTML com informações sobre os genes encontrados	Arquivos contendo o número de arquiteturas detectadas, e informações sobre cada gene detectado
Permite a adição de regiões flanqueadoras ao elemento extraído	Sim	Não	Não
Uso de múltiplos códigos genéticos	Sim	Não	Não
Uso de HMMs de perfil derivados de proteínas inteiras	Sim	Não	Sim
Uso de HMMs de perfil construídos com sequências parciais	Sim	Não	?
Uso de múltiplos HMMs de perfil por gene	Sim	Não	Não
Uso de valores de corte de pontuação personalizados para cada HMM de perfil	Sim	Não. Permite definir valores de cobertura e identidade mínima para as buscas com o BLAST	Não. Permite definir um valor mínimo de <i>e-value</i> e de cobertura do alinhamento.
Tamanho mínimo do elemento	Sim	Não	Não
Máxima distância intergênica	Sim, medida em pares de bases.	Sim, medida em pares de bases.	Sim, medida em números de genes sem similaridade com os HMMs de perfil

9 Uma abordagem integrada para a prospecção e classificação de sequências virais

Os programas desenvolvidos nesse projeto podem ser utilizados isoladamente ou em combinação para detectar e discriminar sequências conhecidas ou remotamente relacionadas. Em conjunto com o montador progressivo GenSeed-HMM (Alves *et al.*, 2016) desenvolvido pelo nosso grupo, esses programas permitem estabelecer uma abordagem integrada para a detecção, classificação e descoberta de sequências mais distantes em conjuntos de dados, sejam eles compostos de leituras brutas ou dados montados (Figura 17).

Partindo-se de um ou mais alinhamentos de múltiplas sequências de grupos de sequências de interesse, pode-se construir HMMs de perfil com o TABAJARA. Com esses modelos, é possível fazer uma prospecção inicial de conjuntos de dados de sequências com o programa HMM-Prospector, de forma a se identificar quais organismos estão presentes nas amostras e quais modelos apresentam melhores resultados, para que os mesmos possam ser utilizados em análises posteriores. Com os HMMs de perfil selecionados, a abordagem pode seguir dois caminhos: se o conjunto de dados for composto de sequências não montadas, pode-se utilizar os modelos como sementes em montagens progressivas alvo-específica com o GenSeed-HMM. Caso o conjunto seja de sequências montadas, é possível buscar elementos multigênicos em contexto sintênico com o programa e-Finder.

A abordagem integrada proposta aqui é utilizada em várias aplicações em nosso laboratório. Por exemplo, nos estudos de caso executados para a validação do e-Finder, os modelos utilizados nas prospecções foram todos construídos com o TABAJARA e posteriormente utilizados para a prospecção de dados em conjuntos de sequências montadas. Quando estamos buscando por sequências virais em conjuntos de sequências não montadas, como metagenomas, HMMs de perfil construídos pelo TABAJARA para diferentes grupos virais podem ser utilizados em uma prospecção inicial com o HMM-prospector. A partir dos resultados, é possível identificar quais vírus estão presentes nas amostras e quais modelos devem ser utilizados em montagens alvo-específicas com o programa GenSeed-HMM para a reconstrução dos genomas desses vírus.

Figura 17 – Fluxo de trabalho da abordagem integrada de produção e uso de HMMs de perfil. A partir de um alinhamento de múltiplas sequências HMMs de perfil podem ser construídos com o programa TABAJARA (A). Esses modelos podem então ser utilizados para prospectar conjuntos de sequências com o HMM-Prospector. Os modelos selecionados podem ser usados como sementes em montagem progressiva com o programa GenSeed-HMM a partir de dados não montados (B) ou, ainda, para a busca de elementos multigênicos com o programa e-Finder em sequências genômicas montadas (C).

10 Conclusões e trabalhos futuros

O objetivo inicial deste projeto (como descrito no item 4) foi o desenvolvimento de metodologias e ferramentas para a construção de HMMs de perfil e a aplicação desses modelos na prospecção de sequências em conjuntos de dados genômicos e metagenômicos. Embora a motivação para o desenvolvimento dessas ferramentas tenha sido a utilização das mesmas para a identificação e classificação de sequências virais, essas ferramentas podem ser aplicadas a qualquer conjunto de sequências, independentemente do organismo. Como os HMMs de perfil são modelos probabilísticos capazes de representar a diversidade de um conjunto de sequências representados por meio de um alinhamento múltiplo de sequências, a aplicação desses modelos na identificação de sequências virais em conjuntos de dados tem se mostrado mais eficiente que buscas de similaridade pareadas, uma vez que esses modelos permitem que sequências mais distantes das conhecidas sejam detectadas.

Como discutido anteriormente, existem algumas bases de HMMs de origem viral disponíveis publicamente, bem como algumas metodologias para a construção desses modelos. Contudo, embora possam ser utilizados, os modelos disponíveis apresentam importantes limitações. Nesse trabalho, descrevemos o desenvolvimento da ferramenta TABAJARA para o desenho racional de HMMs de perfil, sejam eles conservados ou discriminativos. Como pudemos observar nos diferentes testes, os modelos gerados por essa ferramenta são suficientemente sensíveis e específicos para um determinado grupo de sequências, podendo associar com excelentes resultados correlações taxonômicas.

Os modelos construídos com o TABAJARA foram utilizados nesse trabalho em diferentes abordagens, desde a prospecção de sequências de proteínas até a busca de elementos multigênicos em genomas de bactérias e arqueias. Vimos também que quando trabalhamos com conjuntos mais divergentes, como os fagos da família *Microviridae*, é preferível se utilizar múltiplos modelos curtos do que um modelo construído a partir de sequências inteiras de proteínas. Essa abordagem garante que somente regiões específicas do grupo de sequências sejam utilizadas nas buscas e evita que sequências não relacionadas sejam detectadas.

Durante esse trabalho foram também desenvolvidas duas ferramentas para a prospecção de dados, HMM-prospector e e-Finder. A primeira delas é amplamente utilizada em nosso laboratório, uma vez que permite uma rápida varredura de conjuntos de sequências com HMMs de perfil, de forma a identificarmos quais organismos estão presentes nas amostras e quais modelos foram capazes de detectar sequências. Como

vimos nos estudos de caso do capítulo 9, o *e-Finder* foi capaz de detectar novos elementos tanto para casposons como para profagos de *Microviridae*, comprovando seu enorme potencial de aplicação. Por fim, as ferramentas desenvolvidas nesse trabalho em conjunto com o montador progressivo GenSeed-HMM, constituem um conjunto integrado de ferramentas, que podem ser utilizados para a detecção de novos vírus e elementos genéticos móveis, bem como em qualquer outra tarefa relacionada à detecção e/ou discriminação de subgrupos de famílias protéicas.

Como perspectivas futuras, esperamos em breve finalizar e disponibilizar para a comunidade científica a *Viral MinionDB*, uma base de HMMs de perfil derivados de proteínas virais, com relações taxonômicas específicas. Essa base de HMMs de perfil será a primeira do mundo a conter HMMs de perfil construídos especificamente para todas as famílias virais descritas no NCBI. Uma outra aplicação para o futuro seria a criação de um serviço na *web* para a classificação de sequências genômicas e metagenômicas. Utilizando os modelos da base de dados *Viral MinionDB* e o programa *HMM-prospector*, o serviço poderia permitir a realização de buscas de similaridade entre sequências-consulta e o conjunto dos HMMs de perfil, retornando ao usuário uma tabela com a lista das famílias virais encontradas no seu conjunto de dados.

11 Referências bibliográficas

- Abby, S.S. Néron, B., Ménager, H, Touchon, M. & Rocha, E.P.C. (2014). MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS ONE*, **9(10)**: e110726.
- Aiweesakun, P. & Simmonds, P. (2018A) The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. *Microbiome*, **6**:38.
- Aiweesakun, P., Adriaenssens, E.M., Lavigne, R., Kropinski, A.M. & Simmonds, P. (2018B) Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. *Journal of General Virology*, **99**:1331–1343.
- Arlot, S & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistical Surveys*, **4**: 40-79.
- Alves, J.M., de Oliveira, A.L., Sandberg, T.O., Moreno-Gallego, J.L., de Toledo, M.A., de Moura, E.M., Oliveira, L.S., Durham, A.M., Mehnert, D.U., Zanotto, P.M., Reyes, A. & Gruber A. (2016) GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in *Alpavirinae* Viral Discovery from Metagenomic Data. *Front Microbiol.* **7**: 269.
- Altschul, S. F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215** (3).
- Arndt, D., Grant, J., Marcu, A., Sajed, T., Pon, A., Liang, Y., Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.*, **44(W1)**:W16-21
- Baltimore. D. (1971) Expression of Animal Virus Genomes. *Bacteriological Reviews*, **35**(3), 235-241.
- Brenner, S.E., Chothia, C. & Hubbard, T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.*, **95**: 6073-6078.
- Burrell, C.J., Howard, C. R. and Murphy, F. A. (2017) Chapter 36 – *Flaviviruses*. In: *Fenner and White's Medical Virology*. cap. 36, p. 493-518.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.
- Capra, J.A. & Singh, M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**: 1875-1882.

- Carstens, E.B. (2012) Part I: Introduction, Ninth Report of the International Committee on Taxonomy of Viruses.
- Cherwa, J.E. & Fane A. (2011) *Microviridae: Microviruses and Gokushoviruses*. Encyclopedia of Life Sciences
- Chibani, C.M., Farr, A., Klama, S., Dietrich, S. & Liesegang, H. (2019A) Classifying the Unclassified: A Phage Classification Method. *Viruses*, **11**(2), 195.
- Chibani, C.M., Meinecke, F., Dietrich, A.F.S. & Liesegang, H. (2019B) ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks. bioRxiv doi: 10.1101/558171.
- Clokier, M.R.J., Millard, A.D., Letarov, A.V. & Heaphy, S. (2011) Phages in nature. *Bacteriophage* **1**:1, 31-45.
- Cohen J. (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas*, **20**:37-46
- Coutinho, F.H., Silveira, C.B., Gregoracci, G.B., Thompson, C.C., Edwards, R.A., Brussaard, C.P.D, Dutilh, B.E. & Thompson, F.L. (2017) Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications*, **8**: 15955.
- Cover, T.M. and Thomas, J.A. (2006) Elements of information theory. John Wiley & Sons, New York.
- Creasy, A., Rosario K., Leigh B.A., Dishaw, L.J. & Breitbar, M. (2018) Unprecedented Diversity of ssDNA Phages from the Family *Microviridae* Detected within the Gut of a Protochordate Model Organism (*Ciona robusta*). *Viruses*, **10**: 404.
- Daep, C.C, Muñoz-Jordán, J.L. & Eugenin, E. A. (2014) *Flaviviruses*, an expanding threat in public health: focus on *Dengue*, West Nile, and Japanese encephalitis virus. *J Neurovirol.*; **20**(6): 539-560.
- Drake, J. W. (1993) Rates of spontaneous mutation among RNA viruses. Proceedings of the National Academy of Sciences, 90(9): 4171-4175.
- Doore, S.M. & Fane, B.A. (2016) The *Microviridae*: Diversity, assembly, and experimental evolution. *Virology*, 491: 45-55.
- Durham, A.M., Kashiwabara, A.Y., Matsunaga, F.T.G., Ahagon, P.H., Rainone, F., Varuzza, L., Gruber, A. (2005) EGene: a configurable pipeline generation system for automated sequence analysis. *Bioinformatics*. **21**(12):2812-3.
- Dutilh, B.E., Reyes, A., Hall, R.J. & Whiteson, K.L (2017) Editorial: Virus Discovery by Metagenomics: The (Im)possibilities Bas. *Front Microbiol.*, **8**: 1710.

- Eddy, S. R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7(10)**: e1002195.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32(5)**: 1792–1797.
- Enright, A.J.; Van Dongen, S.; Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002, 30, 1575–1584
- Fancello, L., Raoult, D. & Desnues, C. (2012) Computational tools for viral metagenomics and their application in clinical research. *Virology*, **434**: 162-174.
- Fancello, L., Trape, S., Robert, C., Boyer, L., Popgeorgiev, N., Raoult, D. & Desnues, C. (2013) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *The ISME Journal*, **7**: 359–369.
- Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A, Tate, J. & Bateman, A. (2015) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*, **44(D1)**: D279-85.
- Flint, J., Racaniello, V.R., Rall, G.F., Skalka, A.M., Enquist, L.W. (2015) Principles of Virology. ASM Press, Washington, DC.
- Grazziotin AL, Koonin EV & Kristensen DM. (2017) Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res*, **45(Database issue)**: D491–D498.
- Güemes, A.G.C, Youle, M., Cantú, V.A., Felts, B., Nulton, J. & Rohwer, F. (2016). Viruses as Winners in the Game of Life. *Annu. Rev. Virol.*, **3**:197–214.
- Harris, E., Holden, K. L., Edgil, D., Polacek C., Clyde K. (2006) Molecular biology of *Flaviviruses*. *Novartis Found Symp.*, **277**: 23–39.
- Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., Jensen, L.J., Mering, C.V. & Bork, P. (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, **44**: D286–D293
- Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool". *Genome Research*. **12** (4): 656–664.
- Krishnamurthy, S.R & Wang, D. (2017) Origins and Challenges of Viral Dark Matter. *Virus Research*, 239:136-142.

- Kristensen, D.M., Cai, X., Mushegian, A. (2011) Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *J Bacteriol.*, **193**(8):1806–1814.
- Kristensen, D.M., Waller, A.S., Yamada, T., Bork, P., Mushegian, A.R., Koonin E.V. (2013) Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J Bacteriol.*, **195**(5):941–950.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Hausder, D (1994) Hidden Markov Models in Computational Biology Applications to Protein Modeling. *J. Mol. Biol*, **235**: 1501-1531
- Krupovic, M. & Forterre, P. (2011) *Microviridae* Goes Temperate: Microvirus-Related Proviruses Reside in the Genomes of Bacteroidetes. *PLoS ONE*, **6**(5): e19893
- Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D. & Koonin, E.V. (2014) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biology*, 12:36
- Krupovic, M., Shmakov, S., Makarova, K.S., Forterre, P. & Koonin, E. V. (2016). Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity. *Genome Biol. Evol.*, **8**(2): 375–386.
- Krupovic, M., Béguin, P. & Koonin, E.V. (2017) Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Current Opinion in Microbiology*, **38**:36–43.
- Kullback, S.; Leibler, R.A. (1951) On information and sufficiency. *Annals of Mathematical Statistics*. **22** (1): 79–86.
- Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**: 357 – 359.
- Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J, Siddell, S.G. & Smith, D.B. (2018) Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res*, **46**(Database issue): D708–D717.
- Leplae, R., Hebrant, A., Wodak, S.J. & Toussaint, A. (2004) ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* 2017, **32**(Database issue): D45-D49.
- Li, Y., Wang, H., Nie, K., Zhang, C., Zhang, Y., Wang, J., Niu, P. & Ma, X. (2016) VIP: an integrated pipeline for metagenomics of virus identification and Discovery. *Scientific Reports*, **6**:23774.
- Li, W.; Godzik, A. (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13): 1658–1659.

- Lima-Mendez, G., Helden, J.V., Toussaint, A. & Leplae, R. (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, **24**: 6.
- Lin, J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, **37**: 145–151
- Lindenbach, B.D. & Rice, C.M. (2003) Molecular biology of *Flaviviruses*. *Adv. Virus Res.*, **59**: 23–61.
- Liu, J., Glazko, G., Mushegian, A. (2006) Protein repertoire of double-stranded DNA bacteriophages. *Virus Res.*, **117(1)**:68-80.
- Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Tasneem, A., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Bryant, S.H. (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**: D205–D210.
- Manriquea, P., Bolducb, B., Walka, S.T., Oostc, J.V.D, Vosc W.M., & Young, M.J. (2016) Healthy human gut phageome. *PNAS*, **113 (37)**: 10400-10405
- Medema, M.H., Takano, E. & Breitling, R. (2013) Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. *Mol. Biol. Evol.*, **30(5)**:1218–122.
- Meyer, F., Paarmann, D., Souza, M. D., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening J. & Edwards, R.A (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**:386.
- Mokili, J.L., Rohwer, F. & Dutilh, B. E. (2012) Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, **2(1)**: 63-77.
- Moore, N.E., Wang, J., Hewitt, J., Croucher, D., Williamson, D.A., Paine, S., Yen, S., Greening, G.E., Halla, R.J (2015) Metagenomic Analysis of Viruses in Feces from Unsolved Outbreaks of Gastroenteritis in Humans. *Journal of Clinical Microbiology*, **53**: 1.
- O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D., Astashyn A., Badretdin A., Bao Y., Blinkova O., Brover V., Chetvernin V., Choi J., Cox E., Ermolaeva O., Farrell C.M., Goldfarb T., Gupta T., Haft D., Hatcher E., Hlavina W., Joardar V.S., Kodali V.K., Li W., Maglott D., Masterson P., McGarvey K.M., Murphy M.R., O'Neill K., Pujar S., Rangwala S.H., Rausch D., Riddick L.D., Schoch C., Shkeda A., Storz S.S., Sun H.,

- Thibaud-Nissen F., Tolstoy I., Tully R.E., Vatsan A.R., Wallin C., Webb D., Wu W., Landrum M.J., Kimchi A., Tatusova T., DiCuccio M., Kitts P., Murphy T.D., Pruitt K.D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44(D1)**: D733-45.
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., Chothia, C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, **284**: 4.
- Peck, K. & Luring, A.S. (2018) Complexities of Viral Mutation Rates. *J Virol.*, **92**: e01031-17.
- Pirovano, W., Feenstra, K.A. and Heringa, J. (2006) Sequence comparison by sequence harmony identifies subtype specific functional sites. *Nucleic Acids Res.*, **34**: 6540–6548.
- Pickett, B.E., Sadat, E.L, Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C.N., Dietrich, J., Klem, E.B. & Scheuermann, R.H. (2011) ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.*, **40 (Database issue)**: D593-8.
- Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, **5(3)**: e9490.
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E., Finn, R.D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47(D1)**: D427-D432.
- Rampelli, S., Soverini, M., Turrone, S., Quercia, S., Biagi, E., Brigidi P. & Candela, M. (2016) ViromeScan: a new tool for metagenomics viral community profiling. *BMC Genomics*, **17**:165.
- Reyes, A.; Alves, J.M.P.; Durham, A.; Gruber, A. (2017) Use of profile hidden Markov models in viral discovery: Current insights. *Adv. Genom. Genet.*, **7**: 29–45.
- Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16(6)**: 276–277.
- Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L. & Prosperi, M (2016) Challenges in the analysis of viral metagenomes. *Virus Evolution*, **2(2)**: vew022.
- Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime- Ngando, T. & Debroas, D. (2012A) Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *Plos One*, **7(3)**: e33641.

- Roux S, Krupovic M, Poulet A, Debroas D & Enault F. (2012B) Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One*, **7(7)**: e40418.
- Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics*, **15**: 76.
- Roux, S., Enault, F., Hurwitz, B.F. & Sullivan, M.B. (2015), VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**: e985.
- Sanjuán, R & Domingo-Calap, P. (2016) Mechanisms of viral mutation. *Cell. Mol. Life Sci.*, **73**:4433–4448.
- Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**: 544–548.
- Segobola, J., Adriaenssens E., Tsekoa, T., Rashamuse, K. & Cowan, D. (2018) Exploring Viral Diversity in a Unique South African Soil Habitat. *Scientific Reports* **8**:111.
- Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27**: 379–423, 623–656.
- Squires R.B., Noronha J., Hunt V., García-Sastre A., Macken C., Baumgarth N., Suarez D., Pickett B.E., Zhang Y., Larsen C.N., Ramsey A., Zhou L., Zaremba S., Kumar S., Deitrich J., Klem E., Scheuermann R.H. (2012) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and other respiratory viruses*, **6**: 404–416.
- Skewes-Cox, P., Sharpton, T.J., Pollard, K.S. & DeRisi, J.L. (2014) Profile hidden Markov Models for the detection of viruses within metagenomic sequence data. *PLoS One*, **9(8)**: e105067.
- The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45(Database issue)**: D158–D169.
- Tikhe, C.V. & Husseneder, C. (2018) Metavirome Sequencing of the Termite Gut Reveals the Presence of an Unexplored Bacteriophage Community. *Front Microbiol.*, **8**: 2548.
- Tithi, S.S., Aylward, F.O., Jensen, R.V. & Zhang, L. (2018) FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data Saima. *PeerJ*, **6**: e4227.
- Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., Machi D., Mao, C., Nordberg, E.K., Olson, R., Overbeek, R., Pusch, G.D., Shukla, M., Schulman, J., Stevens, R.L., Sullivan, D.E.,

- Vonstein, V., Warren, A., Will, R., Wilson, M.J.C., Yoo, H.S., Zhang, C., Zhang Y. & Sobral B.W. (2014) PATRIC, the bacterial bioinformatics database and analysis
- Wommack, K. E., Bhavsar, J., Polson, S.W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S. & Nasko, D.J. (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*, **6**:421-43.
- Ye, Y., Choi J. & Tang, H. (2011) RAPSearch: a fast protein similarity search tool for short reads. *BMC Bioinformatics*, **12**:159.
- Zerbino, D.R., Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
- Zhang, Y., Shi, M. & Holmes, E.C (2018) Using Metagenomics to Characterize an Expanding Virosphere. *Cell*, **172**: 1168–1172.
- Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39 (Web Server issue)**: W347–W352.