Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós Graduação em Biologia Molecular

Estudo da variabilidade genética da CDR H3 em anticorpos anti-DNA no contexto das famílias murinas VH10 e VH4

Maria Beatriz Walter Costa

Brasília, 4 de julho de 2012

Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós Graduação em Biologia Molecular

Estudo da variabilidade genética da CDR H3 em anticorpos anti-DNA no contexto das famílias murinas VH10 e VH4

Dissertação apresentada no Departamento de Biologia Celular do Instituto de Ciências Biológicas da Universidade de Brasília como requisito parcial para a obtenção do grau de Mestre em Biologia Molecular

Maria Beatriz Walter Costa

Orientador:
Marcelo Brígido

Co-orientadoras:
Tainá Raiol
Andrea Maranhão

Brasília, 4 de julho de 2012

Dissertação de mestrado sob o título *"Estudo da variabilidade genética da CDR H3 em anticorpos anti-DNA no contexto das famílias murinas VH10 e VH4"* defendida por Maria Beatriz Walter Costa no dia 6 de Junho de 2012 em Brasília, pela banca examinadora constituída pelos doutores Antônio Francisco Araújo, Wanessa Carvalho e Sônia Freitas.

Marcelo Brígido (CEL-IB/UnB)
orientador

Tainá Raiol e Andrea Maranhão
(CEL-IB/UnB)
co-orientadoras

Antônio Francisco Araújo (CEL-IB/UnB)
Examinador interno

Wanessa Carvalho (IB/UnB)
Examinadora externa

Sônia Freitas (CEL-IB/UnB)
Suplente efetiva da banca examinadora

Dedico este trabalho à minha mãe

Agradecimentos

Em primeiro lugar gostaria de agradecer aos meus pais, João e Mia, pelo apoio incondicional ao longo de toda essa jornada, que não durou somente dois anos, mas sim vinte e quatro. Só foi possível cumprir mais essa etapa após longos anos de estudo, dedicação e preparação. Agradeço profundamente a orientação e o apoio que me deram, e acima de tudo os exemplos, não só de trabalho árduo, mas também do prazer em aprender. Estou extremamente feliz com essa jornada, especialmente de tê-la cumprida ao lado de vocês. A convivência com vídeos de cirurgias e com grafos deve ter algum efeito curioso na mente de uma criança, e esta deve se consistir no trauma ou na inspiração (ou em ambos...). Bem, no caso em questão, acho que me fez despertar para a Ciência. Muito obrigada!

Um agradecimento muito especial aos meus queridos pais científicos Marcelo, Tainá e Andrea, que me ensinaram muito mais do que técnicas de Biologia Molecular ou de Bioinformática, mas sim em como pensar. Tenho muito orgulho de ter sido formada por mentes tão iluminadas e por pessoas com tanta paixão e dedicação à Biologia.

Agradeço profundamente aos meus avós, Tutu, Walter *im memoriam*, Bia e Geraldo que me deram muito carinho e valiosíssimas e felizes memórias. Cada um participou da minha vida de maneira única e tornou a minha existência muito mais rica e completa.

Meus queridos e inestimáveis irmãozinhos, muito obrigada! Luluzinha e Kikão, vocês fizeram a minha vida muito mais divertida e tumultuada. Com certeza verei grandes feitos de vocês por aí. Um grande beijo e amo vocês.

Querido Mikael Lemos, te agradeço muitíssimo pelo apoio, pelo carinho e por todo o resto. Você foi fundamental para mim durante todo esse processo e espero que possamos realizar também outras coisas juntos. *Ich liebe dich, meine liebe.*

Agradeço muito às minhas amigas/irmãs queridas Galé e Hofmann, que tomaram um lugar especial no meu coração desde que apareceram na minha vida no parquinho do Leonardo, há quase 15 anos.

Meus queridos padrinhos tia Marisa e tio Tõe, muito obrigada por todo o amor que me deram e também por dividirem comigo suas grandes paixões, foram muitos *patchworks* e muitas poesias, que certamente deram um colorido à minha mente "exata".

Muito obrigada ao "macro" tio Bruno e à "micro" Nana, que estiveram presentes nos meus primeiros passos científicos e foram incentivadores fundamentais para que eu seguisse a carreira biológica. Apesar de estar no mundo "micro", no qual fui iniciada em Goiânia pela minha querida prima, jamais esquecerei das saídas de campo "macro" através do cerrado em busca de Tabebuias com meu querido tio floresteiro.

Agradeço aos meus companheiros de lab e amigos da Bio e da Computação, Halian, meu companheiro de jornada, Daniel Saad, Paulo Alvarez, mediador Jedi do Galaxy, Felipe Lessa, Túlio, Ruben, Guttinho e Ina, Bia Ma, Tay, Galina, Adriane, Rafa Burtet,

Kelly Simi, Mariany, Camillo, Bárbara, Fernanda, Flávia, Herdson, Yuri.

Agradeço a todos os professores e mestres que me ajudaram nessa busca sem fim do conhecimento, especialmente ao professor Waldenor, por quem tenho imenso respeito e admiração, ao Chico, Sônia, Wanessa, Werner, Roberto Togawa, Georgios e Natália.

Muito obrigada também aos funcionários do IB que me ajudaram muito ao longo de todo o processo: Ana, secretária da Pós-Graduação, dona Ivonildes, dona Fátima, Thompson e Fernanda, do Laboratório de Biologia Molecular.

Agradeço aos meus queridos tios, por quem tenho imenso carinho e que sempre me deram muito apoio: Tt, Walmir *im memoriam*, Linto *im memoriam*, Inez, Nelson, Gil, Débora, Carlinhos, Izabel, Luiz Henrique, Márcia, Taís, Aloísio. Muito obrigada também aos primos Júlia, Gui, Alex, Rafa (meu companheiro de Mestrado), Bernardo, Ariana e Mônica (meus companheiros de Rock in Rio), André, Marcelo, Marcinha, Belinha, Aninha, Felipe, Nô, Henrique, Laila e Thea, e os muitos L's: Léo e Leandro Azevedo, Luciano, Leandro Walter, Lia, Léo Tostes, Lea e Liliana Werner.

Muito obrigada também à linda família Guimarães, Hélia e Altamir, Júnior, Dani, Iran, Gabriel, dona Elza e senhor Raimundo, que me receberam de braços abertos e sempre com muito carinho e cuidado.

Finalmente agradeço à minha queridíssima amiga Sandriele, infelizmente também *im memoriam*. A pessoa que tinha respostas para todas as minhas perguntas, e com quem eu dividia diversos sonhos. *Danke schön für alle fröhliche Momente!*

Sumário

Resumo

O conhecimento sobre as interações moleculares entre anticorpos com afinidade a ácidos nucléicos e seus alvos ainda é muito limitado. Trabalhos anteriores do grupo de Imunologia Molecular da Universidade de Brasília de busca em bancos de dados apontavam para uma provável tendência de reconhecimento de ácidos nucléicos dos anticorpos pertencentes à família murina de imunoglobulina de cadeia variável pesada dez, ou VH10. Portanto, para melhor entender esse comportamento, decidiu-se analisar bibliotecas de regiões determinantes de complementaridade três de cadeia pesada, CDR H3, em contextos das famílias murinas VH10 e VH4, antes e após essas bibliotecas terem sido selecionadas contra DNA fita simples. Desta forma, procurou-se compreender o comportamento dessas duas famílias e de certos determinantes da CDR H3 no reconhecimento de anticorpos a DNA. As bibliotecas de estudo foram submetidas a seqüenciamento de alto-desempenho, utilizando-se a metodologia Illumina, e foram então analisadas por meio de um *pipeline* de Bioinformática. Esse *pipeline* consistiu na preparação das seqüências FASTQ recebidas e na análise posterior, sendo esta última composta por estudos de variabilidade das bibliotecas, enriquecimento de seqüências após a seleção contra DNA, divergência de Kullback-Leibler, comparação de padrões e composição da CDR H3. A análise revelou que a os padrões de seqüências selecionadas positivamente e negativamente são muito semelhantes em VH10, e que este arcabouço é menos estringente que VH4 em relação a exigir seqüências de CDR H3 mais específicas a DNA. Portanto, concluiu-se que a família VH10 apresenta tendência intrínseca à formação de anticorpos anti-DNA, o que não ocorre com a família VH4.

Abstract

We still possess very little knowledge of the molecular interaction between anti-nucleic acids and their targets. Previous research in databases made by the Molecular Immunology group of the University of Brasilia has indicated a probable tendency of nucleic acids' recognition of the antibodies of the murine heavy chain immunoglobulin family ten, VH10. Therefore, to better understand this behaviour, we analysed libraries of heavy chain complementarity determinig region three, CDR H3, in the context of the murine families VH10 and VH4, before and after these libraries have been selected against DNA single strand. In this manner, we sought to understand the behaviour of these families and of certain CDR H3 features in the antibody's recognition of DNA. The libraries were submitted to deep-sequencing Illumina platform and then analysed by a Bioinformatics pipeline. The pipeline was comprised in the preparation of the FASTQ Illumina sequences and in an afterwards analysis, which consisted in studies of library variability, sequence enrichment, Kullback-Leibler divergence, pattern comparison and CDR H3 composition. Analysis revealed that the patterns of the positive and negative selected groups are similar in VH10 and that this framework is more stringent than VH4 in demanding CDR H3 sequences that are more DNA-specific. Therefore, it was concluded that the VH10 family presents an intrinsic tendency to form anti-DNA antibodies, which does not occur with the VH4 family.

"But I don't want to go among mad people", Alice remarked.
"Oh, you can't help that", said the Cat: "we're all mad here. I'm mad. You're mad."
"How do you know I'm mad?", said Alice.
"You must be", said the Cat, "or you wouldn't have come here."
Alice's Adventures in Wonderlands, Lewis Carroll

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

A	Adenina
C	Citosina
^{o}C	Grau Celsius
CDR	Região determinante de complementariedade
C-terminal	Extremidade carboxi-terminal
cr	Complemento reverso
dH2O	Água destilada
DNA	Ácido desoxirribonucléico
EDTA	Ácido etilenodiaminotetracético
ELISA	Ensaio de ligação imunoenzimática
Fab	Fragmento de anticorpo de ligação ao antígeno
Fc	Fragmento de anticorpo cristalizável - porção constante
FR	Arcabouço (*Framework*)
Fv	Fragmento variável do anticorpo
G	Guanina
g	Grama
g	Força gravitacional
GB	Giga byte
GHz	Giga Hertz
h	Hora
HTML	*HyperText Markup Language*
Ig	Imunoglobulina
kb	Kilobase
L	Litro
M	Molar
mA	Miliampère
mg	Miligrama
min	Minuto
mL	Mililitro
mM	Milimolar
MM	Massa Molecular
ms	Milisegundo
μg	Micrograma
μL	Microlitro
μM	Micromolar
nc	Nucleotídeo

ng	Nanograma
ηm	Nanômetro
OD	Densidade óptica
p	Peso
pb	Par de base
PCR	Reação de polimerase em cadeia
pH	Potencial hidrogeniônico
pt	Proteína
RAM	*Random-access memory*
rpm	Rotações por minuto
RNA	Ácido ribonucléico
scFv	Fragmento variável de anticorpo de cadeia única
T	Timina
Tris	Tri (hidroximetil) aminometano
UTR	Região não traduzida do gene
v	Volume
VH	Domínio variável da cadeia pesada de um anticorpo
VL	Domínio variável da cadeia leve de um anticorpo

1 Introdução

O conhecimento sobre as interações moleculares entre anticorpos anti-ácidos nucléicos e seus alvos ainda é muito limitado. A literatura sugere que a família variável murina de cadeia pesada de imunoglobulina VH10 apresenta tendência intrínseca de ligar-se a ácidos nucléicos (Guedes, 2009; Maranhão, 2001; Brígido e Stollar, 1991). Tal tendência foi observada anteriormente pelo grupo de Imunologia Molecular da Universidade de Brasília em pesquisa realizada em banco de dados, onde verificou-se que cerca de 59,5% dos anticorpos que continham VHs pertencentes à família VH10 eram descritos como ligantes a algum tipo de ácido nucléico. Em contrapartida, tal descrição não foi encontrada para nenhum anticorpo presente em bancos de dados que utilizava a família VH4 (Guedes, 2009). Tal observação suporta a idéia de que há determinantes estruturais na seqüência dos segmentos gênicos variáveis da família VH10 que asseguram o reconhecimento de ácidos nucléicos.

Dentre as regiões determinantes de complementaridade das imunoglobulinas, a região três de cadeia pesada, CDR H3, é a porção mais hipervariável e contribui de maneira decisiva na ligação com o antígeno (Zemlin *et al.*, 2003). Desta forma, este trabalho tem como objetivo analisar o comportamento de ligação a DNA de anticorpos contendo bibliotecas de CDR H3 nos contextos das famílias murinas VH10 e VH4. Essas bibliotecas, geradas previamente por meio da técnica de *phage display* e submetidas à seleção pela capacidade de ligação a DNA fita simples, foram seqüenciadas na plataforma Illumina de seqüenciamento de alto-desempenho. A análise destas seqüências foi feita por meio de um *pipeline* computacional de Bioinformática. Pela comparação entre as bibliotecas, foi caracterizada a importância das famílias VH10 e VH4 na ligação de anticorpos a DNA.

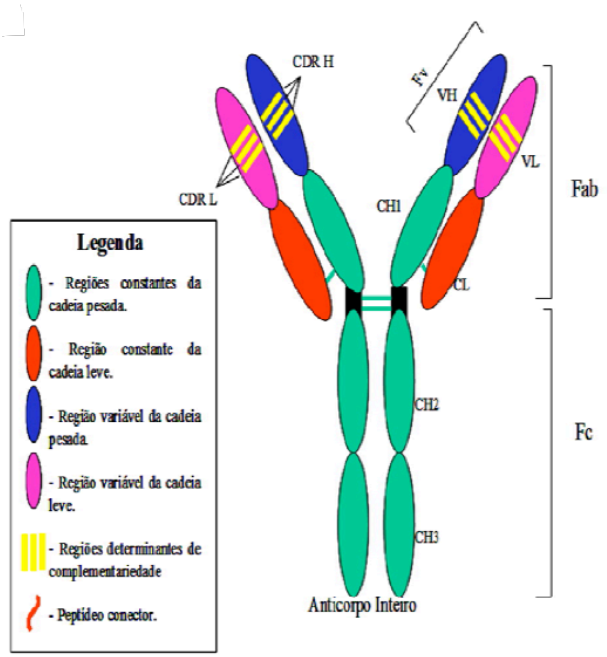
1.1 Sistema imune: respostas inata e adaptativa

O sistema imunológico é constituído por células e moléculas que geram uma resposta coordenada e interativa contra o aparecimento de substâncias estranhas ao corpo. Existem diversos tipos de respostas imunes, sendo que estas são moduladas por diferentes fatores, como por exemplo a natureza do patógeno. Também são levados em consideração se o mesmo patógeno já infectou o corpo anteriormente, se este está confinado a determinado tecido ou se já chegou à corrente sangüínea (Cohen, 2007).

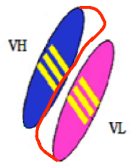
Em vertebrados, a defesa contra patógenos é mediada por reações imediatas da imunidade inata, e por reações mais tardias da imunidade adaptativa. A imunidade inata é caracterizada por mecanismos geneticamente programados que reconhecem padrões invariáveis do invasor, porém não distinguem diferenças finas entre eles, alguns componentes dessa resposta são, por exemplo, os epitélios e células fagocíticas. Por outro lado, a imunidade adaptativa detecta padrões quase invariáveis de um microorganismo de uma dada classe, e faz isso por meio do reconhecimento de antígenos do invasor por meio de receptores celulares. Antígeno é uma substância que induz respostas imunes específicas ou que são alvos dessas respostas. Os principais componentes da resposta imune adaptativa são os linfócitos T e B e seus produtos de secreção, tais como os anticorpos (Iwasaki e Medzhitov, 2010).

1.2 Anticorpos

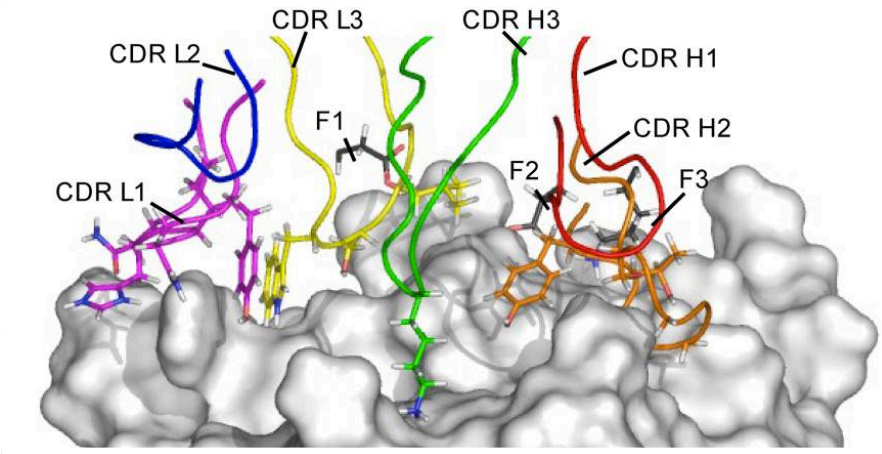
Os anticorpos, ou imunoglobulinas, são secretados exclusivamente pelos linfócitos B após a sua ativação, que ocorre após o encontro e ligação do seu receptor celular com o antígeno. Estes são produzidos em bilhões de formas, cada uma com uma seqüência de aminoácidos e um sítio diferente de ligação ao antígeno. A sua estrutura é composta basicamente de duas cadeias leves, L, e duas pesadas, H. Há uma dupla funcionalidade relacionada à essa estrutura, sendo que a extremidade amino-terminal se liga ao antígeno, enquanto que a extremidade C-terminal é constante e é reconhecida por células efetoras do sistema imune (figura 1a). A porção ligante ao antígeno, denominada paratopo, é a fração variável da molécula e é formada pelos domínios variáveis de ambas as cadeias leve, L, e pesada, H. Por manipulações genéticas, essa porção pode ainda ser expressa na forma de scFv, fragmento variável de cadeia única (figura 1b) (Woof e Burton, 2004; Maranhão e Brígido, 2001).



(a)



(b)



(c)

Figura 1: **Representação esquemática e composição da molécula de imunoglobulina, com destaque para as regiões determinantes de complementaridade.** (a) As regiões de ligação ao antígeno, Fab, são compostas pelas cadeias leve, L, e pesada, H, enquanto que a região efetora, Fc, é composta somente da cadeia pesada (adaptado de Maranhão e Brígido, 2001). (b) Fragmento variável de cadeia única de um anticorpo, scFv, obtido por meio de manipulação genética (adaptado de Maranhão e Brígido, 2001). (c) As regiões determinantes de complementaridade cadeias leve e pesada, CDR L1 a 3 e CDR H1 a 3 de um scFv estão representadas na forma de alças coloridas, enquanto que o antígeno de ligação está representado em cinza. Os resíduos do scFv que participam da interação com o antígeno estão representados na forma de modelo de bastão, com destaque para os resíduos dos arcabouços um, dois e três, F1, F2 e F3 (adaptado de Wilkinson *et al.*, 2009).

Modelos tri-dimensionais mostram que o domínio variável da imunoglobulina apresenta uma estrutura estendida alternada com alças, ou *loops*, em inglês. Existem seis alças no total, sendo que três estão contidas na cadeia leve e três na pesada. As alças se projetam para o solvente e contribuem com a grande maioria das interações moleculares com o antígeno. A maior parte das CDRs estão co-localizadas nestas alças, onde está concentrada a diversidade entre as moléculas de imunoglobulinas. As CDRs são ainda delimitadas por regiões relativamente conservadas, os arcabouços, em inglês *frameworks* (figura 1c), (Wilkinson *et al.*, 2009).

As CDRs 1 e 2 possuem codificação germinal e dependem da família do gene V para obter sua variabilidade. Por outro lado, a CDR 3 é a mais hipervariável, pois é formada a partir da recombinação dos segmentos gênicos do *locus* de imunoglobulina. A CDR 3 da cadeia leve é formada a partir dos dois segmentos V e J, enquanto que a CDR 3 de cadeia pesada é formada a partir dos três segmentos V, D e J. Além da recombinação desses segmentos, a variação da CDR H3 também é causada pela diversidade juncional gerada pela remoção de nucleotídeos nas junções entre os segmentos V e D e D e J. Essa variação de formas é essencial para o reconhecimento dos diferentes antígenos e provêm essencialmente dos processos de recombinação V(D)J e da hipermutação somática (Li *et al.*, 2003; Maranhão e Brígido, 2001).

1.2.1 Geração da diversidade em anticorpos

As células B são as células secretoras de anticorpos do sistema imune e possuem em suas membranas celulares receptores que têm o mesmo sítio de ligação ao antígeno que os anticorpos secretados. Os *loci* que codificam os receptores dessas células são os da cadeia pesada de imunoglobulina, IgH, da cadeia leve, IgL, Igκ e Igλ. Eles devem ser produzidos de forma que apresentem afinidade e especificidade aos múltiplos antígenos dos patógenos invasores, e para isso, o sistema imune desenvolveu níveis hierárquicos de evolução molecular, sendo que os principais são a recombinação V(D)J e a hipermutação somática. Estes são eventos que geram a diversidade necessária para o reconhecimento de cada possível antígeno que possa entrar em contato com o organismo (Sun, Earl e Deem, 2005).

A recombinação V(D)J é o evento responsável pela composição da região variável dos genes de receptor e envolve múltiplos segmentos germinais dos três genes, V, D e J. Há uma etapa inicial de quebra da dupla-fita de DNA feita pela endonuclease RAG, a qual é formada pelos genes RAG1 e RAG2 (abreviações de genes de ativação de recombinação

um e dois). O complexo RAG realiza essas quebras no DNA se guiando em locais de sinalização nas bordas entre dois segmentos codificadores. Estes locais de sinalização possuem bases conservadas separadas por dois espaçamentos que possuem respectivamente 12 e 23 pares de bases menos conservados. Posteriormente é feito ainda um processamento nas extremidades das fitas quebradas, que são reparadas por proteínas de junção de extremidades ubiquitariamente expressas na célula (Jung *et al.*, 2006).

O *locus* murino de cadeia pesada de imunoglobulina, IgH, tem aproximadamente 3 Mb e se localiza próximo à região telomérica do cromossomo 12. Cerca de 150 segmentos gênicos compõem as 14 famílias murinas VH, que se distribuem em cerca de 2,7 Mb a montante dos segmentos gênicos DH (figura 2a). Na recombinação V(D)J da cadeia pesada de imunoglobulina, primeiro os segmentos D e J se unem, e por último o segmento V se une ao DJ, formando V(D)J (figura 2b). A maior parte do éxon variável IgH é formado pelo segmento germinal VH, que comporta as regiões determinantes de complementaridade um e dois, CDR H1 e CDR H2. Já CDR H3 é codificada pela região juncional de VH, DH e JH, o que gera maior variabilidade (figura 2c). A hipermutação somática, por sua vez, ocorre após os eventos de recombinação gênica, expressão das cadeias do receptor e ligação com o antígeno. Neste evento, nucleotídeos randômicos sofrem mutação, para aumentar a especificidade do receptor ao antígeno (Sun, Earl e Deem, 2005).

A especificidade do receptor, e também do anticorpo, ao antígeno depende de sua estrutura terciária, que será determinada pela seqüência primária de bases existente no DNA da célula. É possível portanto, estudar a especificidade de um anticorpo, ou grupo de anticorpos, a um determinado antígeno, por meio do estudo da seqüência de DNA. O seqüenciamento é uma técnica bastante estabelecida na Biologia Molecular, e permite a elucidação do código contido nesta molécula. Existem diversas metodologias diferentes que realizam o seqüenciamento, cada uma com suas particularidades. Exemplos são a tradicional técnica Sanger e a mais recente técnica de alto-desempenho.

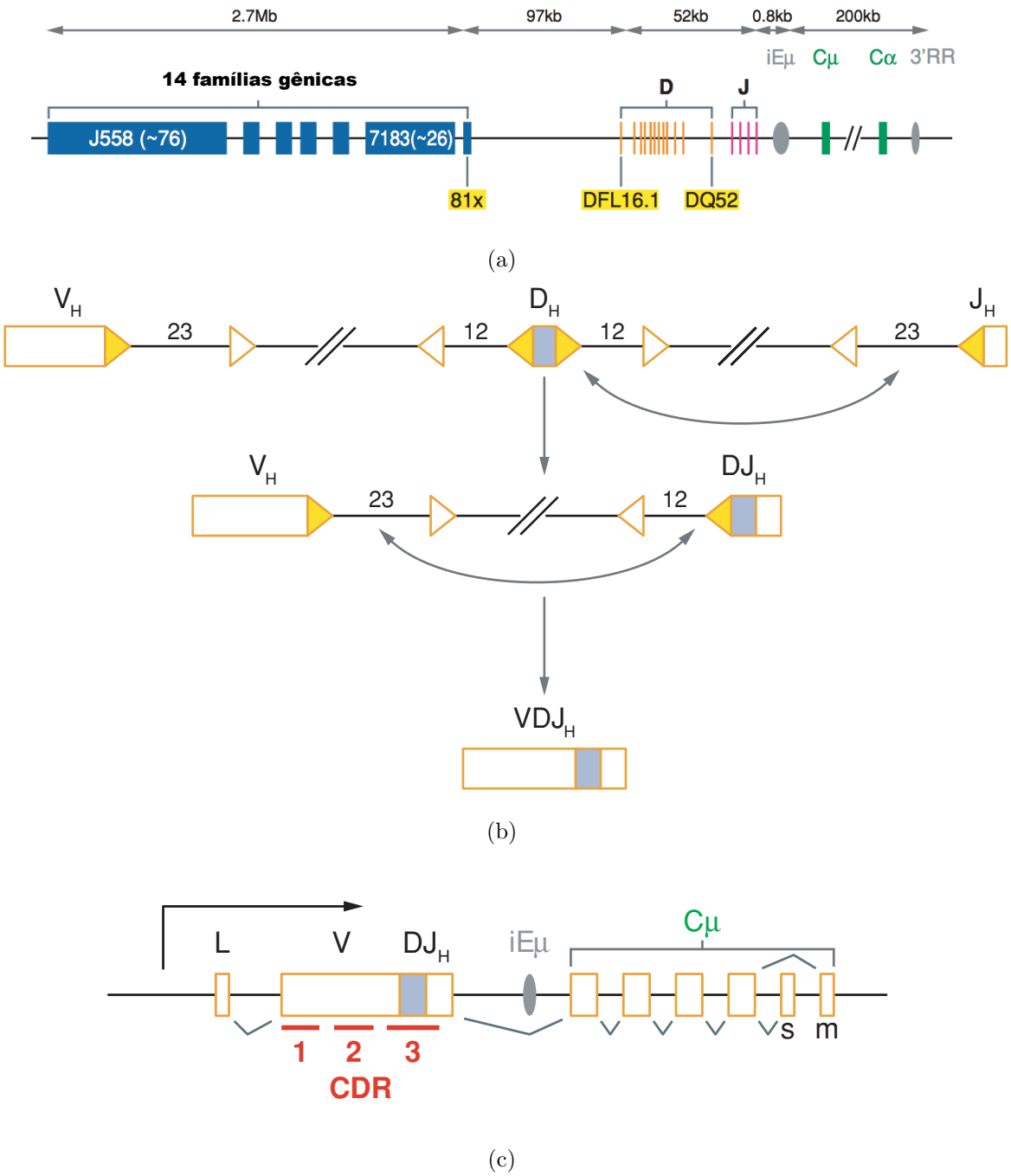


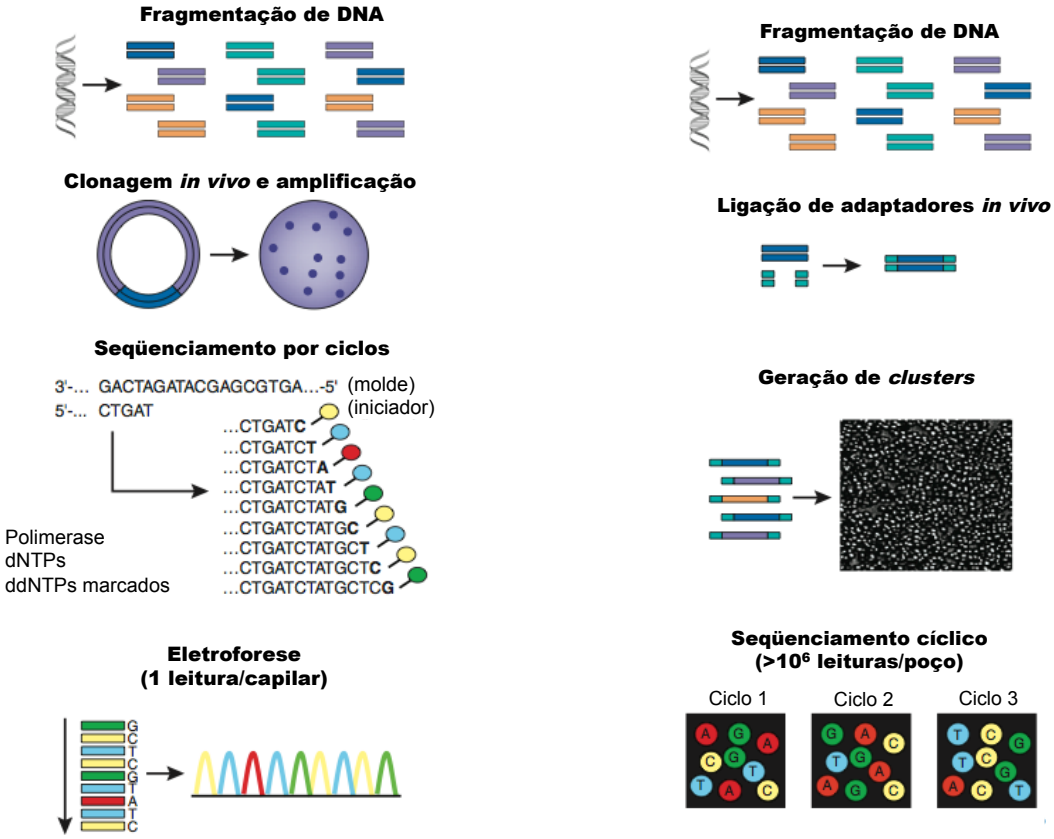
Figura 2: **Representação esquemática do *locus* IgH murino e formação dos genes de receptor de antígeno de células B (adaptado de Jung, 2006).** (a) Esquematização do *locus* IgH murino. (b) Montagem e expressão dos genes IgH murinos por meio da recombinação V(D)J. (c) Segmentos gênicos responsáveis pela codificação das três regiões determinantes de complementaridade.

1.3 Seqüenciamento de alto-desempenho

Na final da década de 1990, a identificação de seqüências de DNA era feita principalmente pela tecnologia de seqüenciamento semi-automático Sanger, baseado em capilares (Swerdlow *et al.*, 1990). Exemplos de aparelhos dessa técnica são o ABI e o MegaBACE. Nesta tecnologia é preciso preparar as amostras por meio de um laborioso processo de clonagem de DNA, sendo que o seqüenciamento é feito em placas de 96 ou 384 amostras, limitando o processo de paralelização. O aparelho gera um cromatograma de detecção de espectros fluorescentes emitidos durante o seqüenciamento (figura 3a). Posteriormente, um *software* interpreta o cromatograma e gera uma seqüência de DNA juntamente com probabilidades de erro associadas à leitura de cada base. Após três décadas de melhorias, a tecnologia Sanger determina seqüências de até 1.000 pb e tem uma acurácia de até 99,999%, calcula-se ainda que o preço de seqüenciamento de um Mb se aproxime de 500 dólares (Shendure e Ji, 2008).

Os *softwares* que fazem a leitura de bases são fundamentais em qualquer processo de seqüenciamento. O *Phred* realiza essa função na tecnologia Sanger e designa um *score* de qualidade para cada leitura de base. Esses *scores* variam de 4 até 60, com valores maiores correspondendo a uma melhor qualidade. Qualidade Phred igual a 10, por exemplo, representa uma probabilidade 1 em 10 de que determinada base seja lida incorretamente, ou seja, a acurácia da leitura é de 90%. Qualidade Phred igual a 20, representa probabilidade de 1 em 100 que a base seja lida incorretamente, sendo a acurácia neste caso igual a 99%, e assim por diante (Phred-*homepage*, 2012; Ewing *et al.*, 1998).

Em comparação com o seqüenciamento Sanger, a recente tecnologia de seqüenciamento de nova geração, ou alto-desempenho, é considerada muito mais prática e competitiva em termos de custo-benefício. Ela pode ser resumida como ciclos de manipulações enzimáticas e geração de coleções de imagens. Existem diversas técnicas bioquímicas usadas para esse tipo de seqüenciamento, e conseqüentemente, diversos aparelhos, como o Illumina, o 454 da Roche, o SOLiD da Applied Biosystems e o HeliScope da Helicos, entre outros. Entretanto, o fluxo de funcionamento desses aparelhos é conceitualmente semelhante (figura 3b). Considerando a quantidade de seqüências geradas, os tamanhos de fragmentos seqüenciados e o custo, estes variam bastante de acordo com a plataforma utilizada. Por exemplo, o 454 gera de 1,5 a 2 milhões de fragmentos de 300 a 600 pares de base a um custo médio de 60 dólares por megabase, enquanto que o seqüenciamento Illumina gera de 15 a 250 milhões de fragmentos de 36 a 150 pares de base a um custo médio de 2 dólares por megabase (Shendure e Ji, 2008).



(a) (b)

Figura 3: **Fluxo de funcionamento das tecnologias de seqüenciamento Sanger convencional e de alto-desempenho (adaptado de Shendure e Ji, 2008).** (a) No seqüenciamento tradicional Sanger, o DNA de interesse é clonado num vetor plasmidial. Para cada reação de seqüenciamento, uma única colônia é escolhida e o DNA plasmidial é isolado. Em cada ciclo de seqüenciamento, gera-se uma série de produtos marcados com bases ddNTP, os quais são submetidos a uma eletroforese de alta-resolução para separação dos produtos. As bases fluorescentes são lidas por um detector que gera um cromatograma de leitura de bases. (b) Na tecnologia de alto-desempenho, o DNA é fragmentado e adaptadores são adicionados às seqüências. Essas milhões de seqüências são imobilizadas numa placa, sendo que cada seqüência se transforma num conjunto diferente, em inglês: *cluster*. Espectros fluorescentes são emitidos a cada etapa de incoporação de nucleotídeos e a cada etapa é feita uma imagem. Cada posição na imagem representa uma base de um conjunto de seqüência diferente, e a ordem das imagens é utilizada para gerar a seqüência de bases.

O seqüenciamento Illumina é feito pela técnica da PCR em ponte, na qual adaptadores são adicionados às seqüências que se deseja identificar e fixados num substrato sólido, para que cada produto da PCR em ponte se mantenha imobilizado dentro da canaleta do seqüenciamento. Milhões de conjuntos de seqüências são amplificados, sendo que cada conjunto representa uma seqüência diferente dentro da mesma canaleta. As fitas de DNA são linearizadas e a DNA polimerase faz a extensão de um nucleotídeo por vez. Esses nucleotídeos são modificados para possuírem uma marcação química fluorescente que determina a identidade do nucleotídeo incorporado por meio do comprimento de onda emitido. Os comprimentos de onda são captados por uma câmara CCD, *charge-coupled device*, e transformados em imagens. Cada ponto fixo na imagem representa um conjunto de seqüência, de forma que a seqüência nucleotídica pode ser lida a partir da série de imagens geradas (Mardis, 2008).

Uma das vantagens dessa técnica é a imensa quantidade de seqüências geradas, o que fornece uma excelente cobertura das amostras seqüenciadas. Porém, isso gera novos problemas, como o processamento, a análise e o armazenamento da imensa quantidade de dados obtida. Portanto, são necessárias ferramentas computacionais muito mais sofisticadas e maior poder de processamento (Trapnell e Salzberg, 2009).

1.4 Bioinformática

A Bioinformática é um campo interdisciplinar que envolve as áreas de Biologia Molecular, Estatística, Matemática e Ciência da Computação, entre outras, e tem como objetivo a análise de dados biológicos, entre eles seqüências nucleotídicas ou protéicas e a predição de estruturas e funções de macromoléculas. Por ser uma área interdisciplinar, a Bioinformática requer habilidades de diferentes campos de pesquisa. Predições feitas *in silico*, ou no ambiente computacional, devem ser verificadas *in vivo* ou *in vitro* no ambiente de laboratório, ou então, dados adquiridos *in vivo* podem ser analisados *in silico* (Setúbal e Meidanis, 1997).

A Bioinformática pode ser definida como (i) desenvolvimento de métodos computacionais para o estudo de função, evolução e estrutura de proteínas, genes e genomas e (ii) desenvolvimento de métodos para manutenção e análise de informações biológicas advindas de experimentos de genômica e seqüenciamento. Esta é uma disciplina bem estabelecida e lida com informações muito heterogêneas como imagens, diagramas, esquemas e especialmente textos, que podem ser estruturados ou não-estruturados. Os arquivos-texto

de Bioinformática podem ter diversos formatos, como por exemplo, o FASTA, que pode conter seqüências de DNA, RNA ou proteína. O formato que armazena as seqüências advindas de seqüênciamento Illumina é o FASTQ, que é uma extensão simples do FASTA, no qual além das seqüências, são armazenadas também pontuações de qualidade associadas a cada base, em inglês *quality scores* (Cock *et al.*, 2010; Higgs e Attwood, 2005).

As informações geradas por projetos de Bioinformática ou Biologia Molecular são armazenadas em repositórios, ou bancos de dados disponíveis na internet, para disseminação do conhecimento para a comunidade científica. Além disso, atualmente muitas ferramentas estão sendo desenvolvidas para lidar com a revolução que ocorreu nos últimos anos na geração de informações biológicas, causada principalmente pelos seqüenciadores de alto-desempenho (Romano, Giugno e Pulvirenti, 2011).

1.4.1 *Pipelines* de Bioinformática

Um procedimento usual para a análise de amostras obtidas a partir de seqüenciamento de alto-desempenho é a execução de um *pipeline* computacional. *Pipeline* é um conceito utilizado em Ciência da Computação quando os resultados de uma fase constitutem as entradas da próxima fase. Um exemplo de *pipeline* para projetos de alto-desempenho é constituído pelas fases: (i) filtragem, (ii) montagem, (iii) mapeamento e (iv) anotação. Durante cada uma dessas etapas são utilizadas ferramentas adequadas à cada problema. No exemplo citado anteriormente, os problemas são a filtragem das seqüências por qualidade, a montagem de seqüências maiores a partir das originais, o mapeamento das seqüências da montagem em um genoma de referência e a análise. No entanto, é de suma importância frisar que o *pipeline* é extremamente flexível, e deve ser cuidadosamente desenvolvido para se adequar às necessidades de cada projeto. Exemplos de *pipelines* de Bioinformática com propósitos diversos seguem abaixo.

Para fazer um mapeamento de sítios de *splicing* alternativo ao longo de todo o genoma da espécie vegetal *Arabidopsis thaliana*, transcritomas de diferentes tecidos desse organismo em diferentes condições de cultivo foram submetidos a seqüenciamento de alto-desempenho Illumina. Primeiramente, retirou-se os adaptadores das seqüências, as quais foram então filtradas por qualidade. Depois, as seqüências foram alinhadas no genoma de referência. Aquelas que não puderam ser alinhadas diretamente foram comparadas com seqüências previamente anotadas. Quando não era encontrada correspondência, a seqüência era submetida a um novo alinhamento no genoma de referência por um outro programa que detecta novas regiões de *splicing*. Finalmente os modelos gênicos foram

gerados e visualizados (Filichkin *et al.*, 2010).

Um mapeamento de regiões de *splicing* alternativo também foi feito para células humanas, sendo essas de rim, HEK-293T, e de linfócitos B, Ramos B, por meio de seqüenciamento de transcritoma em plataforma Illumina. As seqüências foram mapeadas no genoma de referência, de forma a avaliar níveis de expressão, que foram comparados com outras informações experimentais. Por fim, testes estatísticos foram realizados para validação dos modelos propostos para a expressão gênica dessas células (Sultan *et al.*, 2008).

Técnicas de Bioinformática associadas ao seqüenciamento de alto-desempenho também possibilitam a análise simultânea de mais de um genoma no mesmo experimento, como por exemplo, no caso de simbiontes bacterianos e de seus hospedeiros eucariotos. Após a etapa de filtragem de seqüências, é feita a montagem, que é seguida da separação de seqüências por organismo e posterior anotação (Kumar e Blaxter, 2012).

2 Resultados Anteriores

Neste capítulo, será feito um histórico acerca dos estudos de anticorpos anti-DNA feitos pelo grupo de pesquisa do Laboratório de Imunologia Molecular da Universidade de Brasília. As quatro bibliotecas de peptídeos analisadas na presente Dissertação foram criadas anteriormente por meio de tais pesquisas.

2.1 Estudos de anticorpos anti-DNA no grupo de Imunologia Molecular

O grupo de pesquisa do Laboratório de Imunologia Molecular da Universidade de Brasília possui extensa experiência com o estudo de anticorpos ligantes a ácidos nucléicos. O principal objetivo desses estudos é a elucidação dos princípios que governam tal interação. Além disso, os conhecimentos gerados podem ainda ajudar a esclarecer qual é a contribuição desses complexos imunológicos para a patogênese de certas doenças autoimunes, como lúpus eritematoso sistêmico, onde já foi relatada a presença de auto-anticorpos que reagem contra DNA (Maranhão e Brígido, 2000; Brígido e Stollar, 1991).

Camundongos são modelos bastante estabelecidos em tais pesquisas (Jang e Stollar, 2003), e por isso foram escolhidos para estes estudos. Existem em camundongos 14 famílias que compõem o *locus* da cadeia variável pesada de imunoglobulina, e, para saber qual é a influência dessas famílias na construção de anticorpos anti-ácidos nucléicos, foi feito pelo grupo um levantamento em bancos de dados. Nesse levantamento, descobriu-se que dentre todas as famílias murinas, a VH10 é a que possui maior representação em anticorpos descritos como ligantes a ácidos nucléicos (Guedes, 2009). De todas as seqüências dessa família depositadas em bancos de dados, 59,5% possuíam descrição de ligação a algum tipo de ácido nucléico. Por outro lado, a família VH4 não possuía nenhum representante com tais descrições (tabela 1).

Tabela 1: Análise das seqüências de segmentos gênicos VH murinos depositados no banco de dados GenBank anotados quanto à especificidade (retirado de Guedes, 2009).

Família	Seqüências depositadas[1]	Freqüência de seqüências das famílias[2]	Média do número de germinais[3]	Freqüência das famílias/ n^o de germinais	Freqüência de germinais[4] (%)	Número de seqüências ligantes a ácidos nucléicos[5]	Freqüência de seqüências ligantes em cada família[6] (%)	Freqüência de ligantes/ seqüência de germinais
VH1	2348	66,6	100,0	0,7	56,7	422	18,0	0,2
VH2	237	6,7	15,0	0,4	8,5	96	40,5	2,7
VH3	148	4,2	6,5	0,6	3,7	19	12,8	2,0
VH4	32	0,9	2,0	0,5	1,1	0	0,0	0,0
VH5	327	9,3	12,0	0,8	6,8	120	36,7	3,1
VH6	101	2,9	11,0	0,3	6,2	14	13,9	1,3
VH7	170	4,8	3,0	1,6	1,7	45	26,5	8,8
VH8	44	1,2	8,5	0,1	4,8	5	11,4	1,3
VH9	66	1,9	6,0	0,3	3,4	6	9,1	1,5
VH10	37	1,0	3,5	0,3	2,0	22	59,5	17,0
VH11	9	0,3	3,5	0,1	2,0	0	0,0	0,0
VH12	1	0,0	1,0	0,0	0,6	0	0,0	0,0
VH13	1	0,0	1,0	0,0	0,6	0	0,0	0,0
VH14	4	0,1	3,5	0,0	2,0	0	0,0	0,0
Total	3525	100,0	176,5	-	100,0	749	-	-

[1]Número de seqüências depositadas em bancos de dados contendo CDR H3 de cada família
[2]Percentual de seqüências de cada família na amostra analisada
[3]Média do número estimado de segmentos germinais levando em conta as diferentes linhagens de camundongos
[4]Número de segmentos germinais de cada família dividida pelo repertório total de VH
[5]Número de seqüências descritas como ligantes a ácidos nucléicos
[6]Número de seqüências ligantes a ácidos nucléicos dividido pelo número total de seqüências de cada família

Com base nesse estudo, a hipótese levantada pelo grupo foi que a família VH10 apresentaria uma tendência intrínseca de ligação a ácidos nucléicos, ao contrário da família VH4. Essa hipótese foi formulada a partir da premissa de que quanto maior a importância do VH na interação com o DNA, menos restritivo será o universo de regiões determinantes de complementaridade três de cadeia pesada, CDR H3, que pode acompanhá-lo. Por outro lado, um VH menos adaptado a reconhecer DNA dependerá de um universo mais restrito e mais específico de CDR H3 para se ligar a essa molécula.

Para abordar essa hipótese, o grupo construiu bibliotecas de peptídeos anti-DNA por meio da técnica de *phage display*, que será explicitada a seguir.

2.2 Expressão de peptídeos em fagos filamentosos: *phage display*

Bacteriófagos filamentosos, ou fagos, são vírus que infectam bactérias e são extremamente úteis para diversas aplicações de técnicas de Biologia Molecular. Por manipulações de engenharia genética e a utilização de vetores bacterianos denominados fagomídeos, é possível se expressar uma proteína de fusão, contendo um peptídeo de interesse fusionado à uma proteína estrutural do fago. Essa técnica é comumente chamada de *phage display* e serve para diversas finalidades, tanto de pesquisa básica quanto de pesquisas aplicadas. Por meio dessa metodologia, pode-se construir bibliotecas de peptídeos com diferentes estruturas apresentadas na superfície de fagos, de modo que é possível selecionar dentre as diferentes formas criadas as mais específicas a um determinado antígeno.

A incorporação da proteína de interesse na superfície dos fagos filamentosos é feita fusionando-se esses peptídeos com proteínas estruturais do fago, sendo que as proteínas virais mais utilizadas para este fim são a p3 e a p8, dos genes III e VIII. A p3 compõe a extremidade proximal do capsídeo viral, sendo que de três a cinco cópias dela são expressas no vírus, enquanto que a p8 compõe o corpo do capsídeo, sendo que são expressas cerca de 2.800 cópias dessa proteína no fago selvagem (Maranhão e Brígido, 2002).

O protocolo básico da tecnologia de *phage display* pode ser resumido nos seguintes passos: (1) cria-se uma biblioteca de formas randômicas expressas na superfície de um bacteriófago, (2) imobiliza-se a molécula-alvo numa placa, (3) coloca-se o alvo em contato com a biblioteca de formas randômicas, (4) lava-se a placa para retirar os fagos que não se ligaram ao alvo, (5) elui-se os fagos que se ligaram, (6) amplifica-se os fagos ligantes por meio de infecção em bactérias, (7) repete-se os passos (3) a (6) para se enriquecer

a biblioteca e obter mais clones ligantes ao alvo, e (8) amplifica-se os fagos ligantes, de modo a escolher alguns clones para seqüenciamento. Os processos de ligação e lavagem explicitados anteriormente são chamados de ciclos de seleção, e são repetidos algumas vezes para se obter uma biblioteca final enriquecida em peptídeos ligantes à molécula de interesse (Huang, Ru e Dai, 2011; Menendez e Scott, 2005).

2.3 Geração de bibliotecas de peptídeos anti-DNA por meio de técnicas de *phage display*

Relatos da literatura mostram que o domínio das regiões determinantes de complementaridade três de cadeia pesada, CDR H3, contribui de maneira decisiva no reconhecimento de anticorpos a DNA (Jang *et al.*, 1998). Com base nesse conhecimento, o grupo de Imunologia Molecular da Universidade de Brasília decidiu gerar na superfície de fagos filamentosos bibliotecas de scFvs contendo CDR H3 variáveis nos contextos das famílias murinas de cadeia variável pesada de imunoglobulina VH10 e VH4 (Guedes, 2009; Maranhão, 2001). Isso foi feito por meio da técnica de *phage display*, sendo que o antígeno escolhido para a seleção dos peptídeos foi DNA fita simples. Essa abordagem foi feita para se testar a hipótese explicitada anteriormente, no tópico 2.1 deste capítulo, e consistiu na seleção de formas ligantes a DNA e no seqüenciamento Sanger (figura 4).

Como resultado dessa abordagem, foi observada uma maior diversidade de seqüências CDR H3 no universo da família VH10 quando comparado com a família VH4, sendo que nove seqüências diferentes de CDR H3 foram encontradas em VH10 após a seleção e três foram encontradas em VH4 (tabela 2, Guedes, 2009). Um fato relevante é que foi observado nessa pesquisa um clone extremamente freqüente, o qual estava presente em todos os ciclos de seleção e nas amostras das duas famílias. A preponderância dessa seqüência dentre os clones analisados reduziu o universo amostral comprometendo conclusões com embasamento estatístico.

Tabela 2: Diversidade de seqüências de peptídeos ligantes a DNA. As biliotecas analisadas foram obtidas por meio da técnica de *phage display* (adaptado de Guedes, 2009).

Clones seqüenciados		Clones analisados[1]		Diversidade de seqüências[2]	
VH10	VH4	VH10	VH4	VH10	VH4
48	48	31	20	9	3

[1]Clones analisados segundo os critérios de qualidade e correto alinhamento da CDR H3
[2]Seqüências distintas quanto à composição de resíduos de aminoácidos

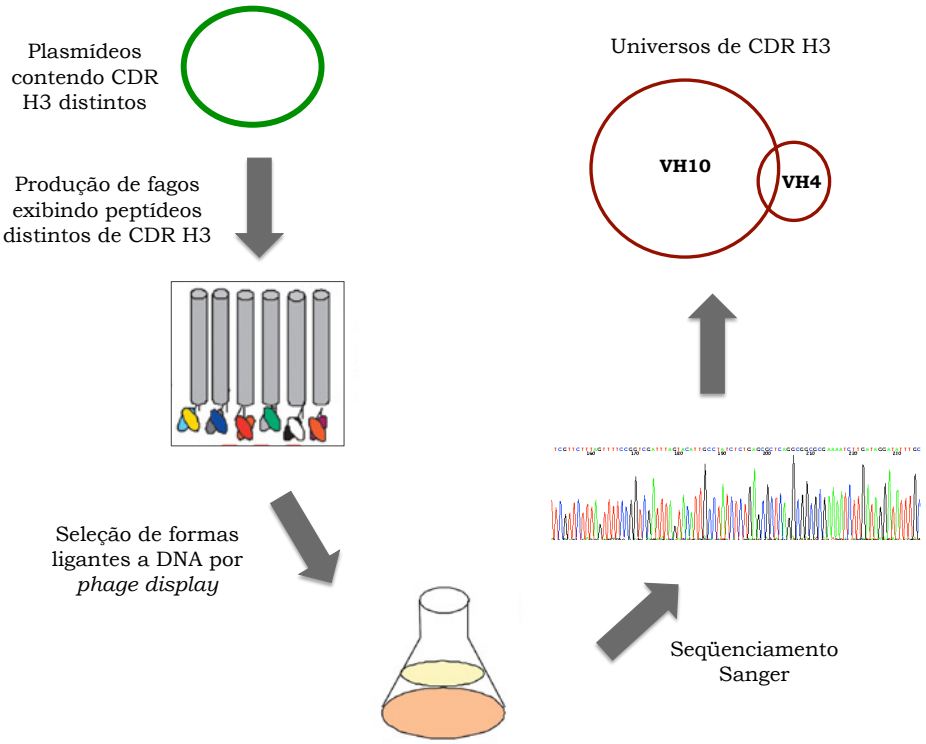


Figura 4: **Abordagem experimental utilizada na pesquisa sobre anticorpos anti-DNA feita no Laboratório de Imunologia Molecular.** Para se estabelecer a contribuição da família VH10 em peptídeos com comportamento de ligação a ácidos nucléicos, foram criadas bibliotecas de CDR H3 variáveis em contexto das famílias VH10 e VH4. Após a seleção das bibliotecas contra DNA, foi feito seqüenciamento Sanger (adaptado de Guedes, 2009).

Portanto, se tornou necessária a busca de outros métodos que permitissem maior profundidade de análise. Recentemente, pesquisadores mostraram que a associação do método de *phage display* com a técnica de seqüenciamento de alto-desempenho fornece um alcance muito superior de cobertura da variabilidade quando se compara com outros métodos tradicionais experimentais, como clonagem seguida de seqüenciamento Sanger (Ravn *et al.*, 2010). Assim, o seqüenciamento de nova geração das bibliotecas de CDR H3 construídas nesta pesquisa parece fornecer uma alternativa adequada, de modo que se possa obter maior profundidade de análise.

3 Objetivos

3.1 Objetivo Geral

• Confirmação da existência de tendência intrínseca à formação de anticorpos com afinidade por DNA dos segmentos gênicos da família murina de cadeia variável pesada dez de imunoglobulina, VH10, e estudo da contribuição da família murina de cadeia variável pesada quatro de imunoglobulina, VH4, na formação desses anticorpos.

3.2 Objetivos Específicos

• Elaboração de códigos de identificação, *barcodes*, para a distinção *in silico*, após o seqüenciamento de alto-desempenho, das quatro bibliotecas de estudo: VH4 ciclo zero, VH10 ciclo zero, VH4 ciclo quatro e VH10 ciclo quatro;

• Seqüenciamento de alto-desempenho em plataforma Illumina das quatro bibliotecas de estudo, citadas anteriormente, feito em uma única canaleta;

• Desenvolvimento de um *pipeline* computacional de análise e comparação das seqüências obtidas a partir do seqüenciamento Illumina;

• Análise de variabilidade e enriquecimento das bibliotecas de estudo;

• Análise de padrões de seqüências de CDR H3 em contexto das famílias VH10 e VH4 que tiveram enriquecimento positivo e negativo após seleção das bibliotecas iniciais, do ciclo zero, contra DNA.

4 Material e Métodos

Este trabalho foi composto a partir de duas metodologias diferentes e complementares, sendo a primeira experimental, feita para a preparação de amostras para seqüenciamento de alto-desempenho Illumina, e a segunda computacional, feita para o tratamento e análise das seqüências obtidas a partir desta plataforma. Este capítulo foi, portanto, dividido em duas partes, a primeira relacionada à abordagem experimental e a segunda à análise por Bioinformática.

4.1 Abordagem experimental

Em um trabalho anterior, como explicitado no capítulo 2, foram geradas bibliotecas de peptídeos contendo regiões correspondentes à CDR H3 em contexto das duas famílias VH10 e VH4 (Guedes, 2009). Quatro dessas bibliotecas foram escolhidas para este estudo, sendo estas as iniciais, pertencentes ao ciclo zero de seleção, e as enriquecidas em peptídeos com afinidade a DNA, pertencentes ao ciclo quatro de seleção. As bibliotecas do ciclo quatro estavam disponíveis sob a forma de células bacterianas infectadas com fagos armazenadas no glicerol a -20^{o}C, enquanto que as bibliotecas do ciclo zero estavam disponíveis sob a forma de fagos armazenados a -4^{o}C (figura 5).

Primeiramente, foi feita a reamplificação dos fagos do ciclo zero, já que o DNA dessas amostras se encontrava degradado. Posteriormente, foi feita a extração de DNA das células bacterianas infectadas, seguido da amplificação da região de interesse por meio de reações de PCR. Ao mesmo tempo foram inseridos nestas amostras códigos de identificação, *barcodes*, para posterior classificação após o seqüenciamento de alto-desempenho. Em seguida, foram feitos experimentos de clonagem seguidos de seqüenciamento automático Sanger ABI, para verificação da integridade desses códigos de identificação. Finalmente, as amostras foram preparadas e seqüenciadas na plataforma de alto-desempenho Illumina (figura 5).

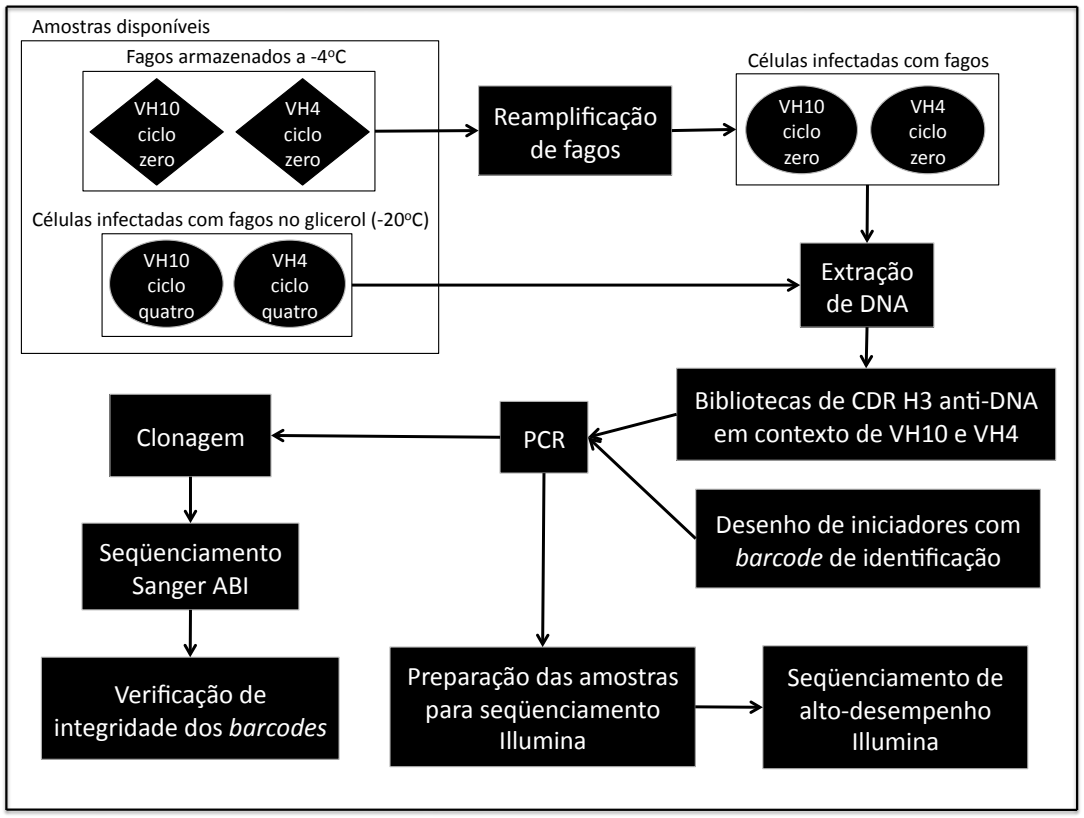


Figura 5: **Representação esquemática da abordagem experimental utilizada neste trabalho.** As etapas experimentais visaram a preparação das amostras para seqüenciamento de alto-desempenho Illumina.

4.1.1 Material da abordagem experimental

4.1.1.1 Bibliotecas de estudo

VH4 ciclo zero, **VH10 ciclo zero**, **VH4 ciclo quatro** e **VH10 ciclo quatro** - as bibliotecas do ciclo zero eram referentes a fagos armazenados a -4^{o}C, e as bibliotecas do ciclo quatro eram referentes a bactérias infectadas com fagos armazenadas no glicerol a -20^{o}C. Cada uma das bibliotecas continha DNA correspondente a regiões de CDR H3 variáveis em contexto das famílias murinas VH4 ou VH10.

4.1.1.2 Linhagem bacteriana

XL1-Blue (Stratagene®) - *rec*A1 *end*A1 *gyr*A96 *thi*-1 *hsd*R17 *sup*E44 *rel*A1 *lac* [F'*pro*AB *lac*IqZ M15Tn10 (Tet®)] (Sambrook *et al.*, 2001).

Essa linhagem bacteriana foi utilizada na reamplificação de fagos do ciclo zero e na transformação bacteriana.

4.1.1.3 Vetor utilizado

pGEM®-T easy Vector Systems - para clonagem de produtos de PCR (Promega®, número de catálogo A1360).

4.1.1.4 Bacteriófago auxiliar

VCSM13 - Derivado do bacteriófago M13 com gene III mutado: origem de duplicação plasmidial derivada do p15 e gene de resistência à canamicina (Stratagene®, número de catálogo 2000251).

Esse bacteriófago auxiliar foi utilizado na reamplificação de fagos do ciclo zero.

4.1.1.5 Oligonucleotídeos utilizados para amplificação das bibliotecas

Os oligonucleotídeos foram sintetizados e posteriormente solubilizados em água MiliQ para concentração de uso de 10 ρmoles/μL. Os produtos de PCR gerados a partir desses iniciadores possuíam tamanho de 77 pares de base.

Tabela 3: Oligonucleotídeos sintéticos utilizados na amplificação das quatro bibliotecas de estudo. *Barcodes* estão marcados em negrito.

Oligonucleotídeo e biblioteca a ser amplificada	Seqüência
Senso de VH4 ciclo zero	5' CAA CAG CC**G** TTT ATT ACT GCG T 3'
Senso de VH4 ciclo quatro	5' CAA CAG CCC **A**TT ATT ACT GCG T 3'
Senso de VH10 ciclo zero	5' ACA **G**AG CCA TGT ATT ACT GCG T 3'
Senso de VH10 ciclo quatro	5' ACA C**T**G CCA TGT ATT ACT GCG T 3'
Reverso comum[1]	5' TGA GGT TCC TTG ACC CCA AT 3'

[1]O iniciador reverso utilizado foi o mesmo para a amplificação de todas as bibliotecas

4.1.1.6 Meios de cultura e soluções para bactérias

Após dissolver os reagentes em água destilada, todos os meios de cultura eram autoclavados a 120^{o}C por 15 minutos.

Meio LB (Luria-Bertani)

Peptona de caseína 1,0% (p/v)

Extrato de levedura 0,5% (p/v)

NaCl 1,0% (p/v)

pH 7,0.

Meio LB ágar

Meio LB adicionado de ágar bacteriológico a 1,4% (p/v).

Meio LB top ágar

Meio LB adicionado de ágar bacteriológico a uma concentração final de 0,7% (p/v).

Meio SB (Super Broth)

Peptona de caseína 3,0% (p/v)

Extrato de levedura 2,0% (p/v)

MOPS 1,0% (p/v)

pH 7,0.

Meio SOB

Bacto-triptona 2,0% (p/v)

Extrato de levedura 0,5% (p/v)

NaCl 0,06% (p/v)

KCl 0,002% (p/v)

pH 7,0.

Meio SOC

Meio SOB 98 mL

Solução estoque de Mg^{2+} 2 M 1 mL

Solução estoque de glicose 2 M 1 mL.

Solução estoque de glicose 2 M

Esterilizada por filtração e estocada a 4^{o}C.

Solução estoque de Mg 2 M

$MgCl^2$ 1 M

$MgSO_4$ 1 M

Esterilizada por filtração e estocada a 4^{o}C.

4.1.1.7 Soluções e material para preparo de células competentes e transformação bacteriana

Cubetas de eletroporação (Gene Pulser/MicroPulser Cuvettes, Biorad®, número de catálogo 165-2086).

Glicerol 10% (v/v)

Esterilizado por filtração e estocado a 4^{o}C.

4.1.1.8 Soluções e reagentes para eletroforese em gel de agarose

GelRed^{TM}Nucleic Acid Gel Stain - para corar amostras de ácidos nucléicos em géis de agarose, dissolvido em água (Biotium®, número de catálogo 41003).

Tampão de corrida TEB 10X

Trizma base 0,89 M

Ácido Bórico 0,89 M

EDTA 0,02 M

dH_2O q.s.p. 1 L

pH 8,0.

Tampão de corrida TAE 50X

Tampão Tris-Acetato 2 M

Trizma-base 242 g

Ácido Acético Glacial 57,10 mL

EDTA pH 8,0 0,05 M

dH_2O q.s.p. 1 L.

4.1.1.9 Marcadores moleculares para DNA

1 kb plus DNA Ladder (Invitrogen®, número de catálogo 10787-026) Fragmentos de DNA em pb: 100; 200; 300; 400; 500; 650; 850; 1.000; 1.650; 2.000; 3.000; 4.000; 5.000; 6.000; 7.000; 8.000; 9.000; 10.000; 11.000; 12.000.

Low Mass DNA Ladder (Invitrogen®número de catálogo 10068-013) Mistura equimolar de fragmentos de DNA em pb de 2.000; 1.200; 800; 400; 200 e 100. Utilizando 2 μL do marcador, os fragmentos correspondem à massa de 100; 60; 40; 20; 10 e 5 ng, respectivamente.

100 bp DNA Ladder (Invitrogen®número de catálogo 15628-019) Fragmentos de

DNA em pb: 100; 200; 300; 400; 500; 600; 700; 800; 900; 1.000; 1.100; 1.200; 1.300; 1.400; 1.500; 2.072.

4.1.1.10 *Kits* comerciais

QIAprep Spin Miniprep Kit (250) - para preparação plasmidial em pequena escala (Qiagen®, número de catálogo 27106);

Qubit Fluorometer - para quantificação de ácidos nucléicos. Invitrogen (número de catálogo Q32860).

4.1.1.11 Enzimas

T4 DNA ligase – Invitrogen®(número de catálogo 15224-017);

Platinum®Taq DNA polimerase – Invitrogen®(número de catálogo 10966-018);

Taq DNA polimerase – Invitrogen®(número de catálogo 10342-053).

4.1.2 Métodos da abordagem experimental

4.1.2.1 Desenho de iniciadores com *barcodes* de identificação

Primeiramente foram desenhados os iniciadores de amplificação da biblioteca VH10 ciclo zero (tabela 3), e para isso foram utilizados os programas Primer 3, PerlPrimer e IDT (Primer3-*homepage*, 2011; PerlPrimer-*homepage*, 2011; IDT-*homepage*, 2011). Os parâmetros avaliados no desenho foram a temperatura de desnaturação do sistema, o conteúdo de GC dos iniciadores, a estabilidade de possíveis estruturas de grampo e a probabilidade da ocorrência de auto-anelamento. A partir do desenho desse par de iniciadores, foram desenhados analogamente os das outras três bibliotecas, sendo que o iniciador reverso foi mantido o mesmo, ou seja, foi comum a todas as bibliotecas. Além disso, a posição do iniciador senso também foi mantida. Portanto, a região amplificada possui o mesmo tamanho em todas as bibliotecas e corresponde a 77 pb.

O pequeno tamanho dos produtos de PCR foi almejado visando-se o seqüenciamento de alto-desempenho Illumina, considerando que as seqüências geradas por essa plataforma

são igualmente pequenas, de até 150 pb. O objetivo era que cada seqüência gerada cobrisse todo o produto da PCR, de modo que não só a CDR H3 pudesse ser lida, mas também o código indicativo de cada biblioteca, *barcode*. Dessa forma, as quatro bibliotecas poderiam ser seqüenciadas na mesma canaleta, com o objetivo de se reduzir o custo desse serviço.

4.1.2.2 Amplificação das bibliotecas por PCR

O sistema de amplificação das bibliotecas se consistiu em cerca de 30 ng de DNA molde, 1 μL de cada iniciador, senso e reverso, numa concentração de 100 μM, 5 μL de tampão da enzima numa concentração de 10X, 2,5 μL de *mix* de dNTP 10 mM, 3 μL de $MgCl_2$ 50 mM, 0,2 μL da enzima Taq DNA polimerase Platinum e H_2O miliQ para completar o sistema até um volume final de 50 μL.

As condições da reação de PCR foram as seguintes: um ciclo inicial de 5 minutos a 94^{o}C seguido de cinco ciclos de 94^oC por 45 segundos, 55^oC por 30 segundos e 72^oC por 45 segundos, seguidos de 30 ciclos de 94^oC por 45 segundos, 54^oC por 30 segundos e 72^oC por 45 segundos, seguidos de uma extensão final de 72^oC por 10 minutos.

4.1.2.3 Análise de DNA em gel de agarose

A agarose foi preparada com concentração de 2,5% ou 4,0% em tampão TEB 1X ou TAE 1X. As amostras de DNA foram preparadas com tampão de amostra GelRed concentração 1X, aplicadas no gel e submetidas a eletroforese em tampão TEB ou TAE 1X (Sambrook *et al.*, 2001). Para visualização do DNA, incidia-se luz ultravioleta no gel por meio de um transluminador (Pharmacia-LKB®), e a imagem era digitalizada em aparato de fotodocumentação (Video Graphic Printer UP- 895 CE, Sony®).

4.1.2.4 Ligação de produtos de PCR

Dois sistemas de ligação foram feitos, um para a ligação efetiva de produtos de PCR com o vetor e o outro para teste de controle negativo, sem inserto. As amostras foram incubadas a temperatura ambiente por pelo menos 8 horas, e em seguida todo o sistema de ligação foi utilizado para transformar alíquotas de 100 μL de células XL1-Blue eletrocompetentes. Diluições das culturas obtidas pelos dois sistemas foram semeadas em placa de meio LB ágar contendo ampicilina a uma concentração de 200 μg/mL, para a determinação da eficiência de transformação. O sistema controle não apresentou nenhum clone, enquanto que o sistema de ligação dos produtos de PCR com o vetor demonstrou

comportamento padrão de ligação efetiva.

O sistema de ligação dos produtos de PCR se consistiu em 50 ng de vetor pGEM®-T *easy*, 3 μL de produto da PCR, 2 μL de tampão da enzima numa concentração de 5X, 1 μL da enzima T4 DNA ligase e H_2O miliQ para completar o sistema até um volume final de 10 μL. Por sua vez, o sistema de ligação controle, ou seja, aquele sem inserto, se consistiu em 50 ng de vetor pGEM®-T *easy*, 2 μL de tampão da enzima numa concentração de 5X, 1 μL da enzima T4 DNA ligase e H_2O miliQ para completar o sistema até um volume final de 10 μL.

4.1.2.5 Análise da clonagem

Após a ligação dos produtos de PCR no vetor pGEM®-T *easy*, verificou-se a eficácia do experimento por meio de PCR de colônia utilizando-se os iniciadores do próprio vetor, pUC/M13 senso e reverso. Colônias eram escolhidas randomicamente, inoculadas em 100 μL de meio LB com ampicilina em concentração de 50 μg/mL. Dez microlitros desse inóculo eram adicionadas à reação como DNA molde.

O sistema de amplificação para a análise da clonagem se consistiu em cerca de 10 μL de DNA molde a uma concentração de cerca de 30 ng/μL, 0,5 μL de cada iniciador, senso e reverso, numa concentração de 100 μM, 3 μL de tampão da enzima numa concentração de 10X, 1,5 μL de *mix* de dNTP 10 mM, 1,5 μL de $MgCl_2$ 50 mM, 0,2 μL da enzima Taq DNA polimerase Platinum e H_2O miliQ para completar o sistema até um volume final de 30 μL.

As condições da reação de PCR foram as seguintes: um ciclo inicial de 5 minutos a 94^{o}C seguido de 30 ciclos de 94^oC por 30 segundos, 55^oC por 30 segundos e 72^oC por 1 minuto, seguidos de uma extensão final de 72^oC por 10 minutos.

4.1.2.6 Preparação de células eletrocompetentes

A preparação de células eletrocompetentes neste trabalho foi feita de acordo com o protocolo descrito abaixo, o qual foi adaptado (Rader, Steinberger e Barbas, 2004).

1. Uma colônia isolada de XL1-Blue é inoculada em 10 mL de meio SB, contendo tetraciclina a uma concentração final de 30 μg/mL. Esse pré-inóculo é então incubado durante a noite sob agitação de 250 rpm a 37^{o}C.

2. Seis Erlenmeyers de 1 L contendo 200 mL de meio SB são inoculados com 1,5 mL

do pré-inóculo do item 1, cada um contendo 5 mL de glicose 20% (p/v) e 5 mL de MgCl2 1 M (não se adiciona antibiótico nesta etapa). Os frascos são então incubados sob agitação de 250 rpm a 37^{o}C, até os inóculos atingirem uma densidade óptica (OD) a 600 ηm igual a 0,7 (a densidade óptica normalmente chega a esse valor após 2 horas e meia de incubação).

3. Depois da densidade óptica atingir o valor assinalado no item 2, os frascos são resfriados em gelo, assim como os frascos utilizados na etapa posterior de centrifugação.

4. Centrifuga-se as culturas a 3.000 x g por 20 min a 4^{o}C, sendo o sobrenadante descartado em seguida.

5. Os sedimentos são ressuspendidos em 25 mL de glicerol 10% (v/v) gelado por meio da utilização de pipetas pré-resfriadas. Os frascos são combinados dois a dois e mais glicerol é adicionado até um volume final de 150 mL.

6. Centrifuga-se as amostras a 3.000 x g por 20 min a 4^{o}C, sendo o sobrenadante descartado em seguida.

7. Os sedimentos são ressuspendidos em 25 mL de glicerol 10% (v/v) gelado por meio da utilização de pipetas pré-resfriadas. Então, os frascos são transferidos para tubos de centrífuga de 50 mL e novamente centrifugados nas condições citadas no item 6.

8. O sobrenadante é descartado e as células são ressuspendidas no volume residual de glicerol (usualmente 1/2 mL). Posteriormente, alíquotas de 150 μL são distribuídas em microtubos novos e estéreis e em seguida congeladas em banho de álcool/gelo seco.

9. As alíquotas são estocadas a -80^{o}C e podem ser utilizadas tanto para eletroporação quanto para pré-inóculo em experimentos de clonagem.

4.1.2.7 Preparação de fago auxiliar e reamplificação de bibliotecas de fagos

O fago auxiliar VCSM13 e a reamplificação das bibliotecas de fagos do ciclo zero foram feitos de acordo com protocolos retirados da literatura (Rader, Steinberger e Barbas, 2004).

4.1.2.8 Seqüenciamento de DNA Sanger ABI e análise de seqüências

Clonagens contendo as amostras de PCR foram seqüenciadas utilizando-se o aparelho automático Sanger ABI 3130 XL Genetic Analyzer (Applied Biosystems®). Foram utilizados 250 ηg de DNA, o oligonucleotídeo M13 senso e reverso do vetor pGEM®-T *easy* e o *kit* ABI Prism BigDye Terminator v.3.1 Cycle Sequencing.

As seqüências obtidas foram analisadas utilizando-se as ferramentas Phred e CAP3 disponíveis no PHPH da página: `www.biomol.unb.br` (Togawa e Brigido, 2003). Após análise de qualidade, as seqüências foram manipuladas e analisadas no programa BioEdit Sequence Aligment Editor (Hall, 2007).

4.1.2.9 Preparação de amostras para seqüenciamento Illumina

Foram quantificados os produtos de PCR de amplificação das quatro bibliotecas de estudo. Quantidades iguais de cada biblioteca, 250 ng, foram colocadas num único microtubo eppendorf®, de modo que ao final, foi obtida uma mistura totalizando 1 μg de DNA, a qual foi seca a vácuo em aparelho SpeedVac.

4.1.2.10 Seqüenciamento de alto-desempenho Illumina

A amostra contendo as quatro bibliotecas foi enviada para o Instituto *Scripps Research* localizado na cidade de San Diego/Califórnia nos Estados Unidos (`http://www.scripps.edu/california/research/ngs/technology.html`). Lá as amostras foram purificadas, para que os produtos de PCR de 77 pb fossem separados de outros componentes desinteressantes para a reação de seqüenciamento, como DNA molde, restos de iniciadores e enzimas.

As seqüências foram geradas pelo aparelho Illumina®HiSeq 2000 utilizando-se a técnica *paired-end* para o seqüenciamento das amostras em somente uma canaleta da placa, com a geração de seqüências de 150 pares de base. Foi utilizado para isso o Sequencing Kit V3, Paired-end 2x150 + 7 indexes. A plataforma de identificação das bases seqüenciadas foi a CASAVA 1.8, sendo que o formato dos arquivos gerados foi o FASTQ versão CASAVA 1.8. Este formato utiliza a codificação de caracteres ASCII + 33.

4.2 Abordagem de Bioinformática

O *pipeline* de Bioinformática desenvolvido neste trabalho foi dividido em três etapas principais, (i) filtragem, (ii) classificação e (iii) análise (figura 6). Na figura 6a, estão descritos os passos detalhados das duas primeiras etapas. Na primeira, (i) filtragem, as seqüências FASTQ recebidas do aparelho Illumina foram filtradas por qualidade. Já na segunda, (ii) classificação, as seqüências filtradas foram classificadas entre as quatro bibliotecas de estudo, sendo que somente a parte de 27 nucleotídeos correspondente à codificação da CDR H3 foi mantida. Essas seqüências nucleotídicas foram então traduzidas, e finalmente, as seqüências únicas foram contadas, tanto as de nucleotídeos quanto as de aminoácidos. Ao final de toda essa preparação, foram obtidos arquivos referentes a cada uma das quatro bibliotecas contendo o número de aparecimento de cada seqüência. Tais arquivos foram utilizados em todos os estudos subseqüentes e possuíam duas colunas, a primeira com a seqüência e a segunda com a contagem.

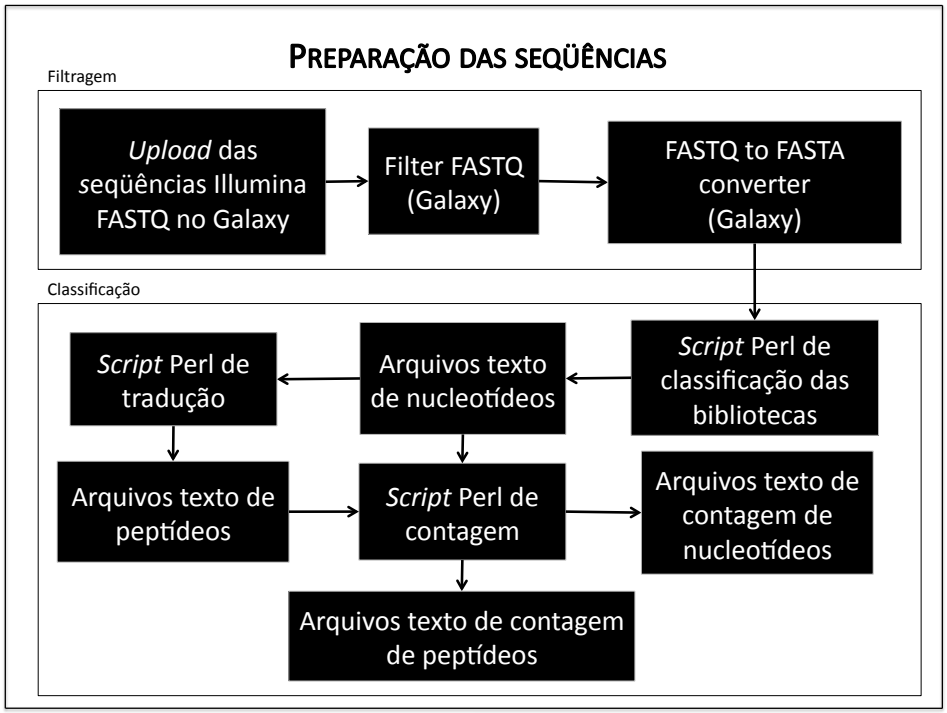
Posteriormente, esses arquivos de contagem foram trabalhados na etapa final do *pipeline*, (iii) análise (figura 6b). Os arquivos de contagem de nucleotídeos foram utilizados para estudos de composição da CDR H3 e divergência de Kullback-Leibler. Já os arquivos de contagem de peptídeos foram utilizados não só para esses dois estudos, mas também para estudos de enriquecimento, comparação de padrões e variabilidade.

4.2.1 Preparação dos dados

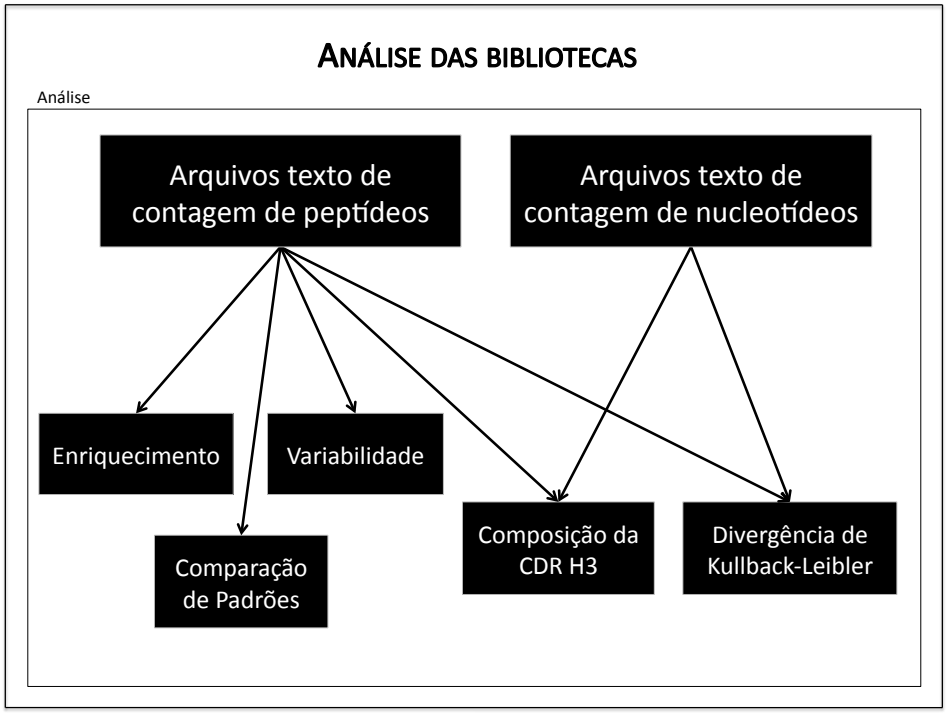
4.2.1.1 Filtragem por qualidade

Primeiramente, foi feita uma avaliação geral de qualidade das seqüências Illumina recebidas com a utilização do programa FastQC instalado localmente. Posteriormente, para o processamento inicial dos arquivos, instalou-se também localmente o *framework* Galaxy em sistema operacional Ubuntu server 10.04.4 LTS, numa servidora HP ProLiant DL380 G7 de 30 GB de memória RAM oito núcleos, com dois processadores Intel Xeon E5506 quad-core, cada núcleo de 2.13 GHz. Foi feito o carregamento dos arquivos Illumina no Galaxy e o formato destes foi classificado como FASTQSANGER, que é o equivalente ao FASTQ Illumina versão CASAVA 1.8, como especificado no item 4.1.2.10.

Em seguida, foi feita a filtragem das seqüências por meio do programa Filter FASTQ no Galaxy. Os parâmetros utilizados na filtragem foram o tamanho mínimo da seqüência sendo de 115 pares de base, qualidade Phred mínima de 20.0 por base, com no máximo 30



(a)



(b)

Figura 6: *Pipeline* **de Bioinformática desenvolvido para análise das seqüências Illumina referentes às bibliotecas de estudo: VH4 e VH10 ciclos zero e quatro.** (a) Preparação das seqüências Illumina, que consiste nas seguintes etapas do *pipeline*: (i) filtragem e (ii) classificação. (b) Análise das seqüências das bibliotecas, que consiste na etapa final do *pipeline*: (iii) análise.

bases com qualidade abaixo do valor mínimo estabelecido. O valor mínimo de qualidade escolhido reflete um mínimo de acurácia de 99% de que a base lida esteja correta, como explicado no item 1.3 do capítulo 1. Já o parâmetro escolhido de 30 bases com qualidade abaixo do valor mínimo foi escolhido por causa dos resultados obtidos a partir da análise do programa FastQC, explicitados no item 5.2 do capítulo 5, nos quais se observa que entre as posições 120 e 150 das seqüências a qualidade média cai (esses resultados podem ser vistos na figura 11).

Após essa etapa de filtragem, foi utilizado o programa FASTQ to FASTA, também no Galaxy, o qual fez a conversão de formatos FASTQ para FASTA, para facilitar a manipulação subseqüente dos arquivos. Em todas as etapas seguintes, foi utilizada a linguagem Perl de programação, e estas foram desenvolvidas em sistema operacional Mac OS X 10.6.8 em um MacBook Pro Apple 4 GB de memória RAM, com processador Intel Core 2 Duo de 2.4 GHz.

4.2.1.2 Classificação das seqüências

Para classificar as seqüências filtradas FASTA entre as quatro bibliotecas de estudo, foi desenvolvido um *script* Perl que utilizasse os códigos de identificação, *barcodes*, em expressões regulares. Estas basearam-se nos *barcodes* de cada biblioteca em posições iniciais, seguidos de porções invariáveis inerentes a cada biblioteca original, uma seqüência variável, *string*, de 27 nucleotídeos correspondente à CDR H3, seguidos de um códon TAT invariável na porção final da seqüência (tabela 4).

Tabela 4: Expressões regulares utilizadas no desenvolvimento do *script* Perl para a identificação dos *barcodes* e classificação das seqüências nas quatro bibliotecas de estudo. A expressão (\w{27}) se refere à parte de 27 nucleotídeos correspondente à codificação da CDR H3, a qual é mantida na saída do *script* Perl. Os *barcodes* estão marcados em negrito.

Biblioteca	Expressão regular
VH4 ciclo zero	/**G**TTTATTACTGCGTACGAGAA(\w{27})TAT/
VH4 ciclo zero - *cr*[1]	/ATA(\w{27})TTCTCGTACGCAGTAATAAAC/
VH4 ciclo quatro	/C**A**TTATTACTGCGTACGAGAA(\w{27})TAT/
VH4 ciclo quatro - *cr*[1]	/ATA(\w{27})TTCTCGTACGCAGTAATAATG/
VH10 ciclo zero	/**G**AGCCATGTATTACTGCGTACGAGAA(\w{27})TAT/
VH10 ciclo zero - *cr*[1]	/ATA(\w{27})TTCTCGTACGCAGTAATACATGGCTC/
VH10 ciclo quatro	/C**T**GCCATGTATTACTGCGTACGAGAA(\w{27})TAT/
VH10 ciclo quatro - *cr*[1]	/ATA(\w{27})TTCTCGTACGCAGTAATACATGGCAG/

[1]Abreviatura de complemento reverso

As expressões regulares foram feitas ainda de modo a não permitir modificações, ou *mismatches*. Como o aparelho Illumina pode seqüenciar as amostras em ambos os sentidos da fita de DNA, o *script* foi feito de modo a reconhecer as expressões regulares tanto no sentido 5' - 3' quanto no seu complemento reverso, de modo a abranger todas as seqüências recebidas. Se alguma dessas expressões regulares é reconhecida na seqüência, a expressão variável entre parênteses, (\w{27}), é mantida e impressa numa das quatro saídas do programa, sendo cada saída representativa de uma biblioteca. Cada arquivo de saída mantem, portanto, somente as seqüências de 27 pb codificadoras da CDR H3, uma a cada linha.

Também por meio de *scripts* Perl foram feitas a tradução e a contagem das seqüências únicas. Ao final foram obtidos arquivos com duas colunas, uma com a seqüência e a outra com sua contagem ordenada de aparecimento.

Foram observadas seqüências de nucleotídeo que continham bases "N", ou seja, base não identificada no seqüenciamento. Neste caso, a seqüência inteira foi excluída da análise. Além disso, foram tratadas seqüências peptídicas que continham códons âmbar, simbolizados pelo caractere "*", os quais vieram originalmente de seqüências nucleotídicas TAG. Esse símbolo indica normalmente códon de terminação, no entanto, a linhagem de *Escherichia coli* utilizada na preparação das amostras desse trabalho, a XL1-Blue supE44, possui a característica de poder interpretar tais códons também como o aminoácido Glutamina, ou "Q" (Singaravelan, Roshini e Munavar, 2010). Levando isso em consideração, decidiu-se substituir esses caracteres por Glutamina, feito também por meio de *scripts* Perl.

4.2.2 Análise dos dados

4.2.2.1 Variabilidade das bibliotecas

Para o estudo de variabilidade, foram utilizadas duas abordagens. A primeira consiste no cálculo do número de seqüências únicas de CDR H3 em cada biblioteca, enquanto que a segunda consiste no cálculo da entropia.

Na primeira abordagem foi feita a contagem do número de seqüências peptídicas únicas de cada biblioteca, seguida de uma normalização. Esse último passo foi necessário devido a uma grande divergência observada na classificação das seqüências Illumina por biblioteca, o que foi discutido na seção 5.3 do capítulo 5. Essa normalização consistiu na divisão do número de seqüências encontrado, tabela 6, pela freqüência de seqüências

classificadas na biblioteca em questão, tabela 5, seguido da multiplicação por 0,25, ou 25%. A multiplicação foi feita porque o esperado era se encontrar uma freqüência de 25% de seqüências classificadas em cada uma das quatro bibliotecas.

Por sua vez, o cálculo da entropia total de cada grupo de seqüências foi feito por meio de um *script* Perl, no qual foi implementada a equação da entropia de Shannon, que segue abaixo (Ueltschi-*homepage*, 2012).

$$S = \sum_{n=1}^{N} p_n \log_2 p_n$$

S: entropia total da seqüência

p_n: freqüência observada do símbolo n

N: número de símbolos distintos na seqüência, quatro para DNA e 20 para proteínas

4.2.2.2 Composição da CDR H3 e enriquecimento de seqüências

Para o estudo de padrões em anticorpos ligantes a ácidos nucléicos, os grupos com enriquecimento positivo, negativo e neutro de VH10 e VH4 foram comparados. Para tal, esses grupos foram submetido à plataforma WebLogo instalada localmente. Cálculos de parâmetros de composição de peptídeos foram feitos pelo programa disponível na internet ProtParam (Gasteiger *et al.*, 2005), e gráficos foram elaborados no programa Microsoft Excell. Já o enriquecimento de uma determinada seqüência após a seleção desta contra DNA foi definido como a divisão do número de vezes que esta era observada no ciclo quatro do seqüenciamento pelo número de vezes que era observada no ciclo zero, como na fórmula que segue abaixo.

$$E = N_4/N_0$$

E: enriquecimento

N_4*: número de vezes que a seqüência foi observada no ciclo quatro*

N_0*: número de vezes que a seqüência foi observada no ciclo zero*

Condições:

Se $N_4 \neq 0$ e $N_0 = 0$, Então $N_0 = 1$

Se $N_4 = 0$ e $N_0 \neq 0$, Então $N_4 = 1$

4.2.2.3 Cálculo da divergência de Kullback-Leibler

A divergência de Kullback-Leibler é uma equação da Teoria da Informação que quantifica a proximidade de duas distribuições probabilísticas, em outras palavras, quanto uma distribuição observada se aproxima de um modelo esperado (Shlens, 2007). Essa equação foi utilizada para se comparar a freqüência observada dos nucleotídeos ou aminoácidos no seqüenciamento de alto-desempenho com a freqüência esperada de aparecimento desses elementos na seqüência de CDR H3, de modo a se avaliar o quanto as amostras biológicas se aproximaram do que era esperado. A implementação desse algoritmo foi feita em linguagem Perl. Esta é uma aplicação da equação de entropia de Shannon e segue explicitada abaixo.

$$D_{KL}(P \parallel Q) = \sum_{i} P(i) \log_2 \frac{P(i)}{Q(i)}$$

$D_{KL}(P \parallel Q)$: divergência de Kullback-Leibler, que deve ser igual a zero se as distribuições forem equivalentes ou não negativa se forem divergentes

$P(i)$: distribuição observada

$Q(i)$: distribuição modelo esperada

Foram feitas duas implementações dessa equação, a primeira para se avaliar o quanto as amostras biológicas referentes ao ciclo zero se aproximaram do desenho de construção das bibliotecas primordiais. Nessa implementação, $P(i)$ foi considerado como as probabilidades observadas no ciclo zero enquanto que $Q(i)$ foi considerado como as probabilidades do desenho de construção, que se encontram explicitadas na figura 7.

Já a segunda implementação foi feita para se avaliar como ocorreu a seleção de peptídeos do ciclo zero para o ciclo quatro no contexto de cada segmento germinal. Neste caso, $P(i)$ foi considerado como as probabilidades observadas no ciclo quatro do segmento germinal enquanto que $Q(i)$ foi considerado como as probabilidades observadas no ciclo zero do segmento germinal em questão.

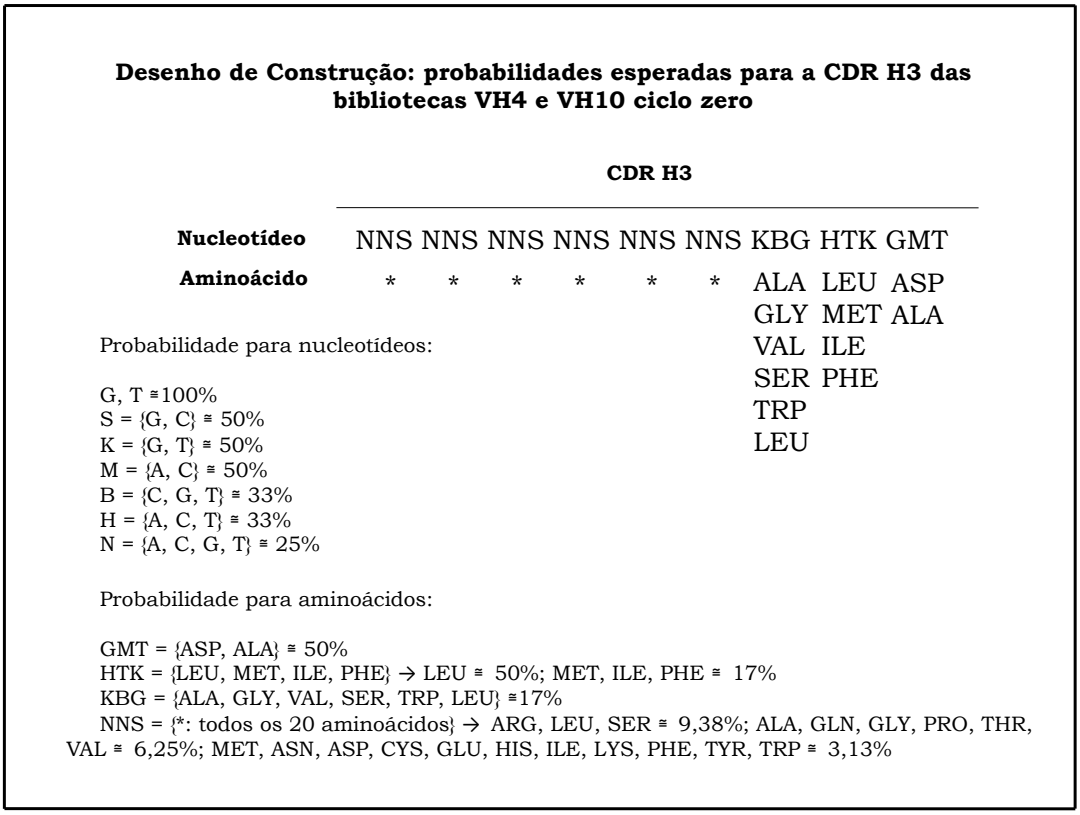


Figura 7: **Probabilidade esperada de aparecimento de nucleotídeos e aminoácidos por posição na construção da CDR H3 das bibliotecas VH4 e VH10 ciclo zero.** A probabilidade de aparecimento esperada para cada nucleotídeo e cada aminoácido está explicitada na figura e depende da sua posição ao longo da CDR H3. Exemplo: G e T são as bases que devem aparecer 100% das vezes em suas respectivas posições na seqüência de nucleotídeo. Já para as seqüências de aminoácidos, o códon GMT pode indicar ASP ou ALA, os quais devem aparecer cada um 50% das vezes.

A título de exemplificação segue abaixo o cálculo da divergência de Kullback-Leibler para a primeira posição de nucleotídeos da seqüência codificadora de CDR H3 da biblioteca VH4 ciclo zero. Na figura 7, observa-se que na primeira posição de nucleotídeos da CDR H3 temos N, o qual pode gerar as bases A, C, G ou T, e cada uma deve aparecer 25% das vezes ($N = \{A, C, G, T\} \cong 25\%$), portanto, a distribuição modelo esperada de cada base é 0,250. A distribuição modelo esperada, Q(x), e a distribuição observada de cada nucleotídeo, P(x), podem ser descritos como segue abaixo. Desta forma:

Distribuição modelo esperada:

Q(A): 0,250, Q(T): 0,250, Q(C): 0,250, Q(G): 0,250

Distribuição observada:

P(A): 0,058, P(T): 0,510, P(C): 0,063, P(G): 0,369

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)}$$

$$D_{KL}(P \parallel Q) = P(A) \log_2 \frac{P(A)}{Q(A)} + P(T) \log_2 \frac{P(T)}{Q(T)} + P(C) \log_2 \frac{P(C)}{Q(C)} + P(G) \log_2 \frac{P(G)}{Q(G)}$$

$$D_{KL}(P \parallel Q) = (0,058) \log_2 \frac{0,058}{0,250} + (0,510) \log_2 \frac{0,510}{0,250} + (0,063) \log_2 \frac{0,063}{0,250} + (0,369) \log_2 \frac{0,369}{0,250}$$

$$D_{KL}(P \parallel Q) = -0.123 + 0,525 + (-0,125) + 0,207$$

$$D_{KL}(P \parallel Q) = 0,484 \text{ bits}$$

4.2.2.4 Demais ferramentas utilizadas

O programa FastQC, escrito em linguagem Java, faz um controle de qualidade inicial das milhões de seqüências geradas por seqüenciadores de alto-desempenho. Ele é extremamente útil para ser utilizado num primeiro momento de avaliação, pois gera diversos relatórios que identificam problemas no seqüenciamento ou na geração de bibliotecas. Pode ser executado em interface gráfica ou em linha de comando, e realiza o processamento sistemático de um grande número de arquivos. Ele recebe como entrada arquivos de formato FASTQ e gera como saída relatórios HTML com os resultados dos módulos executados (FastQC-*homepage*, 2011).

O Galaxy é uma plataforma aberta que contém diversos programas para pesquisas na área biológica e médica. É possível se construir um *pipeline* utilizando as muitas ferramentas disponíveis, já que ele é um *framework*, e guarda as entradas e saídas das ferramentas utilizadas na forma de um histórico para cada usuário. Há uma seção extensa para manipulação de seqüências de alto-desempenho, inclusive filtros de qualidade, conversão de formatos de arquivos, entre outros. Esse *framework* pode ser utilizado na versão disponível na internet ou na versão local, no caso de ser utilizada grande quantidade de dados (Galaxy-*homepage*, 2012).

A linguagem de programação Perl é uma das mais utilizadas na comunidade da Bioinformática. É uma linguagem orientada ao objeto, é *open source* e tem mais de 24 anos de existência, tendo sido criada em 1987 por Larry Wall. É extremamente adequada para manipulação de arquivos texto, tais como os formatos FASTA e FASTQ de seqüências biológicas, e também para programas curtos, que podem ser utilizados numa única linha de comando. O Perl pode ser utilizado tanto em sistemas operacionais UNIX quanto em Windows (Perl-*homepage*, 2012).

O programa WebLogo gera representações gráficas de padrões de alinhamentos, e pode ser utilizado na versão disponível na internet ou na versão local (WebLogo-*homepage*, 2012). Com ele é possível obter uma descrição precisa da similaridade entre seqüências. As representações gráficas são feitas na forma de logos que se consistem em pilhas de letras, sendo estas compostas por aminoácidos ou nucleotídeos, dependendo do arquivo de entrada. A interpretação dos logos produzidos por este programa é a mesma, tanto para arquivos de nucleotídeos quanto para arquivos de aminoácidos. O tamanho da pilha existente em cada posição do logo indica a conservação da seqüência em *bits*. Quanto maior a altura da pilha, mais conservada é a seqüência naquela posição, por conseguinte, quanto menor é a pilha, menos conservada é a seqüência nesta posição. Por sua vez, o

tamanho de cada letra dentro da pilha reflete a sua freqüência relativa naquela posição, quanto maior a letra, maior é sua freqüência. A conservação da seqüência numa determinada posição do alinhamento é a diferença entre a entropia de Shannon máxima para aquela posição e a entropia da distribuição observada, como descrito na equação a seguir (Crooks *et al.*, 2004).

$$R_{seq} = S_{max} - S_{obs}$$

$$R_{seq} = \log_2 N - (-\sum_{n=1}^{N} p_n \log_2 p_n)$$

R_{seq}*: conservação da seqüência numa determinada posição*

S_{max}*: entropia máxima*

S_{obs}*: entropia observada*

p_n*: freqüência observada do símbolo n*

N*: número de símbolos distintos para dado tipo de seqüência, quatro para DNA/RNA e 20 para proteínas, conseqüentemente:* S_{max} *para DNA/RNA é* $\log_2 4 = 2$ *bits e* $\log_2 20 = 4.32$ *bits para proteínas*

5 *Resultados*

Neste capítulo serão descritos os resultados que obtivemos nesta pesquisa, por meio das abordagens experimental e computacional. Os tópicos trabalhados serão a preparação das amostras para seqüenciamento de alto-desempenho com a introdução de *barcodes* de identificação, o seqüenciamento com plataforma Illumina e a análise computacional.

5.1 Preparação de amostras com a introdução de *barcodes* de identificação

O primeiro passo realizado para análise das bibliotecas de estudo foi a construção de iniciadores de amplificação que possuíssem marcações, de modo que fosse possível seqüenciar todas as amostras na mesma corrida de seqüenciamento Illumina. Tal método de marcação por *barcoding* já foi descrito por diversos grupos de pesquisa que mostraram que isso reduziu drasticamente os custos do seqüenciamento (Tu *et al.*, 2012; Hamady *et al.*, 2008), o que foi um motivo importante para o grupo no momento de escolha dessa metodologia. Neste trabalho, os códigos indicativos de cada biblioteca, ou *barcodes*, foram introduzidos por meio da modificação de iniciadores da PCR de amplificação das amostras, um método que já foi descrito anteriormente (Smith *et al.*, 2010). Obtivemos com essa técnica regiões amplificadas de 77 pares de base, de forma que os produtos de PCR pudessem ser completamente identificados nas seqüências geradas a partir da plataforma Illumina.

Os *barcodes* desenvolvidos neste trabalho consistem em mutações no iniciador senso de amplificação de cada biblioteca em relação ao segmento gênico do arcabouço correspondente, segmento germinal de VH10 ou de VH4 (figura 8). Dessa forma é possível se distinguir uma biblioteca da outra por meio de dois nucleotídeos, no mínimo. Para se distinguir um ciclo do outro, existem os dois *barcodes*, e para se distinguir um arcabouço do outro, por exemplo, a biblioteca VH4 ciclo zero da biblioteca VH10 ciclo zero,

existem, além dos dois *barcodes*, as diferenças intrínsecas aos segmentos germinais. Já foram relatados na literatura *barcodes* com uma resolução de quatro nucleotídeos para a diferenciação de 48 bibliotecas (Parameswaran *et al.*, 2007), o que demonstra que a resolução de dois nucleotídeos utilizada neste trabalho é suficiente para a identificação das quatro bibliotecas. Essa metodologia de construção dos iniciadores foi usada com sucesso para preparação de amostras para a plataforma Illumina em outros projetos do grupo de Biologia Molecular da Universidade de Brasília.

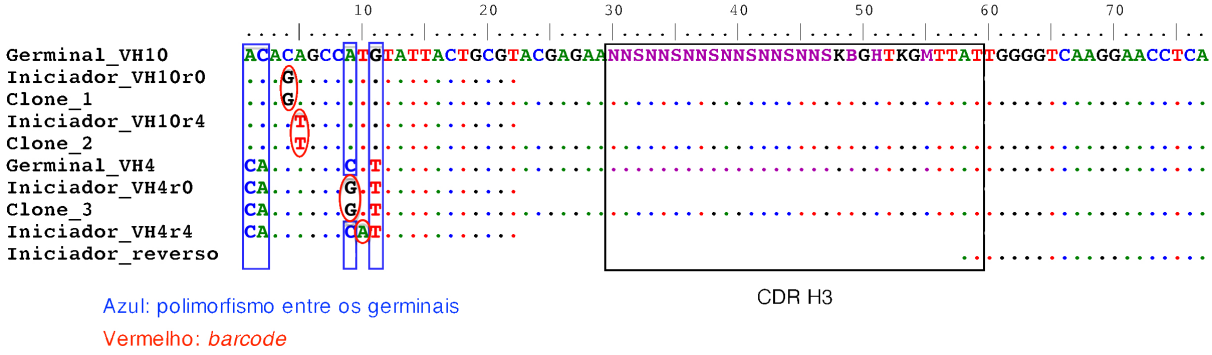


Figura 8: **Alinhamento dos segmentos germinais de VH10 e VH4, iniciadores com *barcodes* e clones positivos do seqüenciamento ABI.** Neste alinhamento, o consenso escolhido foi o VH10 germinal, sendo que nucleotídeos iguais são simbolizados por pontos, e mutações por nucleotídeos. O iniciador reverso é comum a todas as bibliotecas.

Primeiramente, o DNA contendo as bibliotecas de estudo foi obtido, o qual foi então, utilizado como substrato em reações de PCR. Dessas reações foram obtidos produtos de tamanho adequado, considerando que os fragmentos foram planejados para possuir 77 pb (figura 9), o que indicou que a amplificação da porção correspondente à CDR H3 foi bem sucedida. Esta amplificação foi feita ainda com a enzima Platinum®Taq DNA polimerase da Invitrogen, que assegura seis vezes mais fidelidade do produto de polimerização à fita molde do que enzimas Taq DNA polimerases convencionais (Platinum-*homepage*, 2012). Isso foi feito para diminuir a taxa de erros de incorporação de bases, especialmente no caso dos *barcodes*, para que as bibliotecas pudessem ser corretamente identificadas no *pipeline* computacional.

Para certificar que essa estratégia realmente possibilitaria a distinção de uma biblioteca da outra *in silico*, foram feitos procedimentos de clonagem seguidos de seqüenciamento Sanger ABI. Três bibliotecas foram seqüenciadas dessa forma, sendo estas VH4 ciclo zero, VH10 ciclo zero e VH10 ciclo quatro. No alinhamento dos clones representativos de cada uma dessas três bibliotecas com os iniciadores e os segmentos germinais (figura 8), é possível perceber que os *barcodes* correspondentes às famílias supracitadas

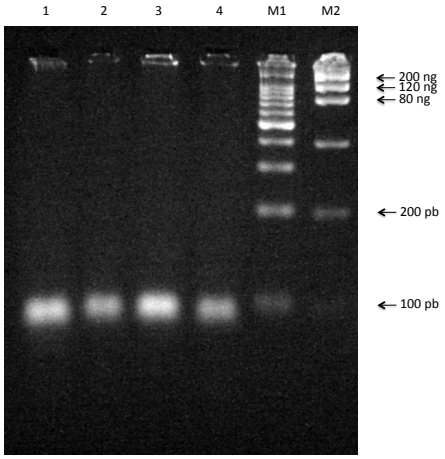


Figura 9: **Gel de agarose 4% da reação de amplificação das bibliotecas de estudo.** Ordem do gel: 1) VH4 ciclo zero, 2) VH10 ciclo zero, 3) VH4 ciclo quatro, 4) VH10 ciclo quatro, M1) marcador 100 bp Promega®, M2) marcador *low mass* DNA *ladder* Promega®.

foram corretamente inseridos, e que as bibliotecas puderam ser facilmente identificadas. Apesar de a biblioteca VH4 ciclo quatro não ter sido analisada nestes experimentos, os resultados das outras bibliotecas indicaram condições apropriadas para o seqüenciamento em plataforma de alto-desempenho.

5.2 Panorama geral das seqüências Illumina

Foram obtidas do seqüenciamento Illumina um total de 193.208.120 seqüências de 150 pares de base em formato FASTQ. Estas seqüências foram recebidas na forma de dez arquivos compactados, sendo metade deles pertencentes a um grupo de seqüenciamento e a outra metade pertencente a outro, cada grupo contendo 96.604.060 seqüências. Esses dois grupos, denominados R1 e R2, foram criados devido à técnica de seqüenciamento utilizada (*paired-end*), especificada no item 4.1.2.10 do capítulo 4, na qual cada seqüência é lida duas vezes em ambos os sentidos senso e reverso (Illumina-*homepage*, 2012).

Como os produtos das PCRs estavam completamente inseridos nas seqüências produzidas, foram obtidos ao final duas cópias da mesma seqüência, cada uma vinda de um dos sentidos do seqüenciamento. Portanto, as seqüências dos dois grupos se complementaram, de forma que a baixa qualidade na seqüência de um dos grupos pôde ser compensada pela seqüência correspondente do segundo grupo. Dessa forma foi possível aumentar ainda mais o espaço amostral das bibliotecas, e conseqüentemente, aumentar a confiabilidade dos resultados.

As seqüências recebidas foram submetidas a uma avaliação geral inicial de qualidade com o programa FastQC. Esta etapa foi sugerida num *pipeline* de análise de RNA-seq, mas também pode ser adequada a qualquer outro *pipeline* de análise de seqüências de plataforma Illumina (Young, 2011). Por meio dessa análise foi possível observar a composição de qualidade geral das seqüências, e também planejar os parâmetros de filtragem mais adequadamente (tais parâmetros se encontram descritos no item 4.2.1.1 do capítulo 4). A título de exemplificação, temos a qualidade média das seqüências do primeiro arquivo do grupo R1 (figura 10). Percebe-se neste caso que a maioria das seqüências tem qualidade Phred acima do padrão estabelecido neste trabalho, que foi de 20, o que corresponde a uma acurácia de leitura de bases de 99%. Isso indica que o seqüenciamento Illumina das amostras desse trabalho foi bem sucedido.

Por sua vez, para o primeiro arquivo do grupo R2 em relação à qualidade por posição na janela de 150 bases (figura 11), percebe-se que até a posição 120, a qualidade está acima do padrão, porém, no intervalo de 120 a 150, a qualidade média decresce consideravelmente. Essa queda não trouxe prejuízo para as análises, já que as seqüências de interesse se situavam antes da posição 100 da janela. De fato, esse resultado auxiliou na escolha de um dos parâmetros da filtragem, o de 30 bases no máximo com qualidade abaixo do valor mínimo estabelecido (item 4.2.1.1 do capítulo 4). Esse parâmetro possibilitou a inclusão das seqüências com queda de qualidade no final e não causou problemas de confiabilidade nos dados, já que as regiões de interesse, CDR H3, se situavam antes da posição 100 na janela, onde não foi observada queda de qualidade.

As seqüências de cada um dos dez arquivos foram filtradas dessa forma, sendo que de um total de 96.604.060 seqüências do grupo R1, 83.303.391 passaram pelo filtro, o que representa 86,23% do total. Já para o grupo de seqüências R2, de um total de 96.604.060 seqüências, 67.891.324 passaram pelo filtro, o que representa 70,28%. Após a filtragem, passou-se à próxima etapa do *pipeline* de Bioinformática, correspondente à classificação.

5.3 Classificação das seqüências entre as bibliotecas

A classificação das seqüências entre as bibliotecas de estudo foi feita na forma da identificação dos códigos indicativos de cada biblioteca, os *barcodes*, por meio de expressões regulares de forma a não permitir nenhuma modificação, ou *mismatch* (como descrito no item 4.2.1.2 do capítulo 4). Por meio dessa abordagem, foi calculado que apenas 6,34% das seqüências filtradas do grupo R1 e 5,95% das do grupo R2 não puderam ser

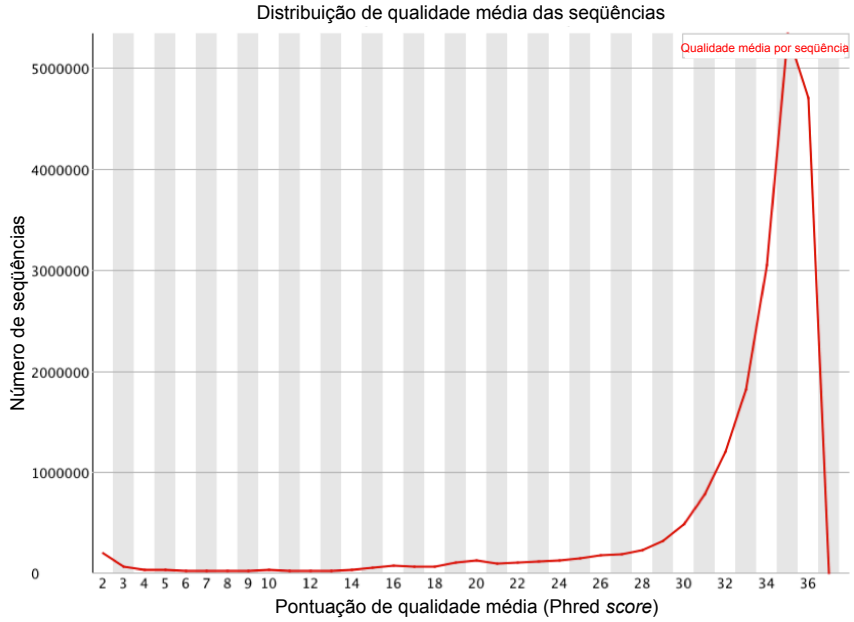


Figura 10: **Qualidade Phred média das seqüências do primeiro arquivo do grupo R1 do seqüenciamento Illumina.** O padrão estabelecido neste trabalho foi o de qualidade Phred acima de 20. Figura adaptada da saída do programa FastQC.

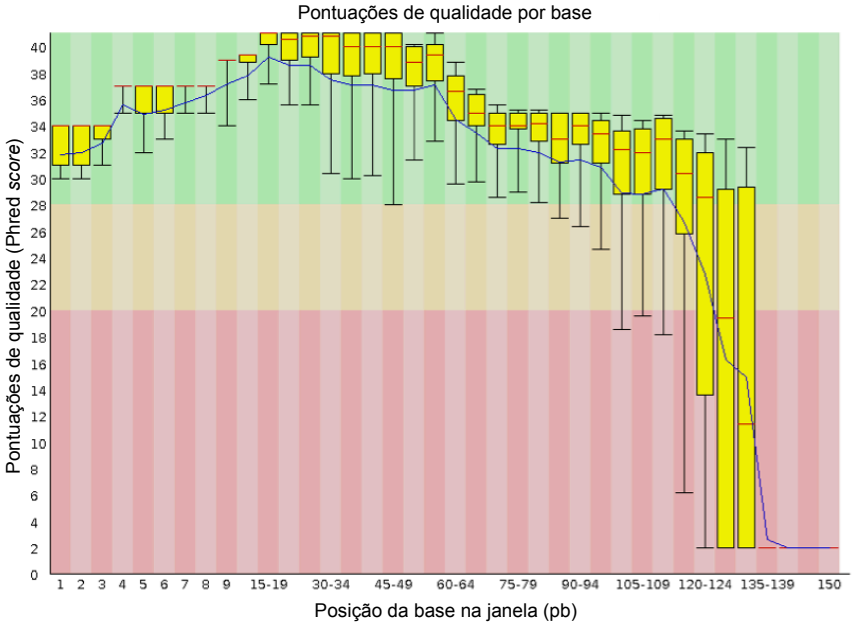


Figura 11: **Qualidade Phred por base ao longo da janela de 150 bases do primeiro arquivo do grupo R2 do seqüenciamento Illumina.** A linha central vermelha é a mediana da qualidade, as caixas amarelas representam a variação de qualidade dentro do quarto de 25 a 75% das seqüências, as linhas pretas, a variação entre 10 e 90% das seqüências e a linha azul representa a média da qualidade. Figura adaptada da saída do programa FastQC.

classificadas entre as quatro bibliotecas de estudo (tabela 5), ou seja, mais de 93% das seqüências mantidas após a filtragem foram classificadas. Foi visto também que a maioria das seqüências foram classificadas como VH10, ciclo zero ou quatro. No entanto, isso não trouxe prejuízo, pois a distribuição dos dados obtidos proporcionou condições adequadas para se analisar como ocorreu a seleção de peptídeos contra DNA, dentre outras análises.

Tabela 5: Contagem de seqüências classificadas em cada biblioteca após o filtro de qualidade e freqüência destas relativa ao total de seqüências mantidas após o filtro, para os grupos R1 e R2 do seqüenciamento Illumina.

Classificação das seqüências entre as bibliotecas de estudo	Número de seqüências classificadas por grupo		Freqüência de seqüências classificadas por grupo	
	R1	R2	R1	R2
VH4 ciclo zero	12.307.946	9.696.207	14,78%	14,28%
VH4 ciclo quatro	7.833.979	6.229.236	9,40%	9,18%
VH10 ciclo zero	26.759.708	22.120.439	32,12%	32,58%
VH10 ciclo quatro	31.117.811	25.804.866	37,35%	38,01%
Não-classificadas	5.283.947	4.040.576	6,34%	5,95%
Total de seqüências mantidas	83.303.391	67.891.324	100,0%	100,00%

Comparando-se as seqüências únicas de CDR H3 dos grupos R1 e R2, foi visto que algumas foram identificadas em somente um grupo. Portanto, para aumentar a quantidade de informação para análise, aquelas só identificadas no grupo R2 foram acrescidas ao R1. O total de seqüências foi então contabilizado, tanto na forma de nucleotídeos quanto na forma de peptídeos (tabela 6). O aumento em relação às seqüências identificadas somente no grupo R1 foi de 10,9% em VH4 ciclo zero, 7,8% em VH4 ciclo quatro, 4,4% em VH10 ciclo zero e 4,9% em VH10 ciclo quatro (tabela 6).

Tabela 6: Número de seqüências únicas de CDR H3 em cada uma das quatro bibliotecas de estudo considerando os grupos R1 e R2 do seqüenciamento Illumina.

Biblioteca	R1 - total de seqüências	R2 - total de seqüências	Únicas de R2[1]	Seqüências (nc)[2]	Seqüências (pt)[3]
VH4 ciclo zero	34.839	10.084	3.798	38.637	22.733
VH4 ciclo quatro	21.157	4.547	1.654	22.811	12.459
VH10 ciclo zero	87.527	13.624	3.858	91.385	35.778
VH10 ciclo quatro	90.233	15.964	4.377	94.610	36.162

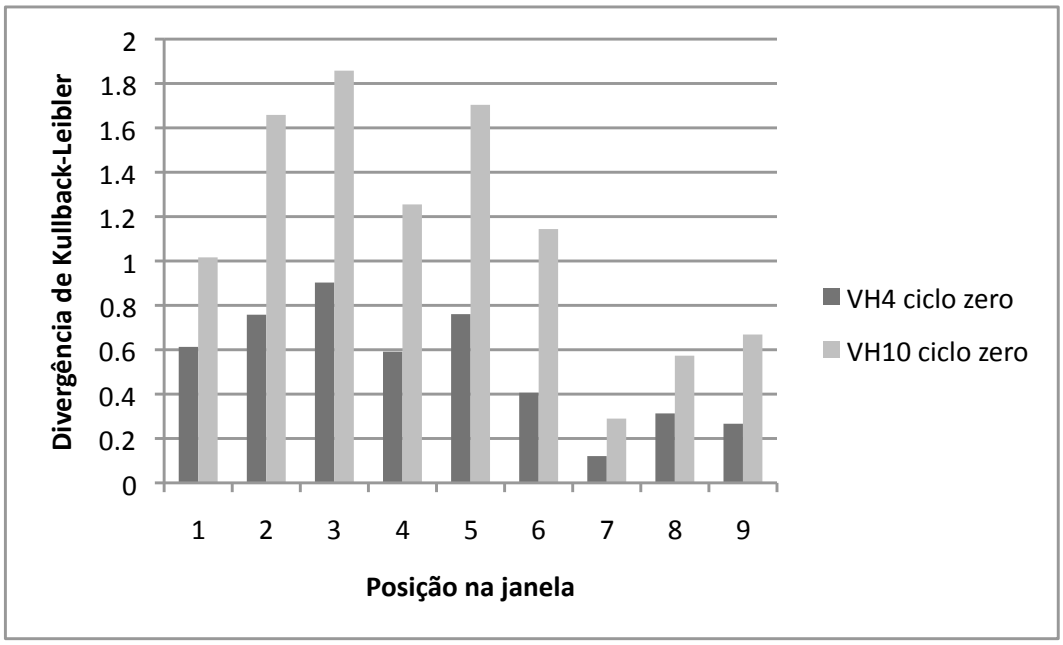
[1]Número de seqüências únicas de R2, que não foram identificadas em R1
[2]Total de seqüências de nucleotídeos de cada biblioteca após a inclusão das seqüências únicas de R2 nas listadas em R1
[3]Total de seqüências de aminoácidos de cada biblioteca

5.4 Análise de composição da região correspondente à CDR H3

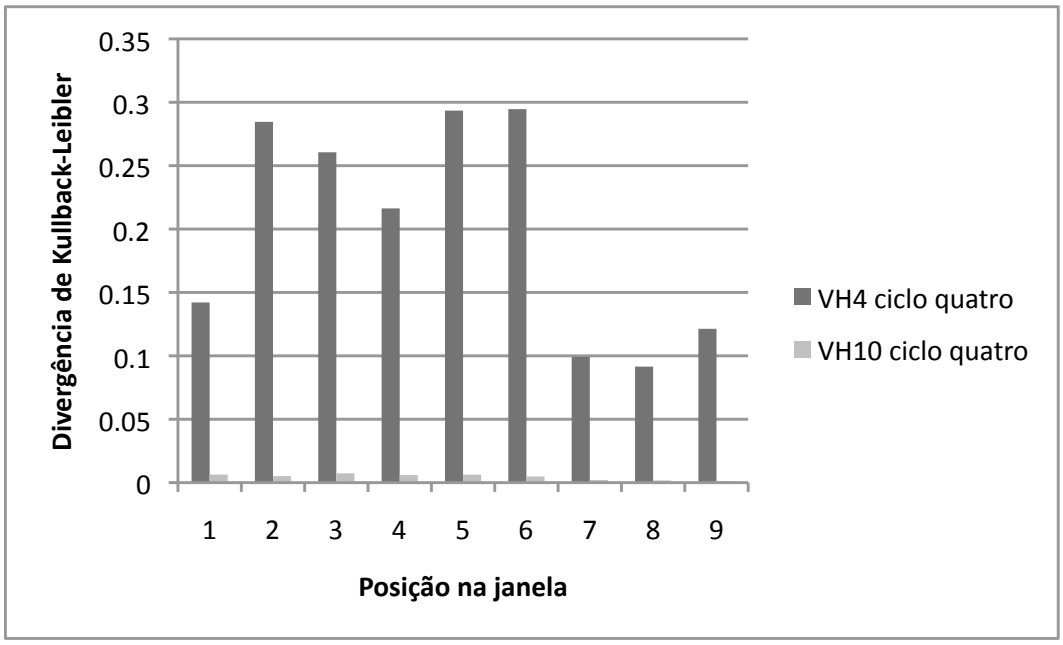
A divergência de Kullback-Leibler pode ser utilizada nas mais diversas análises nas áreas de Biologia Molecular e Genômica, como por exemplo, para a compreensão do processo de transferência horizontal de genes em bactérias (Bohlin *et al.*, 2012), onde as seqüências de estudo foram comparadas com seqüências randômicas. O mesmo princípio foi utilizado neste trabalho ao se comparar as distribuições previstas com as distribuições observadas dos nucleotídeos e aminoácidos que compõem a CDR H3 das quatro bibliotecas de estudo. Por meio dessa abordagem, foi possível obter uma caracterização mais ampla das bibliotecas, de forma a analisar o quão próximas as do ciclo zero chegaram à diversidade almejada e de que forma ocorreu a seleção das bibliotecas do ciclo quatro.

Tomando as probabilidades observadas nas bibliotecas do ciclo zero e comparando-as com as probabilidades esperadas no desenho de construção (figura 7), temos que a biblioteca VH4 se aproximou mais à construção almejada do que VH10 (figura 12a). Levando em consideração que as distribuições de aminoácidos divergem sensivelmente nessas duas bibliotecas iniciais, para se avaliar como ocorreu a seleção das bibliotecas do ciclo quatro, foram comparadas as probabilidades observadas nessa etapa de seleção com as probabilidades esperadas do ciclo zero de cada arcabouço, como especificado no item 4.2.2.3 do capítulo 4. Dessa forma, a comparação é feita entre os ciclos do mesmo arcabouço, e é possível observar como ocorreu a seleção dentro desse contexto. O que pode ser observado com essa abordagem é que VH4 ciclo quatro se afastou consideravelmente do modelo previsto em VH4 ciclo zero, e que a distribuição de probabilidades observada em VH10 ciclo quatro se manteve extremamente semelhante à distribuição esperada no modelo previsto no ciclo zero desse arcabouço (figura 12b).

Outra abordagem utilizada para a caracterização das bibliotecas foi a produção de logos pelo programa WebLogo. Os logos são representações bastante informativas, que possibilitam a visualização de padrões de alinhamentos e do estado de conservação da seqüência em cada posição da janela (Kumar, Ravunny e Chakraborty, 2011), como descrito no item 4.2.2.4 do capítulo 4. Por meio dessa metodologia, foi possível observar a distribuição de bases da região correspondente à CDR H3 das bibliotecas primordiais e selecionadas (figura 13). No geral pode ser observado que a biblioteca VH4 ciclo zero é menos conservada que as outras (figura 13b), e que as distribuições de nucleotídeos das bibliotecas VH4 ciclo quatro e VH10 ciclos zero e quatro são extremamente semelhantes (figuras 13c, 13d e 13e).

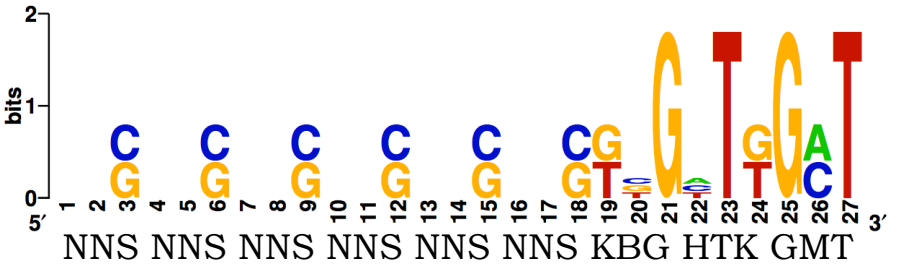


(a)

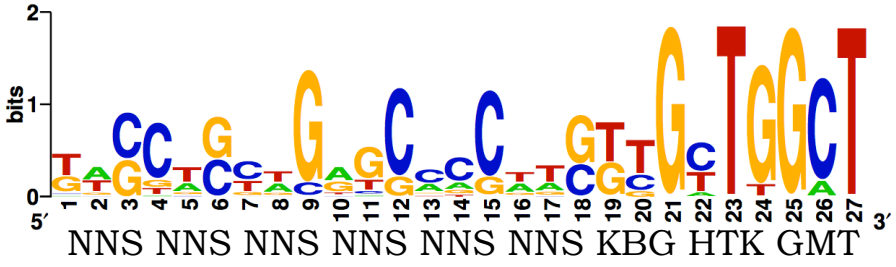


(b)

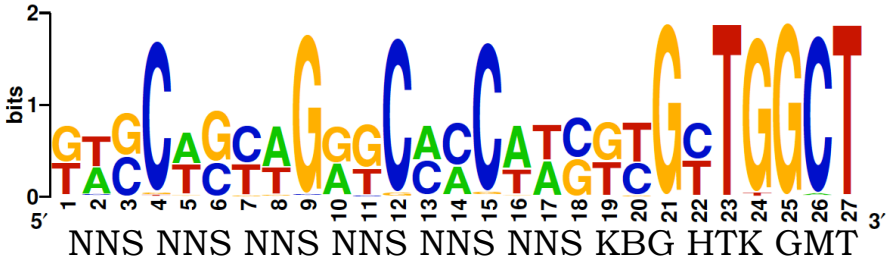
Figura 12: **Divergência de Kullback-Leibler de cada posição na janela de aminoácidos para as quatro bibliotecas de estudo.** (a) Divergência de Kullback-Leibler para os aminoácidos de VH4 ciclo zero e VH10 ciclo zero utilizando como probabilidades previstas as do desenho de construção, que se encontram descritas na figura 7. (b) Divergência de Kullback-Leibler para os aminoácidos de VH4 ciclo quatro e VH10 ciclo quatro utilizando como probabilidades previstas as do ciclo zero de cada arcabouço.



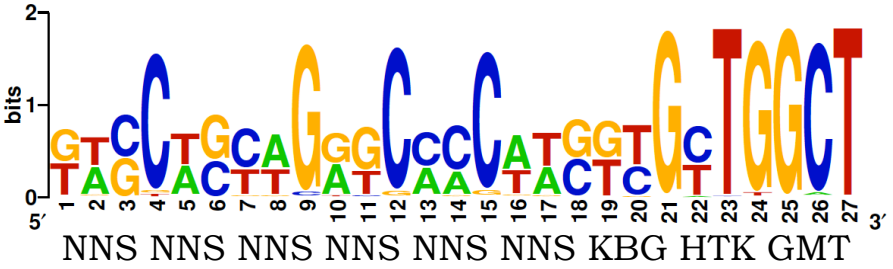
(a)



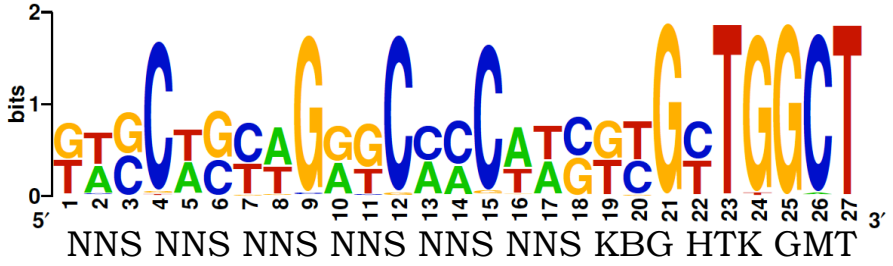
(b)



(c)



(d)



(e)

Figura 13: **Composição de nucleotídeos da região correspondente à CDR H3 das bibliotecas estudadas.** Composição de nucleotídeos (a) do modelo planejado para as bibliotecas do ciclo zero (representado na figura 7), e das famílias (b) VH4 ciclo zero, (c) VH10 ciclo zero, (d) VH4 ciclo quatro e (e) VH10 ciclo quatro. N representa as bases A, C, G ou T, S representa G ou C, K representa G ou T, B representa C, G ou T, H representa A, C ou T e M representa A ou C.

Além dos logos das quatro bibliotecas de estudo, foi criado também um logo, figura 13a, para a distribuição planejada de nucleotídeos dos ciclos zero (figura 7). Essa distribuição planejada pode ser percebida nos logos das bibliotecas correspondentes ao ciclo zero (figuras 13b e 13c), ainda que VH10 seja mais conservada. O modelo foi construído de forma a possuir bases fixas nas posições 21, 23, 25 e 27, o que pôde ser observado em todas as bibliotecas (figuras 13b, 13c, 13d e 13e). Em adição, até a posição 18 da seqüência, foram desenhados no modelo seis códons NNS, sendo N representativo de quatro bases e S de duas. A partir disso, era esperado que, nas bibliotecas do ciclo zero, posições S fossem mais conservadas e compostas de Gs e Cs, enquanto que as outras posições deveriam ser menos conservadas e possuir composição mais distribuída das quatro bases (figura 13a). Isso foi confirmado nas bibliotecas do ciclo zero, apesar de VH10 possuir maior conservação do que o esperado nas posições N da janela.

Outra observação é que na posição quatro de todas as bibliotecas, C prevaleceu em detrimento das outras três bases, especialmente em VH10 ciclo zero, VH10 ciclo quatro e VH4 ciclo quatro (figuras 13c, 13d e 13e). Essa prevalência de C nessa posição pode explicar a predominância dos resíduos Leucina, Glutamina e Histidina na segunda posição da CDR H3 das bibliotecas primordiais, já que estes três aminoácidos são codificados pelos códons CUN e CAN. Essa predominância de resíduos na posição dois será refletida na distribuição da CDR H3 das seqüências selecionadas, especialmente em VH10, o que será discutido no item 5.7 deste capítulo.

5.5 Variabilidade das bibliotecas de estudo

A hipótese de que a família VH10 apresenta tendência intrínseca de ligação a ácidos nucléicos não pôde ser confirmada anteriormente por trabalhos do grupo de Imunologia Molecular da Universidade de Brasília (Guedes, 2009), pois uma amostragem muito reduzida de clones foi obtida, como discutido na seção 2.3 do capítulo 2. Com a nova técnica de seqüenciamento empregada neste trabalho, foi possível se abranger esses resultados anteriores, o que possibilitou análises mais profundas, como a de variabilidade das bibliotecas de estudo. Isso foi feito a partir de duas abordagens, o cálculo de seqüências únicas de CDR H3 em cada biblioteca, e o cálculo da entropia de cada biblioteca.

Na primeira abordagem (figura 14), pode ser visto que o número de seqüências únicas no ciclo zero é superior ao número de seqüências únicas no ciclo quatro, tanto em VH4 quanto em VH10. Isso vai de acordo com o esperado, já que as bibliotecas do ciclo zero so-

freram seleção durante o procedimento de *phage display*. Por outro lado, para se avaliar a real variabilidade de cada uma das construções, foi calculada a entropia total (figura 15). Por meio dessa abordagem, pode ser visto que VH4 ciclo quatro é menos variável que VH4 ciclo zero e que VH10 ciclo quatro é tão variável quanto VH10 ciclo zero, o que vai de acordo com a hipótese inicial do trabalho. Todavia, é importante frisar que VH4 ciclo zero já foi construída sendo mais variável que VH10 ciclo zero, o que vai contra o almejado de que essas construções fossem igualmente variáveis. Finalmente, as variabilidades das bibliotecas VH4 ciclo zero, VH10 ciclo zero e VH10 ciclo quatro são semelhantes.

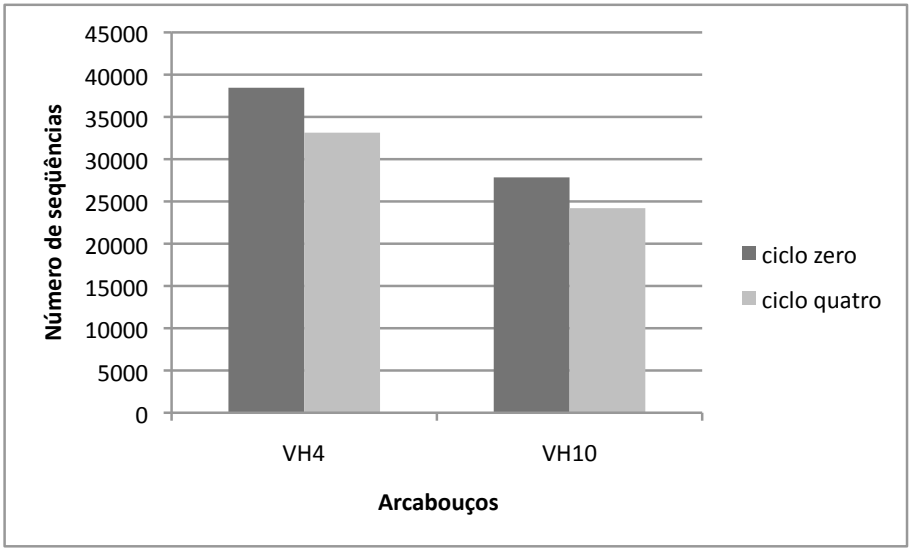


Figura 14: **Número de seqüências peptídicas únicas de CDR H3 em cada biblioteca estudada: VH4 ciclos zero e quatro e VH10 ciclos zero e quatro.**

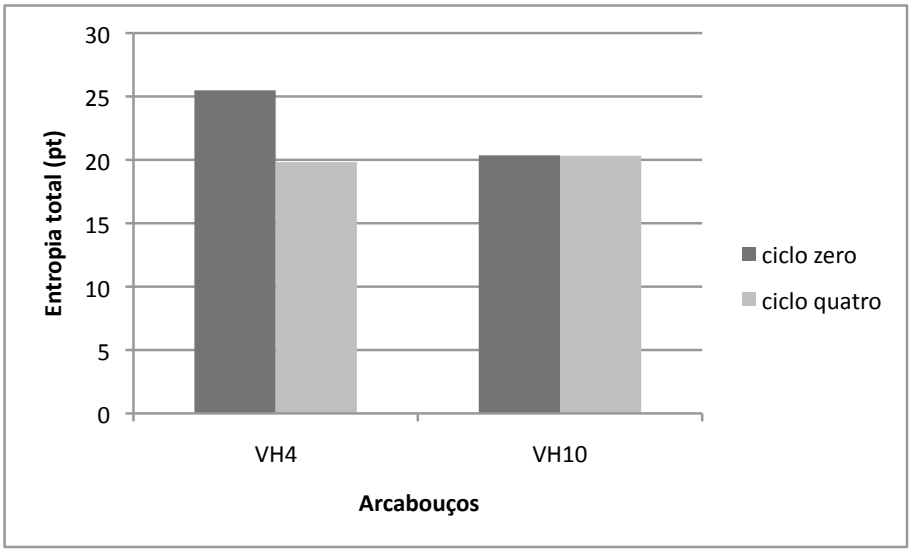
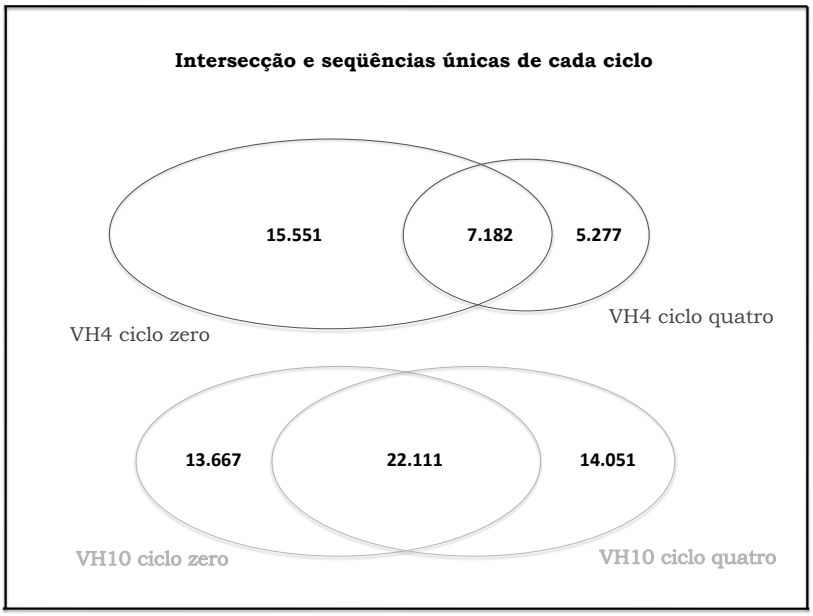


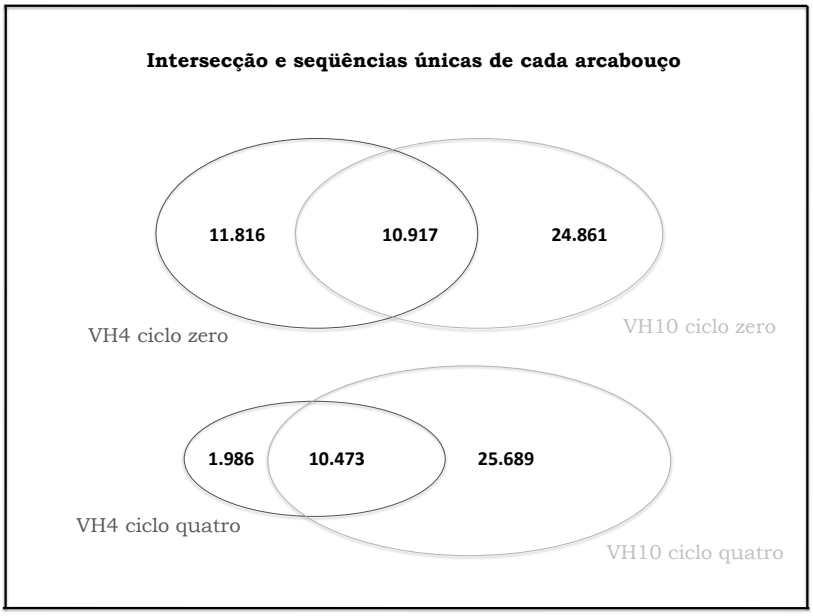
Figura 15: **Entropia total de cada biblioteca de estudo, calculada a partir das seqüências peptídicas de CDR H3.**

Uma outra análise foi feita para avaliar a profundidade alcançada no seqüenciamento de alto-desempenho e consistiu na comparação do número de seqüências únicas e em comum entre as bibliotecas de estudo (figura 16). Foi encontrado a partir desse experimento uma intersecção pequena entre os peptídeos dos ciclos zero e quatro, tanto para VH4 quanto para VH10 (figura 16a). O mesmo pode ser observado em relação à intersecção dos ciclos zero de ambos os arcabouços (figura 16b). Sabe-se que a síntese das bibliotecas do ciclo zero foi feita da mesma forma, o que significa que os universos de seqüências desses grupos são basicamente iguais. Além disso, as bibliotecas de peptídeos do ciclo quatro tiveram origem nas do ciclo zero, após seleção contra DNA. Levando tudo isso em consideração, pode-se concluir, que o seqüenciamento, apesar de todo o seu potencial, não pôde abranger completamente a diversidade de seqüências existente nas amostras de estudo.

Quando se compara os dois arcabouços no ciclo quatro (figura 16b), percebe-se que o número de seqüências únicas de VH10 é muito superior ao de VH4 e também que existe um considerável número de seqüências em comum. Essa perspectiva foi prevista na abordagem experimental inicial desta pesquisa (figura 4), quando se presumiu que haveria uma quantidade maior de CDRs H3 compatíveis com VH10 (Guedes, 2009), tópico que foi abordado no itens 2.1 e 2.3 do capítulo 2.



(a)



(b)

Figura 16: **Comparação de seqüências únicas e em comum entre as quatro bibliotecas de estudo.** (a) Seqüências únicas e em comum entre os ciclos de seleção de VH4 e VH10. (b) Seqüências únicos e em comum entre os arcabouços nos ciclos zero e quatro.

5.6 Seleção de peptídeos contra DNA

Os próximos experimentos feitos neste trabalho tiveram como objetivo a análise da seleção de seqüências. Ravn e colaboradores utilizaram em 2010 o critério de freqüência na biblioteca final para estabelecer quais seqüências foram mais selecionadas. Porém esse parâmetro não seria adequado nesta pesquisa, pois existem duas seqüências que já predominavam as amostras ainda no ciclo zero, tanto em VH4 quanto em VH10. Levando isso em consideração, foi utilizado o critério de enriquecimento para estabelecer o grau de seleção de uma determinada seqüência. O enriquecimento foi definido como o número de vezes que uma seqüência era vista no ciclo quatro dividido pelo número de vezes que a mesma seqüência era vista no ciclo zero, como explicitado no item 4.2.2.2 do capítulo 4.

Algumas seqüências só apareciam em um dos ciclos, e para não excluí-las da análise, decidiu-se considerar o valor de não-aparecimento como sendo igual a um. Isso foi feito partindo-se da premissa de que todos as seqüências do ciclo quatro foram originadas do ciclo zero e de que todos as seqüências do ciclo zero sofreram seleção, sendo esta negativa, positiva ou neutra. Considerou-se que a seqüência que não apareceu num dos ciclos não foi revelada por falta de resolução, mesmo com o potencial da técnica de seqüenciamento de alto-desempenho empregada. Por meio dessa abordagem, foi possível aumentar com sucesso a amostragem de seqüências com enriquecimento positivo e negativo (tabela 7), o que foi muito importante para a avaliação mais completa do conjunto de CDRs. Analisando esses dados juntamente com o perfil dos peptídeos que tiveram enriquecimento não-neutro de VH4 e VH10 (tabela 7 e figura 17), observamos que VH4 possui um grupo maior de seqüências selecionadas negativamente e um grupo menor de seqüências selecionadas positivamente, quando comparado com os peptídeos da família VH10.

Tabela 7: Contagem de seqüências com enriquecimento positivo, superior a 2,0, e negativo, inferior a 0,5, para VH4 e VH10. No experimento estringente, só foram consideradas as seqüências que apareceram nos dois ciclos. Já no inclusivo, as seqüências com aparecimento nulo em um dos ciclos foram considerados como possuindo valor igual a um.

Biblioteca	VH4 estringente		VH4 inclusivo		VH10 estringente		VH10 inclusivo	
Enriquecimento (e)	$e < 0,5$	$e > 2,0$	$e < 0,5$	$e > 2,0$	$e < 0,5$	$e > 2,0$	$e < 0,5$	$e > 2,0$
Número de seqüências	2.862	745	3.835	1.135	1.582	2.451	2.075	3.090

Além disso, é possível analisar mais objetivamente a variabilidade dos grupos com

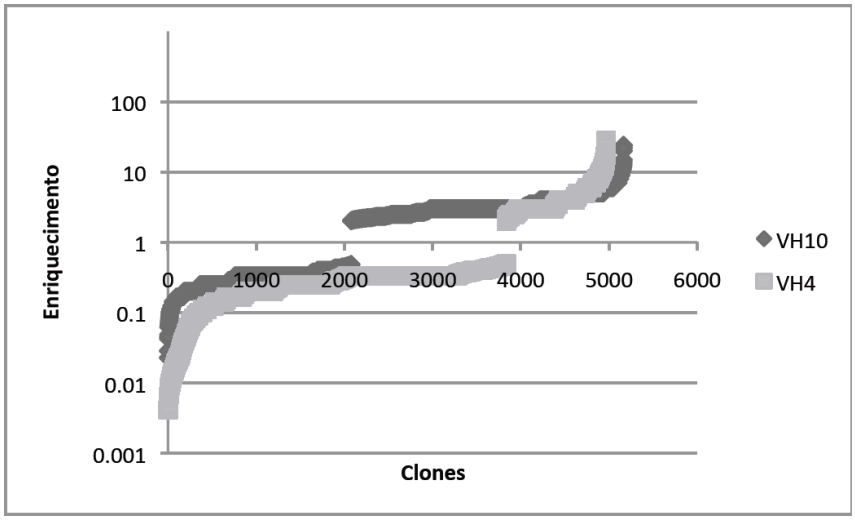


Figura 17: **Enriquecimento de seqüências após seleção contra DNA.** Foi feita uma sobreposição de todas as seqüências de VH4 e de VH10 que têm enriquecimento positivo, acima de 2,0, e negativo, abaixo de 0,5, tendo sido excluídas as seqüências com enriquecimento neutro.

enriquecimento não-neutro de VH4 e VH10 por meio da observação dos valores de entropia (figura 18). Percebe-se a partir disso que o grupo com enriquecimento negativo de VH4 é tão variável quanto a biblioteca VH4 ciclo zero, e que o grupo com enriquecimento positivo é menos variável que ambos. Por sua vez, o grupo negativo de VH10 também é tão variável quanto a biblioteca VH10 ciclo zero, porém, o grupo positivo de VH10 é quase tão variável quanto ambos.

Outro ponto de análise foi relacionado às seqüências com maior aparecimento nas bibliotecas, as seqüências S(1) e S(2), correspondentes respectivamente aos clones VQQVNNALA e YLLSPLLLA. A seqüência S(2) foi aquela que teve a maior freqüência dentre as analisadas por Guedes em 2009. Ela estava presente em todos os ciclos, tanto no contexto de VH4 quanto no de VH10. Esse clone foi visto igualmente em grande freqüência em todas as bibliotecas também neste trabalho, assim como o S(1) (figura 19), o qual também foi observado por Guedes em 2009, porém não com tanta freqüência quanto S(2).

Ambas essas seqüências dominantes formaram scFvs que se mostraram ligantes a DNA em experimentos de *phage ELISA* feitos por Guedes em 2009, tanto em contexto da família VH10 quanto em contexto da família VH4. Entretanto, apesar de aparecer em grande freqüência e também de ter se mostrado ligante a DNA pelos cálculos de enriquecimento, foi observado que S(2) teve seleção neutra nos contextos das duas famílias (tabela 8). Por outro lado, o clone S(1) em VH4 sofreu seleção positiva.

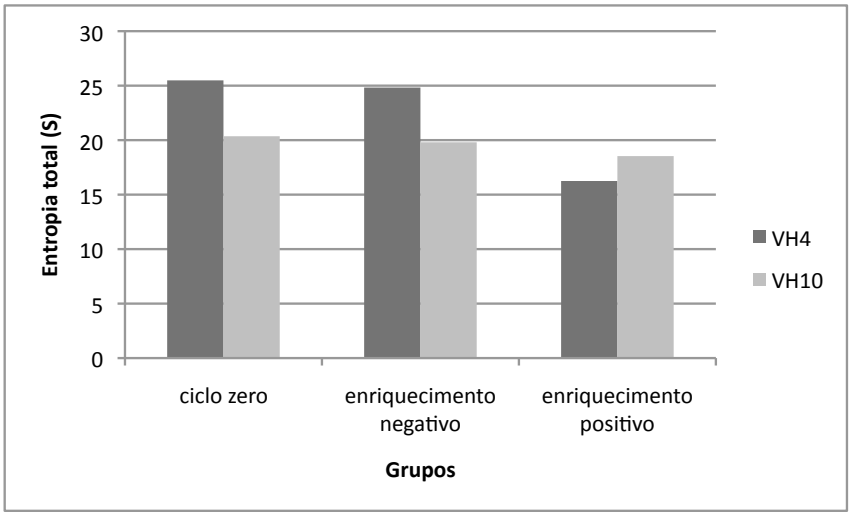


Figura 18: **Entropia total das bibliotecas do ciclo zero de VH4 e VH10 e dos grupos de seqüências com enriquecimento positivo e negativo de cada arcabouço.**

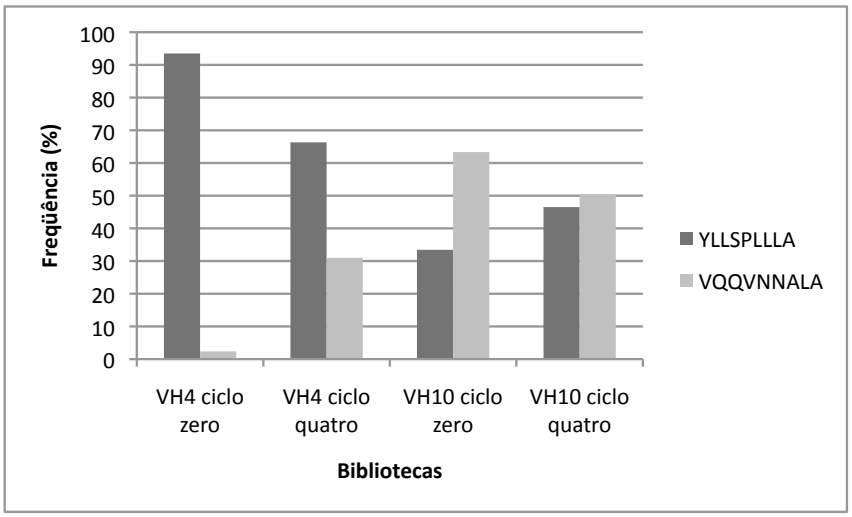


Figura 19: **Distribuição de freqüências de aparecimento dos dois clones dominantes nas bibliotecas de estudo: S(1), correspondente a VQQVNNALA, e S(2), correspondente a YLLSPLLLA.**

Tabela 8: Enriquecimento dos dois clones mais aparentes nas bibliotecas estudadas.

Clone/Biblioteca	VH4	VH10
S(1) VQQVNNALA	8,29	0,93
S(2) YLLSPLLLA	0,45	1,61

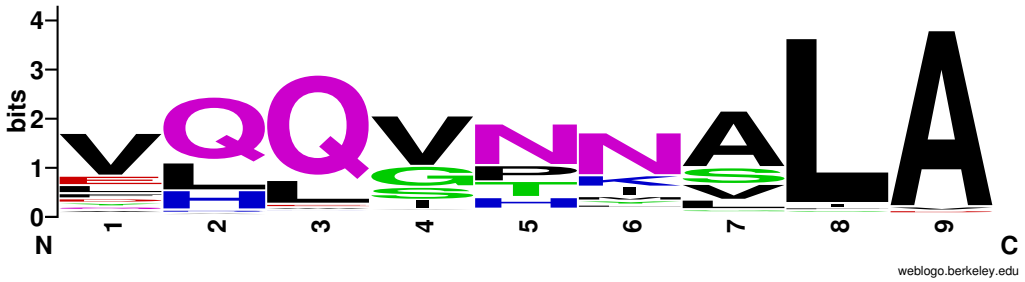
5.7 Análise de padrões em anticorpos ligantes a ácidos nucléicos

Os resultados obtidos a partir dos cálculos de divergência de Kullback-Leibler (figura 12b), juntamento com aqueles obtidos a partir dos estudos de enriquecimento (figuras 17 e 18), mostraram que o segmento germinal VH4 apresentou um universo mais restrito de CDR H3 no grupo selecionado contra DNA. Portanto, o grupo de peptídeos com enriquecimento positivo em VH4 pode trazer alguma informação acerca de seqüências de CDR H3 que apresentem características de ligação a DNA. Os logos também foram utilizados nessa análise, na qual foram investigados os padrões das seqüências que tiveram seleções positiva, negativa e neutra em ambos os contextos dos segmentos germinais.

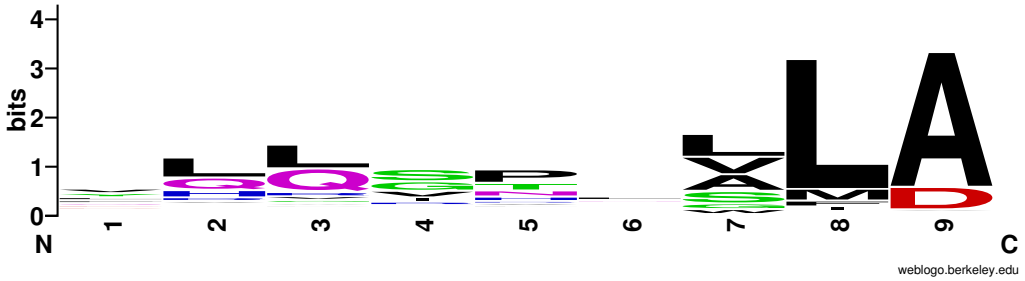
Alguns padrões se destacam ao se analisar a composição do grupo com seleção positiva em VH4, sendo o mais proeminente o padrão VQQVNNALA (figura 20a), cuja seqüência coincide com a de um dos dois clones preponderantes vistos ao longo dessa pesquisa, o clone S(1), o que foi previamente abordado no item 5.6 deste capítulo (tabela 8). Em adição, esse resultado corrobora observações anteriores do grupo de Imunologia Molecular, nas quais o scFv formado por essa CDR H3 em contexto de VH4 se mostrou ligante a DNA (Guedes, 2009).

Como o DNA possui carga negativa, também era esperado se encontrar resíduos positivos no padrão de seleção positiva desse segmento germinal. Porém, foi visto somente sinais pouco freqüentes dos resíduos positivos Histidina, H, nas posições dois e cinco, e Lisina, K, na posição seis, não tendo sido observados sinais relevantes de Arginina, R, um resíduo presente em anticorpos anti-DNA (Herron *et al.*, 1991). Uma outra observação em relação aos padrões de VH4, é que há diferença substancial entre os padrões de seleção positiva e os padrões de seleção negativa e neutra (figura 20), o que mostra que o universo de CDR H3 em contexto de VH4 após seleção contra DNA se torna mais restrito.

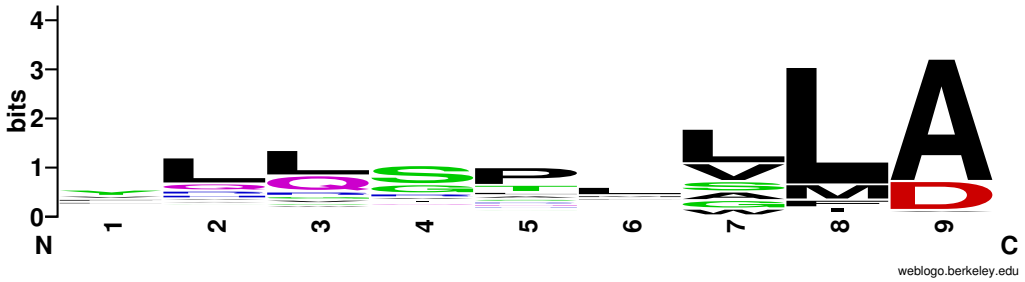
Por sua vez, não podem ser observadas em VH10 diferenças significativas entre os padrões de seleção positiva, negativa e neutra (figura 21). A diferença entre os grupos não é tão pronunciada, o que mostra que VH10 é um segmento germinal menos estringente que VH4 em relação à dependência da CDR H3 para formar scFvs que apresentem comportamento de ligação a moléculas de DNA.



(a)

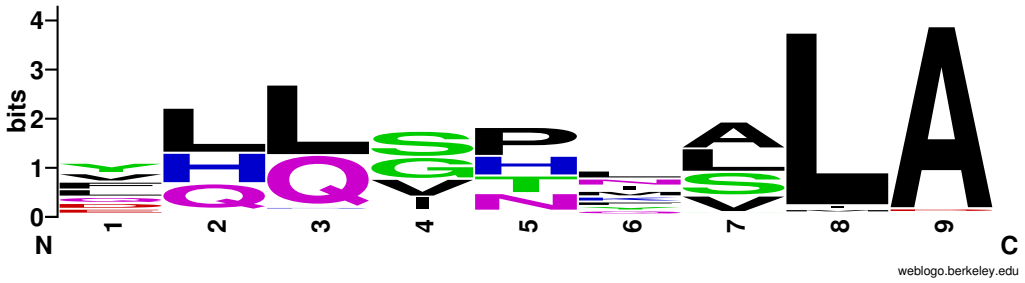


(b)

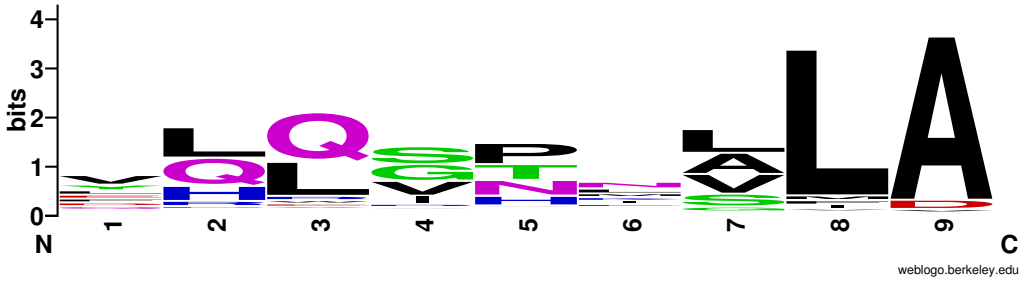


(c)

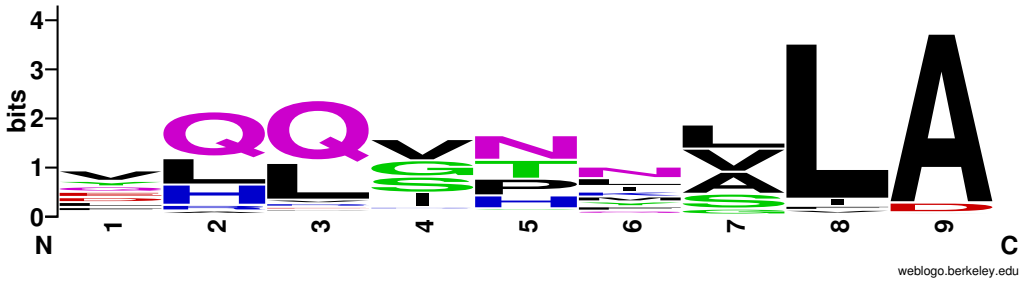
Figura 20: **Distribuição de aminoácidos que compõem a CDR H3 de grupos de seleção contra DNA para VH4.** (a) Seqüências de VH4 selecionadas positivamente. (b) Padrão de seqüências de VH4 não selecionadas. (c) Seqüências de VH4 selecionadas negativamente.



(a)



(b)



(c)

Figura 21: **Distribuição de aminoácidos que compõem a CDR H3 de grupos de seleção contra DNA para VH10.** (a) Seqüências de VH10 selecionadas positivamente. (b) Padrão de seqüências de VH10 não selecionadas. (c) Seqüências de VH10 selecionadas negativamente.

6 Discussão

A variabilidade das bibliotecas originais de VH4 e VH10, correspondentes ao ciclo zero, foram construídas na ordem de 10^7 cada uma (Guedes, 2009). A resolução alcançada com o seqüenciamento Illumina neste trabalho foi também na ordem de 10^7, porém, apesar de a tecnologia de alto-desempenho ser extremamente abrangente, ainda assim não foi possível abranger a totalidade da variabilidade das amostras. Isso pode ser explicado pois, além da grande diversidade inerente às amostras, as quatro bibliotecas foram seqüenciadas juntas e foram encontradas ainda muitas repetições entre as seqüências. Todavia, já foi demonstrado na literatura que o seqüenciamento de nova-geração é uma técnica adequada para esse tipo de estudo, pois gera uma cobertura muito maior de bibliotecas de *phage display* do que métodos experimentais tradicionais. Neste estudo, viu-se que quatro das dez seqüências mais freqüentes identificadas na plataforma Illumina e que reconheciam o antígeno estudado não foram identificados pelo método de seqüenciamento tradicional Sanger (Ravn *et al.*, 2010).

Esse aumento da cobertura também foi observado no presente trabalho. Com a utilização da plataforma Illumina, foram identificadas dezenas de milhares de seqüências para cada biblioteca de estudo (tabela 6). Pode ser dito que a análise de cobertura das bibliotecas foi de fato elevada a um novo patamar, considerando que anteriormente tinham sido obtidos, por meio de clonagem e seqüenciamento Sanger tradicional, somente nove clones de VH10 ciclo quatro e três de VH4 ciclo quatro (tabela 2, Guedes, 2009).

Foi observado que as bibliotecas do ciclo zero, representadas nas figuras 13b e 13c, não são tão diversas quanto o pretendido no desenho inicial, representado na figura 13a. Isso pode ter ocorrido durante a síntese de iniciadores que deu origem às bibliotecas do ciclo zero, procedimento feito em um trabalho anterior do grupo de Imunologia Molecular, ou então durante o processo de incorporação das seqüências nas partículas virais, por meio da infecção em bactérias, também feito anteriormente (Guedes, 2009). Ao longo do processo de expressão do scFv na superfície dos fagos, alguns paratopos se tornam menos estáveis que outros, então, determinados padrões de seqüências podem ter sido preteridos,

se perdendo ao longo de todas essas etapas experimentais.

Na figura 15, foi visto que a biblioteca VH4 ciclo zero é mais variável que as outras três bibliotecas de estudo, VH4 ciclo quatro e VH10 ciclos zero e quatro, e que essas três últimas são igualmente variáveis. Isso mostra que VH4 já foi criada sendo mais variável que VH10 ciclo zero, o que vai contra o almejado na construção dessas bibliotecas. Portanto, essa abordagem não pode ser utilizada para tirar conclusões acerca do processo de seleção de peptídeos contra DNA nos contextos das famílias VH4 e VH10.

Já analisando as seqüências dominantes do trabalho, S(1) VQQVNNALA e S(2) YLLSPLLLA, pode-se dizer que S(2) é caracterizada por muitos resíduos hidrofóbicos enquanto que S(1) é caracterizada por resíduos mais polares, porém ainda assim forma um peptídeo não carregado. Era esperado que estes peptídeos tivessem carga mais positiva, já que os scFvs formados por eles se mostraram ligantes a DNA, uma molécula de carga negativa, em trabalho anterior do grupo de Imunologia (Guedes, 2009). É sugerido portanto que essas seqüências podem favorecer de alguma maneira a montagem de partículas virais no *phage display*, tornando-as mais estáveis. Tal estabilidade pode ter provocado ainda a grande freqüência desses peptídeos observada nas amostras de estudo (figura 19).

Ainda analisando a seqüência S(1), esta correspondeu ao padrão mais proeminente observado nas seqüências com seleção positiva em VH4: VQQVNNALA (figura 20a). Isso sugere que essa CDR H3 pode possuir propriedades de reconhecimento a DNA. No entanto, é importante frisar que toda a abordagem experimental produzida anteriormente por Guedes em 2009 foi feita com o scFv fusionado ao fago via gene VIII, onde diversos outros parâmetros também atuam, além da afinidade do scFv ao antígeno. O gene VIII codifica a proteína de capsídeo do fago, e existem milhares de cópias das proteínas de fusão desse gene com a seqüência de interesse. Portanto, há mais chance do fago contendo as proteínas de fusão se ligar ao antígeno. Levando isso em consideração, uma investigação mais profunda seria necessária, para se avaliar a real afinidade a DNA do scFv formado pela CDR H3 VQQVNNALA.

Trabalhos anteriores trazem a caracterização de peptídeos com comportamento de ligação a ácidos nucléicos, como por exemplo, o 3D8, um scFv que reconhece e hidroliza DNA. Sua estrutura foi determinada e comparada com os modelos BV04-01 e DNA-1, os quais reconhecem DNA fita simples. Regiões de folhas-β se sobrepuseram entre os peptídeos 3D8 e BV04-01, porém foram encontradas diferenças significativas entre as CDRs. Foi visto também na estrutura do 3D8 que resíduos de Tirosina, Y, da CDR H3

eram críticos para a interação com a molécula de DNA, e que esta era feita na forma da alternância entre esses resíduos e as bases timinas do antígeno (Kim *et al.*, 2006). Outros trabalhos também mostraram que esses resíduos entram diretamente em contato com o ligante (Jang e Stollar, 2003). Apesar de não haver resíduos de Tirosina aparentes nos padrões selecionados positivamente em VH4 neste trabalho, a seqüência S(2) YLLSPLLLA, observada em grande freqüência tanto em VH4 quanto em VH10 (figura 8), possui um resíduo de Tirosina na primeira posição, o qual pode auxiliar na interação com moléculas de DNA.

Ao se estudar anticorpos com afinidade a ácidos nucléicos, é importante se analisar a estrutura do modelo BV04-01, um Fab proveniente de um auto-anticorpo que reconhece DNA fita simples. Cristais desse peptídeo foram obtidos na ausência e presença de um trinucleotídeo de timinas, de forma a compreender a interação desse complexo. Além de ajustes sutis nas orientações dos domínios VL e VH, foram observadas modificações conformacionais na terceira alça hipervariável da cadeia pesada, o que aumentou a interação do Fab com o antígeno. O trinucleotídeo estava numa conformação extendida, de forma que bases, açúcares e fofatos estivessem disponíveis para ligação. Resíduos de Serina, S, e Asparagina, N, estabilizaram a interação por meio de ligações de hidrogênio com uma molécula de fosfato, assim como uma Arginina, R, por meio da ligação iônica com outro grupo fosfato. Os contatos dominantes para a ligação foram a interposição entre a Timina central e os anéis de um Triptofano, W, e uma Tirosina, Y (Herron *et al.*, 1991).

Em relação ao estudo de padrões, explicitado no item 5.7 do capítulo 5, foi observada baixa freqüência de resíduos positivos no padrão de seqüências selecionadas positivamente em VH4 (figura 20a), o que vai contra o esperado, já que a seleção foi feita contra uma molécula de carga negativa, DNA. Isso pode ser explicado por dados da literatura, nos quais foi sugerido que a CDR H3 interage com as bases nitrogenadas das seqüências de ácidos nucléicos, e não com a cadeia negativa de fosfato (Jang e Stollar, 2003). Além disso, já foi observado também que diversos aminoácidos contribuem para a ligação a DNA, sem haver a dominância de um resíduo específico (Jang *et al.*, 1998).

Já na análise de enriquecimento, observamos que, quando comparado com VH4, VH10 possui um grupo menor de seqüências selecionadas negativamente e um grupo maior de seqüências selecionadas positivamente (figura 17). Além disso, o grupo positivo em VH4 é menos variável que o grupo negativo, sendo este último tão variável quanto a biblioteca VH4 ciclo zero (figura 18). Já em relação à VH10, temos que o grupo positivo apresenta variabilidade semelhante ao grupo negativo e à biblioteca do ciclo zero de VH10

(figura 18). Em adição, existem os resultados obtidos a partir do cálculo de divergência de Kullback-Leibler aplicado aos ciclos quatro de ambas as famílias comparados com os respectivos ciclos zero (figura 12b). Esses resultados mostraram que, em comparação com a família VH10, o ciclo quatro de VH4 se afastou mais do modelo do ciclo zero. Todas essas informações obtidas por meio dessas abordagens, de enriquecimento e divergência de Kullback-Leibler, indicam que o processo de seleção contra DNA aplicado às bibliotecas do ciclo zero produziu universos diferentes de CDR H3 no contexto das famílias VH4 e VH10. O segmento germinal VH10 se mostrou dependente de um universo menos restrito de CDR H3 para reconhecer moléculas de DNA, o que vai de acordo com a hipótese inicial da pesquisa, na qual foi sugerido que VH10 apresenta reconhecimento a ácidos nucléicos *per se*.

Ao contrário da família VH4, não foram vistas em VH10 diferenças significativas nos padrões de CDR H3 das seqüências com seleção positiva, negativa e neutra (figura 21). Por outro lado, as diferenças entre os grupos de VH4 são mais pronunciadas (figura 20). Isso vai de acordo com o esperado levando-se em conta os resultados dos experimentos de enriquecimento e divergência de Kullback-Leibler (figuras 17, 18 e 12b), que mostraram que VH10 depende de um universo menos restrito de CDR H3 para reconhecer moléculas de DNA. Como este segmento germinal tem se mostrado menos seletivo, então, esses padrões mostrados na figura 21 podem ter sido causados puramente por uma flutuação natural do sistema, e não pela seleção restritiva dos scFvs contra DNA, como ocorreu com VH4.

Os primeiros anticorpos com comportamento de ligação a ácidos nucléicos experimentalmente induzidos pertenciam à família murina VH10, o que já indicava que ela poderia ter alguma implicação no reconhecimento de anticorpos a ácidos nucléicos (Brígido e Stollar, 1991). Todos os estudos realizados neste trabalho, especialmente os de enriquecimento, divergência de Kullback-Leibler e análise de padrões, demonstram que essa família de fato tem relação com o reconhecimento dos scFvs a DNA. Propõe-se portanto que a região correspondente ao segmento germinal VH10 confere aos anticorpos comportamento de reconhecimento de ácidos nucléicos, o que não ocorre com a região correspondente ao segmento germinal VH4.

7 Conclusões e Perspectivas

Neste capítulo final, serão explicitadas as conclusões tiradas a partir dos resultados obtidos neste trabalho, os quais foram mostrados e discutidos respectivamente nos capítulos 5 e 6. Serão listadas também as perspectivas de continuidade desse trabalho.

7.1 Conclusões

Foi observado que a biblioteca VH4 ciclo zero é mais variável que as outras três bibliotecas de estudo, VH4 ciclo quatro, VH10 ciclo zero e VH10 ciclo quatro, e que essas últimas são igualmente variáveis. Por sua vez, foram vistas mais seqüências com enriquecimento negativo e menos seqüências com enriquecimento positivo no contexto da família VH4, o que se mostrou o contrário quando no contexto de VH10. Por sua vez, os experimentos de Kullback-Leibler mostraram que, em comparação com a família VH10, o ciclo quatro de VH4 se afastou mais do modelo do ciclo zero. Já em relação aos clones S(1) VQQVNNALA e S(2) YLLSPLLLA, com S(1) correspondendo ao padrão selecionado positivamente em VH4, pode-se dizer que são padrões de CDR H3 que podem apresentar comportamento de reconhecimento a DNA. Em adição, os padrões de seqüências selecionadas positivamente e negativamente são muito diferentes em VH4 e muito semelhantes em VH10. Pode-se dizer ainda que o segmento germinal VH4 é muito mais estringente do que VH10, e exige um universo mais restrito de seqüências de CDR H3 para formar scFvs que se liguem a DNA. Por fim, conclui-se que a região peptídica correspondente ao segmento germinal VH10 apresenta tendência intrínseca a reconhecer DNA, o que não ocorre no caso do segmento germinal VH4.

7.2 Perspectivas

Como continuidade deste trabalho é proposto:

• Busca de determinantes específicos no segmento germinal VH10 que se liguem a moléculas de ácidos nucléicos ou que confiram propriedades de ligação ao anticorpo.

• Realização de um estudo estrutural do peptídeo de CDR H3 VQQVNNALA no contexto das famílias murinas VH4 e VH10, e estudo da interação tanto da CDR H3 quanto do arcabouço com a molécula de DNA.

• Disponibilização dos *scripts* desenvolvidos e publicação das seqüências obtidas neste trabalho num domínio da internet.

Referências

BOHLIN, J. *et al.* Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. **BMC genomics**, BioMed Central Ltd, v. 13, n. 1, p. 66, 2012.

BRÍGIDO, M.; STOLLAR, B. Two induced anti-z-dna monoclonal antibodies use vh gene segments related to those of anti-dna autoantibodies. **The Journal of immunology**, Am Assoc Immnol, v. 146, n. 6, p. 2005, 1991.

COCK, P. *et al.* The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. **Nucleic acids research**, Oxford Univ Press, v. 38, n. 6, p. 1767–1771, 2010.

COHEN, I. Real and artificial immune systems: computing the state of the body. **Nature Reviews Immunology**, Nature Publishing Group, v. 7, n. 7, p. 569–574, 2007.

CROOKS, G. *et al.* Weblogo: a sequence logo generator. **Genome research**, Cold Spring Harbor Lab, v. 14, n. 6, p. 1188–1190, 2004.

EWING, B. *et al.* Base-calling of automated sequencer traces usingphred. i. accuracy assessment. **Genome research**, Cold Spring Harbor Lab, v. 8, n. 3, p. 175–185, 1998.

FASTQC-*HOMEPAGE*. 2011. Sítio Web: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, acessado em 20/11/2011.

FILICHKIN, S. *et al.* Genome-wide mapping of alternative splicing in Arabidopsis thaliana. **Genome research**, Cold Spring Harbor Lab, v. 20, n. 1, p. 45, 2010. ISSN 1088-9051.

GALAXY-*HOMEPAGE*. 2012. Sítio Web: `http://galaxy.psu.edu/`, acessado em 3/4/2012.

GASTEIGER, E. *et al.* Protein identification and analysis tools on the expasy server. **The proteomics protocols handbook**, Springer, p. 571–607, 2005.

GUEDES, L. **Contribuição do Segmento VH e da CDRH3 no Reconhecimento Antigênico de Ácidos Nucleicos**. Disserta (Mestrado) — Universidade de Brasília, 2009.

HALL, T. **BioEdit Sequence Alignment Editor for Windows 95/98/NT/XP**. [S.l.]: Carlsbad: Ibis Biosciences, 2007.

HAMADY, M. *et al.* Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. **Nature methods**, Nature Publishing Group, v. 5, n. 3, p. 235–237, 2008.

HERRON, J. *et al.* An autoantibody to single-stranded dna: Comparison of the three-dimensional structures of the unliganded fab and a deoxynucleotide–fab complex. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 11, n. 3, p. 159–175, 1991.

HIGGS, P. G.; ATTWOOD, T. K. **Bioinformatics and Molecular Evolution**. [S.l.]: Blackwell Publishing, 2005.

HUANG, J.; RU, B.; DAI, P. Bioinformatics resources and tools for phage display. **Molecules**, Molecular Diversity Preservation International, v. 16, n. 1, p. 694–709, 2011.

IDT-*HOMEPAGE*. 2011. Sítio Web: `http://www.idtdna.com/analyzer/applications/oligoanalyzer/`, acessado em 6/4/2011.

ILLUMINA-*HOMEPAGE*. 2012. Sítio Web: `http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn`, acessado em 10/4/2012.

IWASAKI, A.; MEDZHITOV, R. Regulation of adaptive immunity by the innate immune system. **Science**, American Association for the Advancement of Science, v. 327, n. 5963, p. 291, 2010.

JANG, Y. *et al.* The structural basis for dna binding by an anti-dna autoantibody. **Molecular immunology**, Elsevier, v. 35, n. 18, p. 1207–1217, 1998.

JANG, Y.; STOLLAR, B. Anti-dna antibodies: aspects of structure and pathogenicity. **Cellular and molecular life sciences**, Springer, v. 60, n. 2, p. 309–320, 2003.

JUNG, D. *et al.* Mechanism and control of v (d) j recombination at the immunoglobulin heavy chain locus. **Annu. Rev. Immunol.**, Annual Reviews, v. 24, p. 541–570, 2006.

KIM, Y. *et al.* Heavy and light chain variable single domains of an anti-dna binding antibody hydrolyze both double-and single-stranded dnas without sequence specificity. **Journal of Biological Chemistry**, ASBMB, v. 281, n. 22, p. 15287–15295, 2006.

KUMAR, S.; BLAXTER, M. Simultaneous genome sequencing of symbionts and their hosts. **Symbiosis**, Springer, p. 1–8, 2012.

KUMAR, S.; RAVUNNY, R.; CHAKRABORTY, C. Conserved domains, conserved residues, and surface cavities of c-reactive protein (crp). **Applied biochemistry and biotechnology**, Springer, p. 1–9, 2011.

LI, Y. *et al.* X-ray snapshots of the maturation of an antibody response to a protein antigen. **Nature structural biology**, New York, NY: Nature Pub. Co., c1994-c2003., v. 10, n. 6, p. 482–488, 2003.

MARANHÃO, A.; BRÍGIDO, M. Expression of anti-z-dna single chain antibody variable fragment on the filamentous phage surface. **Brazilian Journal of Medical and Biological Research**, SciELO Brasil, v. 33, n. 5, p. 569–579, 2000.

MARANHÃO, A. Q. **Utilização de Bibliotecas Apresentadas em Fagos para a Seleção de Anticorpos Ligantes a Ácidos Nucléicos**. Tese (Doutorado) — Universidade de Brasília, 2001.

MARANHÃO, A. Q.; BRÍGIDO, M. M. Anticorpos humanizados. **Biotecnologia Ciência e Desenvolvimento**, 2001.

_____. Bibliotecas apresentadas em fagos. **Biotecnologia Ciência e Desenvolvimento**, 2002.

MARDIS, E. Next-generation dna sequencing methods. **Annu. Rev. Genomics Hum. Genet.**, Annual Reviews, v. 9, p. 387–402, 2008.

MENENDEZ, A.; SCOTT, J. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. **Analytical biochemistry**, Anal Biochem, v. 336, n. 2, p. 145, 2005.

PARAMESWARAN, P. *et al.* A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. **Nucleic Acids Research**, Oxford Univ Press, v. 35, n. 19, p. e130, 2007.

PERL-*HOMEPAGE*. 2012. Sítio Web: `http://www.perl.org/`, acessado em 3/4/2012.

PERLPRIMER-*HOMEPAGE*. 2011. Sítio Web: `http://perlprimer.sourceforge.net/index.html`, acessado em 6/4/2011.

PHRED-*HOMEPAGE*. 2012. Sítio Web: `http://www.phrap.com/phred/`, acessado em 3/4/2012.

PLATINUM-*HOMEPAGE*. 2012. Sítio Web: `http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn`, acessado em 10/4/2012.

PRIMER3-*HOMEPAGE*. 2011. Sítio Web: `http://primer3.sourceforge.net/`, acessado em 6/4/2011.

RADER, C.; STEINBERGER, P.; BARBAS, C. Phage display: a laboratory manual. In: _____. [S.l.]: Cold Spring Harbor Laboratory Pr, 2004. cap. Selection from antibody libraries.

RAVN, U. *et al.* By passing in vitro screening - next generation sequencing technologies applied to antibody display and in silico candidate selection. **Nucleic Acids Research**, 2010.

ROMANO, P.; GIUGNO, R.; PULVIRENTI, A. Tools and collaborative environments for bioinformatics research. **Briefings in Bioinformatics**, Oxford Univ Press, 2011.

SAMBROOK, J. *et al.* **Molecular Cloning: a Laboratory Manual**. [S.l.]: Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY:, 2001.

SETÚBAL, J.; MEIDANIS, J. **Introduction to Computacional Molecular Biology**. [S.l.]: PWS Publishing Company, 1997.

SHENDURE, J.; JI, H. Next-generation dna sequencing. **Nat Biotechnol**, v. 26, n. 10, p. 1135–1145, 2008.

SHLENS, J. **Notes on Kullback-Leibler Divergence and Likelihood Theory**. 2007. Documento pdf disponível no sítio Web: `http://www.snl.salk.edu/~shlens/kl.pdf`, acessado em 5/1/2012.

SINGARAVELAN, B.; ROSHINI, B.; MUNAVAR, M. Evidence that the supe44 mutation of escherichia coli is an amber suppressor allele of glnx and that it also suppresses ochre and opal nonsense mutations. **Journal of bacteriology**, Am Soc Microbiol, v. 192, n. 22, p. 6039, 2010.

SMITH, A. *et al.* Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. **Nucleic acids research**, Oxford Univ Press, v. 38, n. 13, p. e142–e142, 2010.

SULTAN, M. *et al.* A global view of gene activity and alternative splicing of the human transcriptome. **Science**, 2008.

SUN, J.; EARL, D.; DEEM, M. Glassy dynamics in the adaptive immune response prevents autoimmune disease. **Physical review letters**, APS, v. 95, n. 14, p. 148104, 2005.

SWERDLOW, H. *et al.* Capillary gel electrophoresis for dna sequencing:: Laser-induced fluorescence detection with the sheath flow cuvette. **Journal of Chromatography A**, Elsevier, v. 516, n. 1, p. 61–67, 1990.

TOGAWA, R.; BRIGIDO, M. Phph: Web based tool for simple electropherogram quality analysis. In: **International Conference on Bioinformatics and Computational Biology**. [S.l.: s.n.], 2003.

TRAPNELL, C.; SALZBERG, S. How to map billions of short reads onto genomes. **Nature biotechnology**, Nature Publishing Group, v. 27, n. 5, p. 455–457, 2009.

TU, J. *et al.* Pair-barcode high-throughput sequencing for large-scale multiplexed sample analysis. **BMC genomics**, BioMed Central Ltd, v. 13, n. 1, p. 43, 2012.

UELTSCHI-*HOMEPAGE*. 2012. Documento pdf disponível no sítio Web: `http://www.ueltschi.org/teaching/chapShannon.pdf`, acessado em 5/1/2012.

WEBLOGO-*HOMEPAGE*. 2012. Sítio Web: `http://weblogo.berkeley.edu/`, acessado em 3/2/2012.

WILKINSON, I. *et al.* High resolution nmr-based model for the structure of a scfv-il-1β complex. **Journal of Biological Chemistry**, ASBMB, v. 284, n. 46, p. 31928, 2009.

WOOF, J.; BURTON, D. Human antibody–fc receptor interactions illuminated by crystal structures. **Nature Reviews Immunology**, Nature Publishing Group, v. 4, n. 2, p. 89–99, 2004.

YOUNG, M. D. **Guide to analyzing RNA-seq data**. 2011. Documento pdf disponível no sítio web: `http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn`, acessado em 5/1/2012.

ZEMLIN, M. *et al.* Expressed murine and human cdr-h3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. **Journal of molecular biology**, Elsevier, v. 334, n. 4, p. 733–749, 2003.