UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Milene Barbosa Carvalho

DESENVOLVIMENTO DE ESTRATÉGIAS DE ANÁLISE DE INTERAÇÕES ANTÍGENO-ANTICORPO E VISUALIZAÇÃO DE ANTICORPOS EM LARGA ESCALA

Belo Horizonte
2019

Milene Barbosa Carvalho

Desenvolvimento de estratégias de análise de interações antígeno-anticorpo e visualização de anticorpos em larga escala

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais

Orientadora: Liza Figueiredo Felicori Vilela
Coorientador: Franck Molina

Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Bioinformática da UFMG
Belo Horizonte, 2019

043 Carvalho, Milene Barbosa.
Desenvolvimento de estratégias de análise de interações antígeno-anticorpo e visualização de anticorpos em larga escala [manuscrito] / Milene Barbosa Carvalho. – 2019.

170 f. : il. ; 29,5 cm.

Orientadora: Liza Figueiredo Felicori Vilela. Coorientador: Franck Molina.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática - Teses. 2. Anticorpos. 3. Alinhamento de Sequência. 4. Complexo Antígeno-Anticorpo**.** 5. Base de dados. I. Vilela, Liza Figueiredo Felicori. II. Molina, Franck. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela Biblioteca do Instituto de Ciências Biológicas

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"Desenvolvimento de estratégias de análise de interações antígeno-anticorpo e visualização de anticorpos em larga escala"

Milene Barbosa Carvalho

Tese aprovada pela banca examinadora constituída pelos Professores:

Profa. Liza Figueiredo Felicori Vilela - Orientadora
UFMG

Prof. Franck Molina - Coorientador
CNRS

Profa. Silvia Beatriz Boscardin
USP

Prof. Lucas Bleicher
UFMG

Profa. Sabrina de Azevedo Silveira
UFV

Prof. Alberto Felix Antonio da Nobrega
UFRJ

Belo Horizonte, 22 de julho de 2019.

AGRADECIMENTOS

Agradeço à minha família pelo constante apoio, principalmente meus pais, a Tia Petite, a Nat e o Luís.

À Valéria por ser minha fiel escudeira, presente em todos os momentos.

Aos meus orientadores, Liza e Franck, pelas oportunidades proporcionadas e discussões que levaram ao meu desenvolvimento na área de Bioinformática.

Aos colegas do grupo de Biologia Computacional que me acompanharam no início do doutorado, em especial a Elany, a Vivi, a Camila, a Déia, o Tiago e a Taciana, aos que eu conheci depois, principalmente a Naiá, o Luan e o Igor e aos "novatos", em especial o Bruno que me ajudou com as revisões da tese. Além desses, gostaria de agradecer à Marcellinha e à Stephanie pelas discussões técnicas, desabafos e amizade.

Ao pessoal do Sys2diag, principalmente o Chico e a Pri que me deram um pouco de lar na França e a Laetitia, a Sandrine e a Camille que me entenderam mesmo não falando a mesma língua que eu.

Aos amigos do CAFEC e dos balzaquianos, principalmente a Mairinha, a Mandinha, a Lud e o Gago, pela ajuda para que eu relaxasse um pouco.

Aos amigos do Departamento de Ciência da Computação da UFSJ, em especial a Carol, o Guidoni, a Fernanda, o Vinícius, o Flávio e o Elder, que me apoiaram para que eu pudesse me afastar e realizar o doutorado.

Aos profissionais que cuidaram da minha mente e do meu corpo: Miriã, Rodrigo, Ana Carolina e Flávio.

Aos professores do Programa de Pós-Graduação em Bioinformática da UFMG, em especial, a Rafaela, o Lucas Bleicher, a Raquel e o Miguel, pelo suporte.

Ao pessoal da secretaria do Programa de Pós-Graduação em Bioinformática da UFMG, principalmente a Sheila e o Tiago, pela disponibilidade e atenção para responder meus questionamentos.

Aos professores do passado, principalmente a Cristiane e o Carlos Augusto, que serviram de inspiração para eu seguir este caminho.

RESUMO

Devido à capacidade dos anticorpos de se ligaram de maneira específica e com alta afinidade a substâncias estranhas ao organismo, eles são utilizados em terapias de doenças, diagnósticos e como ferramentas para pesquisas. Com o crescente interesse na utilização dessas proteínas, diversas estruturas de anticorpos vêm sendo resolvidas e um grande número de sequências é descoberto a cada experimento de sequenciamento de repertório de anticorpos. Apesar do sucesso da utilização de anticorpos e dos diversos estudos já realizados para se tentar identificar as propriedades que determinam a interação com o antígeno, ainda não são claras as regras que governam o reconhecimento entre anticorpo e antígeno e sua especificidade. Portanto, neste trabalho, foram desenvolvidas estratégias de análise e visualização em larga escala das cadeias dos anticorpos e suas interações com o antígeno. Nesse contexto, foi desenvolvida a plataforma Yvis, que permite a visualização de um alinhamento de milhares de sequências de anticorpos em uma única representação (*Collier de Diamants*). Para explorar as propriedades de antígeno e anticorpo e suas interações, foi implementado o banco de dados Ydb, que armazena dados da composição e das propriedades físico-químicas das cadeias de complexos antígeno-anticorpo, além da caracterização de suas interações e da interface desses complexos utilizando diferentes critérios de definição. Essas características diferenciam o Ydb dos demais bancos de dados que contêm informações de estruturas de anticorpos. Análises de parte desses dados destacaram as diferenças entre as definições de interface, além das posições mais conservadas ou divergentes nas sequências de anticorpos. Além disso, foi possível visualizar as posições que geralmente realizam interações com o antígeno e os tipos de interação presentes em cada posição. Em conjunto, a plataforma Yvis, o Ydb e a representação *Collier de Diamants* oferecem um ambiente para a análise de anticorpos que auxilia na formulação de hipóteses sobre os principais resíduos e interações presentes nos complexos e ajuda no entendimento das propriedades dos anticorpos.

PALAVRAS-CHAVE: Anticorpos. Visualização de alinhamento de sequências em larga escala. Interações antígeno-anticorpo. Plataforma Yvis. *Collier de Diamants*. Banco de dados Ydb.

ABSTRACT

Due to the ability of antibodies to bind with high affinity and specificity to substances recognized as foreign to the body, they are used in various disease therapies, diagnostics, and as research tools. With the growing interest in the use of these proteins, several antibody structures have become available, and antibody repertoire sequencing experiments discover a large number of sequences. Despite the successful use of antibodies in diverse areas and the various studies already conducted to try to identify the properties that determine the antigen interaction, the rules governing the antigen-antibody recognition and their specificity are still unclear. Therefore, in this work, strategies for antibody high-density analysis and visualization and their interactions with antigen were developed. In this context, the Yvis platform was developed, which allows the visualization of an alignment of thousands of antibody sequences in a single representation (*Collier de Diamants*) proposed here. To explore the antigen and antibody properties and their interactions, we implemented the Ydb database, which stores data on composition and physicochemical properties of antigen-antibody complex chains, as well as the characterization of the interface of these complexes using different definition criteria and their interactions. These characteristics differentiate Ydb from other databases that contain antibody structure information. Analyzes of some of these data highlighted the differences between distinct interface definitions as well as the most conserved and most divergent positions in antibody sequences. Moreover, it was possible to visualize positions that generally interact with the antigen and the types of interaction present in each position. Together, the Yvis platform, the Ydb database, and the *Collier Diamants* representation provide an antibody analysis environment that assists in formulating hypotheses about the key residues and interactions present in the complexes and help in understanding the antibody properties.

KEYWORDS: Antibodies. High-density alignment visualization. Antibody-antigen interactions. Yvis plataform. *Collier de Diamants*. AntibodY database (Ydb).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A	Alanina
AAIF	*Antigen Antibody Interaction Finder*
AbDb	*Antibody Structure Database*
AgAbDb	*Antigen-Antibody Interactions Database*
AIRR	*Adaptative Immune Receptor Repertoire*
AIRR-Seq	*Adaptative Immune Receptor Repertoire Sequencing*
Ala	Alanina
Arg	Arginina
Asn	Asparagina
Asp	Ácido aspártico
BCRs	*B Cell Receptors*
BLAST	*Basic Local Alignment Search Tool*
bNAbs	*broadly-Neutralizing Antibodies*
C	Cisteína
CDRs	*Complementarity-Determining Regions*
CH1-3	*Constant domains of heavy chain*
CL	*Constant domain of light chain*
CSV	*Comma-separated values*
Cys	Cisteína
D	Ácido aspártico
DNA	*DeoxyriboNucleic Acid*
DOI	*Digital Object Identifier*
E	Ácido glutâmico
EMA	*European Medicines Agency*
F	Fenilalanina
Fab	*Antigen-binding fragment*
Fc	*Crystallizable fragment*
FDA	*Food and Drug Administration*

FR	*Framework Regions*
G	Glicina
Gln	Glutamina
Glu	Ácido glutâmico
Gly	Glicina
gp120	*Glycoprotein 120*
GRAVY	*GRand AVerage of hYdropathy*
H	Histidina
His	Histidina
HIV	*Human Immunodeficiency Virus*
HMM	*Hidden Markov Models*
HTML	*HyperText Markup Language*
HTS	*High-throughput Sequencing*
I	Isoleucina
IEDB	*Immune Epitope Database*
Ig	Imunoglobulina
IgA	Imunoglobulina A
IgD	Imunoglobulina D
IgE	Imunoglobulina E
IgG	Imunoglobulina G
IGH	*Immunoglobulin heavy locus*
IGHV	*Immunoglobulin heavy variable gene*
IGK	*Immunoglobulin kappa locus*
IGKV	*Immunoglobulin kappa variable gene*
IGL	*Immunoglobulin lambda locus*
IGLV	*Immunoglobulin lambda variable gene*
IgM	Imunoglobulina M
Ile	Isoleucina
IMGT	*International ImMunoGeneTics information system*

JSON	*JavaScript Object Notation*
K	Lisina
L	Leucina
Leu	Leucina
Lys	Lisina
M	Metionina
Met	Metionina
MHC	*Major Histocompatibility Complex*
MSA	*Multiple Sequence Alignment*
N	Asparagina
NMR	*Nuclear Magnetic Resonance spectroscopy*
OAS	*Observed Antibody Space*
P	Prolina
PCR	*Polymerase chain reaction*
PDB	*Protein Data Bank*
PDF	*Portable Document Format*
Phe	Fenilalanina
PHP	*PHP: Hypertext Preprocessor*
PICCOLO	*Protein Interaction Collection Online*
PIRD	*Pan Immune Repertoire Database*
PMID	*PubMed Identifier*
PNG	*Portable Network Graphics*
Pro	Prolina
Q	Glutamina
R	Arginina
RNA	*RiboNucleic acid*
S	Serina
SAbDab	*Structural Antibody Database*
SACS	*Self-maintaining database of antibody crystal structure information*

scFv	*Single-chain variable Fragment*
Ser	Serina
SQL	*Structured Query Language*
SRA	*Sequence Read Archive*
SVG	*Scalable Vector Graphics*
T	Treonina
TCRs	*T Cell Receptors*
Thr	Treonina
Trp	Triptofano
Tyr	Tirosina
V	Valina
Val	Valina
VH	*Variable domain of heavy chain*
VL	*Variable domain of light chain*
W	Triptofano
Y	Tirosina
Ydb	*AntibodY Database*
Yvis	*AntibodY Visualization platform*

SUMÁRIO

1 Introdução

1.1 Anticorpos

Os anticorpos ou imunoglobulinas (Igs) são proteínas do sistema imunológico de vertebrados, produzidas pelas células B (ou linfócitos B). Essas moléculas são capazes de se ligarem, de maneira específica e com alta afinidade, a substâncias reconhecidas como estranhas ao organismo (antígenos). Por meio dessa ligação, os anticorpos podem neutralizar agentes tóxicos ou desencadear uma resposta imune contra o antígeno (FRENZEL et al., 2017; SELA-CULANG; KUNIK; OFRAN, 2013).

Os anticorpos geralmente são constituídos por duas cadeias leves idênticas e duas cadeias pesadas idênticas, unidas por ligações dissulfeto. As interações entre essas cadeias fazem com que a molécula tenha a forma da letra Y, como apresentada pela molécula IgG na Figura 1.1. Em humanos e camundongos, por exemplo, essa configuração é encontrada com maior frequência, no entanto, formas multiméricas, compostas por mais de uma dessas unidades, também podem ser encontradas (isotipos IgM e IgA) (TEPLYAKOV et al., 2016). Além disso, algumas espécies como camelos, lhamas e tubarões, podem produzir anticorpos formados apenas por um par de cadeias pesadas (GONZALEZ-SAPIENZA; ROSSOTTI; TABARES-DA ROSA, 2017). No entanto, no restante deste texto, o termo anticorpo será utilizado para se referir à molécula mais comumente encontrada, composta por duas cadeias leves e duas pesadas formando uma molécula em Y.

As cadeias leves de um anticorpo contêm um domínio variável (VL) e um constante (CL), enquanto as cadeias pesadas possuem um domínio variável (VH) e três constantes (CH1, CH2 e CH3), como apresentado na molécula IgG da Figura 1.1. Além dos domínios das cadeias que compõem os anticorpos, é possível reconhecer três fragmentos nestas moléculas, o Fc (*crystallizable fragment* – fragmento cristalizável) e dois Fabs (*antigen-binding fragments* – fragmentos de ligação ao antígeno) idênticos (Figura 1.1). A digestão de um anticorpo pela enzima papaína gera esses três fragmentos. O Fc é composto

por 2 ou 3 domínios constantes dependendo do isotipo do anticorpo. Os isotipos IgG, IgA e IgD possuem 2 fragmentos enquanto os isotipos IgM e IgE possuem 3. Além de diferenciar os tipos de isotipos de anticorpos, o Fc é responsável pela interação com receptores e proteínas do sistema imunológico, o que pode desencadear uma resposta desse sistema. Cada Fab é composto pela cadeia leve (VL e CL) e os domínios VH e CH1 da cadeia pesada. Os domínios variáveis das duas cadeias (VL e VH) possuem três regiões de alça (*loop*) hipervariáveis, conhecidas como CDRs (*Complementarity-Determining Regions* – regiões determinantes de complementariedade) (Figura 1.1 e 1.2). Tais regiões contêm a maioria dos resíduos do anticorpo que interagem com os antígenos (FINLAY; ALMAGRO, 2012; TILLER; TESSIER, 2015). No entanto, alguns resíduos dos CDRs não fazem parte do paratopo, isto é, da região do anticorpo que está em contato com o antígeno. A região do antígeno que está em contato com o anticorpo é conhecida como epitopo. Além disso, alguns resíduos do epitopo fazem interações com resíduos do anticorpo que não pertencem aos CDRs, conhecidas como regiões de *framework* (FR – *framework regions*). Essas regiões estão representadas em verde na Figura 1.2.

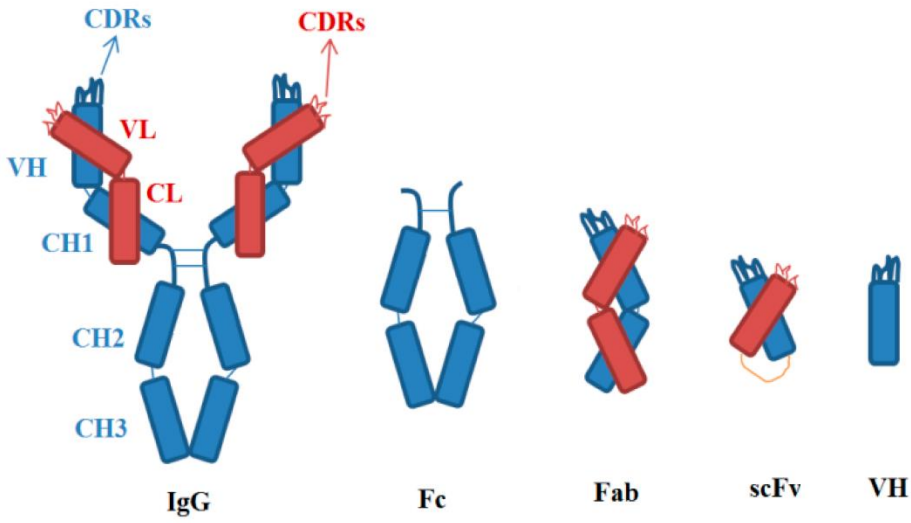


Figura 1.1: Representação de um anticorpo (IgG) e os domínios da cadeia leve (VL e CL) e pesada (VH, CH1-3). Os fragmentos Fc (*crystallizable fragment*) e Fab (*antigen-binding fragment*), presentes no IgG estão também apresentados separadamente. A molécula scFv (*single-chain variable Fragment*) é formada pela conexão do VH e do VL através de um *linker*. Figura adaptada de Li et al. (2016).

Os domínios variáveis das cadeias dos anticorpos compartilham similaridades funcionais e estruturais, mesmo comparando-se cadeias leves e pesadas. O domínio variável de cada uma dessas cadeias, apresentado na Figura 1.2, é composto de 9 fitas beta (A, B, C, C', C'', D, E, F e G) ligadas por 5 voltas beta e 3 *loops*, formando um sanduíche de duas folhas (ABED e GFCC'C'') empacotadas uma contra a outra através de interações hidrofóbicas. Estas folhas são mantidas unidas através de uma ponte dissulfeto entre duas cisteínas conservadas, uma em cada folha (LEFRANC, 2014). Os *loops* existentes entre as fitas B e C, F e G e entre C' e C", correspondem às regiões de CDRs (1, 2 e 3, respectivamente), enquanto as demais porções correspondem às regiões de *framework*.

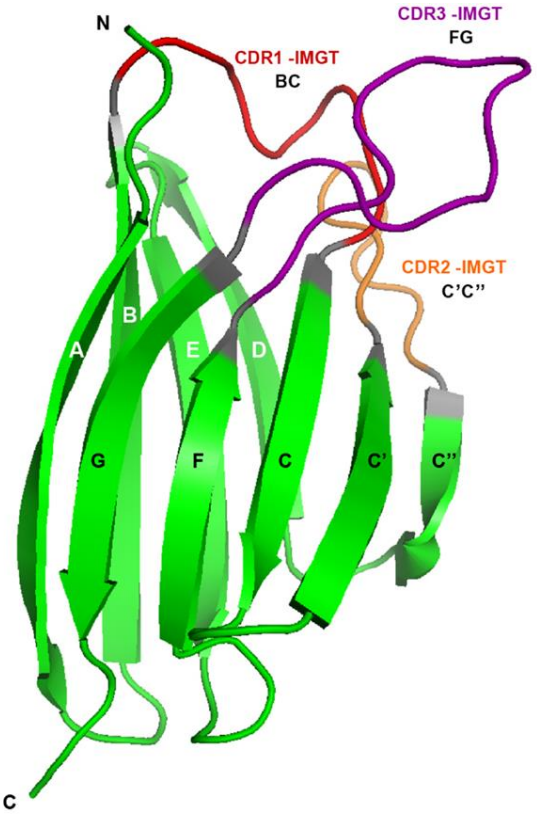


Figura 1.2: Representação (*ribbon*) da estrutura tridimensional do domínio variável da cadeia pesada de um anticorpo. O domínio variável da cadeia leve também apresenta esta estrutura. Letras identificam as fitas e cores das letras diferenciam as fitas de cada uma das duas folhas beta da estrutura. As regiões de *framework*, compostas pelas fitas e alças que as unem, estão representadas em verde e com outras cores são representados os 3 *loops* que formam os CDRs. Fonte: Adaptado de Lefranc (2014).

A Figura 1.3 apresenta o alinhamento de 605 regiões de *framework* de domínios varáveis das cadeias pesadas de anticorpos. Nessa figura foram omitidos os CDRs, pois eles apresentam variações tanto no número e composição de resíduos quanto na estrutura. Os círculos representam as regiões de início e fim de cada CDR. Nessa figura é possível visualizar a similaridade que existe entre as regiões de *framework* dos domínios variáveis das cadeias dos anticorpos. Essa similaridade permite a utilização de esquemas de numeração para comparar sequências de domínios variáveis de anticorpos, mesmo sem o conhecimento da estrutura de um anticorpo.

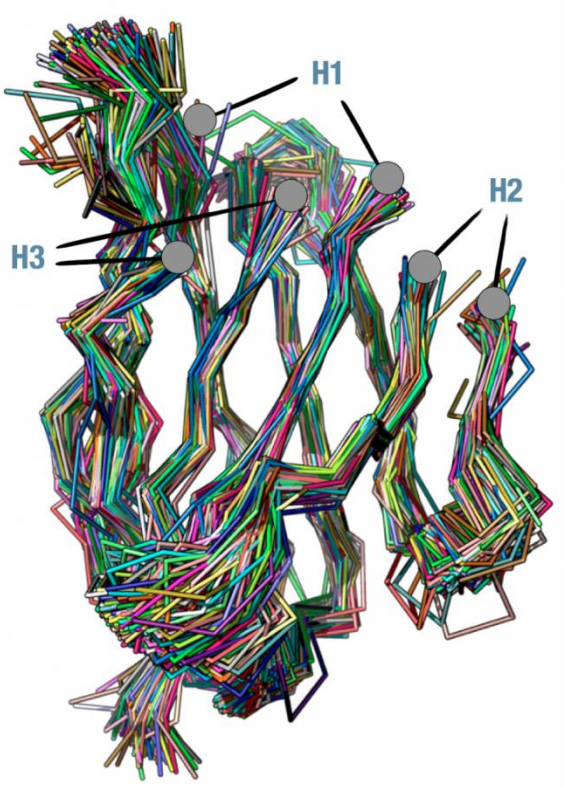


Figura 1.3: Alinhamento estrutural do domínio variável de 605 cadeias pesadas de anticorpos com, no máximo, 95% de identidade. As regiões de CDR (H1, H2, H3) foram omitidas e substituídas por 2 círculos cada (início e final de cada CDR). Fonte: Figura obtida do *Oxford Protein Informatics Group*, acessível em https://goo.gl/dYGjUr.

1.2 Esquemas de numeração de domínios variáveis de cadeias de anticorpos

Para permitir a comparação de diferentes cadeias de proteínas, é necessária a identificação de resíduos com funções semelhantes ou estruturalmente equivalentes (HONEGGER; PLÜCKTHUN;

PLUCKTHUN, 2001). Para isso, antes mesmo da caracterização de um grande número de estrutura de anticorpos, Kabat e seus colaboradores alinharam 77 sequências de porções variáveis de anticorpos e outras proteínas da superfamília das imunoglobulinas (WU; KABAT, 1970). A partir desse alinhamento, os pesquisadores definiram um esquema de numeração que leva em consideração alguns resíduos conservados nas diversas sequências analisadas e adiciona *gaps* nas sequências a fim de manter esses resíduos sempre nas mesmas posições. Outra característica observada nesse estudo foi a existência de regiões com variabilidade de tamanho e resíduos, os CDRs. Posteriormente, Chothia e colaboradores fizeram algumas alterações no esquema de numeração, principalmente no posicionamento dos *gaps* no CDR1, para adaptá-lo às estruturas cristalográficas de anticorpos existentes na época (CHOTHIA; LESK, 1987).

Tanto o esquema de numeração proposto por Kabat quanto o proposto por Chothia utilizam definições diferentes para cadeias leves e pesadas. Posteriormente, Lefranc e colaboradores propuseram um esquema de numeração unificado, não somente para cadeias leves e pesadas de anticorpos, mas também para receptores de células T (LEFRANC, 2014; LEFRANC et al., 2003). Esse esquema de numeração leva em consideração as sequências e também a estrutura das moléculas, além das definições das regiões de alça hipervariáveis, os CDRs. Estes e outros esquemas de numeração foram anteriormente analisados e o mapeamento das posições definidas em cada um deles pode ser encontrado em (DONDELINGER et al., 2018). O esquema de numeração do IMGT define 5 posições que possuem aminoácidos conservados entre as sequências do domínio variável de cadeias de anticorpos: posições 23 e 104 contêm cisteínas, a posição 41 contém um triptofano, a posição 89 contém um aminoácido hidrofóbico e a posição 118 contém uma fenilalanina ou um triptofano. Essas posições conservadas foram destacadas na Figura 1.4, que apresenta as regiões definidas pelo esquema de numeração: *frameworks* e CDRs. As posições de 1 a 26 correspondem ao *framework* 1, de 27 a 38 ao CDR1, de 39 a 55 ao *framework*

2, de 56 a 65 ao CDR2, de 66 a 104 ao *framework* 3, de 105 a 117 ao CDR3-IMGT e as posições de 118 a 128 correspondem ao *framework* 4. *Gaps* nos CDRs são sempre inseridos no centro dessas regiões e caso o CDR3 tenha mais que 13 aminoácidos, posições adicionais são inseridas entre as posições 111 e 112 (112.1, 111.1, 112.2, 111.2).

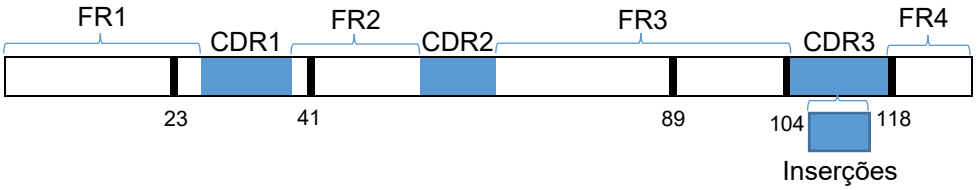


Figura 1.4: Representação do domínio variável da cadeia de um anticorpo segundo o esquema de numeração do IMGT. As 4 regiões de *framework* (FRs) e os 3 CDRs são exibidos e as 5 posições que têm resíduos conservados são destacadas. Nas cadeias onde o CDR3 possui mais do que 13 aminoácidos, posições são inseridas entre as posições 111 e 112.

A partir da aplicação de um esquema de numeração sobre sequências de domínios variáveis de anticorpos é possível obter uma correlação entre sequência e a estrutura do domínio. Dessa maneira, mesmo sem se ter a estrutura de um anticorpo, é possível identificar as regiões de *framework* e CDRs, e comparar diferentes sequências considerando suas características estruturais.

A conservação da estrutura das regiões de *framework* e a variabilidade existente nos CDRs são características dos anticorpos que são alcançadas por meio de mecanismos que ocorrem no desenvolvimento e maturação de todas as células B, mas também através dos diferentes contatos que cada organismo tem ao longo da vida.

1.3 Diversidade do repertório de anticorpos

A diversidade de anticorpos produzidos por um indivíduo é alcançada através de alguns processos que ocorrem durante o desenvolvimento e maturação das células B. A seguir é descrito, de maneira simplificada, o mecanismo de desenvolvimento e maturação destas células em humanos, que é similar ao da maioria dos demais vertebrados. Em humanos, o domínio variável da cadeia pesada é codificado pelos genes do *locus* IGH e da leve pelos genes do *locus* IGK (*kappa*) ou do IGL (*lambda*). Em células não-

linfoides são encontrados diversos segmentos gênicos V (*variable*) e J (*joining*) nos *loci* de cadeias leves e pesadas, além de diversos segmentos gênicos D (*diversity*) no *locus* IGH. Comumente estes segmentos gênicos são chamados somente de genes e, no restante desse texto, essa denominação mais comum será utilizada. O *locus* IGH humano é apresentado na Figura 1.5, destacando os genes V, D e J da região variável e os genes da região constante (C). Além dos genes, são também representados os diversos alelos já conhecidos para cada um desses genes (quadrados sobre os genes na Figura 1.5). Os diversos genes V, D e J de células não-linfoides são conhecidos como *germlines*. Na formação de cada célula B, ocorre o processo de recombinação somática em que alguns destes genes são removidos e apenas um gene V, um J e um D (em cadeias pesadas) são unidos no DNA (*DeoxyriboNucleic Acid* - ácido desoxirribonucleico) para produzir um éxon completo da região variável de cada cadeia (Figura 1.5). Durante o processo de junção dos genes, podem ocorrer ainda a inserção ou remoção de nucleotídeos, como apresentado na Figura 1.5 (EHRENMANN et al., 2010; GEORGIOU et al., 2014). Portanto, este processo de recombinação aliado às inserções e remoções de nucleotídeos, gera um repertório de células B capazes, a princípio, de reconhecer qualquer antígeno com uma afinidade baixa ou média através de receptores codificados pelos genes recombinados (FINLAY; ALMAGRO, 2012; WATSON; GLANVILLE; MARASCO, 2017).

As células B que já sofreram o processo de recombinação dos genes V(D)J (genes V e J no *locus* da cadeia leve e genes V, D e J no *locus* da cadeia pesada) são conhecidas como células maduras e, se ainda não se ligaram a um antígeno, são chamadas de virgens ou *naïves*. Células B maduras possuem receptores em sua superfície que se ligam aos antígenos, conhecidos como BCRs (*B Cell Receptors*). Esses receptores são imunoglobulinas ligadas à membrana das células B e, consequentemente, possuem a mesma estrutura dos anticorpos secretados por estas células, exceto por um domínio transmembranar que os mantêm na

superfície celular. Portanto, a região variável dos BCRs e dos anticorpos secretados por uma célula B são codificados pelos mesmos genes V(D)J (HOEHN et al., 2016).

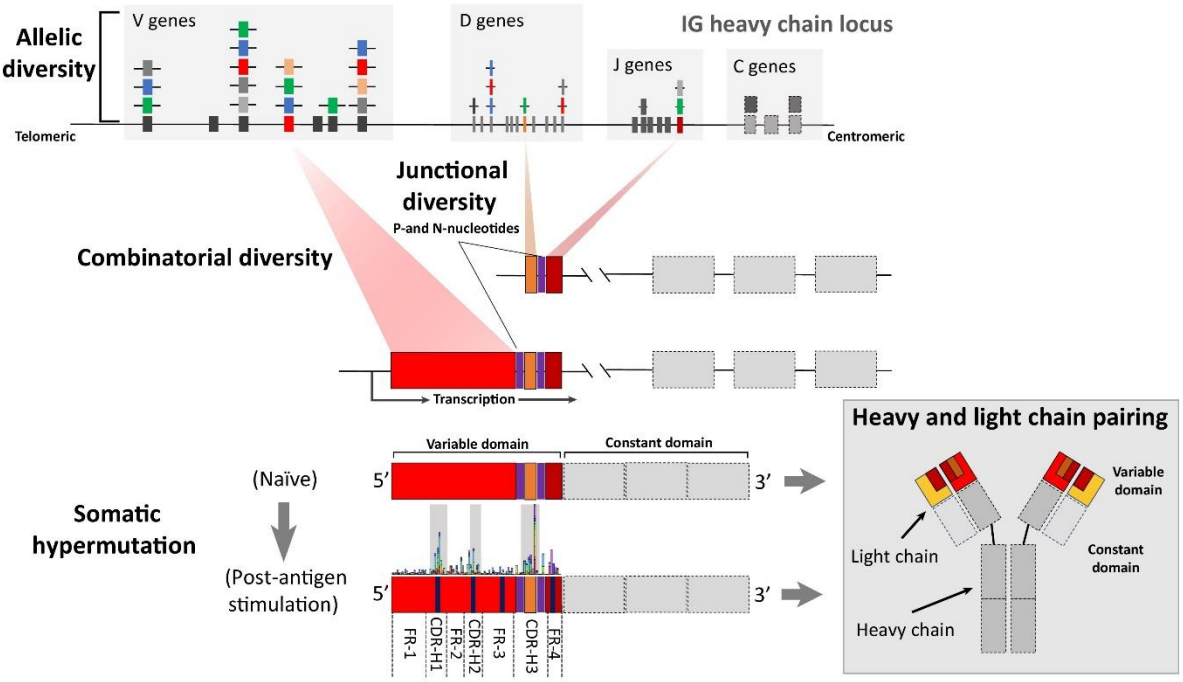


Figura 1.5: Principais elementos que determinam a diversidade de anticorpos. Nessa figura é apresentado um desenho esquemático dos processos de recombinação dos genes V, D e J da cadeia pesada e hipermutação somática que ocorrem durante a formação e expansão clonal de anticorpos. A diversidade alélica em um indivíduo é representada pelos diversos genes V, D e J presentes em seu genoma e sobrepostos são representados genes pertencentes a outros indivíduos. As regiões de *frameworks* (FRs) e CDRs são destacadas nos transcritos apresentados antes e após a hipermutação somática (após estímulo pelo antígeno) (WATSON; GLANVILLE; MARASCO, 2017)

Após a interação entre um antígeno e um receptor de uma célula B e a ativação dessa célula, ocorre a expansão clonal da mesma e as células geradas a partir dela diferenciam-se em células capazes de produzir anticorpos codificados, a princípio, pelos mesmos genes V(D)J. No entanto, após a ativação de uma célula B, as células produzidas podem sofrer hipermutação somática, isto é, quando ocorrem mutações pontuais nos genes rearranjados, diversificando-os em relação às sequências *germlines* (Figura 1.5). Neste processo, podem surgir sequências que codificam anticorpos (e BCRs) que podem se ligar com maior afinidade ao antígeno. Por competição entre as células B para se ligarem ao antígeno, aquelas com BCRs mais específicos e que se ligam com maior afinidade, são selecionadas e seguem o processo de

expansão clonal. Dessa maneira. essas células podem produzir anticorpos mais adequados ao combate contra o antígeno (GEORGIOU et al., 2014; GONZÁLEZ-MUÑOZ et al., 2012).

1.4 Sequenciamento do repertório de receptores de células B

A diversidade gerada pelos processos que ocorrem durante a recombinação dos genes dos *loci* de imunoglobulinas, juntamente com a hipermutação somática ocorrida a partir da ativação de células B, começou a ser explorada há algumas décadas, por exemplo, por meio de técnicas de sequenciamento de Sanger. No entanto, o advento das tecnologias de sequenciamento em larga escala (HTS – *High-throughput Sequencing*) possibilitou a análise do repertório de receptores de células B (ou de anticorpos) com uma profundidade até então não obtida (GEORGIOU et al., 2014). Chamado de Rep-Seq (BENICHOU et al., 2012), Ig-Seq (GEORGIOU et al., 2014), BCR-Seq (LAVINDER et al., 2015) ou AIRR-Seq (*Adaptative Immune Receptor Repertoire Sequencing*) (NIELSEN; BOYD, 2018), o sequenciamento de repertório de receptores de células B permite a análise da resposta de células B de organismos saudáveis, infectados, vacinados ou com alguma desregulação no repertório, causada por idade, alergia ou doença autoimune, por exemplo (BOYD; CROWE, 2016; ICHINOHE et al., 2018). Esses estudos viabilizam um melhor entendimento da dinâmica do sistema imunológico, a descoberta de novas moléculas baseadas em anticorpos tanto para diagnóstico quanto para terapia e o desenho de moléculas que permitem o desenvolvimento de vacinas (MIHO et al., 2018).

Uma típica análise de repertório através de Rep-Seq é iniciada com o isolamento de células B do sangue do organismo alvo ou de algum outro tecido que contenha esse tipo de célula. O RNA (*RiboNucleic acid*) das células B é extraído e o DNA complementar ao RNA é gerado. Posteriormente, a região variável (ou parte dela) de um ou mais *loci* de imunoglobulina é amplificado por PCR (*Polymerase chain reaction*). Os *amplicons* dos *loci* de imunoglobulinas são isolados e a preparação da biblioteca de sequenciamento é realizada. Após essa etapa, o material é sequenciado utilizando equipamentos de HTS e as sequências

obtidas devem então ser processadas e analisadas com o auxílio de ferramentas de bioinformática. Este é um exemplo de *workflow* para a realização de um sequenciamento de repertório de receptores de células B, porém existem diversas outras alternativas (BOYD; CROWE, 2016; GEORGIOU et al., 2014; IMKELLER; WARDEMANN, 2018). Nesse processo, diversas fontes de viés podem levar a análises incorretas do repertório, como, por exemplo, a incorporação de nucleotídeos incorretos durante PCR ou erros de sequenciamento. Algumas dessas fontes podem ser conhecidas e minimizadas, utilizando-se técnicas biológicas e de bioinformática (FRIEDENSOHN; KHAN; REDDY, 2017; WARDEMANN; BUSSE, 2017).

O processamento e análise das sequências geradas por Rep-Seq envolvem diversos passos. Alguns deles, comuns a diversos estudos deste tipo, compreendem o pré-processamento das *reads*, atribuição de genes V(D)J, definição de clones e linhagens e, por fim, análise do repertório (GREIFF et al., 2015; LÓPEZ-SANTIBÁÑEZ-JÁCOME; AVENDAÑO-VÁZQUEZ; FLORES-JASSO, 2019; MIHO et al., 2018; YAARI; KLEINSTEIN, 2015). Na etapa de pré-processamento, as *reads* são montadas (se necessário) gerando sequências e alguns filtros de qualidade são aplicados. Posteriormente, são definidos os possíveis genes que deram origem às sequências por meio do alinhamento dessas sequências com as sequências dos genes *germline* do organismo sequenciado. A partir destes dados, pode ser gerada uma árvore de linhagem dos anticorpos ou receptores sequenciados. Além disso, geralmente, é possível analisar a diversidade do repertório, os modelos e taxas de mutação e a frequência de utilização de cada gene V(D)J. Em alguns trabalhos é possível ainda a análise da evolução do repertório ou a resposta do mesmo a uma infecção ou vacinação.

Atualmente, uma análise de repertório de receptores de células B pode gerar milhares a bilhões de sequências (KOVALTSUK et al., 2018). Esse grande número de sequências e os possíveis vieses gerados no experimento fazem com que ainda sejam demandadas ferramentas para a análise e visualização de

dados de repertórios de receptores de células B. Apesar disso, as análises já realizadas permitiram um avanço no entendimento do sistema imune e na utilização deste entendimento para a definição de terapias e vacinas.

1.5 Seleção e produção de anticorpos monoclonais

A habilidade dos anticorpos de se ligar de maneira específica e com alta afinidade a, virtualmente, qualquer superfície é uma das chaves do sistema imunológico e também uma ferramenta para pesquisas e diagnósticos, além de permitir que sejam utilizados na terapia para diversas doenças (SELA-CULANG; KUNIK; OFRAN, 2013). A terapia baseada em anticorpos representa hoje a classe de terapias biológicas com maior sucesso clínico. Até o final do ano de 2017, 57 anticorpos e 11 biossimilares estavam em uso clínico, aprovados pelas agências de controle de medicamentos dos Estados Unidos ou da União Europeia, FDA (*Food and Drug Administration*) e EMA (*European Medicines Agency*), respectivamente (GRILO; MANTALARIS, 2019). Somente em 2018, 12 novos anticorpos foram aprovados por pelo menos uma destas agências (KAPLON; REICHERT, 2019). Além dos anticorpos, outras moléculas baseadas em suas estruturas vêm sendo produzidas e utilizadas comercialmente, como regiões Fab, ou estão em fase de teste, como os scFvs (STROHL, 2018).

Para a utilização de anticorpos como uma ferramenta de diagnóstico ou para terapia, normalmente, é necessária a produção de uma grande quantidade de anticorpos idênticos (anticorpos monoclonais), que podem ser originados de uma mesma célula B ou de clones idênticos desta célula ou ainda produzidos por um sistema de expressão de proteínas. A técnica de hibridoma foi a primeira técnica que impulsionou a produção de anticorpos monoclonais. Nessa técnica, após a imunização de um animal, células B são extraídas do baço e fusionadas a uma linhagem celular tumoral (KOHLER; MILSTEIN, 1975). Os hibridomas, linhagens celulares obtidas através dessa fusão, podem ser cultivados gerando novas células

que produzem um mesmo anticorpo. Um hibridoma que produz um anticorpo de interesse pode então ser isolado e utilizado para a produção de anticorpos monoclonais.

Apesar do sucesso da técnica de hibridoma e de sua ampla utilização, anticorpos produzidos por outros animais podem ser reconhecidos como proteínas estranhas pelo sistema imunológico humano, desencadeando uma resposta indesejada. Para reduzir este tipo de problema, técnicas de engenharia genética são utilizadas para produzir genes que codificam anticorpos que se ligam a um antígeno específico, como os descobertos pela imunização de animais, mas com características similares as dos anticorpos humanos. Os genes podem ser inseridos/transfectados em um sistema de expressão, como bactérias, leveduras e células mamíferas (APGAR et al., 2016; JONES et al., 1986; KOHLER; MILSTEIN, 1975; MORRISON et al., 1984). Nas primeiras tentativas para reduzir a imunogenicidade de anticorpos oriundos de outros organismos, os genes que codificavam as imunoglobulinas eram modificados por engenharia genética para conter a região variável de um anticorpo produzido pela imunização de um animal e a região constante de um anticorpo humano (MORRISON et al., 1984). Nesta técnica, são gerados anticorpos quiméricos, já que eles contêm partes originárias de diferentes organismos. Posteriormente, surgiu a técnica de humanização de um anticorpo, na qual somente algumas porções do anticorpo produzido por outros organismos, principalmente aquelas que se ligam ao antígeno, são mantidas no gene sintetizado, enquanto as demais porções são substituídas por porções encontradas em anticorpos humanos (APGAR et al., 2016; JONES et al., 1986). Desta maneira, o sucesso da produção de anticorpos quiméricos ou humanizados está associado ao conhecimento das regiões que interagem com o antígeno e das sequências e estruturas que definem um anticorpo humano.

Apesar das técnicas de descoberta de anticorpos *in vivo* utilizarem o processo de seleção natural do sistema imunológico, em que o animal imunizado produz anticorpos com características desejáveis para uma resposta a um determinado antígeno, outra técnica tem sido utilizada para a descoberta de anticorpos

que se ligam a um determinado antígeno: a seleção *in vitro* de anticorpos. As técnicas de *display*, que utilizam fagos (*phage display*), bactérias, leveduras, células mamíferas ou ribossomos, permitem a seleção *in vitro* de anticorpos que se ligam a um determinado antígeno alvo (CONROY et al., 2017). Para isso, uma biblioteca de sequências que codificam a região variável de anticorpos é gerada. Essa biblioteca pode ser obtida a partir de doadores humanos, imunizados ou não, ou pode ser sintética ou semissintética, gerada a partir do conhecimento que se tem sobre os anticorpos humanos. Abordagens sintéticas abrem possibilidades para a descoberta de anticorpos que podem não ser encontrados nas amostras de doadores. No *phage display*, a sequência é ligada ao gene de uma proteína de superfície do fago e é expressa na superfície do vírus (fase *display*). Bilhões de clones diferentes são colocados em contato com o antígeno alvo (fase *panning*) e, aqueles que se ligaram ao antígeno têm sua sequência elucidada (fase *screening*) (FRENZEL et al., 2017). Estas etapas podem ser repetidas para se obter sequências com maior afinidade, por exemplo. As sequências selecionadas podem, então, ser inseridas em células que expressarão o anticorpo monoclonal. As demais técnicas de *display* possuem fases semelhantes a estas descritas para o *phage display*. A seleção de anticorpos utilizando técnicas de *display*, ou seja, *in vitro*, possibilitam um maior controle da exposição do antígeno aos anticorpos, mas pode ser trabalhosa e demorada, levando meses para produzir um anticorpo com as características desejadas (SORMANNI; APRILE; VENDRUSCOLO, 2018). Além disso, o sucesso das técnicas de *display* está associado às bibliotecas utilizadas. Por exemplo, na concepção de bibliotecas sintéticas e semissintéticas, é necessário utilizar um conjunto de sequências que possam ser expressas, que não sejam reconhecidas pelo sistema imunológico como invasores e que se liguem com alta afinidade ao antígeno.

Os métodos para identificação e geração de anticorpos a partir de imunização (*in vivo*) ou métodos de *display* (*in vitro*), descritos anteriormente, contribuíram substancialmente para o desenvolvimento de terapias baseadas em anticorpos (TILLER; TESSIER, 2015). No entanto, anticorpos selecionados *in vivo*

ou *in vitro*, normalmente têm como alvo epitopos imunodominantes, isto é, aqueles que mais facilmente podem ser reconhecidos. Esse tipo de reconhecimento pode atrapalhar a descoberta de anticorpos que se ligam a outros epitopos, mas com menor afinidade. Assim, pode ser que um anticorpo neutralizante que se liga a um epitopo diferente dos imunodominantes, não seja selecionado utilizando as técnicas anteriormente descritas. Além disso, essas técnicas de seleção ainda apresentam problemas para a obtenção de anticorpos contra alguns alvos, como algumas proteínas de membrana e proteínas ou peptídeos que formam agregados (SORMANNI; APRILE; VENDRUSCOLO, 2018). Para contornar estes problemas e para produzir uma molécula com propriedades funcionais e biofísicas adequadas, pode-se aplicar alguma técnica de engenharia de anticorpos ou desenho racional de anticorpos (ROY et al., 2017; SORMANNI; APRILE; VENDRUSCOLO, 2018).

As técnicas aplicadas na engenharia de anticorpos são diversas e variam desde métodos baseados em conhecimento obtido de resultados de mutagênese, por exemplo, a métodos computacionais baseados em princípios básicos da estrutura e composição dos anticorpos. Com o uso destes métodos, é possível redesenhar ou otimizar anticorpos já existentes ou ainda criar anticorpos *de novo* (TILLER; TESSIER, 2015). Geralmente, o processo de engenharia do anticorpo é guiado pelo conhecimento prévio da estrutura dos anticorpos e de suas interações com o antígeno (TEPLYAKOV et al., 2016). Mas, como muitas vezes a estrutura de um anticorpo específico é desconhecida, bem como suas interações, outros métodos devem ser utilizados para guiar o processo de engenharia, como a modelagem de anticorpos e o atracamento (*docking*) molecular. No entanto, o sucesso dessas técnicas está diretamente relacionado ao conhecimento das estruturas dos anticorpos, a relação entre as sequências das cadeias dos anticorpos e sua estrutura, e das interações que ocorrem em um complexo antígeno-anticorpo (KRAWCZYK; DUNBAR; DEANE, 2017; SIRCAR; KIM; GRAY, 2009; WEITZNER et al., 2017).

1.6 Disponibilização de dados de sequências e estruturas de anticorpos e complexos antígeno-anticorpo

O interesse e o uso de imunoglobulinas como ferramentas de pesquisa, diagnóstico e terapia levaram à disponibilização de um crescente número de estruturas e sequências de anticorpos em banco de dados de proteínas em geral. Esta seção apresenta as principais bases de dados de informações estruturais de anticorpos e, criadas mais recentemente, as bases de dados de experimentos de sequenciamento de repertório de imunoglobulinas.

1.6.1 Bancos de dados com informações estruturais de anticorpos

O principal repositório de estrutura de proteínas é o *Protein Data Bank* (PDB) (BERMAN, 2000) e, aproximadamente 2,1% das estruturas depositadas são de anticorpos (FERDOUS; MARTIN, 2018). Além disso, esse número tem crescido com o aumento de depósitos a cada ano. Devido ao interesse em relação às estruturas de anticorpos e de complexos antígeno-anticorpo, além da necessidade de informações relativas às estruturas disponíveis no PDB, diversos bancos de dados dedicados a anticorpos foram desenvolvidos nos últimos anos. Dentre esses bancos de dados, destacam-se: *Immune Epitope Database* (IEDB) 3.0 (VITA et al., 2015), abYsis (SWINDELLS et al., 2017), SACS: *Self-maintaining database of antibody crystal structure information* (ALLCORN; MARTIN, 2002), *IMGT/3Dstructure-DB* (EHRENMANN; KAAS; LEFRANC, 2010), *Antigen-Antibody Interactions Database* (AgAbDb) (KULKARNI-KALE et al., 2014), *The Structural Antibody Database* (SAbDab) (DUNBAR et al., 2014), *Antibody Structure Database* (AbDb) (FERDOUS; MARTIN, 2018) e, mais recentemente, AppA (NGUYEN; VERMA; ZHONG, 2019) que serão descritos a seguir.

O IEDB é um banco de dados de epitopos curado manualmente que armazena dados experimentais de ligação de epitopos com receptores de célula T, anticorpos e moléculas de MHC (*Major Histocompatibility Complex*) (VITA et al., 2015). Por ser um banco de dados de epitopos, o IEDB possui

diversos registros sem a estrutura tridimensional do complexo antígeno-anticorpo. No entanto, quando um epitopo do banco de dados tem sua estrutura disponível no PDB (*Protein Data Bank*)(BERMAN, 2000), os dados referentes a ela são disponibilizados. Quando a estrutura tridimensional é disponível, o IEDB apresenta uma lista de resíduos correspondente ao epitopo, uma lista de resíduos correspondente ao paratopo e a área de contato das duas moléculas. Estas informações são apresentadas em duas seções, quando possível: "contatos curados" e "contatos calculados". A seção "contatos curados" apresenta as informações existentes na publicação relacionada à estrutura tridimensional. A seção "contatos calculados" apresenta os dados de epitopo e paratopo calculados a partir da estrutura tridimensional disponível no PDB, considerando os resíduos do antígeno e do anticorpo que estão a uma distância máxima de 4Å (PONOMARENKO et al., 2011). Apesar de listar os resíduos que compõem o paratopo e o epitopo, a atual versão do IEDB não apresenta os dados referentes aos tipos de contatos existentes entre antígeno e anticorpo.

O abYsis é um sistema *web* que inclui um banco de dados de sequências e estruturas de anticorpos integrado com um conjunto de ferramentas de análise destes dados (SWINDELLS et al., 2017). Nesse banco de dados, as sequências das cadeias dos anticorpos são automaticamente numeradas segundo os esquemas de numeração de Kabat e Chothia, mencionados anteriormente na seção 1.2. Além disso, são disponibilizadas informações como CDRs, pareamento de cadeias leves e pesadas e identificação de resíduos não usuais, isto é, pouco frequentes em relação às sequências armazenadas pelo abYsis. A pesquisa na base de dados pode ser realizada de diversas maneiras, como pelos nomes dos antígenos, organismo, referência bibliográfica, estruturas canônicas de CDRs ou *motifs*. O abYsis apresenta os resultados de maneira gráfica e permite a análise de tendências dos dados. Além disso, o abYsis apresenta uma visualização clássica do alinhamento de múltiplas sequências (MSA – *Multiple Sequence Alignment*)

das cadeias dos anticorpos analisados. Apesar dos diversos recursos do abYsis, ele não apresenta dados relacionados às interações entre o antígeno e o anticorpo, ao paratopo e ao epitopo.

O IMGT/3Dstructure-DB é um banco de estruturas tridimensionais, desenvolvido pelo IMGT, que possui dados de anticorpos ou de receptores de imunoglobulinas, receptores de células T e proteína de MHC. Ele fornece uma identificação padronizada que se baseia na ontologia definida pelo IMGT, incluindo dados de *germlines* (genes e alelos para a região variável) e numeração dos resíduos de anticorpos segundo a numeração única do IMGT (LEFRANC et al., 2003). Além dessas numerações, o IMGT definiu o *IMGT/Collier de Perles* (RUIZ; LEFRANC, 2002), uma representação gráfica em duas dimensões baseada na numeração do IMGT (LEFRANC, 2014; LEFRANC et al., 2003). Esta representação permite a visualização de cada sequência do domínio variável de uma cadeia de anticorpo, relacionando-a com sua estrutura tridimensional. O IMGT/3Dstructure-DB também apresenta informações sobre os contatos entre antígeno e anticorpo. Estes contatos são definidos por um programa local escrito em C que considera que dois átomos estão em contato se nenhuma molécula de água pode ser posicionada entre eles. Desta forma, usando como *cutoff* o diâmetro da molécula de água, é possível definir as interações de Van der Waals. Para a identificação das interações de hidrogênio, são considerados critérios de distância entre os átomos doadores e aceptores, além do ângulo entre os átomos doador, aceptor e hidrogênio. Os contatos atômicos são identificados como polar, ligação de hidrogênio, não polar, ou ponte dissulfeto. Além disso, os contatos são categorizados segundo a localização dos átomos na molécula: cadeia lateral ou cadeia principal (EHRENMANN; KAAS; LEFRANC, 2010). As informações de contatos são disponibilizadas na interface do banco de dados como uma tabela que relaciona o par de resíduos envolvidos em um contato com a quantidade e o tipo de interações por eles realizadas. O IMGT/3Dstructure-DB ainda apresenta outra tabela com as mesmas informações resumidas, informando somente a contagem de resíduos e a contagem de interações de cada tipo, para cada cadeia das moléculas.

O AgAbDb é um banco de dados de interações entre antígeno e anticorpo. As interações são definidas a partir da análise de complexos disponíveis no PDB através do programa *Antigen Antibody Interaction Finder* (AAIF). Este programa calcula interações não covalentes (interações de Van der Waals, ligações de hidrogênio, pontes salinas e interações mediadas por água) usando critérios de distância de geometria e, no caso das interações mediadas por água, somente as moléculas presentes na estrutura PDB são utilizadas (KULKARNI-KALE et al., 2014). As interações são listadas identificando o par de átomos envolvidos e o tipo de interação (nível atômico). Porém, quando pares de resíduos que interagem são listados, não há indicação do tipo de interação. Adicionalmente, os resíduos do paratopo são identificados utilizando o esquema de numeração Kabat. Também são apresentadas a estrutura secundária do epitopo, a contagem de interações por resíduo e a área da superfície acessível ao solvente de cada resíduo do epitopo e do paratopo no complexo ou fora dele, calculada através de diagramas de Voronoi (KULKARNI-KALE et al., 2014). Apesar dos recursos disponíveis no AgAbDb, de outubro de 2017 até junho de 2019, não houve alteração no número de complexos depositados, indicando que este banco de dados não tem sofrido atualizações.

O AppA (NGUYEN; VERMA; ZHONG, 2019) é uma plataforma *web* para análise, comparação e visualização de resíduos de complexos antígeno-anticorpo. A partir dos complexos disponibilizados no PDB, o AppA identifica os resíduos que fazem parte da interface e calcula as possíveis interações realizadas entre eles. Uma limitação desse servidor é a análise de somente um registro do PDB por vez. No entanto, o Appa informa para cada resíduo da interface, sua estrutura secundária, o grau de exposição ao solvente (enterrado, intermediário ou exposto) e o número de interações realizadas, além de exibir uma representação da estrutura do complexo. A plataforma ainda permite a comparação de 2 complexos, sobrepondo suas interfaces. Apesar dos recursos disponibilizados e de sua atualização semanal, o modo

de disponibilização dos dados do AppA impede a sua utilização para avaliar um grande número de complexos antígeno-anticorpo.

O AbDb (FERDOUS; MARTIN, 2018) é uma coleção de estruturas de regiões variáveis de anticorpos obtidas a partir do SACS (ALLCORN; MARTIN, 2002), uma base de dados periodicamente atualizada com as estruturas de anticorpos disponíveis no PDB. O AbDb disponibiliza arquivos de estruturas das regiões variáveis de cada anticorpo (descartando a região constante e dividindo múltiplos anticorpos em vários arquivos) no formato do PDB e a numeração dos domínios variáveis das cadeias utilizando os esquemas de numeração de Kabat (WU; KABAT, 1970), Chothia (CHOTHIA; LESK, 1987) e Martin (ABHINANDAN; MARTIN, 2008). Com o uso da interface *web* é possível buscar os registros na base de dados informando o identificador do PDB, a espécie do organismo produtor do anticorpo ou do antígeno, ou ainda informando o nome do anticorpo ou do antígeno. O AbDb também fornece a possibilidade de o usuário baixar alguns conjuntos de dados, como estruturas de anticorpos em sua forma livre, em complexo com proteínas ou em complexo com antígenos não proteicos. É possível também obter uma lista de anticorpos idênticos, em complexo ou em sua forma livre. Apesar de apresentar dados atualizados das estruturas do PDB, o AgDb não fornece informações de contatos entre antígenos e anticorpos.

O SAbDab é um banco de dados atualizado semanalmente com anotações de estruturas de anticorpos presentes no PDB. Para cada estrutura do PDB, são anotados os pares de cadeias leve e pesada de cada anticorpo e a(s) cadeia(s) do antígeno, caso a estrutura seja de um complexo antígeno-anticorpo. Além disso, o SAbDab apresenta informações sobre os organismos relacionados ao antígeno e ao anticorpo, o tipo de cadeia leve, a publicação científica relacionada à estrutura e os dados de afinidade de ligação do complexo antígeno-anticorpo, quando disponíveis (DUNBAR et al., 2014). O SAbDab ainda aplica o esquema de numeração Chothia e apresenta, para cada anticorpo, a porção variável numerada

assim como as sequências dos CDRs. Além disso, quando a estrutura do PDB apresenta um anticorpo com o par de cadeias leve e pesada, são apresentados os dados referentes à orientação das cadeias para a região variável do anticorpo. O SAbDab permite diferentes modos de seleção de resultados: exibir todas as estruturas disponíveis, buscar por uma estrutura especificada por seu identificador ou realizar uma busca avançada por critérios específicos. Esses critérios podem ser, por exemplo, o organismo produtor do anticorpo, o tipo de anticorpo (Fab, scFv, VH ou VL), o tipo de cadeia leve, a forma do anticorpo na estrutura (livre ou em complexo), os dados de afinidade (se forem conhecidos) ou se o anticorpo possui resíduos específicos em determinadas posições. Além disso, o SAbDab tem uma opção de busca não redundante, em que a identidade das sequências é analisada e são exibidos somente os resultados em que a identidade não ultrapassa um valor de *cutoff* selecionado pelo usuário. Para filtrar os registros considerando as identidades, isto é, gerar uma análise sem estruturas redundantes, o SAbDab utiliza o programa cd-hit (LI; GODZIK, 2006) que realiza o agrupamento (*clustering*) das sequências. Nos grupos (*clusters*) gerados, é garantido que não existam sequências que tenham valor de identidade superior a um *cutoff*. A partir de cada um desses grupos é extraída uma sequência representativa do grupo que fará parte do conjunto de sequências não redundantes.

Quando uma busca é realizada no SAbDab, além da exibição de informações na tela e seu possível detalhamento através de *links*, é disponibilizado um arquivo com os resultados dessa busca para *download*, permitindo que esses dados sejam facilmente utilizados em outras implementações. Por extrair os dados de maneira automática, algumas falhas de anotação são encontradas no SAbDab, como por exemplo, o organismo produtor do anticorpo identificado como um vírus. No entanto, o SAbDab disponibiliza um *hyperlink* para que os usuários possam informar problemas de anotação e melhorar a qualidade do banco de dados. Apesar das diversas informações sobre as estruturas do PDB que contêm anticorpos, o SAbDab não apresenta informações sobre as interações entre antígeno e anticorpo.

Cada um dos bancos de dados apresentados nesta seção possui vantagens e desvantagens quando comparados aos demais. Mas, apesar da disponibilidade de diversos bancos de dados de estrutura de anticorpos existentes no PDB, nenhum deles permite a análise das possíveis interações existente em cada complexo antígeno-anticorpo, alinhadas às regiões de *framework* e CDRs das cadeias dos anticorpos. Portanto, para a realização de análises desse tipo, são necessárias buscas em mais de um banco e processamento de resultados em diferentes formatos. Além disso, como a cada dia o número de estruturas depositadas no PDB cresce, é necessário um banco de dados que seja frequentemente atualizado, assim, o banco de dados pode conter informações das descobertas recentes.

1.6.2 Bancos de dados com informações de sequenciamento de repertórios de receptores de células B

Com o sequenciamento em larga escala, um grande conjunto de dados de repertório de anticorpos tem sido gerado. No entanto, a falta de uma padronização na anotação e armazenamento desses dados dificulta a utilização das informações geradas pelos diversos pesquisadores da área. Na tentativa de superar esse problema, a comunidade AIRR tenta desenvolver mecanismos para o compartilhamento de dados de sequenciamento de repertório armazenados em múltiplos repositórios (CORRIE et al., 2018; RUBELT et al., 2017). No entanto, apesar desses esforços, grande parte dos trabalhos de sequenciamento de repertório depositam somente os dados brutos obtidos (*raw reads*) em repositórios como o *Sequence Read Archive* (SRA) (LEINONEN et al., 2011), sem qualquer tipo de anotação em relação às sequências de aminoácidos, codificadas pela montagem das *reads* de nucleotídeos, ou anotações dos CDRs. Com o objetivo de minimizar o trabalho de se executar um conjunto de softwares para a geração das informações a partir de dados brutos de sequenciamento de repertórios, foram criados alguns repositórios como o *Pan Immune Repertoire Database* (PIRD) (ZHANG et al., 2018) e o *Observed Antibody Space* (OAS) (KOVALTSUK et al., 2018), para armazenamento de dados de anotações de sequenciamento de repertório.

O *Pan Immune Repertoire Database* (PIRD) (ZHANG et al., 2018) é um banco de dados criado com o objetivo de coletar e armazenar informações de sequenciamento de repertório de receptores de células B (BCRs) e T (TCRs - *T Cell Receptors*). Além das informações obtidas a partir do sequenciamento, também são armazenadas informações do experimento como: identificação do projeto, tamanho da amostra, tipo do receptor, espécie produtora, artigos publicados, tecido de origem, dados de estudos longitudinais ou pré/pós-imune, etc. Além disso, esse banco armazena diversas informações estatísticas do repertório, como o número total de sequências, o número de sequências únicas totais e de CDR3 (tanto de nucleotídeos quanto de aminoácidos), a utilização de genes V e J, a distribuição de tamanho de CDR3, o pareamento de genes V e J, os 10 clones mais frequentes, etc. Outro recurso oferecido por esse banco é a visualização dessas estatísticas utilizando gráficos, por exemplo, *heatmaps* e histogramas. A busca de dados nesse banco pode ser realizada por informações do projeto ou das amostras, como o tipo de receptor analisado ou espécie produtora, ou ainda os resultados de um alinhamento local (BLAST - *Basic Local Alignment Search Tool*) de uma sequência informada pelo usuário contra a base total das sequências armazenadas.

O *Observed Antibody Space* (OAS) (KOVALTSUK et al., 2018) é um repositório que reúne dados de sequenciamento de repertório de receptores de células B. Esse repositório é formado pela coleta de dados brutos de experimentos armazenados em bases de dados de sequenciamento, como o SRA, e o armazenamento, de maneira padronizada, das informações relativas ao experimento, como, por exemplo, o tipo de célula B, a fonte dessas células e o organismo produtor. Além disso, nesse repositório foram incluídas, após o processamento das sequências brutas obtidas, as sequências de aminoácidos que codificam cada receptor e a numeração delas segundo o esquema de numeração do IMGT (LEFRANC et al., 2003).

Considerando a quantidade de dados gerados em cada sequenciamento de repertório e o número de experimentos atualmente realizados, somado com o esforço da comunidade AIRR e dos bancos de dados apresentados, pode-se esperar um grande número de dados padronizados nos próximos anos que precisarão de ferramentas para auxiliar nas suas análises e visualização.

No entanto, apesar das várias descobertas realizadas nos estudos de sequenciamento de repertório, é necessário alinhar o conhecimento gerado pelas sequências descobertas com as respectivas estruturas dos anticorpos (KOVALTSUK et al., 2017). Como a cristalização e a resolução de uma estrutura é demorada e consome muitos recursos, a modelagem de anticorpos surge como uma opção para, a partir das sequências obtidas pelo sequenciamento de anticorpos, obter-se mais informações das possíveis conformações adotadas pelas sequências e das interações realizadas com os antígenos (DEKOSKY et al., 2016; KRAWCZYK et al., 2018).

1.7 Análise de estruturas de anticorpos

Em busca de elucidar os mecanismos de afinidade e especificidade das interações antígeno-anticorpo, diversos trabalhos foram desenvolvidos para analisar as estruturas desses complexos. Esses trabalhos, iniciados a partir da análise das estruturas obtidas com técnicas de cristalografia, busca encontrar características comuns aos complexos, ou ao menos a grupos de complexos antígeno-anticorpo. Ao longo dos anos, novas estruturas de complexos foram resolvidas e novas análises realizadas, o que confirmou as inferências feitas a partir de grupos menores de estruturas ou que modificou tais inferências. Os trabalhos iniciais, que datam das décadas de 80 e 90, exploraram um conjunto reduzido de complexos (de 3 a 19 complexos) (DAVIES et al., 1990; LO CONTE; JANIN; CHOTHIA, 1999; MACCALLUM; MARTIN; THORNTON, 1996; WEBSTER; HENRY; REES, 1994; WILSON; STANFIELD, 1993).

Uma das características comumente analisada em complexos antígeno-anticorpo é a preferência de aminoácidos nas interfaces do complexo (DALKAS et al., 2014; GONZÁLEZ-MUÑOZ et al., 2012;

KRINGELUM et al., 2013; KUNIK; OFRAN, 2013; NGUYEN et al., 2017; PENG et al., 2014; RAMARAJ et al., 2012; STAVE; LINDPAINTNER, 2013). Essa análise é interessante, pois o enriquecimento de certos tipos de aminoácidos nas interfaces podem revelar interações energeticamente favoráveis que determinam a afinidade e especificidade das interações (PENG et al., 2014). Além disso, as informações obtidas nesses trabalhos são utilizadas para guiar os processos de engenharia de anticorpos (GONZÁLEZ-MUÑOZ et al., 2012). No entanto, existem algumas diferenças entre estes trabalhos, o que faz com que nem sempre suas conclusões estejam totalmente de acordo umas com as outras (KUNIK; OFRAN, 2013). A Tabela 1.1 apresenta alguns dos trabalhos que realizaram a análise da preferência de aminoácidos nas interfaces. Como é possível observar nessa tabela, o número de estruturas utilizadas em cada trabalho é variável. Isso se deve ao número de estruturas disponíveis no momento da realização dos estudos, o método utilizado para obter as estruturas que contêm anticorpos (como busca direta no PDB ou utilização de bancos de dados como os apresentados na seção 1.6.1) e o procedimento para eliminar a redundância do conjunto de dados. Essa última ação é necessária para que as análises não tenham um viés causado por anticorpos ou antígenos muito estudados ou que são utilizados como ferramentas de pesquisa em diversos contextos. Durante o processo de remoção de redundância, é necessário escolher, dentre cada grupo de estruturas semelhantes, uma estrutura representativa do grupo. Na Tabela 1.1 também estão apresentados os diferentes critérios de remoção de redundância e de escolha de estrutura significativa dos trabalhos analisados. A escolha da estrutura representativa pode ser baseada em diferentes critérios, como a resolução do complexo, o tamanho da interface ou definida aleatoriamente. Novamente, escolhas diferentes podem levar a resultados variados.

Outro critério que é consideravelmente variável entre os trabalhos é a forma utilizada para definir os resíduos do antígeno e do anticorpo que fazem parte da interface do complexo. As definições mais usuais são baseadas em critérios de distância entre os resíduos (ou seus átomos) e a área do resíduo

Tabela 1.1: Trabalhos que analisam a preferência de aminoácidos nas interfaces dos complexos antígeno-anticorpo.

Referência	Número inicial de estruturas, banco de dados e restrições	Remoção de redundância (critério e ferramenta)	Número final de estruturas	Definição de estrutura representativa	Definição de interface
(NGUYEN et al., 2017)	? SabDab	Sequências com identidade máxima de 99% (cd-hit)	403	Sequência mais longa	Distância entre átomos ≤ 6Å Variação na área de acessibilidade ao solvente
(DALKAS et al., 2014)	? IEDB-3D Resolução < 2,5 Å Anticorpo com cadeia leve e pesada	Sequências de CDR com identidade máxima de 70% (ClustalW)	105	Melhor resolução	Área de acessibilidade ao solvente (em relação a GXG) > 5%
(PENG et al., 2014)	? Antígenos com ao menos 85 aminoácidos Resolução ≤ 3,8 Å	?	110 (111 anticorpos)	Escolha manual (STAVE; LINDPAINTNER, 2013)	Variação na área de acessibilidade ao solvente (dividido por área no anticorpo livre) > 0,2 ou > 0
(KRINGELUM et al., 2013)	801 estruturas com anticorpo Antígenos com ao menos 20 aminoácidos	Entradas duplicadas ou mutações de um antígeno Interfaces similares (vetores de interações com ângulo < 0,8 rad)	107	?	Distância entre átomos ≤ 4Å
(KUNIK; OFRAN, 2013)	784 estruturas reduzidas a 352 Anticorpo com cadeia leve e pesada Antígeno proteico	Sequências com identidade < 97% e cobertura < 95% (Blastclust)	200	Similaridade das interfaces e maior número de interações	Predição via Paratome, baseado em alinhamento estrutural Distância entre átomos ≤ 6Å
(STAVE; LINDPAINTNER, 2013)	? Antígenos com ao menos 85 aminoácidos Resolução ≤ 3,8 Å	?	110 (111 anticorpos)	Escolha manual igual ao de (PENG et al., 2014)	Distância entre átomos ≤ 4Å
(RAMARAJ et al., 2012)	101	?	53	?	Área de acessibilidade ao solvente quando desassociado > 50Å^2 Distância entre átomos ≤ 5Å

O símbolo de interrogação (?) representa dados que não puderam ser obtidos a partir do trabalho em questão.

acessível ao solvente, ou a variação dessa acessibilidade quando as moléculas estão em complexo ou em sua forma livre. Além disso, considerando-se um mesmo critério, diferentes valores de *cuttoff* podem ser escolhidos. Por exemplo, ao se escolher como critério de definição de interface a distância entre os átomos dos resíduos, é necessário escolher também o valor máximo de distância que será aceito para que um par de resíduos faça parte da interface. Nos trabalhos apresentados na Tabela 1.1, quando utilizado o critério de distância máxima, observa-se valores de 4 a 6Å. Quando o critério utilizado é a área de acessibilidade ao solvente, é possível variar o critério de definição de área de acessibilidade, a métrica utilizada para o cálculo da variação (por exemplo, diferença absoluta ou proporcional à área exposta do resíduo ou a área total do resíduo) e ainda o valor utilizado como corte para se considerar que existe uma variação significativa na área de acessibilidade ao solvente. Além disso, diferentes critérios podem ser combinados para definir os resíduos da interface (NGUYEN et al., 2017; RAMARAJ et al., 2012). Apesar das diferenças nos dados e métodos utilizados para a definição da interface de complexos antígeno-anticorpo, é possível destacar algumas conclusões obtidas nessas análises, apresentadas a seguir.

Os epitopos podem ser divididos em duas categorias: lineares ou conformacionais. Em epitopos lineares, a ligação do anticorpo ocorre apenas em uma sequência contínua de resíduos de aminoácidos do antígeno. Por outro lado, em epitopos conformacionais, os resíduos estão próximos através do enovelamento do antígeno e não necessariamente próximo em sua estrutura primária (FORSSTRÖM et al., 2015). Epitopos têm em média 15±4 resíduos (KRINGELUM et al., 2013), sendo que epitopos lineares e conformacionais têm em média nove e dezoito resíduos, respectivamente (DALKAS et al., 2014). Em geral, os epitopos são ricos em triptofano, tirosina, lisina, arginina, histidina, ácido glutâmico, ácido aspártico, glutamina (DALKAS et al., 2014; RAMARAJ et al., 2012) e asparagina (DALKAS et al., 2014). Aminoácidos menos frequentes nos epitopos são a leucina, a alanina, a cisteína, a isoleucina e a valina (DALKAS et al., 2014). Apesar da diferença entre as frequências de cada aminoácido nos epitopos,

comparando-se a frequência dos resíduos do epitopo com o restante da superfície do antígeno, não é possível distinguir a região de interface das demais regiões expostas ao solvente (KRINGELUM et al., 2013; PENG et al., 2014). Outra característica conhecida dos epitopos é serem, geralmente, compostos por alças (*loops*) (OFRAN; SCHLESSINGER; ROST, 2008; PENG et al., 2014).

O paratopo é rico em tirosina e triptofano (KRINGELUM et al., 2013; KUNIK; OFRAN, 2013; MIAN; BRADWELL; OLSON, 1991; NGUYEN et al., 2017; PENG et al., 2014; RAMARAJ et al., 2012), com ênfase na tirosina (DAVIES; COHEN, 1996). Além destes aminoácidos, são também frequentes a serina (DALKAS et al., 2014; KRINGELUM et al., 2013; KUNIK; OFRAN, 2013; NGUYEN et al., 2017; RAMARAJ et al., 2012), a histidina (CHEN; VAN REGENMORTEL; PELLEQUER, 2009; NGUYEN et al., 2017; RAMARAJ et al., 2012), a asparagina (KUNIK; OFRAN, 2013), a fenilalanina (PENG et al., 2014; RAMARAJ et al., 2012), a isoleucina (RAMARAJ et al., 2012) e o ácido aspártico (DALKAS et al., 2014). Além disso, são menos frequentemente encontrados no paratopo a prolina, a lisina, o ácido glutâmico (DALKAS et al., 2014; PENG et al., 2014), mas também a histidina, glutamina, arginina, alanina, leucina, isoleucina, valina, metionina, prolina e cisteína (PENG et al., 2014). Como a definição de enriquecimento ou depleção de aminoácidos na interface é variável, devido ao conjunto analisado e ao outro conjunto de comparação (interfaces de proteínas em geral ou superfície das moléculas, por exemplo), alguns resultados dos trabalhos analisados podem ser diferentes, como o caso da diferença da frequência da histidina e da isoleucina.

Apesar dos diversos estudos conduzidos para caracterizar a composição de aminoácidos das interfaces de complexos antígeno-anticorpo, nem sempre a frequência dos aminoácidos revela a importância dos mesmos na ligação entre as moléculas (KUNIK; OFRAN, 2013). Por causa disso, outra característica importante na análise dos complexos antígeno-anticorpo é o conjunto de interações que existem na interface. No entanto, os tipos de interações analisadas e suas definições são diferentes entre

as pesquisas desenvolvidas. A Tabela 1.2 apresenta as definições de interface utilizadas em três trabalhos que analisam as interações entre antígenos e anticorpos.

Nguyen e colaboradores (NGUYEN et al., 2017) perceberam o importante papel da tirosina nas interações entre o paratopo e o epitopo, sendo o resíduo mais frequente na interface e também o que possui maior número de interações de hidrogênio e de Van der Waals, além de contatos hidrofóbicos. Além da tirosina, a glicina realiza um número significativo de interações de hidrogênio e de Van der Waals. O número de interações iônicas é menor que as interações de hidrogênio e de Van der Waals, sendo a arginina e o ácido aspártico os resíduos que mais fazem este tipo de interação. Além das interações comumente analisadas em complexos, Nguyen e colaboradores analisaram as moléculas de água presentes na interface e perceberam que elas mediam um número significante de interações entre o antígeno e o anticorpo. Foram encontradas moléculas de água na interface ligadas ao paratopo e ao epitopo através de interações de hidrogênio em mais da metade dos complexos analisados. Os resíduos de tirosina, serina, ácido aspártico e asparagina são os que mais fazem interações com essas moléculas de água.

Dalkas e colaboradores (DALKAS et al., 2014) destacaram que as interações mais frequentes na interface de complexo antígeno-anticorpo são as interações de hidrogênio e os contatos hidrofóbicos. As pontes salinas são, na maioria dos complexos antígeno-anticorpo, formadas por um resíduo carregado positivamente no antígeno e um negativamente no anticorpo. Por outro lado, as interações de hidrogênio são preferencialmente realizadas pelos resíduos de tirosina, serina, treonina e asparagina, do paratopo, com uma predominância da tirosina. Além disso, os autores perceberam que os contatos hidrofóbicos são mais frequentes em epitopo lineares, enquanto em epitopos conformacionais, as interações de hidrogênio, cátion/amino-π e empilhamento π são mais frequentes. Independentemente do tipo de epitopo analisado, a tirosina é o resíduo que mais realiza contatos hidrofóbicos com o epitopo. Esse resíduo também é preferido nas interações de empilhamento π. Dalkas e colaboradores ainda destacam o maior número de

Tabela 1.2: Trabalhos que analisam as interações entre antígeno-anticorpo

Referência	**(NGUYEN et al., 2017)**	**(BURKOVITZ; OFRAN, 2016)**	**(DALKAS et al., 2014)***
Interações de hidrogênio[1]	a(D-H-A) > 120º d(D-A) < 3.5Å	a(D-H-A) ≥ 90º d(D-A) ≤ 3.9Å d(H-A) ≤ 2.5Å	d(D-A) < 3.89Å OU 90º < a(D-H-A) < 270º d(A-H) ≤ 4Å
Contatos Hidrofóbicos	d(átomos da CL de resíduos apolares[2]) < 5Å	-	d(átomos dos resíduos apolares[2]) < 5Å
Interações de Van der Waals	d(átomos) < rvdw(átomo1) + rvdw(átomo2) + 0,5Å	-	-
Interações iônicas / pontes salinas	Interações iônicas: d(átomos da CL de um aa básico e de um aa ácido[3]) < 6Å	Pontes salinas: d(O da CL da Asp ou Glu e N da CL da Arg, Lys ou His) ≤ 4Å	Pontes salinas: não indica os átomos d(átomos) ≤ 4Å
Interações cátion-π, amino-π, empilhamento π	-	*Empilhamento π*[4]*:* d(centroides dos anéis π) ≤ 8Å d(ao menos um átomo de cada anel) ≤ 4,5Å 0 ≤ a(plano de um anel e reta formada pelos centroides) ≤ 60º 0 ≤ a(planos dos anéis) ≤ 30º *Cátion-π*[4]*:* d(cátion da CL da Lys ou Arg e centroide do anel π) ≤ 7Å d(cátion da CL da Lys ou Arg e o pano do anel) ≤ 6Å a(reta, formada pelo cátion da CL da Lys ou Arg e o centroide do anel, e plano do anel) ≤ 45º	*Empilhamento π*[4]: d(átomo de um anel π e um átomo do outro anel) ≤ 5Å Um átomo de um anel deve estar contido em um cilindro (h =5Å, r = 3 x r(outro anel)) com base no outro anel *Amino-π*[4]*:* d(átomo do anel e átomo do grupo amino da Glu ou Asn) ≤ 4,5Å O átomo deve estar contido um cilindro (h =4,5Å, r = 2 x r(anel)) com base no anel *Cátion-π*[4]*:* d(átomo do anel e átomo do grupo positivo da Arg, Lys ou His) ≤ 4,5Å O átomo deve estar contido um cilindro (h =4,5Å, r = 2 x r(anel)) com base no anel
Moléculas de água na interface	Molécula de água que faz interação de hidrogênio com um resíduo do epitopo e um do paratopo	-	-

Abreviações: a: ângulo, d: distância, rvdw: raio de Van der Waals, CL: cadeia lateral, aa: aminoácido, r: raio
[1]Átomos doadores (D): NE, NH1 e NH2 da Arg; ND2 da Asn; NE2 da Gln; ND1 e NE2 da His; NZ da Lys; OG da Ser; OG1 da Thr; NE1 do Trp; OH da Tyr. Átomos aceptores: OD1 da Asn; OD1 e OD2 do Asp; OE1 da Gln; OE1 e OE2 do Glu; ND1 e NE2 da His; OG da Ser; OG1 da Thr; OH da Tyr.
[2] Resíduos apolares: Ala, Ile, Leu, Met, Phe, Pro, Tyr, Trp, Val.
[3] Átomos carregados de aminoácidos básicos (CZ, NE, NH1, NH2 da Arg; CD2, CE1, CG, ND1, NE2 da His; NZ da Lys) e de aminoácidos ácidos (CG, OD1, OD2 da Asp; CD, OE1, OE2 do Glu).
[4] Anel π: anel aromático dos aminoácidos Trp, Tyr, Phe e His
* Como os empilhamentos π entre os resíduos Tyr, Trp e Phe também são identificados como contatos hidrofóbicos, estas interações somente foram contabilizadas no primeiro grupo.

interações cátion/amino-π quando se compara complexos antígeno-anticorpo com outras interfaces entre proteínas, sendo que o número de interações do tipo cátion-π é duas vezes mais frequente que as do tipo amino-π nestes complexos. Novamente, a tirosina é o resíduo do paratopo mais frequente nessas interações. Devido à ocorrência da tirosina em todos os tipos de interação analisadas, exceto nas pontes salinas, e que, consequentemente ela pode realizar diversos tipos de interação, os autores concluíram que ela desempenha um papel importante no reconhecimento e ligação dos anticorpos aos antígenos (DALKAS et al., 2014).

Diferentemente dos trabalhos anteriormente descritos, Peng e colaboradores (PENG et al., 2014) analisaram a interação de antígenos e anticorpos diferenciando os átomos que compõem os resíduos. Eles concluíram, assim como os demais trabalhos, que o paratopo é rico em resíduos aromáticos, principalmente a tirosina. Os autores também concluíram que estes resíduos são circundados por cadeias laterais hidrofílicas, principalmente de ácido aspártico, asparagina, serina, treonina e glicina. Em especial, as cadeias laterais dos resíduos do parotopo são ricas em átomos doadores e aceptores de hidrogênio. Normalmente, as cadeias laterais aromáticas do paratopo interagem, principalmente, com átomos da cadeia principal do antígeno e os carbonos alifáticos das cadeias laterais. Além disso, as cadeias laterais hidrofílicas podem contribuir substancialmente para a especificidade dos anticorpos através das interações de hidrogênio realizada com o epitopo (PENG et al., 2014).

Apesar do conhecimento gerado pelos trabalhos que analisaram as estruturas de complexos antígeno-anticorpo e sua utilização nas áreas de modelagem e engenharia de anticorpos, ainda não são claras as regras que governam o reconhecimento entre antígeno e anticorpo e a especificidade dos anticorpos (MARILLET et al., 2017; PENG et al., 2014; SELA-CULANG; KUNIK; OFRAN, 2013). Além disso, como ressaltado por Burkovitz e Ofran (BURKOVITZ; OFRAN, 2016), o desenvolvimento

de ferramentas para a análise de complexos antígeno-anticorpo ainda está no início de seu desenvolvimento.

Por essas razões, neste trabalho é proposto o desenvolvimento de ferramentas para análise de estruturas de complexos antígeno-anticorpos e para visualização de sequências de anticorpos em larga escala a fim de identificar padrões que gerem um melhor entendimento dos princípios que determinam a afinidade e especificidade dos anticorpos. Estas análises podem contribuir gerando o conhecimento de como mudanças em alguns aminoácidos podem influenciar na estrutura e interação dos anticorpos (KRAWCZYK; DUNBAR; DEANE, 2017). Os padrões identificados podem facilitar o entendimento da imunidade humoral e o desenvolvimento de terapias baseadas em anticorpos (PENG et al., 2014). Além disso, esses padrões podem guiar o processo de engenharia de anticorpos tanto para uso em terapias como também para diagnósticos e pesquisa. Estas informações podem ainda ser utilizadas na geração de bibliotecas sintéticas exploradas por técnicas de *display* e na modelagem de anticorpos (FINLAY; ALMAGRO, 2012; HONEGGER; PLÜCKTHUN; PLUCKTHUN, 2001). Essa análise inclui todas as interações comumente analisadas em complexos antígeno-anticorpo e as diversas estruturas que foram depositadas posteriormente à realização dos demais trabalhos citados anteriormente. Por fim, as ferramentas de análise aqui propostas são semanalmente atualizadas semiautomaticamente, (com intervenção de um curador, se necessário) permitindo, também, a análise de estruturas depositadas no futuro.

2 Objetivos

O objetivo geral deste trabalho é desenvolver estratégias de análise e visualização em larga escala das cadeias dos anticorpos e suas interações com o antígeno que permitam uma melhor compreensão do processo de reconhecimento e ligação de anticorpos a antígenos.

Os objetivos específicos deste trabalho são:

- Definir um conjunto de dados de estruturas de anticorpos e seus complexos com os antígenos que permitam a análise das cadeias de domínios variáveis dos anticorpos e das interações com os antígenos;
- Propor um método de visualização em larga escala de alinhamento de múltiplas sequências de domínios variáveis de anticorpos;
- Implementar uma plataforma que disponibilize ao público em geral a visualização proposta e permita a análise de sequências de domínios variáveis de anticorpos;
- Extrair propriedades estruturais e físico-químicas e calcular as interações moleculares existentes nas estruturas de complexos de antígeno-anticorpo;
- Implementar um banco de dados para armazenar as propriedades e interações calculadas ou extraídas;
- Analisar os dados gerados em busca de uma melhor compreensão do processo de reconhecimento e ligação de anticorpos a antígenos.

3 Métodos

A fim de alcançar os objetivos deste trabalho as seguintes etapas foram desenvolvidas:

a) Proposta de uma representação de visualização do alinhamento de domínios variáveis de cadeias de anticorpos em larga escala;

b) Implementação de uma plataforma *web* que disponibiliza a visualização proposta e permite a análise de um conjunto de sequências de anticorpos obtido a partir de diversas fontes de dados;

c) Extração de dados de estruturas de complexos antígeno-anticorpo disponíveis no PDB (BERMAN, 2000);

d) Cálculo e armazenamento de propriedades das estruturas e das interações dos complexos obtidos na etapa anterior;

e) Análise dos dados obtidos na etapa anterior.

A representação obtida a partir da execução da etapa (a), chamada *Collier de Diamants*, será apresentada na seção 4.1 deste texto. Na seção 4.2.2, são apresentados os materiais e métodos utilizados na realização da etapa (b) que resultou na plataforma Yvis disponível em http://bioinfo.icb.ufmg.br/yvis/. Os materiais e métodos necessários à execução das etapas (c) e (d) são descritos na seção 5.2. Essas etapas geraram o banco de dados Ydb apresentado no capítulo 5. Para a execução da etapa (e), foram utilizados a representação *Collier de Diamants*, a plataforma Yvis e o banco de dados Ydb, além de *scripts* implementados para o processamento dos dados como apresentados na seção 5.4.1.

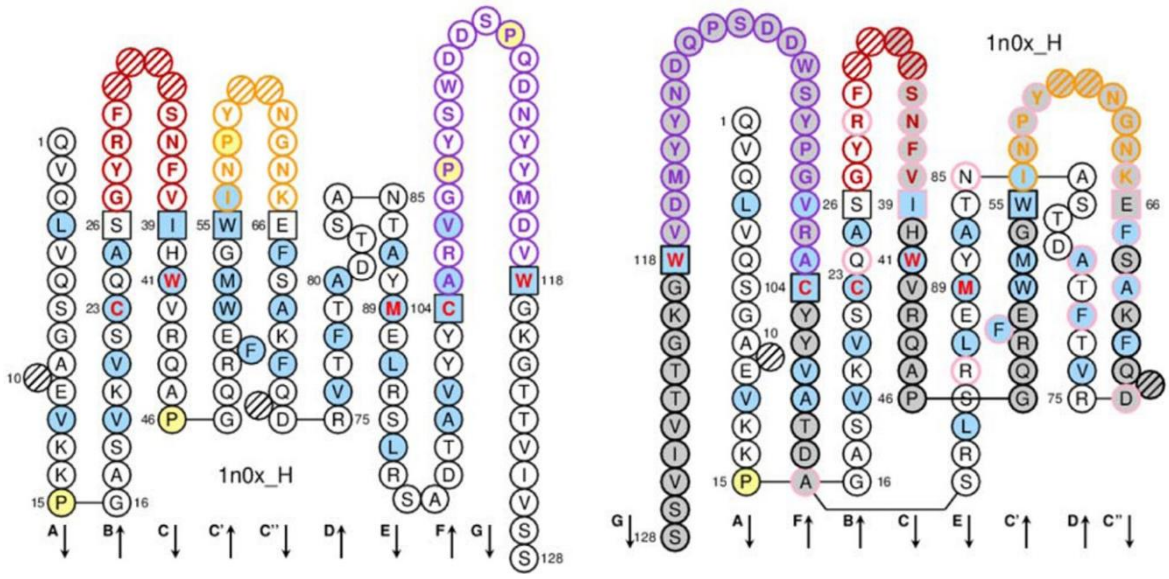
4 Visualização de alinhamento de sequências de anticorpos em larga escala

Como descrito na seção 1.1, anticorpos possuem regiões variáveis com estrutura altamente conservada, apesar das diferenças em suas sequências de aminoácidos. O domínio variável de cada cadeia leve e pesada de um anticorpo é composto de 9 fitas beta (A, B, C, C', C'', D, E, F e G) ligadas por 5 voltas beta e 3 *loops*, formando um sanduíche de duas folhas (ABED e GFCC'C'') empacotadas uma contra outra por meio de interações hidrofóbicas. Estas folhas mantêm-se unidas por uma ponte dissulfeto entre duas cisteínas conservadas, uma em cada folha (LEFRANC, 2014). Essa estrutura conservada permitiu ao IMGT definir, a partir de seu sistema de numeração, descrito na seção 1.2 (LEFRANC, 2014; LEFRANC et al., 2003), representações em duas dimensões dos domínios de anticorpos e receptores de células T, conhecida como IMGT *Collier de Perles* (KAAS; LEFRANC, 2007; RUIZ; LEFRANC, 2002).

A Figura 4.1 apresenta as duas representações do domínio variável de uma cadeia de um anticorpo seja ela leve ou pesada. Como o IMGT *Collier de Perles* é baseado no esquema de numeração do IMGT, a sequência da cadeia representada contém *gaps* inseridos pela aplicação desse esquema e, por meio dele, é possível identificar as regiões de *framework* e CDRs. A representação da Figura 4.1a corresponde à sequência do domínio variável da cadeia, representando suas fitas beta, na mesma ordem da estrutura primária da cadeia. Essa representação é conhecida como *Collier de Perles* apresentada em uma camada. Os CDRs são destacados nas cores vermelho, amarelo e roxo e são delimitados pelos resíduos das regiões de *framework* representadas nos quadrados.

Os resíduos conservados que determinam as posições de cada um dos resíduos no esquema de numeração e, consequentemente, no *Collier de Perles* são destacados com fonte vermelha. Além disso, posições hachuradas representam os *gaps* da sequência em relação ao esquema de numeração. A Figura 4.1b apresenta as mesmas informações, no entanto, as fitas betas são rearranjadas para que fitas próximas

na estrutura tridimensional estejam também próximas nessa representação, chamada *Collier de Perles* apresentada em 2 camadas.



(a) (b)

Figura 4.1: Representações do domínio variável da cadeia pesada (H) do anticorpo da estrutura 1N0X do PDB. (A) *Collier de Perles* em 1 camada. (B) *Collier de Perles* em 2 camadas. Figura adaptada de (LEFRANC, 2014).

Apesar de ser uma representação útil para a visualização da sequência de um anticorpo relacionando-a com sua estrutura tridimensional, a representação do *Collier de Perles* somente permite a análise de sequências individualmente, não permitindo a inspeção de diversas sequências em conjunto. A maneira que existe para comparar diversas sequências é utilizar uma das visualizações de alinhamento de múltiplas sequências, no entanto, a relação com a estrutura tridimensional do anticorpo é perdida.

A visualização do alinhamento de múltiplas sequências, geralmente, é feita por meio de uma matriz em que cada linha representa uma sequência e cada coluna uma posição do alinhamento. Cada célula dessa matriz geralmente contém um resíduo da sequência e cores são utilizadas para classificar estes resíduos ou posições. Essa classificação pode estar associada a propriedades físico-químicas dos resíduos ou a valores quantitativos que, por exemplo, representam hidrofobicidade ou exposição ao solvente (PROCTER et al., 2010). A utilização da classificação através de cores permite ao usuário analisar as

tendências para cada posição ou região de grupos de sequências, além de sequências ou posições divergentes. Existem outras maneiras de apresentar padrões a partir de um alinhamento de múltiplas sequências, como o WebLogo (CROOKS et al., 2004). No entanto, não foram encontradas ferramentas que permitam a análise de um grupo de sequências de anticorpos além de ferramentas tradicionais de visualização de alinhamento de múltiplas sequências.

4.1 Proposta de representação de atributos de múltiplas cadeias de anticorpos

Com a crescente disponibilidade de dados de anticorpos, seja através de novos anticorpos com estruturas resolvidas e disponibilizadas no PDB, seja através do sequenciamento do repertório de receptores de células B (Rep-Seq), surge a necessidade de se desenvolver meios para a análise de um grande conjunto de dados. Um dos gargalos nesse processo é a visualização do alinhamento de múltiplas sequências de cadeias de anticorpos acompanhado de informações da estrutura dessas cadeias. Para suprir essa necessidade, neste trabalho foi proposta uma representação de atributos de sequências de domínios variáveis de cadeias de anticorpos baseada na representação do IMGT/*Collier de Perles* (Colar de pérolas). Nessa proposta, cada uma das posições do *Collier de Perles* apresenta um gráfico relativo a algum atributo dos resíduos das sequências alinhadas segundo o esquema de numeração do IMGT. Exemplos desses atributos são: classificação segundo propriedades físico-químicas dos resíduos, como classificação de resíduos segundo sua cadeia lateral, hidrofobicidade, estrutura secundária ou ainda tipos e número de interações realizadas com o antígeno. Cada gráfico de posição exibe a proporção dos resíduos das cadeias analisadas, segundo a frequência deles, semelhante a um gráfico de setores ou pizza. Os setores correspondem a classes possíveis para o atributo analisado. Como cada "pérola" do "colar de pérolas" foi substituída por uma representação com múltiplas faces, a representação proposta foi chamada de *Collier de Diamants* (Colar de Diamantes).

A Figura 4.2 apresenta a representação proposta para o alinhamento de 113 sequências de domínios variáveis de cadeias de anticorpos tendo como atributo analisado as propriedades físico-químicas dos aminoácidos classificados como apresentado na Tabela 4.1 (CROOKS et al., 2004). Os números do lado esquerdo de cada gráfico de pizza indicam a posição por ele representada. No esquema de numeração do IMGT, quando o CDR3 possui mais de 13 resíduos, as inserções são colocadas entre as posições 111 e 112 e são identificadas por uma dessas posições seguida de um ponto e outro número, na ordem 112.1, 111.1, 112.2, 111.2. Além disso, as posições dos *frameworks* que servem de âncora para cada um dos três CDRs, isto é, uma posição antes e uma posição depois de cada CDR, são destacadas por quadrados. As cores desses quadrados, na representação proposta, identificam os CDRs: verde para o CDR1, laranja para o CDR2 e azul para o CDR3.

Tabela 4.1: Classificação dos resíduos das sequências analisadas segundo suas propriedades químicas

Classe do aminoácido	**Cor**	**Aminoácidos**
Ácido	Vermelho	Ácido aspártico (D) e ácido glutâmico (E)
Básico	Azul	Arginina (R), histidina (H) e lisina (K)
Hidrofóbico	Preto	Fenilalanina (F), prolina (P), valina (V), leucina (L), isoleucina (I), alanina (A), triptofano (W) e metionina (M)
Neutro	Roxo	Asparagina (N) e glutamina (Q)
Polar	Verde	Cisteína (C), serina (S), glicina (G), tirosina (Y) e treonina (T)
Gap	Cinza	-

As sequências apresentadas na Figura 4.2 foram obtidas a partir de estruturas do PDB de anticorpos de lhama (somente cadeia pesada) e foi aplicado um filtro de redundância para garantir que as sequências analisadas tenham no máximo 95% de identidade (os detalhes do processo de filtragem serão descritos nas seções 4.2.1 e 4.2.2). Com a representação proposta, é possível observar as posições hipervariadas, por exemplo a posição 113, posições altamente conservadas, como as posições 23, 41 e 53 e posições variáveis, mas com predominância de alguma classe, como a posição 63. Assim como a representação

Collier de Perles pode ser apresentada utilizando uma ou duas camadas, como descrito anteriormente, a representação proposta neste trabalho também pode ser apresentada dessas duas maneiras.

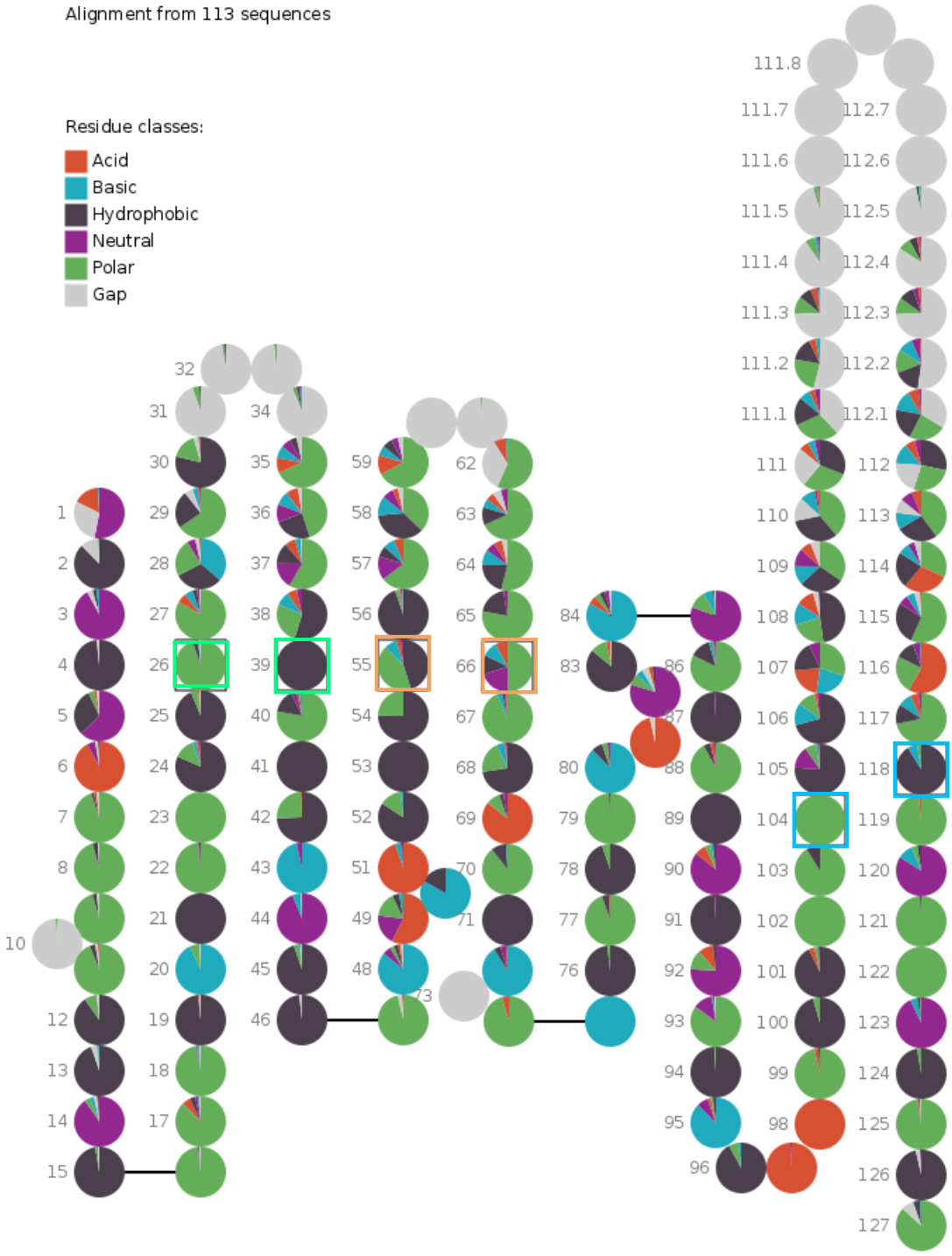


Figura 4.2: Proporção de resíduos em cada posição de 113 sequências de cadeias pesadas de anticorpos de lhama, utilizando a representação proposta neste trabalho através da apresentada em 1 camada.

Ferramentas de visualização de múltiplas sequências, além de utilizar cores para representar propriedades no alinhamento, utilizam outras formas de realce para identificar predominância de resíduos em determinadas posições, comparando o alinhamento com uma sequência consenso ou uma sequência de referência. As ferramentas podem realçar a predominância de diversas maneiras, como não colorindo certas posições, colocando marcas sobre o alinhamento, colocando os resíduos com letras maiúsculas ou minúsculas ou ainda, substituindo-os por outros símbolos (PROCTER et al., 2010). Na representação proposta neste trabalho, o realce de posições de predominância de uma classe para o atributo analisado é feito implicitamente, pois o gráfico de setores permite a visualização direta de classes dominantes em determinadas posições. No entanto, para comparar um grupo de sequências analisadas com uma sequência específica, como uma sequência de referência, consenso ou de um gene *germline*, foi utilizado um recurso adicional. Nesse recurso, cada um dos resíduos da sequência é desenhado no centro do gráfico de setores circundado pela cor correspondente à classe do resíduo segundo o atributo analisado, como mostrado na Figura 4.3. Com essa representação, comparando-se a cor do círculo central com os setores de um gráfico de posição correspondente, é possível perceber quais posições possuem resíduos com a mesma classe do resíduo da sequência especificada e quais são divergentes.

Outro recurso proposto nessa representação é o destaque de posições relevantes à análise considerando um novo atributo além do utilizado na classificação do gráfico de setores. Esse atributo adicional deve ser algum atributo de cada uma das posições que possa ser simplesmente contado. Por exemplo, quando se analisa um grupo de sequências de anticorpos que se ligam a um determinado epitopo, é interessante conhecer quais são as posições das cadeias do anticorpo que mais realizam interações com o antígeno. Dessa forma, um exemplo de atributo adicional que pode ser utilizado nessa análise é o número de sequências em que cada uma das posições faz contato com o epitopo. Outro exemplo, seria o número de sequências em que cada posição tem acessibilidade ao solvente acima de algum valor limite.

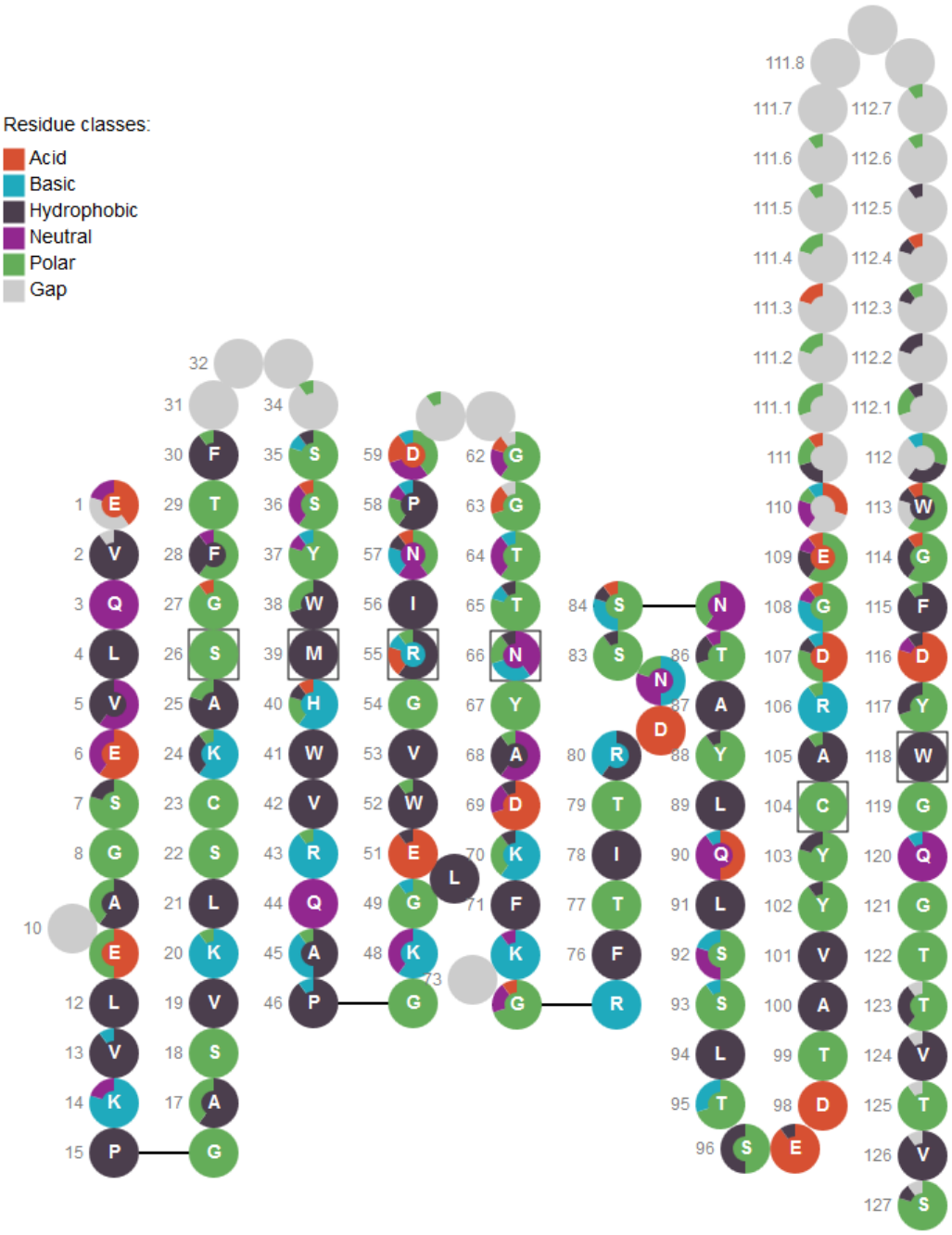


Figura 4.3: Proporção de resíduos (classificados segundo suas propriedades físico-químicas) em cada posição de 10 sequências de cadeia pesada de anticorpos contra zika vírus comparada com a referência consenso de sequências de anticorpos contra dengue (dados não publicados).

O atributo adicional na visualização proposta é representado através de um círculo adicional, visível na parte exterior de cada um dos gráficos de setor, com raio proporcional ao número de sequências que possuem tal atributo, como apresentado na Figura 4.4. Essa figura apresenta parte do CDR3,

destacando, para as estruturas analisadas, diversas posições que podem realizar interações com o epitopo (considerando uma distância máxima de 8Å entre carbonos alfa). Nessa figura, é possível observar que, nas cadeias analisadas, é mais frequente as posições 109 e 110 interagirem com o antígeno que a posições 116 e 117 e que a posição 106 não interage com o antígeno em nenhuma das estruturas analisadas.

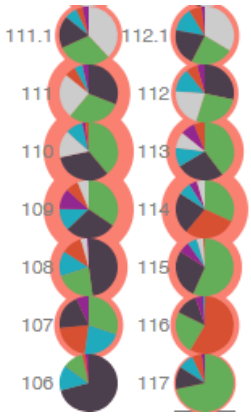


Figura 4.4: Destaque das posições que mais realizam interações com o antígeno no CDR3 para as sequências representadas na Figura 4.2.

4.2 Yvis: plataforma para análise e visualização em larga escala de alinhamento de anticorpos

Neste trabalho foi implementada uma plataforma de análise e visualização de alta densidade de alinhamentos de domínios variáveis de cadeias de anticorpo (*antibod**Y** **Vis**ualization platform*), disponibilizada em um servidor *web* no endereço http://bioinfo.icb.ufmg.br/yvis/. Essa plataforma inclui um banco de dados de estruturas de anticorpos curado e atualizado semanalmente (*Yvis database*) e um conjunto de recursos para a análise de anticorpos, como a visualização proposta, o *Collier de Diamants*, e múltiplas opções para a filtragem das sequências analisadas. Essa plataforma será apresentada na próxima seção e os detalhes de sua implementação serão apresentados na seção 4.2.2.

4.2.1 Funcionalidades da plataforma Yvis

O banco de dados da plataforma Yvis é uma coleção de informações de estruturas de anticorpos presentes no PDB em sua forma livre ou em complexo com um antígeno. As informações de cada cadeia

armazenada nesse banco de dados compreendem: a identificação da estrutura no PDB e da cadeia, os organismos produtores do anticorpo e do antígeno (caso o anticorpo esteja em complexo com um antígeno proteico) ou tipo do antígeno (hapteno, carboidrato ou ácido nucleico), a sequência com *gaps* segundo o esquema de numeração do IMGT, informações dos genes *germlines* V e J (alelo inferido e o valor de identidade) e as posições da cadeia do anticorpo que possivelmente realizam contatos com o antígeno (quando a estrutura do PDB é de um complexo entre anticorpo e antígeno proteico). Os detalhes da geração dessas informações serão apresentados na seção 4.2.2. A Figura 4.5 apresenta uma visão geral da plataforma Yvis.

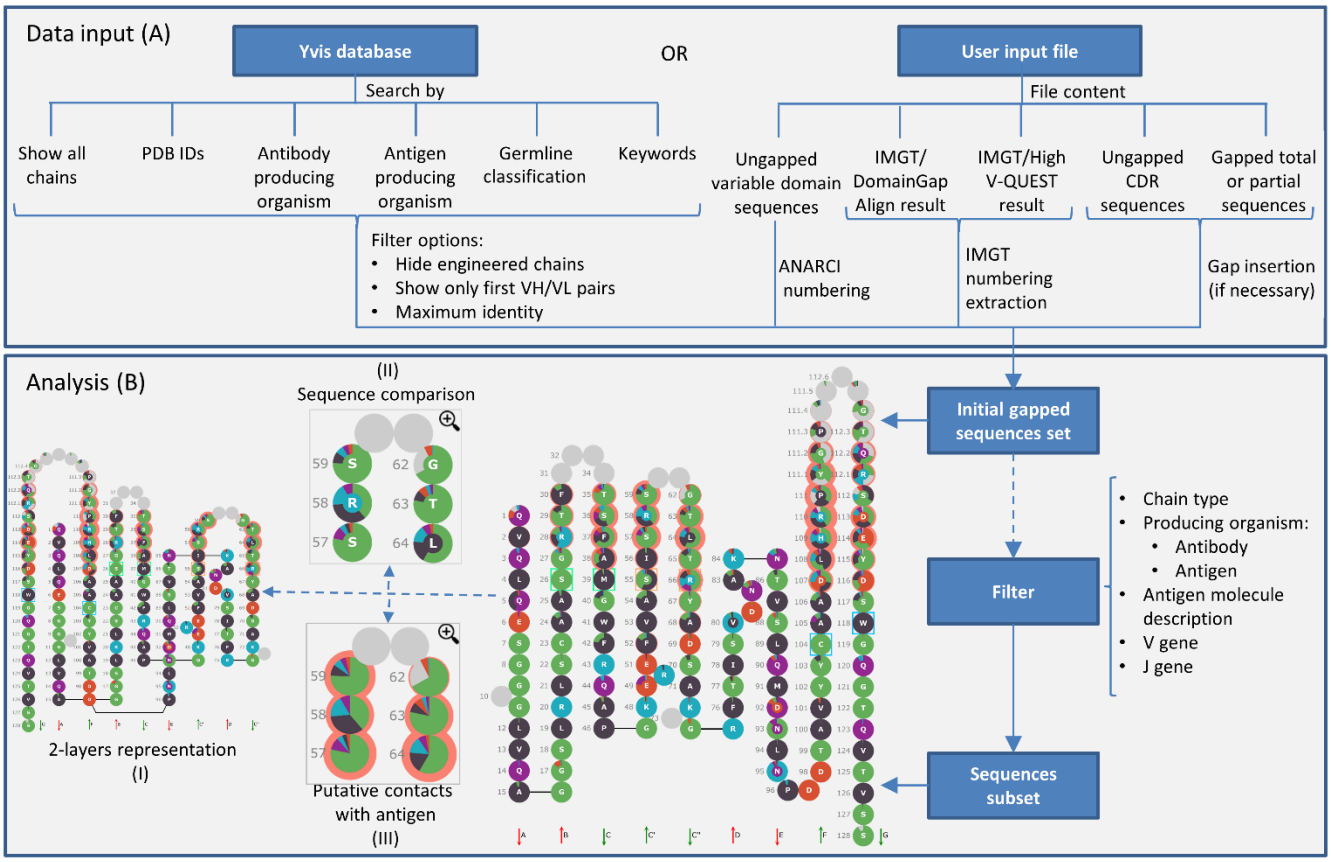


Figura 4.5: Visão geral da plataforma Yvis. A caixa (A) apresenta os tipos de entradas de dados aceitos pelo Yvis. A caixa (B) apresenta o processo de análise dos dados na plataforma e as opções de visualização do alinhamento de múltiplas sequências, as regiões de possíveis contatos e a comparação do alinhamento com uma sequência especificada pelo usuário. Figura retirada do artigo (CARVALHO; MOLINA; FELICORI, 2019).

Ao utilizar a plataforma Yvis, são apresentadas diversas opções para o usuário definir os dados de entrada para o processamento e consequente análise. Essas opções são divididas em duas fontes: sequências armazenadas no banco de dados do Yvis e arquivos do usuário (Figura 4.5A). Para analisar as sequências das estruturas armazenadas na base de dados do Yvis, o usuário deve escolher entre as seguintes opções: analisar todas as estruturas disponíveis na base de dados; indicar uma lista de identificadores das estruturas segundo o PDB ou, ainda, os identificadores da estrutura e das cadeias; selecionar uma ou mais espécies de organismo produtor do anticorpo; selecionar uma ou mais espécies de organismo produtor de antígenos proteicos, ou o tipo de antígeno, ou a opção de anticorpos livres; selecionar um ou mais alelos dos genes V e J, além do organismo de origem do gene; ou dados dos artigos de referência do depósito das estruturas no PDB. As diversas listas apresentadas para a seleção do usuário contêm somente opções referentes ao tipo de seleção escolhida, por exemplo, a lista de organismos produtores de antígeno é diferente da lista de produtores de anticorpos. Já a opção de exibição por dados dos artigos de referência do depósito da estrutura permite ao usuário informar palavras-chave presentes no título e no sumário do artigo, os autores, o ano de publicação e o código de indexação da publicação, nos formatos *Digital Object Identifier* (DOI) e/ou *PubMed Identifier* (PMID). Esses dados podem ser utilizados com os operadores lógicos AND, OR e NOT. Nessa opção, a requisição do usuário é processada consultando as informações dos artigos disponíveis no PDB.

Todas as opções de análise de sequências armazenadas no banco de dados do Yvis fazem com que a ferramenta selecione todas as cadeias leves e pesadas de todas as estruturas disponíveis na base de dados que correspondam ao critério de seleção. No entanto, frequentemente, o usuário não deseja analisar mais de um par de cadeias leve e pesada por estrutura, já que, normalmente, os pares de cadeia leve e pesada de um mesmo anticorpo são iguais, com exceção dos anticorpos bivalentes. Nesse caso, o usuário tem a opção de restringir a opção de exibição indicando se somente um par de cadeias ou todas as cadeias devem

fazer parte da seleção. Outro recurso importante do Yvis é a possibilidade de indicar um valor máximo de identidade entre as sequências analisadas. Caso essa opção seja selecionada, a ferramenta inclui na análise somente as sequências que satisfazem esse critério, considerando o valor máximo de identidade para cada uma das cadeias (leve e pesada). Com essas opções, o usuário pode gerar conjuntos de sequências não redundantes para análise, evitando vieses nos resultados. Outro critério adicional que o usuário tem para as sequências armazenadas no Yvis é a inserção ou não de anticorpos marcados como modificados geneticamente, por meio do campo *engineered*, no arquivo PDB.

Caso o usuário deseje analisar suas próprias sequências de domínios variáveis das cadeias dos anticorpos e não mais as sequências de estruturas conhecidas, a plataforma Yvis disponibiliza diferentes opções de conteúdo e formatos de arquivo: arquivo FASTA contendo sequências de cadeias de anticorpos, com ou sem *gaps* produzidos pela aplicação do esquema de numeração ou não, ou contendo sequências de somente um CDR; arquivo de resultado de uma submissão ao IMGT/HighV-QUEST (ALAMYAR et al., 2012) ou a página de resultado da submissão ao IMGT/DomainGapAlign (EHRENMANN; KAAS; LEFRANC, 2010). Caso seja escolhido o formato FASTA, o usuário pode informar na linha de identificação de cada sequência (iniciada com o símbolo >) suas características, como os organismos produtores do anticorpo e do antígeno, para serem utilizadas em filtros posteriores. A ferramenta IMGT/HighV-QUEST permite ao usuário submeter ao servidor do IMGT um arquivo contendo sequências de nucleotídeos de cadeias de anticorpos. Dentre os arquivos gerados depois do processamento pelo IMGT/HighV-QUEST, existe um arquivo com a tradução das sequências e os *gaps* gerados após a aplicação do esquema de numeração do IMGT. Já a ferramenta IMGT/DomainGapAlign recebe como entrada sequências de aminoácidos e retorna as sequências com os *gaps*.

Após o usuário selecionar as sequências que serão analisadas é apresentada uma tela com diversos painéis para auxiliar a análise das sequências. A Figura 4.6 apresenta esses painéis, identificados por letras.

O painel “A” é o cabeçalho de todas as páginas da plataforma Yvis que pode ser utilizado para a navegação entre as telas de apresentação, análise, estatísticas do banco de dados da plataforma, tutorial e informações referentes aos autores e versões da ferramenta. O painel “B” apresenta as informações referentes aos dados utilizados para a seleção das sequências analisadas. Neste exemplo, foi utilizado a opção de seleção da espécie do organismo produtor do antígeno, restringindo as sequências a anticorpos contra o HIV (*Human Immunodeficiency Virus*) e a apenas um par de cadeias leve e pesada para cada entrada do PDB.

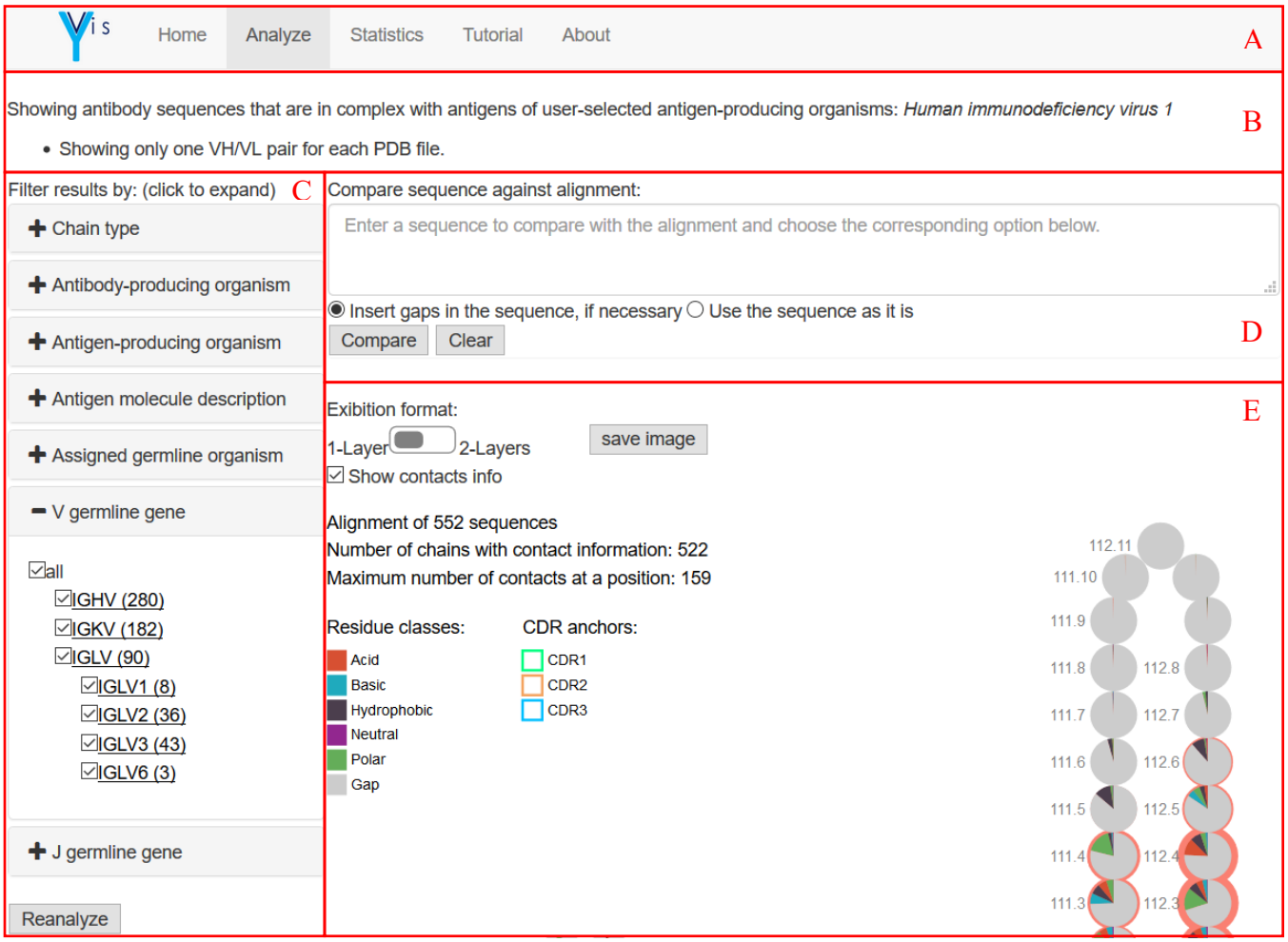


Figura 4.6: Tela de apresentação de resultados e realização de análises sobre um conjunto inicial de dados. (A) Menu da plataforma. (B) Informações sobre o conjunto inicial de dados. (C) Opções para filtragem dos dados. (D) Ferramenta de comparação de uma sequência com o alinhamento. (E) Apresentação do *Collier de Diamants*.

O painel “C” da Figura 4.6 fornece ao usuário filtros que permitem gerar análises e representações mais específicas a partir das sequências inicialmente utilizadas na representação inicial. Para isso, é

possível filtrar os resultados considerando o tipo de cadeia (leve ou pesada), o organismo produtor do anticorpo ou do antígeno, a descrição da molécula do antígeno disponível no PDB, o organismo utilizado na atribuição de *germlines*, e os alelos V e J atribuídos. Além disso, esses filtros podem ser combinados. As opções apresentadas em cada filtro correspondem às opções disponíveis dada a seleção de sequências realizada pelo usuário inicialmente na ferramenta. Ao lado de cada opção de filtro é apresentado o número de sequências que atendem esse critério. A partir de uma visualização filtrada, sempre é possível retornar à primeira visualização, removendo os filtros adicionados.

O painel "D" da Figura 4.6 permite ao usuário inserir uma sequência para ser comparada com o conjunto de sequências apresentados no *Collier de Diamants*, como descrito anteriormente e apresentado na Figura 4.3. O usuário pode optar por deixar a plataforma Yvis aplicar o esquema de numeração à sequência, alinhando-a às demais sequências, ou por fornecer a sequência já com os *gaps* inseridos.

O painel "E" da Figura 4.6 apresenta a visualização *Collier de Diamants*, descrita na seção anterior, exibindo o alinhamento das múltiplas sequências selecionadas pelo usuário. Na plataforma Yvis, cada posição da representação apresenta um gráfico de setores classificando os aminoácidos segundo suas propriedades químicas (Tabela 4.1). Além disso, ela utiliza, como atributo adicional, as posições das sequências de anticorpos que estão próximas ao antígeno (menos de 8Å de distância entre os carbonos alfa), quando a estrutura correspondente à sequência está disponível no PDB. As representações apresentadas nas Figuras 4.2, 4.3 e 4.4 foram produzidas na plataforma Yvis. Além do *Collier de Diamants*, o painel "E" da Figura 4.6 apresenta as ações e informações referentes à representação, como o número total de sequências exibidas, o número de sequências que possuem informações de contatos e o número de sequências que os realizam na posição com o maior número de contatos. Dentre as ações disponíveis, é possível alternar a visualização entre a representação em 1 ou 2 camadas, exibir ou ocultar as informações de contatos e salvar a representação nos formatos PNG (*Portable Network Graphics*) ou

SVG (*Scalable Vector Graphics*). Além disso, posicionando-se o cursor sobre cada gráfico de setores do *Collier de Diamants*, é exibido o número de contatos realizados pelos aminoácidos desta posição. Ao final dessa tela (não exibido na Figura 4.6), o Yvis fornece uma tabela com informações das cadeias analisadas. Essa tabela contém, para cada sequência, os identificadores da estrutura do PDB e da cadeia, a informação se a cadeia foi geneticamente modificada ou não e os organismos relacionados ao anticorpo e ao antígeno. Além disso, ela apresenta a descrição da molécula do antígeno, a sequência do domínio variável da cadeia com e sem *gaps* de acordo com o esquema de numeração do IMGT, a ferramenta utilizada para aplicar o esquema de numeração e as informações referentes aos *germlines* associados à sequência, isto é, o organismo de origem, os alelos dos genes V e J e a porcentagem de identidade entre a sequência e os alelos atribuídos. Para sequências inseridas pelo usuário, essas informações são exibidas quando adicionadas na linha de identificação de cada sequência (iniciada com o símbolo >) dos arquivos FASTA, com exceção do tipo da cadeia que é detectado automaticamente conforme as características conservadas das cadeias dos anticorpos. Os dados disponibilizados nesta tabela podem ser exportados em formatos como PDF (*Portable Document Format*) e CSV (*Comma-separated values*). Na sequência com *gaps*, os CDRs e as posições que possivelmente realizam contatos são destacados. Caso o usuário deseje comparar uma sequência dessa tabela com as demais sequências analisadas, como no recurso da plataforma Yvis descrito anteriormente, basta clicar sobre a sequência correspondente na tabela de dados. Além disso, os alelos apresentados possuem um *link* para a sequência de aminoácidos correspondentes na página do IMGT (GIUDICELLI; CHAUME; LEFRANC, 2005).

Como o *Collier de Diamants* utiliza a classificação de aminoácidos (Tabela 4.1) segundo suas propriedades químicas para representar o conteúdo de cada posição no alinhamento de múltiplas sequências, a partir dessa representação, não é possível conhecer a frequência de cada um dos resíduos em cada posição. No entanto, esse tipo de detalhamento é fornecido pela plataforma Yvis, quando o

usuário clica sobre uma posição, por meio de um gráfico de barras como o apresentado na Figura 4.7. Esse gráfico apresenta o número de cada um dos tipos de aminoácidos na posição selecionada e, para cada uma das barras, é possível detalhar o número de posições que contêm o resíduo referente a ela. As barras do gráfico possuem a cor correspondente à classe do aminoácido representado e podem ser ordenadas pelo tipo de aminoácido ou por sua frequência. Esse segundo gráfico apresentado (gráfico de barras) permite um detalhamento das informações apresentadas pela primeira representação (*Collier de Diamants*).

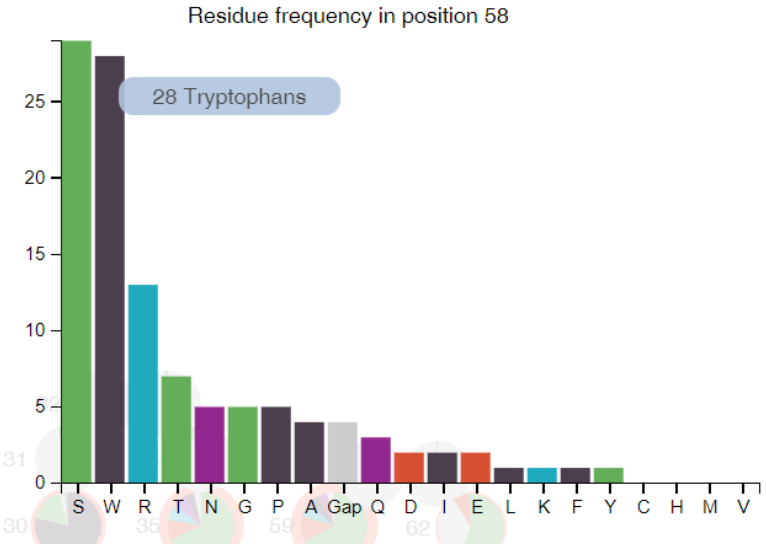


Figura 4.7: Gráfico representando o número de aminoácidos de cada tipo na posição 58 do alinhamento apresentado na Figura 4.2, ordenado pela frequência. As cores das barras representam a classificação dos resíduos segundo a Tabela 4.1. A informação sobre o número de triptofanos na posição 58 é obtida posicionando o cursor sobre a barra correspondente.

Apesar da plataforma Yvis utilizar as propriedades químicas dos aminoácidos para classificá-los e a distância entre resíduos do antígeno e anticorpo para a definição de possíveis contatos, a implementação da plataforma Yvis foi realizada de maneira que seja fácil sua alteração para a análise de outros atributos das cadeias dos anticorpos. Dessa maneira, outras versões da plataforma poderão oferecer ao usuário outras opções de análises.

4.2.2 Implementação da plataforma Yvis: Materiais e métodos

Para a implementação da plataforma Yvis, descrita na seção anterior, foram utilizados PHP 5.5.9 e banco de dados MySQL versão 5.5.52, executados sobre um servidor Apache 2.4.7. O banco de dados MySQL armazena as informações dos domínios variáveis das cadeias de anticorpos disponíveis no PDB e processadas por quatro *scripts* desenvolvidos em Python versão 2.7.6. Semanalmente, esses *scripts* são executados no servidor que armazena os dados da plataforma Yvis.

O primeiro *script* utiliza a biblioteca Mechanize versão 0.2.5 (LEE, 2017) para acessar a página do banco de dados SAbDab (DUNBAR et al., 2014), apresentado na seção 1.6.1, simulando uma requisição para a listagem de todas as estruturas presentes neste banco de dados. Como resposta para tal requisição, é retornada uma página HTML (*HyperText Markup Language*) contendo um *link* para um arquivo com o resumo de todos os registros disponíveis no SAbDab. O *script* encontra o *link* na página HTML e o acessa. Cada linha desse arquivo representa o pareamento de uma cadeia leve e uma pesada (ou apenas 1 das cadeias, se o anticorpo for formado por somente um domínio variável) e, se a estrutura contém um complexo antígeno-anticorpo, o antígeno ao qual as cadeias estão ligadas também são pareados. O SAbDab foi escolhido por conter o pareamento entre cadeias do anticorpo e do antígeno e por ter uma base de dados atualizada semanalmente, fazendo com que o Yvis possa ter informações de estruturas recentemente depositadas.

O segundo *script* implementado é responsável pela aquisição e geração dos dados armazenados no banco de dados do Yvis. Esse *script* recebe como parâmetro o arquivo resumo obtido do SAbDab e, para cada linha do arquivo, obtém as informações de organismo produtor de antígenos (se a estrutura for de um complexo antígeno-anticorpo) e anticorpos e gera a sequência das cadeias dos anticorpos com *gaps* a partir da aplicação do esquema de numeração do IMGT. Além disso, obtém informações dos alelos dos genes V e J associados às cadeias leve e pesada, com seus respectivos valores de identidade e organismo

e, por fim, o *script* obtém o arquivo PDB que contém a estrutura para processar as informações referentes aos contatos entre o anticorpo e o antígeno. O *script* extrai do arquivo resumo os identificadores do PDB de cada complexo e obtém o arquivo da estrutura correspondente no servidor do PDB por meio da biblioteca Biopython (COCK et al., 2009) versão 1.66. Do arquivo resumo também são extraídas as espécies produtoras do anticorpo e do antígeno, no entanto, ocasionalmente essas informações estão ausentes. Nesse caso, o *script* obtém a informação diretamente do arquivo PDB.

O *script* também obtém do PDB um arquivo FASTA correspondente a cada uma das estruturas analisadas contendo a sequência de aminoácidos das cadeias dos anticorpos. Essas sequências são submetidas ao IMGT/DomainGapAlign (EHRENMANN; KAAS; LEFRANC, 2010), acessível através de um servidor *web*, indicando como espécie alvo, a espécie definida anteriormente. Caso essa espécie não esteja na lista utilizada pelo IMGT/DomainGapAlign, esta ferramenta escolhe a espécie que tenha uma sequência *germline* mais próxima da sequência analisada. O *script*, utilizando o pacote Mechanize versão 0.2.5 (LEE, 2017), insere as sequências das cadeias no formulário *web*, realiza sua submissão e captura a página HTML retornada. Na página retornada, o IMGT disponibiliza um alinhamento das sequências submetidas com outras sequências armazenadas no servidor do IMGT já alinhadas segundo o esquema de numeração do IMGT. As sequências alinhadas apresentam *gaps*, quando necessário, para mantê-las no padrão definido pelo esquema de numeração do IMGT. A partir da página de resultados gerada, o *script* implementado gera a sequência com *gaps* e extrai as informações dos genes *germline*: alelos V e J, a identidade da sequência submetida com esses alelos e a espécie correspondente. A sequência com *gaps* do domínio variável de cada cadeia do anticorpo é armazenada no banco de dados do Yvis em dois campos: a sequência sem inserções no CDR3, contendo 128 aminoácidos e *gaps*, e a sequência da inserção do CDR3 caso ele tenha mais de 13 aminoácidos. Essa estratégia foi adotada para permitir o alinhamento de sequências com qualquer número de inserções no CDR3. A partir da página de resultado

do IMGT/DomainGapAlign, o *script* também extrai o tipo de cadeia analisada (leve ou pesada). No caso de scFvs, moléculas que possuem os dois domínios em uma mesma cadeia, o *script* obtém da página de resultados as informações dos dois domínios e as armazena separadamente no banco de dados do Yvis.

Para a definição de possíveis contatos entre anticorpos e antígenos proteicos, o *script* implementado, por meio da biblioteca Biopython, verifica, para cada carbono alfa dos resíduos das cadeias do anticorpo, se há algum carbono alfa dos resíduos da(s) cadeia(s) do antígeno a uma distância menor ou igual a 8Å. Essa distância foi escolhida para acomodar, além das interações de hidrogênio, de Van der Waals, pontes salinas, contatos hidrofóbicos e interações π, as interações mediadas por moléculas de água (NGUYEN et al., 2017; VIART et al., 2016). Se existir um resíduo do antígeno a uma distância menor ou igual a 8Å de um resíduo do anticorpo, a posição ocupada por ele no esquema de numeração é marcada como uma posição que realiza contato.

O segundo *script* armazena as linhas do arquivo resumo do SAbDab que não puderam ser numeradas pelo IMGT/DomainGapAlign em um arquivo de erros. Uma possível causa desse problema é a atribuição incorreta do organismo produtor do anticorpo. Quando isso ocorre, o segundo *script* é executado novamente, mas sem indicar ao IMGT/DomainGapAlign qual é a espécie produtora do anticorpo analisado. Se nessa segunda execução o IMGT/DomainGapAlign conseguir numerar a sequência, a cadeia é inserida com sucesso no banco de dados do Yvis. Caso contrário, um novo arquivo de erros será gerado contendo a linha do arquivo resumo do SAbDab correspondente. Esse arquivo é avaliado por um curador em busca da causa do problema. Em caso de sucesso na nova atribuição, o curador será notificado pelo quarto *script*, responsável por analisar possíveis inconsistências na base de dados.

Após a execução do segundo *script*, o terceiro *script* implementado calcula a identidade entre as cadeias armazenadas no banco de dados do Yvis. Para a análise de identidade das cadeias, utilizada para remover a redundância do conjunto de dados analisado, o *script* verifica as novas cadeias inseridas no

banco pelo *script* descrito anteriormente e calcula a identidade entre elas e as demais cadeias já armazenadas. Para reduzir o número de comparações e valores de identidade armazenados, os dados foram divididos em duas tabelas distintas, uma para as cadeias leves e outra para as cadeias pesadas. No cálculo de identidade, *gaps* e *mismatches* (resíduos diferentes em uma mesma posição do alinhamento) são contabilizados da mesma maneira e o alinhamento das sequências é realizado pela comparação das posições após a aplicação do esquema de numeração.

Um problema encontrado nos bancos de dados analisados na seção 1.6.1, é a falta de padronização nos nomes utilizados na definição das espécies produtoras de anticorpos e, principalmente, na de antígenos. Tal problema dificulta a pesquisa nesses bancos de dados. Na tentativa de reduzir esse problema, um quarto *script* foi implementado. Esse último *script* verifica se os nomes das espécies produtoras armazenados no banco de dados do Yvis estão padronizados segundo a Taxonomia do Uniprot (https://www.uniprot.org/taxonomy/). Para isso, a plataforma Yvis mantém duas tabelas com os nomes padronizados das espécies produtores de anticorpos e antígeno. Novos nomes são inseridos nessas tabelas quando um curador confere sua existência na Taxonomia do Uniprot. Dessa maneira, quando um nome de espécie produtora de antígeno ou de anticorpo não é encontrado pelo quarto *script* na tabela correspondente, um e-mail é enviado ao curador, indicando o nome da espécie e o registro do PDB em que o nome foi encontrado. O curador deve então analisar a causa desse problema, isto é, nome padronizado, mas ainda não inserido na tabela de nomes, ou ainda, nome não padronizado ou ausente no arquivo resumo do SAbDab e estrutura do PDB. No primeiro caso, o curador insere o nome na tabela correspondente. Já no segundo caso, pode ser necessário que o curador consulte a bibliografia relacionada à estrutura depositada no PDB ou verifique a equivalência do nome a algum outro já armazenado nas tabelas da plataforma. Nessa última situação, o curador cria regras indicando nomes sinônimos que são verificadas antes do envio da notificação ao curador. Dessa maneira, o número de intervenções do curador

tende a diminuir uma vez que mais nomes padronizados e regras de equivalência são cadastrados. Após a curadoria, o quarto *script* é executado, pois os dados das cadeias são disponibilizados na interface da plataforma Yvis somente após a análise e aprovação (automática) dos nomes das espécies produtoras do antígeno e do anticorpo. Além do problema de nomes não padronizados, o quarto *script* implementado verifica possíveis inconsistências no banco de dados da plataforma Yvis e também as envia por e-mail ao curador. As inconsistências relatadas são anticorpos com cadeias geradas por espécies diferentes, geralmente causadas por erro de anotação da espécie produtora, e espécie produtora de anticorpo diferente da espécie atribuída ao gene *germline*. Essa última divergência pode ser causada por falha na anotação do organismo produtor do anticorpo ou pela ausência da espécie produtora na lista de espécies reconhecidas pelo IMGT/DomainGapAlign.

Os *scripts* apresentados anteriormente são responsáveis pela geração e armazenamento dos dados referentes às estruturas de anticorpos disponíveis no PDB. Para a implementação da exibição de opções de seleção de sequências para análises e consultas ao banco de dados, foram implementados diversos módulos na linguagem PHP. Esses módulos geram páginas HTML que, utilizando JavaScript, geram a visualização a partir dos dados processados em PHP, os filtros adicionais e a tabela com as informações das cadeias visualizadas. Quando um filtro de identidade é selecionado, um módulo PHP faz uma consulta ao banco de dados e obtém os valores de identidade entre todas as sequências escolhidas pelo usuário. A partir dessa consulta, o módulo PHP utiliza um algoritmo guloso, baseado no algoritmo utilizado pela ferramenta cd-hit (LI; GODZIK, 2006). Esse algoritmo analisa cada uma das sequências e classifica-as como inseridas ou removidas. Para cada nova cadeia analisada, é verificado se já existe alguma no grupo de inseridas que possui identidade superior ao *cutoff* definido pelo usuário. Se existir, essa sequência é marcada como removida, caso contrário, ela é marcada como inserida. Esse algoritmo garante que não existirão sequências com identidade superior ao *cutoff* definido pelo usuário. No entanto, ele não garante

que o maior número de sequências possível seja encontrado. Como o filtro de identidade é executado durante a requisição do usuário, ele não pode demorar muito tempo para ser processado. Portanto, esse algoritmo atende às necessidades considerando as limitações de tempo de resposta ao usuário.

Quando o usuário adiciona suas próprias sequências, elas são enviadas ao servidor para que sejam verificadas e alinhadas, se necessário. O alinhamento de sequências de CDRs é feito particionando a sequência e inserindo *gaps*, quando necessário, no meio da sequência, como definido no sistema de numeração do IMGT (LEFRANC et al., 2003). No entanto, quando o usuário insere sequências correspondentes a todo domínio variável, o alinhamento precisa ser realizado considerando as posições conservadas nas regiões de *framework*. Para realizar esse alinhamento, um módulo PHP submete as sequências fornecidas pelo usuário ao programa ANARCI versão 1.2 (DUNBAR; DEANE, 2015). Esse programa aplica o esquema de numeração do IMGT a partir do alinhamento de cada sequência a um conjunto de modelos ocultos de Markov (HMM – *Hidden Markov Models*) que descrevem as sequências *germlines* de humanos, camundongos, ratos, coelhos, porcos e macacos Rhesus. Devido à restrição das espécies utilizadas na geração desses modelos, as espécies de *germline* atribuídas podem ser diferentes da que produziu as sequências. Após a execução do ANARCI, o módulo PHP processa seu resultado gerando as informações de domínio e alelos *germline*, além das sequências com *gaps* gerados pela aplicação do esquema de numeração.

Quando o usuário submete um arquivo de sequências de cadeias de anticorpos ao IMGT/DomainGapAlign (EHRENMANN; KAAS; LEFRANC, 2010), essa ferramenta aplica o esquema de numeração sobre as sequências e define os alelos dos genes V e J relacionados a elas. O resultado do IMGT/DomainGapAlign é uma página HTML. Para realizar a análise dessas sequências numeradas no Yvis, o usuário deve salvar essa página como um arquivo e submetê-la à plataforma utilizando a opção correspondente. O Yvis processa o arquivo no navegador do usuário utilizando JavaScript sem enviar

esses dados ao servidor da plataforma e extrai as informações necessárias à analise com o processamento das *tags* de HTML e *links* contidas no arquivo.

Quando o usuário analisa um conjunto de sequências de nucleotídeos, obtidos através do sequenciamento do genoma ou transcriptoma de células B utilizando o IMGT/HighV-QUEST (ALAMYAR et al., 2012), diversos arquivos são produzidos por essa ferramenta. Para realizar a análise desses dados com a plataforma Yvis, o usuário deve submeter o arquivo que contém as sequências de aminoácidos e informações relativas à aplicação do esquema de numeração sobre essas sequências (4_IMGT-gapped-AA-sequences.txt). O Yvis processa o arquivo no navegador do usuário utilizando JavaScript sem enviar estes dados ao servidor da plataforma e extrai as informações necessárias à análise na plataforma a partir do processamento desse arquivo texto.

Para produzir a visualização *Collier de Diamants* e permitir a apresentação dos gráficos e as interações com os mesmos, a partir de dados fornecidos pelos módulos PHP no formato JSON (*JavaScript Object Notation*), foi utiliza a biblioteca D3.js versão 3.5.17 juntamente com métodos em JavaScript. Para a exibição da tabela de informações de cadeias, foi utilizado o *plug-in* DataTables versão 1.10.15 da biblioteca jQuery versão 1.12.4. A biblioteca Bootstrap 3.3.7 foi utilizada para o desenvolvimento de uma interface *web* com o usuário. Tanto a tabela, quanto a visualização e os filtros que podem ser aplicados foram implementados utilizando-se JavaScript e suas bibliotecas. Portanto, após o acesso aos dados do conjunto de cadeias a ser analisadas, a lógica da ferramenta é executada no navegador do usuário, liberando o servidor para atender outras requisições. A única exceção é a comparação do alinhamento de múltiplas sequências com uma sequência em que é necessário aplicar o esquema de numeração. Nesse caso, uma requisição ao servidor é gerada, contendo a sequência que é alinhada com o auxílio do programa ANARCI da mesma maneira que as demais sequências fornecidas pelo usuário.

4.2.3 Dados armazenados atualmente na plataforma Yvis

A plataforma Yvis é atualizada semanalmente e, em meados de junho de 2019, armazenava informações de 3.550 estruturas de anticorpos, sendo 186 depositadas nesse ano. Do total de estruturas analisadas, 22 nomes de organismos produtores de antígeno foram definidos, sendo que a grande maioria das estruturas possuem anticorpos produzidos por humanos (1.506) e camundongos (1.395), correspondendo a pouco mais de 81% das estruturas. O número de estruturas de anticorpos em complexo exclusivamente com haptenos (186), ácidos nucleicos (22) e carboidratos (107) é inferior ao número de anticorpos em sua forma livre (1.166). Esse número é ainda menor que o número de estruturas de anticorpos em complexo com pelo menos um antígeno proteico (2.069). Esses antígenos proteicos são produzidos por 158 organismos distintos, com nomes padronizados segundo a taxonomia do UniProt.

Para manter uma base de dados com nomes padronizados segundo a taxonomia do UniProt, foram definidas algumas regras que substituem nomes não padronizados, como nomes sinônimos ou com definição de cepa, por exemplo, para nomes de espécies definidas na taxonomia do UniProt. O Apêndice 1 apresenta uma tabela que lista os nomes não padronizados que foram modificados e o nome padronizado correspondente. Apesar dos diversos nomes alterados por essas regras, elas não são suficientes para padronizar todos nomes. Algumas vezes, informações faltantes ou estranhas foram encontradas, como um vírus como organismo produtor do anticorpo. Nesses casos, foram necessárias a inspeção manual feita por um curador e a modificação manual dos nomes na base de dados do Yvis. No total, foram realizadas alterações manuais nos nomes da espécie produtora do anticorpo de 49 estruturas e, nos nomes da espécie produtora do antígeno de 81 estruturas. Além disso, 12 estruturas apresentavam espécie produtora de anticorpos diferentes entre as cadeias de um mesmo anticorpo. Essas estruturas também tiveram os nomes alterados para garantir que todo anticorpo tenha uma única espécie produtora associada.

4.2.4 Estudos de caso

Nesta seção são apresentados três estudos de caso da utilização da plataforma Yvis para a visualização e análise de sequências e estruturas relacionadas ao vírus HIV-1 (*Human Immunodeficiency Virus* tipo 1). O HIV possui diversos mecanismos para escapar da ação do sistema imunológico, no entanto, alguns indivíduos por ele infectados são capazes de desenvolver anticorpos neutralizantes chamados bNAbs (*broadly-neutralizing antibodies*) depois de diversos anos de infecção. Como esses anticorpos são capazes de reconhecer múltiplas cepas do vírus HIV-1, acredita-se que seu estudo pode gerar o conhecimento necessário para o desenvolvimento de vacinas contra o HIV-1 (SUN et al., 2017; WU et al., 2011). Dessa forma, diversos anticorpos contra o HIV-1 foram isolados nas últimas décadas e algumas estruturas destes anticorpos estão disponíveis no PDB.

4.2.4.1 Estudo de caso 1: Visualização de cadeias de anticorpos contra a proteína gp120 de vírus HIV-1

Para se obter as estruturas de anticorpos anti-HIV disponíveis no PDB, pode-se realizar uma pesquisa através da ferramenta de busca disponibilizada pelo próprio PDB ou através de outros bancos de dados que armazenam informações dessas estruturas. Nesse estudo de caso, realizado no final de 2.017, a busca foi realizada diretamente no PDB e nos bancos de dados apresentados na seção 1.6.1 (que existiam na época) para a comparação das estruturas retornadas por cada pesquisa. Uma exceção é o banco de dados *IMGT/3Dstructure-DB* (EHRENMANN; KAAS; LEFRANC, 2010) que não permite a pesquisa de estruturas por meio da informação do antígeno e, portanto, não foi utilizada neste estudo de caso.

Em uma pesquisa ao PDB, com as palavras "*antibody*" e "*HIV-1*", foram obtidas 511 estruturas. No entanto, diversas dessas estruturas não possuem as cadeias de anticorpos, pois, apesar de estarem associadas a publicações que contêm informações de anticorpos, não necessariamente todas as estruturas relacionadas contêm anticorpos. Para levantar quais dessas estruturas realmente continham

cadeias de anticorpos, foi analisado o relatório detalhado gerado pelo PDB como um arquivo no formato CSV. A partir desse arquivo foi realizada a inspeção manual de todos os registros pelo campo que contém o nome da macromolécula. Com base nessa análise, restaram 461 estruturas do PDB. Esse tipo de análise exige tempo e conhecimento de palavras relacionadas às cadeias de anticorpos, por exemplo: *fab*, *heavy*, *light*, etc. Essas estruturas não foram inspecionadas para verifica se realmente os anticorpos nela presentes fazem algum tipo de interação com o HIV-1 ou proteínas/peptídeos derivados dele. Portanto, ainda podem existir estruturas que não representam anticorpos que interagem com o HIV-1 nesse conjunto. Além disso, algumas estruturas podem não ser encontradas por não possuírem as palavras-chave utilizadas na pesquisa.

No *Immune Epitope Database* (IEDB) 3.0 (VITA et al., 2015), somente é possível navegar pelos registros que contêm estrutura tridimensional a partir da espécie produtora do anticorpo. Para contornar essa restrição, é possível realizar a pesquisa por experimentos associados ao antígeno HIV-1, gerar um relatório com resultados em um arquivo CSV e, posteriormente, filtrar pelos resultados que possuem estruturas do PDB associadas. Com esse método, foram obtidas 289 estruturas do PDB. Dessas, 29 estruturas não estão presentes no resultado gerado pelo PDB, possivelmente por estarem associadas a palavras chave diferentes das utilizadas na busca realizada no PDB. Uma vantagem da pesquisa realizada no IEDB é que, por ser uma base de dados curada manualmente, os nomes dos organismos relacionados ao antígeno são armazenados de maneira padronizada, garantindo que nesse banco de dados não exista outros registros além dos exibidos na busca pelo antígeno especificado.

O *Antigen-Antibody Interactions Database* (AgAbDb) (KULKARNI-KALE et al., 2014) retorna 140 estruturas do PDB que contêm complexos de anticorpos com o HIV-1 ou proteínas/peptídeos derivados dele. Destas estruturas, 15 não estão presentes no conjunto retornado pelo IEDB. Uma diferença entre o resultado da pesquisa do AgAbDb e as demais já apresentadas é que este banco de dados somente

retorna dados referentes a complexos antígeno-anticorpo, portanto, estruturas que contenham somente o anticorpo não são analisadas por esta base de dados.

O *Structural Antibody Database* (SAbDab) (DUNBAR et al., 2014) também não permite a pesquisa através do organismo relacionado ao antígeno, no entanto, é possível obter um arquivo CSV contendo todas as estruturas armazenadas nesse banco de dados. A partir desse arquivo, é possível analisar o campo espécie do antígeno e, manualmente selecionar todas as expressões associadas ao HIV-1. Como os nomes dos organismos associados ao antígeno não são padronizados, esta seleção exige que o usuário percorra toda a lista observando quais são relacionadas ao HIV-1. Nesta pesquisa foram encontradas 206 estruturas, sendo que as estruturas relacionadas a HIV não foram inseridas, pois não se sabe se estão relacionadas ao vírus HIV-1 ou HIV-2. A pesquisa realizada pelo SAbDab tem a mesma restrição da realizada pelo AgAbDb, somente estruturas referentes a complexos antígeno-anticorpo são retornadas, mesmo que no SAbDab existam estruturas de anticorpos que não estão em complexo. No entanto, a informação referente ao antígeno somente é apresentada para os complexos.

O abYsis (SWINDELLS et al., 2017) não permite a pesquisa pelo organismo relacionado ao antígeno de estruturas do PDB, mas é possível pesquisar por palavras-chave no nome do registro anotado pelo sistema. No entanto, diferentemente dos bancos de dados analisados anteriormente, não é possível selecionar somente estruturas que possuem um complexo antígeno-anticorpo. Entretanto, o abYsis possui um recurso que permite a análise de um conjunto de sequências obtidas a partir de uma consulta à base de dados. Uma das formas de visualização dos resultados é o alinhamento das sequências, como apresentado na Figura 4.8. Apesar de ser uma representação clássica de alinhamento de múltiplas sequências, esse tipo de visualização dificulta a análise de um número de sequências superior ao que é exibido na visualização e a comparação de uma sequência com as demais.

Na plataforma Yvis, os nomes dos organismos relacionados aos antígenos são padronizados, permitindo a pesquisa diretamente pelo nome *(Human immunodeficiency virus 1*). Assim como no AgAbDb e no SAbDab, quando a pesquisa por antígeno é realizada, são exibidas somente estruturas de complexo antígeno-anticorpo, mesmo sendo armazenadas informações referentes a estruturas de anticorpos fora de complexos. Outra restrição da pesquisa pelo nome do organismo produtor do antígeno é que, caso o complexo seja formado por antígenos de mais de um organismo, o antígeno associado ao complexo é armazenado no Yvis com o nome "*Multiple organisms*". Mesmo com tais restrições, a pesquisa por HIV-1 retornou 209 estruturas, superando o número de estruturas fornecidas pelas outras ferramentas que possuem restrições de pesquisa por antígeno somente quando o anticorpo está em complexo. Caso o usuário deseje fazer suas análises adicionando também os anticorpos que não estão em complexo, ele pode obter a lista de estruturas de outro banco de dados para a busca, por exemplo, o PDB ou o IEDB e depois utilizar a opção de especificar os identificadores do PDB na plataforma Yvis.

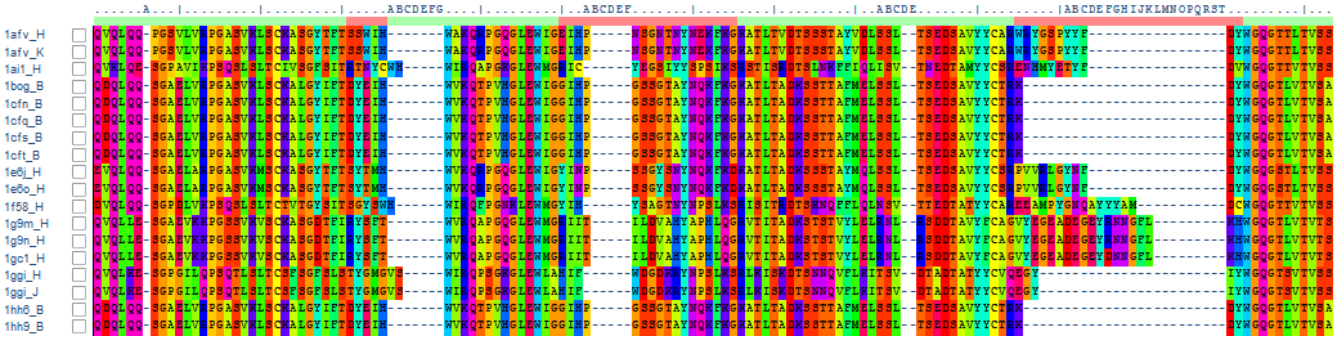


Figura 4.8: Alinhamento de cadeias pesadas de anticorpos de estruturas de HIV-1 realizado pelo abYsis. O esquema de numeração utilizado neste alinhamento é o Chothia (CHOTHIA; LESK, 1987). Fonte: análise utilizando a ferramenta disponível em http://www.abysis.org/

Durante a seleção das estruturas do PDB a serem analisadas na plataforma Yvis, foi utilizado um filtro de redundâncias de sequências, sendo inseridas no conjunto de análise somente aquelas que tivessem no máximo 99% de identidade entre elas. Isto significa que, sequências que não se diferenciavam por pelo menos duas posições não foram inseridas no conjunto de análise. O conjunto de sequências considerando esse filtro de redundâncias possui 180 cadeias (leves e pesadas). A partir desse resultado, novos filtros

puderam ser aplicados. Adicionando o filtro por cadeias pesadas e oriundas de anticorpos humanos obtém-se um conjunto com 74 cadeias. No entanto, esse conjunto possui anticorpos que se ligam a diferentes proteínas do HIV-1. Um dos possíveis alvos dos anticorpos contra HIV-1 é a glicoproteína gp120 do envelope exterior do vírus. Ao adicionar o filtro para moléculas de antígeno que fazem referência explícita a esta proteína, obtém-se 36 cadeias pesadas humanas. A representação destas cadeias gerada pelo Yvis é apresentada na Figura 4.9.

Observando-se a representação da Figura 4.9 é possível reconhecer posições com alto grau de conservação entre os anticorpos contra a proteína gp120 do HIV-1. Como exemplo, destacam-se as posições 21, 22, 26, 42, 43, 49, 94, 121, 124, 126 e 127, sendo que nenhuma destas posições fazem parte do conjunto de posições conservadas utilizadas para a aplicação do esquema de numeração do IMGT. Outra informação importante que se pode extrair da Figura 4.9 é em relação às estruturas que contêm informações de contato entre o antígeno e o anticorpo, obtidas da estrutura do PDB. Pode-se observar que todas as cadeias com informações de interação fazem contato com o antígeno na posição 64 e muitas vezes com outros resíduos do CDR2 (posições 56 a 65), além de alguns resíduos do CDR3 (posições de 105 a 117). Sabe-se que o anticorpo VRC01 realiza a maioria das interações com a proteína gp120 através do CDR2 da cadeia pesada, com mais da metade da superfície de interação pertencente a este CDR (ZHOU et al., 2010). Além disso, diversos anticorpos VRC01-*like* já foram isolados de pacientes infectados por HIV-1 e possuem alta capacidade de neutralização (SUN et al., 2017). Baseado nisso, é esperado que no PDB haja diversas estruturas relacionadas a essa classe de anticorpos.

Outra característica relevante dos anticorpos VRC01-*like* é sua origem gênica, eles são derivados do subgrupo IGHV1 (SUN et al., 2017). Analisando-se a tabela gerada pela plataforma Yvis com as informações complementares das cadeias, percebe-se que os anticorpos analisados têm como *germline*

alelos dos subgrupos IGHV1, IGHV3, IGHV4 e IGHV5. O subgrupo IGHV1 se destaca dos demais, com 22 sequências e os outros com 5, 6 e 3, respectivamente.

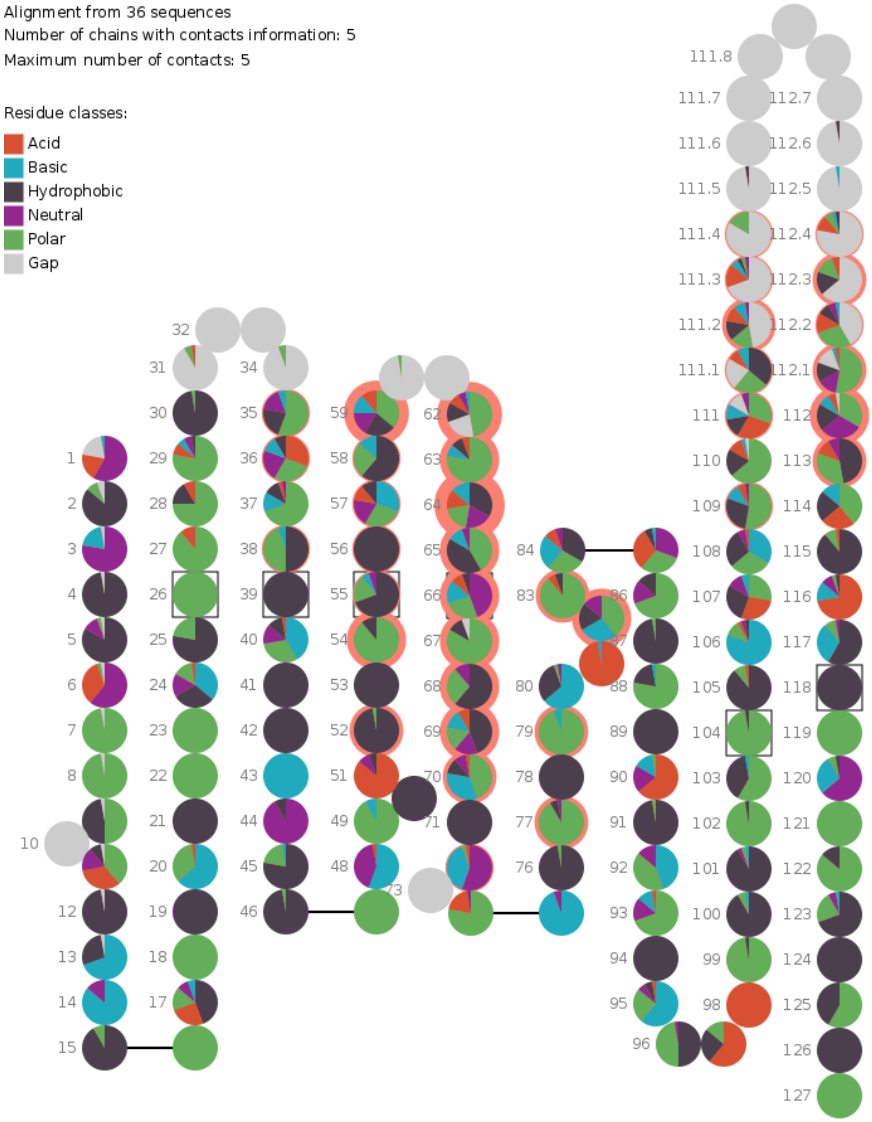


Figura 4.9: Representação do alinhamento de 36 cadeias pesadas humanas de anticorpos contra a proteína gp120 do HIV-1. A informação de contatos foi obtida a partir de 5 sequências.

Em fevereiro de 2.019, o número de estruturas depositadas de anticorpos anti-gp120 era maior que na época da realização deste estudo de caso. A Figura 4.10 apresenta o alinhamento de 119 cadeias pesadas de anticorpos anti-gp120, contra 36 anteriormente apresentadas. As observações feitas na análise anterior continuam válidas neste novo conjunto de dados, sendo que o maior número de estruturas em que foi possível analisar os contatos (114) confirmam o que foi anteriormente analisado com apenas 5 estruturas.

Além disso, analisando-se a tabela gerada pela plataforma Yvis contendo as informações complementares das cadeias, percebe-se que 72 cadeias analisadas têm como *germline* genes do subgrupo IGHV1 e que esse continua sendo o subgrupo mais frequente nos anticorpos anti-gp120.

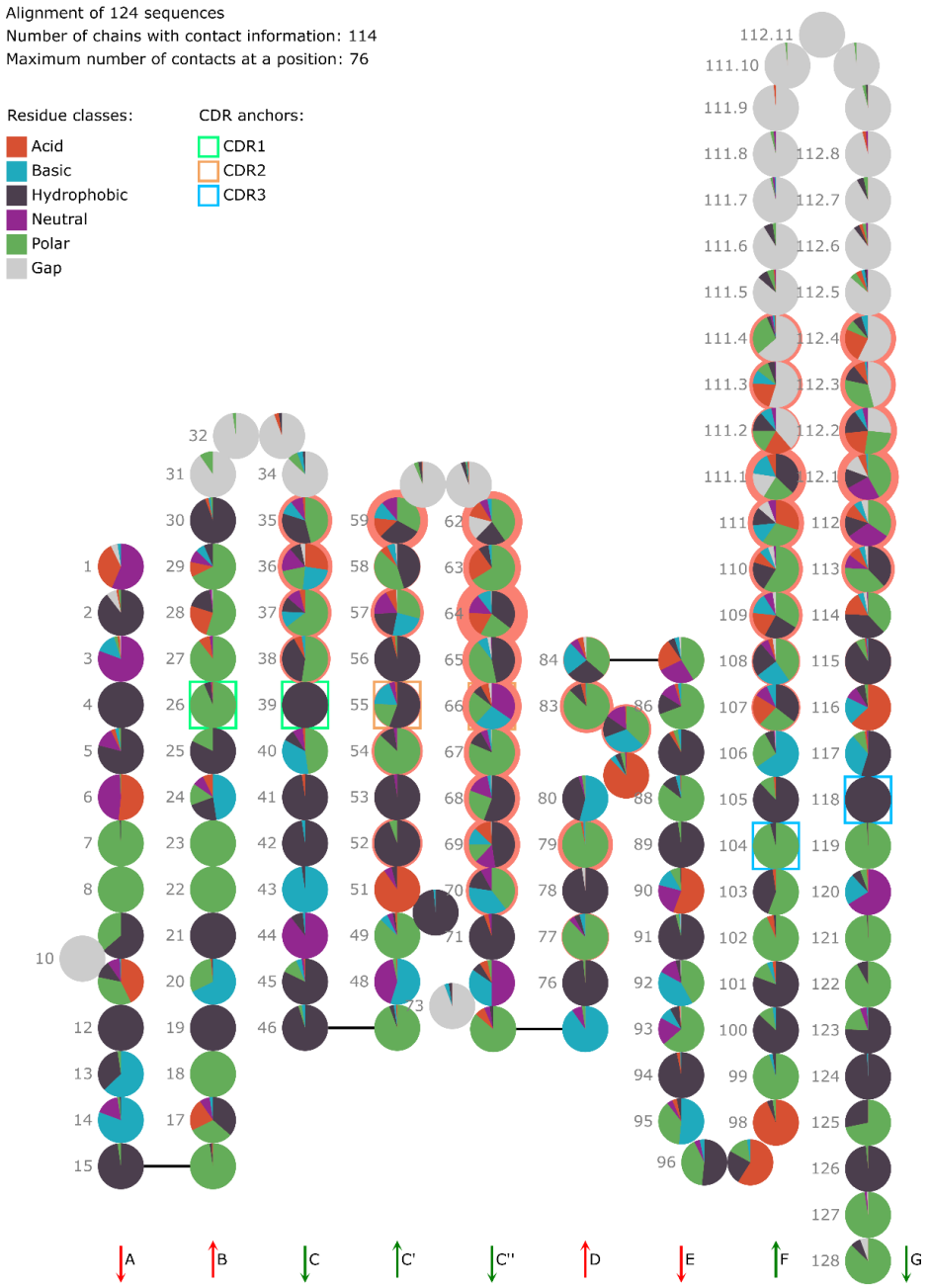


Figura 4.10: Representação do alinhamento de 124 cadeias pesadas humanas de anticorpos contra a proteína gp120 do HIV-1. A informação de contatos foi obtida a partir de 114 sequências.

4.2.4.2 Estudo de caso 2: Análise de dados de sequenciamento de repertório de paciente infectado com o vírus HIV-1

Nesse segundo estudo de caso, foi obtido a partir do SRA (*Sequence Read Archive*) (LEINONEN et al., 2011) um arquivo FASTA utilizando o número de acesso SRR1767440. Esse arquivo contém um conjunto de sequências de cadeias pesadas de anticorpos obtidas a partir de um estudo de sequenciamento de repertório de um paciente infectado pelo vírus HIV-1 (WU et al., 2015). O arquivo de sequências foi submetido ao IMGT/HighV-QUEST (ALAMYAR et al., 2012) e, após seu processamento, o IMGT/HighV-QUEST gerou diversos arquivos de análise das sequências. O arquivo contendo 478.047 sequências traduzidas e anotações da aplicação do esquema de numeração do IMGT foi submetido à plataforma Yvis que, por sua vez, processou o novo arquivo excluindo sequências com aminoácidos ambíguos, geralmente Ns, restando 330.800 sequências. A apresentação do alinhamento das sequências visualizado com o *Collier de Diamants* é exibida na Figura 4.11.

Como nesse paciente infectado foram encontrados ao menos três anticorpos neutralizantes derivado do alelo VH1-2*02, foi utilizado o filtro de genes *germline* nesse estudo de caso, o que restringiu a análise a 97.751 sequências. A visualização *Collier de Diamants* do alinhamento dessas sequências é apresentado na Figura 4.12. Utilizando o recurso de comparação de sequência com um alinhamento, disponível na plataforma Yvis, foi possível comparar as sequências com o alelo do qual elas foram geradas. Para isso, foi inserida uma sequência de comparação formada pelo alelo VH1-2*02 e *gaps* nas posições dos genes D e J. Por meio dessa visualização, é possível analisar o tamanho do CDR3 facilmente. A maioria das sequências analisadas possuem 2 inserções no CDR3 e menos que 10% delas têm mais que 4 inserções. Algumas posições, por exemplo, 36, 66, 92, 93, e 95, apresentam diversas sequências com uma classe de aminoácidos diferente da classe da sequência *germline*. Essa constatação é baseada na comparação dos maiores setores do gráfico de pizza que representam cada posição e a cor do círculo central que representa a classe do resíduo da sequência de comparação.

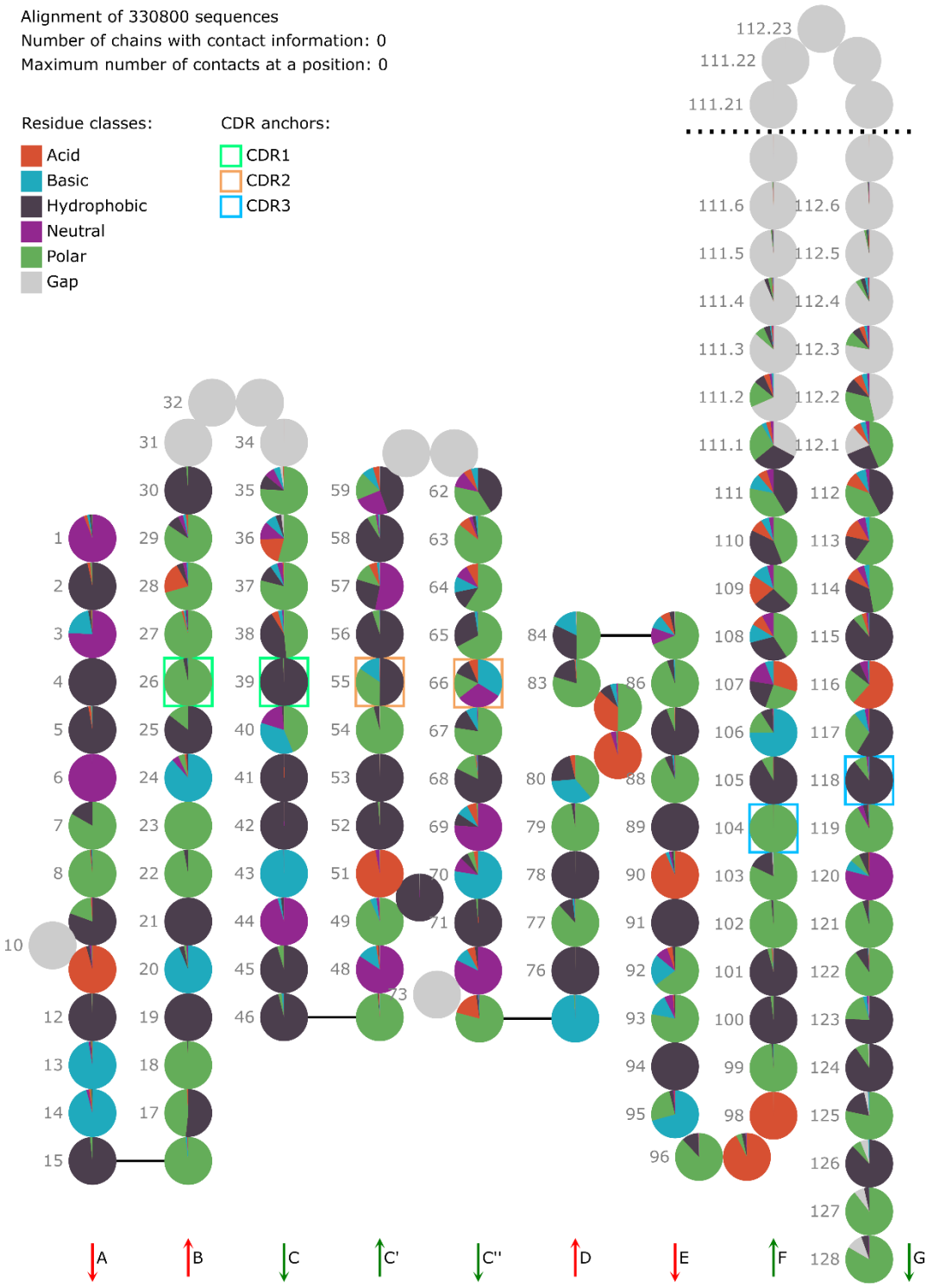


Figura 4.11: *Collier de Diamants* de 330.800 sequências de cadeias pesadas de anticorpos de um paciente infectado por HIV. Setas indicam as fitas dos domínios variáveis das cadeias e suas cores representam as duas folhas que formam esse domínio. As posições de ancoragem dos CDRs (verde-CDR1, laranja-CDR2 e azul-CDR3) são representadas por quadrados. As posições entre 111.7 e 111.20 e as posições entre 112.20 e 112.5 foram omitidas.

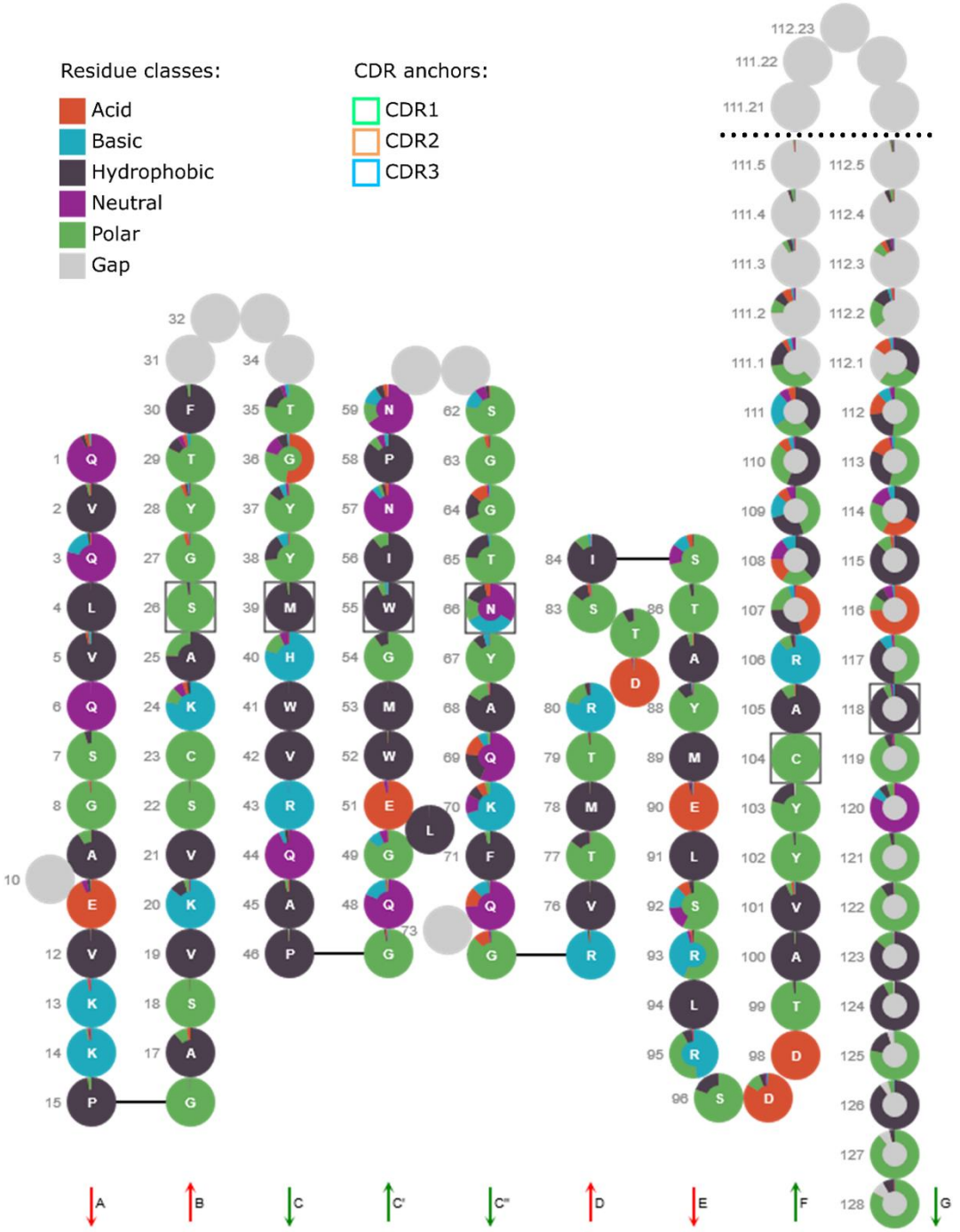


Figure 4.12: Representação *Collier de Diamants* de 97.751 sequências de cadeias pesadas de anticorpos derivadas do alelo IGHV1-2*02 de um paciente infectado por HIV. As posições entre 111.6 e 111.20 e as posições entre 112.20 e 112.6 foram omitidas. A sequência do IGHV1-2*02 seguida de *gaps* (nas posições referentes aos genes D e J) foi utilizada como sequência de comparação.

Para explorar com mais detalhes os dados analisados, é possível abrir o gráfico de barras com o detalhamento de conteúdo de cada posição. Por exemplo, no gráfico de barras da posição 36 (Figura 4.13), é possível observar que essa posição que, na sequência *germline*, continha uma glicina (G) sofreu mutações que levaram à codificação de ácido aspártico em aproximadamente metade das sequências analisadas. Como esse aminoácido possui carga negativa e é localizado em um CDR, provavelmente essa modificação é selecionada positivamente durante a expansão de células B para aumentar a afinidade do anticorpo ao antígeno.

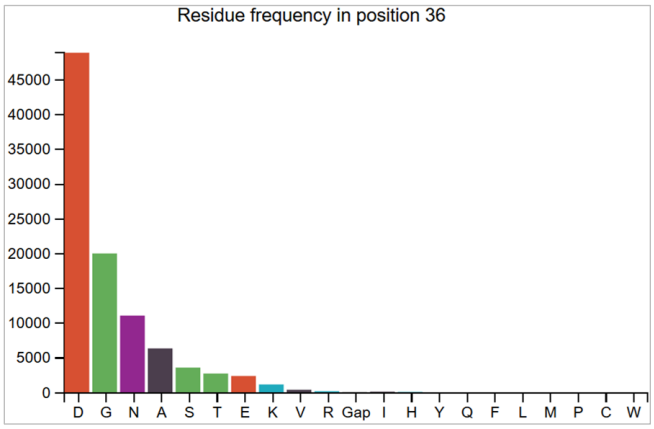


Figura 4.13: Gráfico de barras que apresenta a frequência de aminoácidos na posição 36 do alinhamento de 97.751 sequências de cadeias pesadas de anticorpos derivadas do alelo IGHV1-2*02 de um paciente infectado por HIV.

4.2.4.3 Estudo de caso 3: Visualização de cadeias pesadas de anticorpos neutralizantes anti-HIV-1

Wu e colaboradores fizeram uma análise do repertório de anticorpos de pacientes infectados por HIV-1 e posteriormente clonaram e expressaram alguns dos anticorpos para a testar a neutralização dos mesmos (WU et al., 2011). Após o sequenciamento do DNA complementar, obtido a partir de uma PCR projetada para amplificar RNAs originados de genes do subgrupo IGHV1, as sequências foram analisadas. No artigo, os autores apresentam na figura S18 a frequência dos aminoácidos do domínio variável de 22 cadeias pesadas dos anticorpos VRC01-*like* neutralizantes obtidos do doador 74. Essa figura, produzida com o WebLogo (CROOKS et al., 2004) é reproduzida na Figura 4.14. Nas representações produzidas

pelo WebLogo, a frequência dos aminoácidos em cada posição é determinada pela altura das letras e, como referência, foi colocada a sequência do alelo IGHV1-2*02, sendo que inserções referentes a esse alelo não foram incluídas na análise. Nessa figura, os aminoácidos em vermelho são os idênticos aos do gene IGHV1-2*02.

Figura 4.14: Frequência dos aminoácidos nas cadeias pesadas de anticorpos neutralizantes VRC01-*like* representada pelo WebLogo. A sequência do alelo IGHV1-2*02 é apresentado abaixo da representação. Aminoácidos em vermelho são idênticos aos aminoácidos do alelo. Fonte: (WU et al., 2011).

Apesar de indicar que foram utilizadas 22 sequências para a análise da Figura 4.14, o artigo apresenta 24 sequências neutralizantes para o doador 74 nos dados suplementares. Essas sequências foram alinhadas pelo IMGT/DomainGapAlign (EHRENMANN; KAAS; LEFRANC, 2010) já que o ANARCI apresentou inserções não usuais ao tentar alinhá-las. No entanto, ao submeter a página de resultados contendo as 24 sequências à plataforma Yvis, foi percebido visualmente no *Collier de Diamants* que três sequências pareciam estar "desalinhadas" em relação às outras. Isto pode ser observado quando diversas posições conservadas possuem poucas sequências divergentes e, normalmente, essas sequências têm classe igual a posições vizinhas. Tal constatação foi confirmada inspecionando-se os aminoácidos das sequências alinhadas geradas pelo IMGT/DomainGapAlign. Portanto, essas sequências foram removidas da análise. A Figura 4.15 apresenta a visualização *Collier de Diamants* para as 21 sequências comparadas com o alelo IGHV1-2*02. Diferentemente da Figura 4.14, todo o domínio variável é exibido e, como o

alelo IGHV1-2*02 corresponde somente a posições dos *frameworks* 1, 2 e 3 e aos CDRs 1 e 2 e uma pequena porção do CDR3, as demais posições foram preenchidas com *gaps* na sequência de comparação.

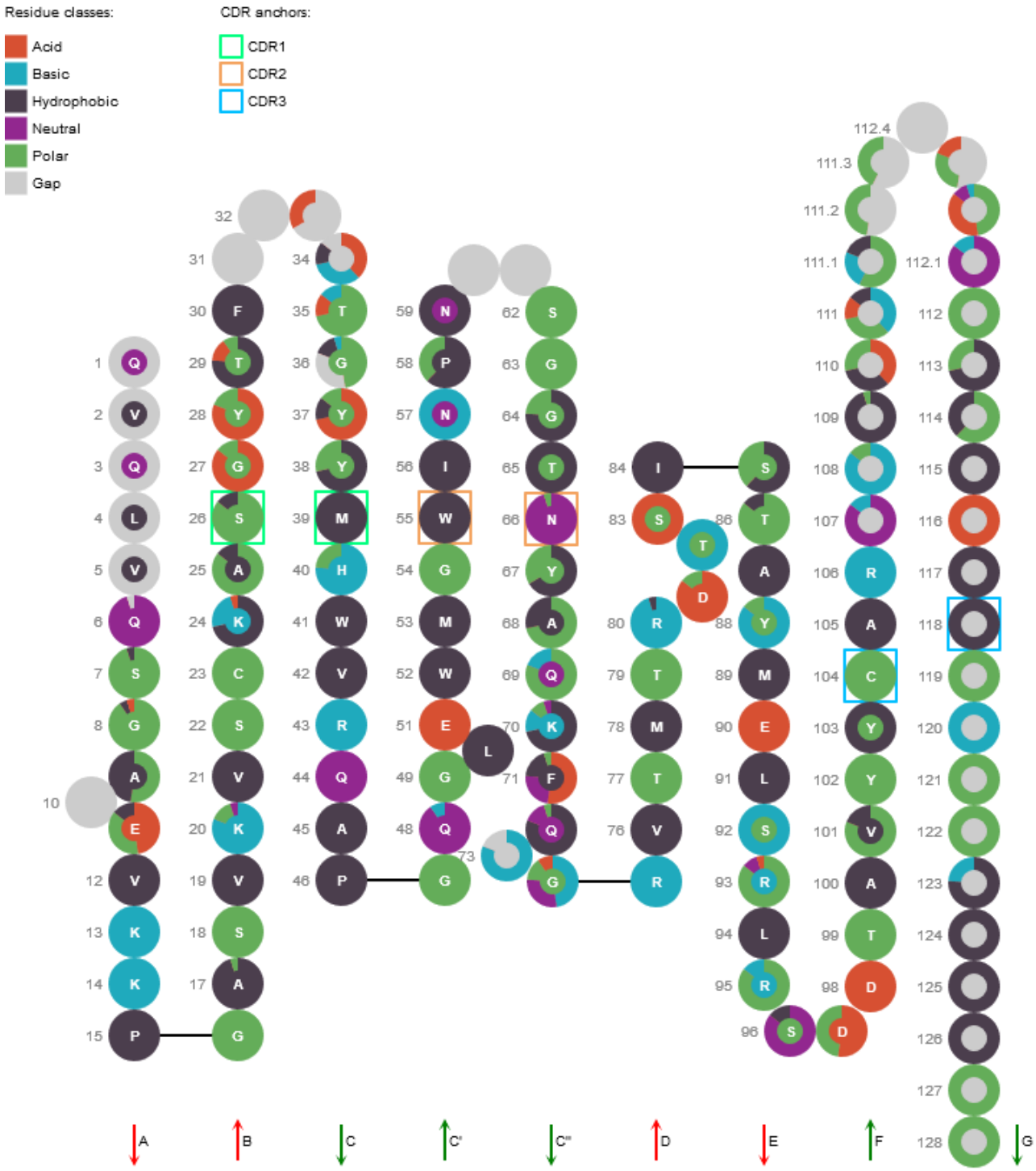


Figura 4.15: Representação de 21 sequências de cadeias pesadas de anticorpos produzidos pelo doador 74 (WU et al., 2011) comparadas com a sequência do alelo IGHV1-2*02. *Gaps* foram inseridos no restante da sequência de comparação, substituindo os genes D e J.

Observando a diferença do círculo central com o conteúdo de cada posição na Figura 4.15, é possível perceber que diversas posições dos anticorpos analisados foram modificadas no processo de expansão clonal das células B, distanciando-se da sequência do alelo IGHV1-2*02. Essa observação confirma a informação de que os anticorpos neutralizantes contra a proteína gp120 possuem um alto grau de mutações somáticas (WU et al., 2011).

Além disso, é possível perceber diversas posições conservadas entre as sequências, muitas delas sendo comuns às posições conservadas das estruturas do PDB analisadas anteriormente (Figura 4.10). Outra informação relevante que pode ser obtida comparando-se as Figuras 4.9, 4.10 e 4.15 é o fato de que muitas das posições em que os aminoácidos foram alterados em relação à sequência *germline*, são posições que estão envolvidas diretamente na interação com o antígeno (círculos exteriores aos gráficos de setores nas Figuras 4.9 e 4.10) das estruturas disponíveis no PDB analisadas. Essa característica é mais um indício de que as hipermutações somáticas ocorridas durante o processo de maturação dos anticorpos, juntamente com a seleção clonal guiam a produção de anticorpos neutralizantes contra o antígeno.

Além da visão geral do alinhamento das sequências, a plataforma Yvis permite também a exploração de detalhes relacionados a cada posição do alinhamento. Como exemplo da utilidade desse recurso, pode-se analisar a posição 36 da Figura 4.15. Essa posição possui resíduos polares (classe representada pela cor verde), em quase metade das sequências. No entanto, o gráfico de detalhamento da posição 36 (Figura 4.16) mostra que diferentes aminoácidos polares estão presentes nessa posição, apesar de, isoladamente, nenhum dos aminoácidos superar o número de *gaps* presentes nela. A presença de diversas sequências com aminoácidos de uma mesma classe nessa posição indica que essa classe pode ter características importantes para a interação com o antígeno. A convergência de características de

anticorpos distintos, oriundos de um mesmo indivíduo ou de diversos indivíduos, é um sinalizador de que características conservadas estão relacionadas à capacidade neutralizante desses anticorpos.

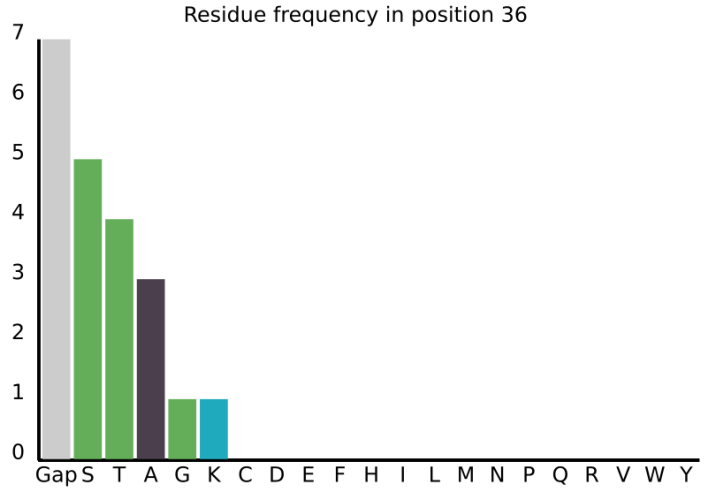


Figura 4.16: Gráfico de barras detalhando os aminoácidos presentes na posição 36 do alinhamento das 21 sequências de cadeias pesadas de anticorpos produzidos pelo doador 74 (WU et al., 2011).

4.3 Conclusões de visualização de alinhamento de sequências de anticorpos em larga escala

Com os estudos de caso apresentados anteriormente foi possível observar que a representação proposta neste trabalho, o *Collier de Diamants*, juntamente com a plataforma Yvis permitem a análise de domínios variáveis de cadeias de anticorpos baseada tanto em estruturas do PDB quanto em sequências obtidas pelo usuário. Além disso, os estudos de caso mostraram que a plataforma pode ser utilizada tanto em estudos de prospecção de estruturas de anticorpos presentes no PDB em que se deseja ter uma ideia do panorama geral das sequências presentes nessa base, quanto em análises de estruturas ou sequências específicas. A plataforma ainda permitiu a análise de características de conjuntos de sequências e a comparação desses conjuntos com outras sequências, nesse caso com sequências de genes *germlines*. Além disso, a plataforma implementada pode ser facilmente alterada para representar outras características das sequências e estruturas de anticorpos que podem ser associadas a posições definidas pelo esquema de numeração do IMGT.

A base de dados do Yvis é atualizada semanalmente, de maneira automática, e armazena dados obtidos do SAbDab e arquivos do PDB (identificadores da estrutura e das cadeias, nomes dos organismos produtores de antígeno e anticorpo e descrição da molécula), dados processados pelo IMGT/DomainGapAlign (sequências numeradas, alelos V e J atribuídos e os respectivos valores de identidade) e os possíveis contatos entre antígeno e anticorpo obtidos a partir do processamento das coordenadas da estrutura. Os nomes dos organismos produtores de antígeno e anticorpo são padronizados seguindo a taxonomia do UniProt, o que facilita as buscas na base de dados.

A plataforma Yvis permite a exploração de sua base de dados de estruturas, pesquisando, principalmente por tipo de antígeno, espécies produtora de antígeno ou anticorpo e genes *germline* atribuídos. Além disso, o usuário pode restringir o conjunto de dados analisados, evitando cadeias modificadas geneticamente, múltiplos anticorpos de uma mesma estrutura ou sequências de anticorpos com identidade superior a um *cutoff* definido durante a busca. Esses diferentes critérios de busca e restrições do conjunto de dados analisados não estão presentes em outros bancos de dados de estruturas de anticorpos como o abYsis (SWINDELLS et al., 2017), o IMGT/3Dstructure-DB (KAAS, 2004) e o SAbDab (DUNBAR et al., 2014).

A plataforma Yvis também fornece a opção de análise de dados fornecidos pelo usuário. Esses dados podem ser sequências de aminoácidos em formato FASTA ou numeradas pelo do IMGT/DomainGapAlign. Além disso, a plataforma permite a análise de dados gerados pelo IMGT/HighV-QUEST, que, geralmente, correspondem a milhares de sequências obtidas a partir de experimentos de sequenciamento de repertório de anticorpos.

A plataforma Yvis foi desenvolvida para facilitar a análise de um grande número de estruturas e sequências de anticorpos que já estão disponíveis e que ainda serão geradas. Ela apresenta uma nova visualização de alta densidade, o *Collier de Diamants*, que permite a visualização de um alinhamento de

dezenas a milhares de sequências de anticorpos em uma única representação, aproximando a sequência de aminoácidos à estrutura tridimensional do domínio variável da cadeia de anticorpos. A plataforma Yvis oferece um ambiente para a análise de sequências de anticorpos que auxilia na formulação de hipóteses sobre os principais resíduos e posições das sequências de anticorpos. Recentemente, a plataforma Yvis foi apresentada à comunidade científica por meio da publicação do artigo (CARVALHO; MOLINA; FELICORI, 2019).

5 Ydb: banco de dados de complexos antígeno-anticorpo

Após a análise dos bancos de dados com informações estruturais de anticorpos apresentados na seção 1.6.1, percebeu-se que nenhum deles permitiria a análise dos anticorpos presentes em estruturas disponíveis no PDB considerando diferentes definições de interface e tipos de interação. Além das informações oriundas das estruturas dos anticorpos, também seria interessante obter informações referentes às propriedades físico-químicas dos anticorpos e a genes *germline* e ainda, a aplicação de um esquema de numeração para facilitar a comparação das cadeias dos anticorpos.

Como nenhum dos bancos de dados apresentados na seção 1.6.1 possui todas as informações referentes aos complexos antígeno-anticorpo desejadas para a realização das análises propostas neste trabalho, optou-se pelo desenvolvimento de um banco de dados a partir de outros bancos de dados e ferramentas disponíveis. O banco de dados implementado, chamado *AntibodY DataBase* (Ydb), contém as informações organizadas de maneira a facilitar o acesso para a realização de diversas análises. O Ydb é um banco de dados implementado para reunir informações físico-químicas e de interações de complexos antígeno-anticorpo com estrutura disponível no PDB, onde o antígeno é uma proteína ou peptídeo.

5.1 Dados disponíveis no Ydb

Os principais dados disponíveis no Ydb referentes às estruturas disponíveis no PDB que contêm complexo antígeno-anticorpo estão esquematizados na Figura 5.1 e os métodos para extração dessas informações são detalhados na seção 5.2. Cada estrutura do PDB analisada possui pelo menos um complexo formado por pelo menos uma cadeia do anticorpo e uma cadeia do antígeno. No entanto, algumas estruturas possuem mais de um complexo, já que um anticorpo pode ter mais de uma região de contato com o antígeno ou pode existir mais de um anticorpo em uma mesma estrutura. Cada cadeia, seja ela de um antígeno ou de um anticorpo, é formada por diversos resíduos de aminoácidos, que, por sua vez, são compostos por átomos. Para cada estrutura armazenada no Ydb, são extraídos os seguintes dados do

arquivo do PDB: uma breve descrição das moléculas presentes, a data de depósito da estrutura, o tipo de experimento utilizado na obtenção da estrutura, sua resolução, as medidas de qualidade do modelo (*r-value* e *r-free*), os autores e o título da publicação científica relacionada à estrutura.

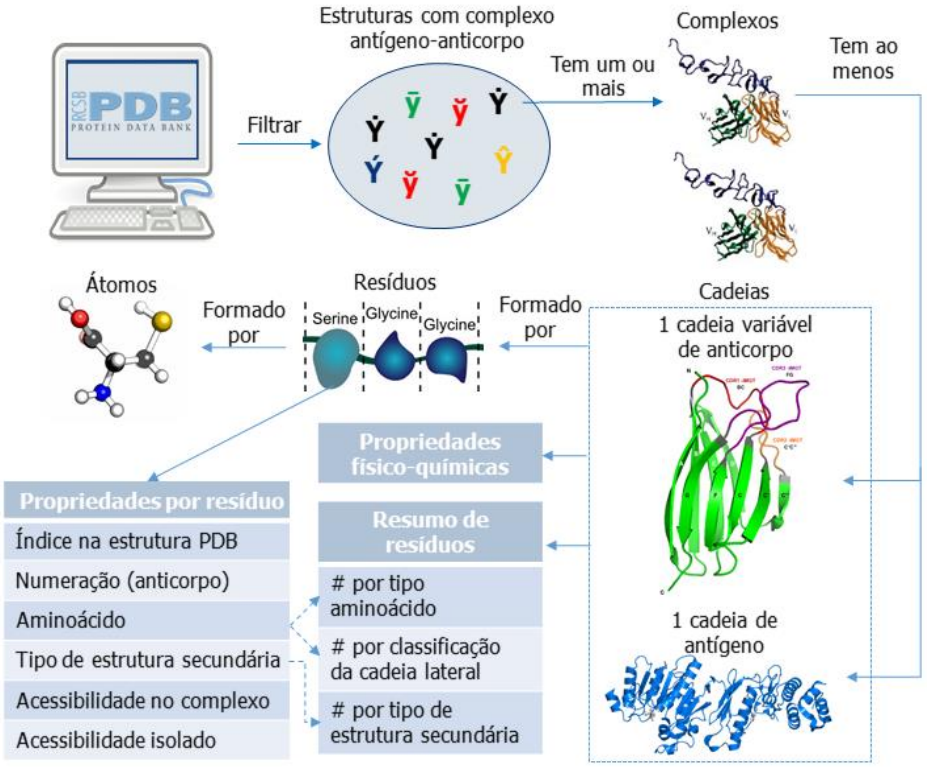


Figura 5.1: Hierarquização dos dados armazenados no Ydb e principais propriedades armazenadas para cadeias e resíduos.

Para cada cadeia armazenada no Ydb, seja ela oriunda de um antígeno ou anticorpo, são armazenados o nome padronizado do organismo produtor da cadeia (como na plataforma Yvis), seu tipo (antígeno ou, se oriunda de um anticorpo, cadeia leve ou pesada), a sequência de aminoácidos da cadeia definida pelos depositantes da estrutura no PDB (campos SEQRES do arquivo PDB) e a sequência reconhecida na estrutura tridimensional (obtida através dos campos ATOM). Algumas vezes, essas duas sequências são diferentes, pois, em algumas estruturas, nem todos os resíduos são reconhecidos. Quando existe uma divergência nessas duas sequências, é armazenada uma mensagem de alerta no Ydb. Para cadeias provenientes de anticorpos, são armazenados os alelos correspondentes aos genes *germline* V e J

relacionados às suas sequências e os respectivos valores de identidade. Além disso, para cada cadeia (Figura 5.1), são calculadas e armazenadas as seguintes propriedades físico-químicas: hidropatia (definida pelo GRAVY - *GRand AVerage of hYdropathy*), ponto isoelétrico teórico, índice alifático e massa molecular. A partir de todos os resíduos que compõem cada cadeia, são armazenados os seguintes valores: número de aminoácidos de cada tipo, número de aminoácidos de cada classe de cadeia lateral (classes apresentadas na Tabela 5.1) e número de aminoácidos de cada classe de estrutura secundária (hélices, pontes, fitas e voltas).

Tabela 5.1: Classificação dos aminoácidos quanto à cadeia lateral. Mesmos critérios utilizados em VIART et al., 2016.

Classe do aminoácido	**Aminoácidos**
Álcool	Serina e treonina
Aromático	Fenilalanina, tirosina e triptofano
Hidrofóbico	Cisteína, isoleucina, leucina, metionina, prolina e valina
Negativo	Ácido aspártico e ácido glutâmico
Pequeno	Alanina e glicina
Polar	Asparagina e glutamina
Positivo	Arginina, histidina e lisina

Para cada resíduo de aminoácido de cada cadeia, são armazenadas as seguintes informações: classe de estrutura secundária a qual ele pertence, a área do resíduo acessível ao solvente no complexo antígeno-anticorpo e a área calculada em uma estrutura que contém somente a molécula do antígeno ou do anticorpo. Além disso, para resíduos do anticorpo, é armazenada a posição do resíduo segundo o esquema de numeração do IMGT (LEFRANC et al., 2003).

Cada complexo antígeno-anticorpo é definido pelas cadeias leve e pesada do anticorpo, sendo que algumas estruturas possuem apenas uma das cadeias, e por uma ou mais cadeias de antígeno. Além dessas informações, são armazenados para cada complexo, se disponível, dados experimentais de cálculo de afinidade. Esses dados compreendem o método utilizado para a análise de afinidade, o valor de afinidade,

a variação na energia livre de Gibbs (ΔG), a temperatura em que a afinidade foi coletada e o identificador da publicação (PMID - *PubMed Identifier*) referente a essa análise no PubMed (NCBI, 2019).

As possíveis interações realizadas entre o antígeno e o anticorpo são classificadas no Ydb como interações de hidrogênio (mediadas ou não por moléculas de água), eletrostáticas, aromáticas, cátion-π, Van der Waals, Van der Waals *clash* e contatos hidrofóbicos. O método utilizado para o cálculo de cada tipo de interação é descrito na seção 5.2. No Ydb, cada par de átomos, sendo um do antígeno e outro do anticorpo, que realiza uma interação é armazenado juntamente com o tipo de interação calculada, bem como a distância entre esses átomos. Além disso, para diminuir a complexidade da análise de interações de cada par de átomos, os dados de interação interatômica foram agrupados considerando os resíduos que interagem. Assim, para cada par de resíduos que possuem átomos que interagem entre si, são armazenados no Ydb o par de resíduos e os tipos e números de interações realizadas por esse par.

Devido às diferentes maneiras existentes para se determinar os resíduos que fazem parte da interface de um complexo antígeno-anticorpo, como apresentado na seção 1.7, para a definição das interfaces dos complexos armazenados no Ydb foram utilizados três diferentes métodos: distância entre um átomo de um resíduo do anticorpo e um átomo de um resíduo do antígeno, variação na área de acessibilidade ao solvente dos resíduos quando anticorpos e antígenos são analisados separadamente e em complexo e, por último, as possíveis interações químicas realizadas entre o anticorpo e o antígeno considerando os tipos de átomos envolvidos e, principalmente, a distância entre eles. Para os dois primeiros métodos, diferentes valores de *cutoff* foram analisados. As distâncias máximas permitidas para que os átomos fossem considerados parte da interface foram de 3, 4, 5 e 6Å, na definição que utiliza a distância como determinante da interface. Para a definição baseada na variação da área de acessibilidade ao solvente foram consideradas: qualquer valor de variação, superior a $10Å^2$ e superior a $50Å^2$. A partir dos critérios anteriormente definidos e seus valores de *cutoff*, para cada complexo antígeno-anticorpo

foram calculadas oito interfaces. Os diferentes métodos utilizados na definição de interface presentes no Ydb permitem a análise do impacto da escolha desse método na caracterização de paratopos e epitopos. Independentemente do método utilizado, para cada interface calculada, foram armazenados os resíduos que as compõem.

A partir dos resíduos presentes em cada interface e das cadeias às quais eles pertencem, foram definidas as regiões de paratopo e epitopo. Para cada epitopo ou paratopo, é gerada uma sequência contendo os resíduos da interface correspondente à cadeia a qual eles pertencem, na mesma ordem em que são encontrados na estrutura primária de cada cadeia. Caso haja descontinuidades, isto é, o paratopo ou epitopo não contém todos os resíduos de um intervalo da cadeia, os resíduos que não fazem parte da interface são ignorados. A partir das sequências definidas para epitopos e paratopos, são calculadas as mesmas propriedades físico-químicas calculadas para as cadeias: hidropatia (definida pelo GRAVY), ponto isoelétrico teórico, índice alifático e massa molecular. Além disso, a partir dos resíduos que compõem essas regiões, são armazenadas as mesmas contagens realizadas para as cadeias: número de aminoácidos de cada tipo, número de aminoácidos de cada classe de cadeia lateral e número de aminoácidos de cada classe de estrutura secundária. Tais valores permitem a comparação de paratopos e epitopos de diferentes complexos e também a comparação dos resíduos que participam da interface de uma cadeia com o restante dessa cadeia. Além desses valores de comparação, para cada paratopo ou epitopo é calculado e armazenado o número de interações realizadas pelos resíduos dessa região.

5.2 Implementação do Ydb: Materiais e métodos

Para a coleta de dados e cálculo das informações armazenadas no Ydb, foi implementado um *script* em Python versão 2.7.6 que mantém uma conexão com um banco de dados relacional MySQL versão 5.5.52 onde os dados apresentados na seção anterior são armazenados. A Figura 5.2 apresenta o diagrama de relacionamento do banco de dados, contendo os atributos de cada tabela e seus relacionamentos.

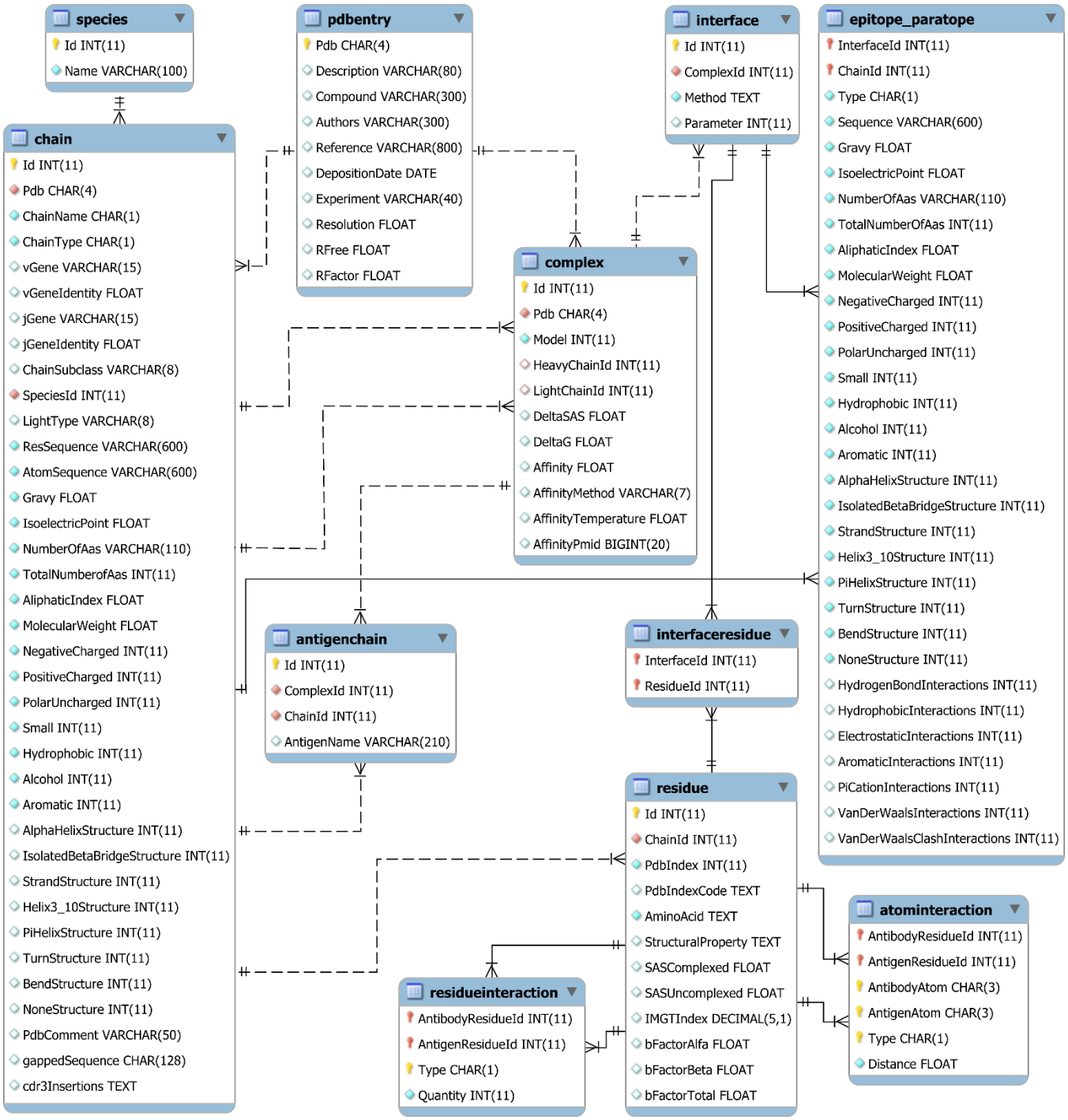


Figura 5.2: Diagrama de entidade e relacionamento do banco de dados Ydb, representando todas as tabelas, atributos e relacionamentos. Diagrama construído pela ferramenta MySQL Workbench 6.3. Os relacionamentos refletem a hierarquização dos dados armazenados no Ydb, como apresentado na seção anterior.

A Figura 5.3 apresenta um diagrama com as principais informações processadas e geradas pelo *script* e os principais *softwares* e bibliotecas utilizados. Inicialmente, o *script,* utilizando o pacote

Mechanize versão 0.2.5 (LEE, 2017), acessa o banco de dados SAbDab (apresentado na seção 1.6.1), da mesma maneira que o *script* da plataforma Yvis, gerando um arquivo resumo. Desse arquivo são extraídas as informações referentes às cadeias da estrutura que formam um complexo, a identificação das cadeias leve, pesada e do antígeno, o pareamento dessas cadeias, os dados referentes ao experimento que gerou a estrutura e, se disponível, os dados de análise de afinidade dos complexos antígeno-anticorpo.

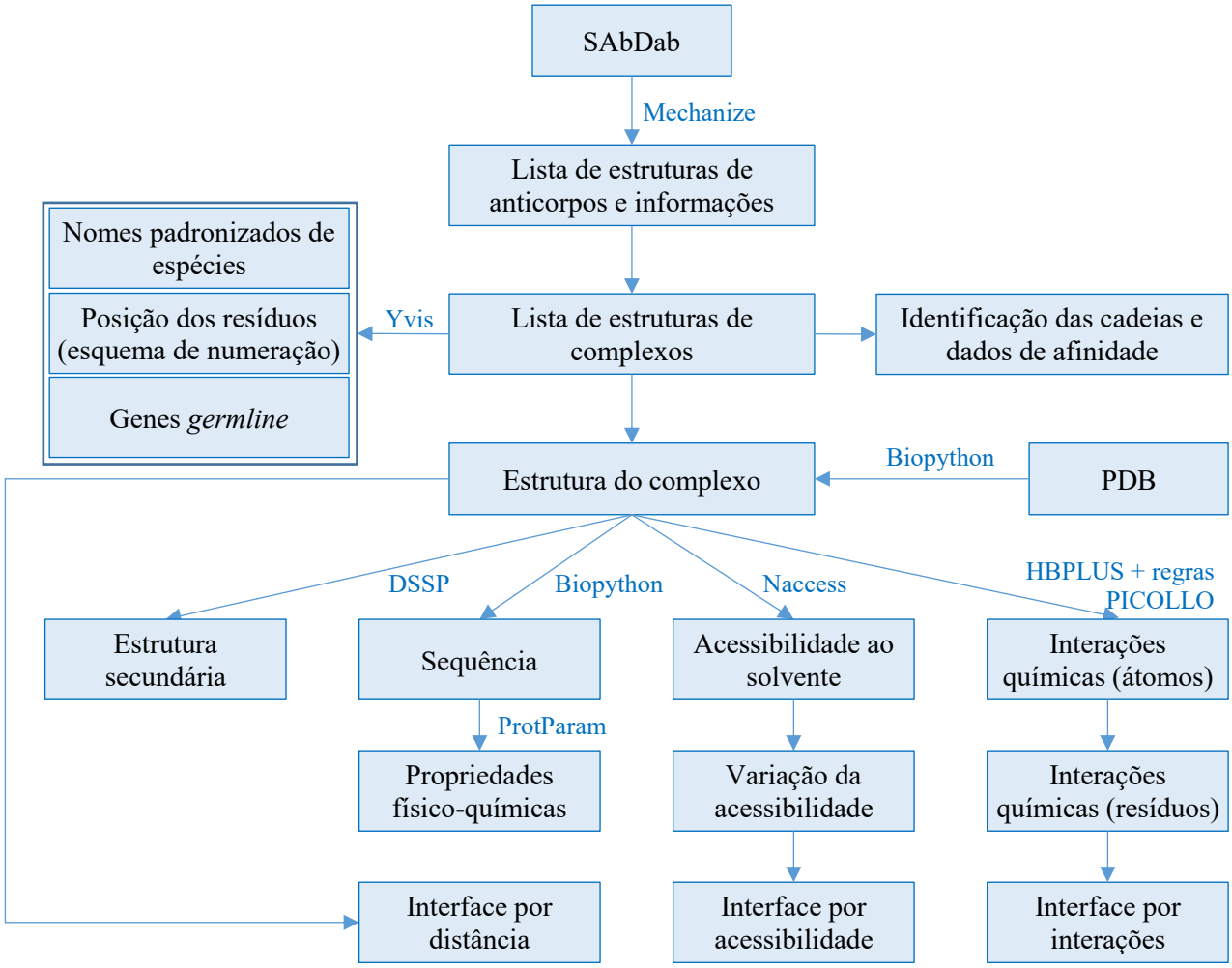


Figura 5.3: Diagrama representando as principais informações processadas e geradas pelo *script* do Ydb. Os principais softwares externos e bibliotecas utilizados são destacados com fonte azul.

A partir das informações disponíveis no arquivo de resumo do SAbDab, o *script* verifica se existem novos complexos entre anticorpos e moléculas proteicas (proteínas ou peptídeos), que não estão armazenados no Ydb. Caso algum complexo não esteja armazenado no Ydb, o *script* implementado acessa

o banco de dados da plataforma Yvis, apresentada no capítulo anterior e obtém os nomes padronizados das espécies dos organismos produtores do antígeno e do anticorpo de cada complexo. Além disso, o *script* obtém as informações armazenadas no banco de dados da plataforma Yvis a partir do processamento da página de resultados do IMGT/DomainGapAlign (EHRENMANN; KAAS; LEFRANC, 2010): a sequência com *gaps* do domínio variável de cada cadeia gerada pela aplicação do esquema de numeração do IMGT (LEFRANC et al., 2003), os alelos dos genes *germline* V e J atribuídos às regiões variáveis das cadeias e os respectivos valores de identidade entre o *germline* e a sequência da cadeia. A partir da sequência com *gaps*, o *script* do Ydb deduz o índice de cada resíduo da cadeia no esquema de numeração.

Além das informações obtidas da plataforma Yvis, o *script* busca a estrutura a ser analisada no servidor do PDB e analisa o arquivo da estrutura utilizando a biblioteca Biopython (COCK et al., 2009) versão 1.66, principalmente o módulo Bio.PDB (HAMELRYCK; MANDERICK, 2003). Para a obtenção das informações referentes à estrutura secundária das cadeias analisadas, o programa DSSP (KABSCH; SANDER, 1983) é executado, a partir da biblioteca Biopython. Essa biblioteca, executa o DSSP e processa seu arquivo de saída, associando a cada resíduo de uma cadeia a estrutura secundária da qual ele faz parte.

O *script* implementado calcula as propriedades físico-químicas (GRAVY, ponto isoelétrico teórico, índice alifático e massa molecular) das cadeias de anticorpos e antígenos armazenadas no Ydb, bem como das regiões de paratopo e de epitopo, respectivamente, utilizando a ferramenta ProtParam do portal ExPASy (WILKINS et al., 1999) acessada pelo módulo SeqUtils da biblioteca Biopython.

Para a obtenção da área de acessibilidade ao solvente de cada resíduo de cada cadeia, o *script* executa o programa Naccess (HUBBARD; THORTON, 1996) por meio da biblioteca Biopython, que faz a análise do arquivo de saída do programa. O Naccess é executado três vezes tendo como entrada três diferentes arquivos: o arquivo original do PDB que contém o complexo antígeno-anticorpo e outros dois arquivos gerados pelo *script,* utilizando o módulo Bio.PDB (HAMELRYCK; MANDERICK, 2003). Um

arquivo contém somente a estrutura do anticorpo e o outro somente a estrutura do antígeno. Por meio do processamento dos arquivos pelo Naccess, o *script* armazena os valores da área de acessibilidade ao solvente de cada resíduo na estrutura com o complexo antígeno-anticorpo e na estrutura isolada, contendo somente o antígeno ou o anticorpo.

Como não foi encontrado um banco de dados que disponibilizasse o cálculo de interações químicas de estruturas do PDB atualizadas na mesma frequência que o SAbDab, foi implementado, no *script* do Ydb, o cálculo das interações entre antígenos e anticorpos. Para isso, foi utilizada a mesma estratégia do banco de dados PICCOLO (*Protein Interaction Collection Online*) (BICKERTON; HIGUERUELO; BLUNDELL, 2011). Inicialmente, átomos de diferentes cadeias que estão a uma determinada distância são selecionados e, para cada par de átomos, critérios de distância, tipo de átomo e, em alguns casos, ângulos entre os átomos são utilizados. Para a primeira parte do processamento, no *script* do Ydb, foi utilizado o módulo NeighborSearch (HAMELRYCK; MANDERICK, 2003) da biblioteca Biopython. Esse módulo, a partir de um átomo e de um valor de distância, retorna todos os átomos de uma cadeia que estão a uma distância máxima do átomo dado como entrada. Esse valor é passado como parâmetro para a função NeighborSearch. Portanto, no *script* do Ydb, para cada átomo pertencente às cadeias do anticorpo, foram selecionados todos os átomos das cadeias de antígeno que estão a, no máximo, 6,05Å de distância de um átomo do anticorpo (BICKERTON; HIGUERUELO; BLUNDELL, 2011). Após a seleção dos átomos candidatos, critérios específicos para cada tipo de interação são avaliados. Esses critérios e os tipos de interações analisadas estão resumidos na Tabela 5.2.

Para a definição de interações de Van der Waals, considerou-se que um par de átomos a uma distância inferior à soma dos raios de Van der Waals desses átomos mais 0,5Å realiza esse tipo de interação. No entanto, caso a distância seja inferior à soma dos raios, a interação é identificada como Van der Waals *clash*. Os valores dos raios de Van der Waals para cada átomo utilizados nesta implementação

são os mesmos utilizados no PICCOLO (BICKERTON; HIGUERUELO; BLUNDELL, 2011). A definição dos demais tipos de interação levam em consideração não somente o critério de distância, mas também o tipo de átomo. A Tabela 5.3 apresenta a classificação dos átomos dos resíduos que realizam interações eletrostáticas, aromáticas, cátion-π e contatos hidrofóbicos utilizada. Interações hidrofóbicas são definidas quando dois átomos são considerados hidrofóbicos e eles estão a uma distância inferior a 5Å. Interações eletrostáticas são definidas, simplificadamente, a partir da carga formal de um grupamento considerando um ambiente com pH 7. Caso um átomo classificado como catiônico esteja a uma distância inferior a 6Å de um átomo classificado como aniônico, considera-se que eles realizam uma interação eletrostática. Para a definição de interações aromáticas, outra simplificação foi realizada. Os ângulos formados pelos planos dos anéis não foram analisados e somente o critério de distância (inferior a 6Å) e a presença do anel são considerados. As interações cátion-π são definidas quando um átomo catiônico está a uma distância inferior a 6Å de um átomo aromático (BICKERTON; HIGUERUELO; BLUNDELL, 2011).

Para a análise das interações de hidrogênio, foi utilizado o programa HBPLUS (MCDONALD; THORNTON, 1994), considerando como parâmetros de distância os mesmos utilizados como padrão na execução do programa LigPlot+ (LASKOWSKI; SWINDELLS, 2011): distância máxima entre o átomo de hidrogênio e o átomo aceptor de 2,70Å e entre os átomos doador e o aceptor de 3,35Å. O HBPLUS, além de analisar os tipos de átomos envolvidos na interação e a distância entre eles, também analisa os ângulos formados pelos átomos. O *script* do Ydb executa o HBPLUS, processa sua saída e coleta as interações entre átomos de antígenos e anticorpos. Além disso, o *script* analisa as interações entre os átomos de antígenos e anticorpo e as moléculas de água, pois elas formam redes que ajudam a interface, fazendo com que antígeno e anticorpo sejam complementares, estabilizando o complexo (BHAT et al., 1994; NGUYEN et al., 2017). Para isso, o *script* analisa as moléculas de água que interagem com cadeias

de antígeno e anticorpo, simultaneamente. Também são analisadas as interações entre 2 moléculas de água e, se cada uma delas faz interação com as cadeias do antígeno e do anticorpo, essa interação também é armazenada. A partir da implementação das regras descritas anteriormente, foi possível calcular as possíveis interações que ocorrem em cada complexo antígeno-anticorpo e armazená-las no Ydb. Essas informações, em nível atômico, foram generalizadas no nível de resíduos, sendo também armazenados no Ydb os pares de resíduos que as realizam, além dos pares de átomos que realizam interações.

Tabela 5.2: Critérios utilizados para a definição de interações entre átomos (a_i e a_j) das cadeias do antígeno e do anticorpo, baseados em (BICKERTON; HIGUERUELO; BLUNDELL, 2011). A distância entre os átomos a_i e a_j é representada por $d(a_i , a_j)$ e o raio de Van der Waals de um átomo é representado por vdw(a). A_{cc} representa o átomo aceptor em uma interação de hidrogênio e a_H o átomo de hidrogênio. O tipo de um átomo de um resíduo é definido na Tabela 5.3.

Tipo de interação	**Tipo do átomo a_i**	**Tipo do átomo a_j**	**Critério de distância**
Van der Waals	Qualquer	Qualquer	$d(a_i, a_j) < vdw(a_i)+vdw(a_j)+0,5Å$
Van der Waals *clash*	Qualquer	Qualquer	$d(a_i, a_j) < vdw(a_i)+vdw(a_j)$
Contato hidrofóbico	Hidrofóbico	Hidrofóbico	$d(a_i, a_j) < 5,0Å$
Eletrostática	Catiônico	Aniônico	$d(a_i, a_j) < 6,0Å$
Aromática	Aromático	Aromático	$d(a_i, a_j) < 6,0Å$
Cátion-π	Catiônico	Aromático	$d(a_i, a_j) < 6,0Å$
Interação de hidrogênio*	Doador	Aceptor	$d(a_i, a_j) < 3,9Å$ e $d(a_H, a_{acc}) < 2,5Å$
Interação de hidrogênio mediada por água*	Doador ou aceptor	Doador ou aceptor	$d(a_i, a_j) < 3,9Å$ e $d(a_H, a_{acc}) < 2,5Å$

*As interações de hidrogênio possuem um critério adicional às demais relacionado ao ângulo formado pelos átomos.

Para a caracterização das interfaces dos complexos antígeno-anticorpo, os resíduos das cadeias do anticorpo e do antígeno, participantes da interface, são definidos como resíduos do paratopo e do epitopo, respectivamente. No entanto, apesar da definição de epitopo e paratopo englobar todas as cadeias de antígeno e anticorpo, na implementação do Ydb, optou-se por manter as informações sobre essas regiões

separadas por cadeias, permitindo assim a comparação dessas regiões específicas com as cadeias que as contêm.

Tabela 5.3: Classificação dos átomos dos resíduos que podem realizar interações eletrostáticas, aromáticas, cátion-π e contatos hidrofóbicos, utilizada neste trabalho baseada em (BICKERTON; HIGUERUELO; BLUNDELL, 2011). Os resíduos que não possuem átomos que realizam esses tipos de interação não foram exibidos.

Resíduo	Átomos de Carbono	Classificação	Demais átomos	Classificação
Alanina	β	Hidrofóbico	-	-
Arginina	β e γ ζ	Hidrofóbico Catiônico	Nitrogênio ε, η1 e η2	Catiônico
Asparagina	β	Hidrofóbico	-	-
Ácido aspártico	β γ	Hidrofóbico Aniônico	Oxigênio δ1 e δ2	Aniônico
Cisteína	β	Hidrofóbico	-	-
Glutamina	β e γ	Hidrofóbico	-	-
Ácido glutâmico	β e γ δ	Hidrofóbico Aniônico	Oxigênio ε1 e ε2	Aniônico
Histidina	β δ2, ε1 e γ	Hidrofóbico Aromático	Nitrogênio δ1 e ε2	Aromático
Isoleucina	β, δ1, γ1 e γ2	Hidrofóbico	-	-
Leucina	β, δ1, δ2 e γ	Hidrofóbico	-	-
Lisina	β, δ e γ	Hidrofóbico	Nitrogênio ζ	Catiônico
Metionina	β, ε e γ	Hidrofóbico	Enxofre δ	Hidrofóbico
Fenilalanina	β δ1, δ2, ε1, ε2, γ e ζ	Hidrofóbico Aromático	-	-
Prolina	β e γ	Hidrofóbico	-	-
Treonina	γ2	Hidrofóbico	-	-
Triptofano	β δ1, δ2, ε2, ε3, γ, η2, ζ2 e ζ3	Hidrofóbico Aromático	Nitrogênio ε1	Aromático
Tirosina	β δ1, δ2, ε1, ε2, γ e ζ	Hidrofóbico Aromático	-	-
Valina	β, δ1, γ1 e γ2	Hidrofóbico	-	-

Para a definição da interface utilizando a distância entre átomos do antígeno e do anticorpo, o *script* analisa a distância entre todos os pares de átomos da estrutura, sendo um átomo do antígeno e outro

do anticorpo. Todos os pares que são encontrados com distância menor do que um *cutoff*, comparados como descrito no cálculo de interações, são considerados parte do epitopo ou do paratopo. Como o Ydb considera diferentes valores de distância (*cutoffs*) para definição de diversas interfaces, os resíduos que fazem parte de uma interface mais restritiva (menor valor de distância) automaticamente são inseridos nas interfaces menos restritivas.

Para a definição de interface baseada na variação na área de acessibilidade ao solvente, o *script* utiliza os valores já armazenados no Ydb calculados pelo Naccess, como descrito anteriormente. O *script* calcula diferença absoluta entre os valores da área de acessibilidade ao solvente de cada resíduo das cadeias da estrutura com o complexo antígeno-anticorpo e de das cadeias das estruturas que contêm essas moléculas isoladas. Os resíduos que têm a diferença na área de acessibilidade maior que os valores de *cutoff* são considerados como parte do epitopo ou paratopo. Novamente, os resíduos que fazem parte de uma interface mais restritiva (valor maior de variação) automaticamente são inseridos nas interfaces menos restritivas.

Para a definição de interface baseada nas interações entre resíduos foram utilizadas as interações químicas calculadas pelo *script* e armazenadas no Ydb. Dessa maneira, todo resíduo que realiza alguma interação faz parte dessa interface.

5.3 Comparação do Ydb com os demais bancos de dados com informações estruturais de anticorpos

A Tabela 5.4 apresenta os principais recursos disponibilizados pelo Ydb e os demais bancos de dados com informações estruturais de anticorpos apresentados na seção 1.6.1. Como pode-se observar nesse quadro, para as análises propostas neste trabalho, o Ydb é mais completo que os demais bancos.

Tabela 5.4: Principais recursos disponibilizados pelo Ydb e demais bancos de dados com informações estruturais de anticorpos apresentados na seção 1.6.1. Alguns bancos possuem não somente registros associados a estruturas, o que exige um filtro adicional nos resultados. A padronização dos nomes das espécies ou organismos associados à antígenos e anticorpos permite uma pesquisa mais exata para a análise, garantindo que todos os registros associados aos organismos analisados sejam retornados. As definições de CDRs são realizadas pelos diferentes esquemas de numeração descritos na seção 1.2. As definições de paratopo e epitopo e das interações entre as cadeias são bem distintas entre os bancos de dados. A possibilidade de análise conjunta de dados indica se é possível realizar a comparação dos dados de um grupo de cadeias ou complexos.

Banco de dados	**Dados somente de estruturas**	**Padronização dos nomes das espécies ou organismos**	**Informações de *germline***	**Definição de paratopo e epitopo**	**Definição das interações entre cadeias**	**Possibilidade de análise conjunta de dados**
abYsis	Não	Não	Sim	Não	Não	Tabela com campos variáveis e alinhamento de cadeias do anticorpo
AgAbDb (desatualizado)	Sim	Sim	Não	Sim	Van der Waals, ligação de hidrogênio, ponte salina e interação mediadas por água	Não
AppA	Sim	Sem informação	Não	Sim	Van der Waals, ligação de hidrogênio, interações hidrofóbicas, iônicas e mediadas por água	Não
IEDB	Não	Sim	Sim	Indicados no artigo original e calculados com critério de distância	Não	Tabela com campos fixos
IMGT/ 3Dstructure-DB	Sim	Não	Sim	Possibilidade de inserção de uma molécula de água entre um par de resíduos	Polar, ligação de hidrogênio, não polar e ponte dissulfeto	Não
SAbDab	Sim	Não	Sim	Não	Não	Não
Ydb	Sim	Sim	Sim	Diferentes critérios baseados em distância, variação na área de acessibilidade ao solvente e interações	Van der Waals, Van der Waals *clash*, contato hidrofóbico, eletrostática, aromática, cátion-π, interação de hidrogênio mediada ou não por água	Sim, atualmente, através de consultas SQL

5.4 Análise dos dados armazenados no Ydb

Em maio de 2.019, o Ydb armazenava dados de 3.785 complexos antígeno-anticorpo, correspondendo a 2.092 estruturas do PDB. A grande maioria dessas estruturas, totalizando 1.933, foram obtidas através de cristalografia, sendo as demais obtidas por meio de ressonância magnética nuclear (NMR - *Nuclear Magnetic Resonance spectroscopy*) ou microscopia eletrônica. Nas próximas seções alguns dados do Ydb coletados em maio de 2.019 serão analisados e os materiais e métodos utilizados nessa análise serão explicitados.

5.4.1 Análise do Ydb: Materiais e métodos

Inicialmente foram caracterizados os complexos armazenados no Ydb. Os atributos armazenados no Ydb, como espécie produtora de anticorpos, por exemplo, foram acessados diretamente através de consultas SQL (*Structured Query Language*). Para a análise da similaridade das cadeias de anticorpos armazenado no Ydb foi implementado um *script* em Python versão 2.7.6 que utiliza as bibliotecas Numpy versão 1.11.3 (VAN DER WALT; COLBERT; VAROQUAUX, 2011) e Pandas versão 0.23.4 para a manipulação dos dados, a biblioteca Scipy versão 1.2.1 (JONES et al., 2.019) para a aglomeração dos dados e as bibliotecas Seaborn versão 0.9.0 e Matplotlib versão 1.5.1 (HUNTER, 2007) para a visualização dos mesmos. Para a definição das matrizes de identidade foram realizadas comparações de todas as sequências de cadeias pesadas e todas leves, contra todas, separadamente. *Gaps* e *mismatches* foram contabilizados com o mesmo peso. A matriz de identidade dos anticorpos foi calculada a partir da matriz das cadeias individuais, levando em consideração o tamanho das cadeias. Uma matriz de distância foi calculada para cada matriz de identidade, onde o valor de uma posição definida pela linha i e a coluna j é definido como 1 menos o valor da posição definida pela mesma linha e coluna da matriz de identidade. Na fase de aglomeração dos dados, foi utilizado o algoritmo de aglomeração hierárquico de minimização

de variância de Ward, implementado na função linkage da biblioteca Scipy, sobre a matriz de distâncias. Os *heatmaps* foram construídos utilizando a função clustermap da biblioteca Seaborn.

Para minimizar o viés causado por estruturas idênticas, foi gerado um conjunto não redundante de complexos antígeno-anticorpo, sendo cada complexo formado por uma cadeia leve, uma pesada e uma ou mais cadeias proteicas de antígeno. Para isso, foi utilizada a função fcluster da biblioteca Scipy que teve como entrada o resultado da função linkage, descrita anteriormente, e um *cutoff* de 0,3 como critério de distância dos elementos de um grupo (dissimilaridade).

Para a caracterização dos complexos filtrados, foram realizadas consultas SQL e a plataforma Yvis para a visualização das classes de aminoácidos em cada posição das cadeias de anticorpos. Para a análise das estruturas utilizadas nos trabalhos correlatos, foram obtidos os identificadores das estruturas e das cadeias analisadas por esses trabalhos. Esses dados foram então utilizados como entrada de um *script* implementado em Python que obtém os dados das interfaces acessando o Ydb. Caso algum identificador de estrutura fosse considerado obsoleto pelo PDB, a estrutura que a substituiu foi considerada na análise. Se essa nova estrutura não continha as cadeias originalmente utilizadas, o complexo correspondente foi removido da análise.

Para a análise das interações entre antígeno e anticorpo, foram utilizadas as interações calculadas e armazenadas no Ydb. Um *script* em Python foi implementado para acessar o Ydb e gerar um arquivo JSON contendo, para cada posição do alinhamento das sequências através da aplicação do esquema de numeração do IMGT, o número de cadeias que realizam algum tipo de interação nessa posição e a contagem e tipos das interações realizadas por essa posição considerando todas as cadeias. O arquivo JSON foi utilizado como entrada de uma implementação da representação *Collier de Diamants*, considerando os tipos de interações como classes representadas em cada posição e o número de sequências que realizam alguma interação como atributo adicional.

5.4.2 Análise do Ydb: Resultados

Os anticorpos armazenados no Ydb foram produzidos por 15 diferentes espécies (*Camelus bactrianus*, *Camelus dromedarius*, *Cricetulus migratorius*, *Homo sapiens*, *Lama glama*, *Macaca fascicularis*, *Macaca mulatta*, *Mus musculoides*, *Mus musculus*, *Oryctolagus cuniculus*, *Pan troglodytes*, *Phoca vitulina*, *Rattus norvegicus*, *Rattus rattus* e *Vicugna pacos*), além de anticorpos quiméricos, sintéticos, obtidos a partir de bibliotecas de sequências de camelídeos ou que não tiveram a espécie determinada.

Como em uma mesma estrutura existem domínios variáveis de cadeias de anticorpos idênticos, para as análises apresentadas nesta seção foi utilizado apenas um dos complexos existentes em cada arquivo de estrutura, exceto quando os domínios não são idênticos, o que ocorre em 85 estruturas que contêm ao menos 2 anticorpos distintos. Após esse processo, o conjunto de dados de análise continha 2.188 complexos, apresentados no Apêndice 2. Desses complexos, 298 possuem somente cadeia pesada, 18 possuem somente cadeia leve e 1.872 possuem um par de cadeias, o que totaliza 2.177 cadeias pesadas e 1.891 leves. A maioria desses complexos é formada por anticorpos produzidos por humanos (~44%) e camundongos (~37%), sendo o restante formado por camelídeos (~12%) e as demais espécies listadas anteriormente.

A Figura 5.4 apresenta três *heatmaps* que representam as matrizes de identidade das sequências das cadeias dos anticorpos presentes no Ydb. Cada linha e cada coluna dessa matriz e, consequentemente, do *heatmap*, representam uma sequência. O valor de cada posição da matriz (m_{ij} – linha *i*, coluna *j*) e do *heatmap* corresponde à identidade entre as sequências *i* e *j*. Como a identidade entre uma sequência e ela mesma (m_{ii}) é igual a 100%, a diagonal das matrizes e dos *heatmaps* é sempre igual a 1. Além disso, como a identidade entre as sequências i e j é a mesma entre as sequências j e i, os valores abaixo da diagonal são iguais aos valores acima dela, apresentados de maneira espelhada.

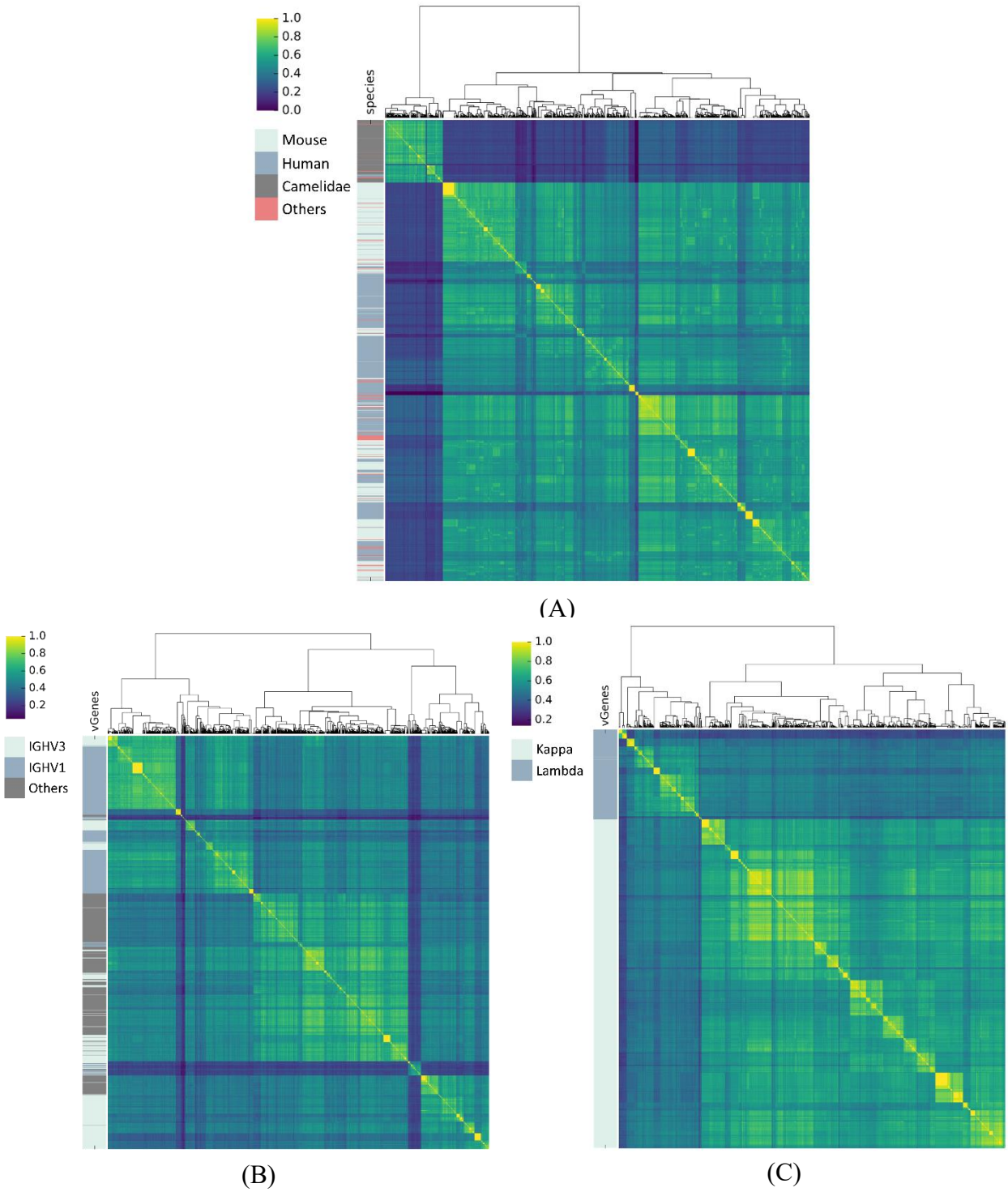


Figura 5.4: *Heatmaps* das matrizes de identidade de sequências dos complexos presentes no Ydb. Os valores representados variam de 0 a 1 (100% de identidade). (A) Sequências de anticorpos de 2.188 pares de cadeia pesada + leve. Caso uma das cadeias estiver ausente, a sequência é preenchida com gaps na região correspondente. A barra lateral apresenta as espécies produtoras do anticorpo. (B) Sequências de 2.177 cadeias pesadas de anticorpos. A barra lateral apresenta os genes V mais frequentes nessas sequências. (C) Sequências de 1.891 cadeias leves de anticorpos. A barra lateral apresenta os *loci* de cadeia leve das cadeias.

Para a definição da ordem de apresentação das sequências em cada *heatmap* foi utilizado um algoritmo de agrupamento hierárquico que forma grupos de sequências tentando minimizar a variância da identidade das sequências que fazem parte de um mesmo grupo (*cluster*). Esse algoritmo agrupa pares de sequências e, iterativamente, unifica os grupos formados até que exista somente um.

Acima de cada *heatmap* é apresentado um dendrograma que representa o agrupamento realizado com a utilização de uma árvore. Devido ao número de sequências analisadas, os níveis mais próximos das folhas da árvore não podem ser visualizados, no entanto, é possível observar a altura das linhas horizontais que representam as últimas unificações de grupos realizadas pelo algoritmo de agrupamento. Essa altura representa a "distância" dos grupos unificados. Quanto mais separadas são essas linhas, maior é a variância do grupo obtido após a unificação.

O *heatmap* apresentado na Figura 5.4A foi desenhado a partir da matriz de similaridade em que cada sequência é formada pelo domínio variável da cadeia pesada seguido do domínio variável da cadeia leve de cada um dos 2.188 complexos analisados. Devido ao número elevado de linhas e colunas dessa matriz, foram omitidos os identificadores dos complexos no *heatmap*. A diagonal do *heatmap* é amarela, indicando que cada complexo tem valor de identidade 1 (100%) quando comparado com ele mesmo. Além da diagonal, é possível observar outras posições representadas em amarelo, o que indica que existem complexos idênticos quando os domínios variáveis de suas cadeias são comparados. Uma barra à esquerda do *heatmap* apresenta a espécie produtora do anticorpo representado em cada linha. É possível perceber que a maioria dos anticorpos é produzida por humanos e camundongos e, em menor número, por espécies da família *Camelidae*, como destacado anteriormente. Esse último conjunto de anticorpos gera duas grandes faixas azuis no *heatmap*, pois esses anticorpos não apresentam cadeia leve e, quando comparado com os demais anticorpos que possuem cadeia leve e pesada, possuem um valor de identidade baixo. Essa característica faz com que esse conjunto de anticorpos seja unificado com os demais somente nas últimas

etapas do agrupamento, o que aumenta o valor da variância do novo grupo formado, como é possível verificar na linha horizontal superior do dendograma. Além disso, é possível perceber por meio da análise das cores do *heatmap* que a maioria dos anticorpos que possuem cadeia leve e pesada têm mais de 40% de identidade entre eles.

O *heatmap* apresentado na Figura 5.4B representa a matriz de identidade das 2.177 cadeias pesadas presentes nos anticorpos analisados. A barra na lateral esquerda desse gráfico representa os principais subgrupos de genes V (IGHV) presentes no conjunto de sequências. O número de sequências originadas a partir de genes V dos subgrupos IGVH3 e IGVH1 destaca-se em relação aos demais, representando, respectivamente, 36% e 33% das cadeias pesadas analisadas. Isso se deve ao fato de que os genes do subgrupo IGVH1 são os mais frequentes em camundongos enquanto que os dos subgrupos IGHV3 e IGVH1 são muito frequentes em humanos (ARNAOUT et al., 2011) e essas duas espécies possuem maior representatividade no conjunto de sequências analisadas.

O *heatmap* apresentado na Figura 5.4C representa a matriz de identidade das 1.891 cadeias leves presentes nos anticorpos analisados. Esse *heatmap* comparado com o *heatmap* da Figura5.4B, apresenta mais posições em tons de amarelo, o que indica que as sequências de cadeias leves são mais similares entre si. A barra ao lado do gráfico representa os genes V dos dois *loci* de cadeias leves, *kappa* (IGKV) e *lambda* (IGLV). É possível perceber uma clara divisão entre as sequências originadas desses genes tanto observando o gráfico de barras quanto o *heatmap*, que apresenta duas faixas azuis correspondentes às sequências originadas a partir do gene IGKV. Além disso, o dendrograma apresenta a unificação das cadeias originadas dos genes IGKV e IGLV na última etapa da aglomeração. O maior número de sequências originadas a partir dos genes IGKV pode ser justificado pelo maior número de genes desse tipo quando comparados com os IGLV e da ordem em que acontecem os rearranjos nos *loci* de cadeias leves, iniciada pelos genes IGKV em humanos e camundongos (COLLINS; WATSON, 2018), espécies

produtoras de anticorpos mais frequentes no conjunto analisado. No entanto, esse número também pode estar enviesado pelos anticorpos mais estudados ou que foram mais facilmente cristalizados. Portanto, para uma análise com menos viés, é necessário reduzir a redundância dos dados analisados.

Para diminuir a redundância dos dados analisados foi aplicado a esse conjunto o mesmo algoritmo de agrupamento empregado na geração dos dados da Figura 5.4. Em seguida, foi definido um nível de corte do dendrograma que representa o agrupamento das sequências formadas pela junção da cadeia pesada e leve de cada anticorpo. Para cada ramo da árvore que se encontra abaixo desse nível de corte, foi atribuído um grupo e os anticorpos pertencentes a ele foram consideradas redundantes. Para a remoção da redundância, foi escolhido um único anticorpo para representar cada grupo, restando 853 anticorpos com identidade máxima de 95% entre eles. Essa escolha foi baseada na resolução da estrutura do anticorpo, sendo escolhidas estruturas com maior resolução. Após essa etapa, foram removidos das análises posteriores os anticorpos que possuíam apenas uma cadeia, o que evitou o viés causado pela forma diferente de interação deles com os antígenos (ZAVRTANIK et al., 2018). Além disso, foram removidas as estruturas que possuem resolução menor que 2,5Å, para que os cálculos de interface sejam mais precisos. Esses filtros limitaram as estruturas analisadas a 385 complexos antígeno-anticorpo. Os anticorpos analisados no restante desta seção estão destacados na tabela apresentada no Apêndice 2.

A Figura 5.5 apresenta o *Collier de Diamants* das cadeias pesadas dos 385 complexos extraídos do Ydb. Nessa representação é possível observar diversas posições com a classe de resíduos conservada, principalmente nas regiões de *framework*, além das posições com resíduos conservados segundo o esquema de numeração do IMGT. Dentre as posições com classe conservada pode-se destacar as posições 15, 16, 22, 42, 94 e 118. Na maioria das sequências o CDR-H1 e o CDR-H2 possuem 8 resíduos e o CDR-H3 possui 12 resíduos. O CDR-H1 possui posições com classes mais conservadas que os demais e o CDR-

H3 é o que possui maior diversidade de tamanho e de composição de classes em cada posição (SHIRAI; KIDERA; NAKAMURA, 1999).

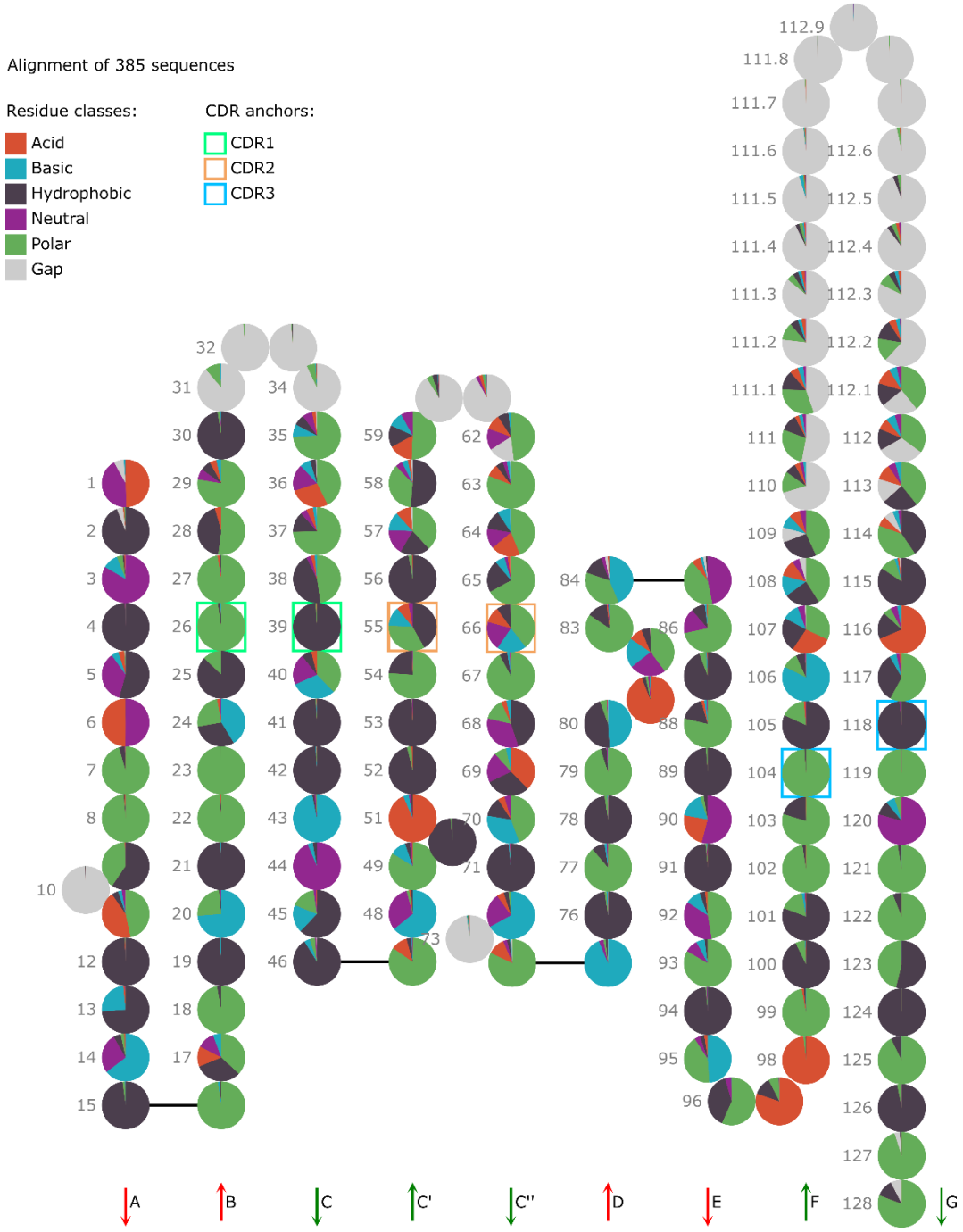


Figura 5.5: *Collier de Diamants* das cadeias pesadas dos 385 complexos extraídos do Ydb.

Das sequências analisadas, aproximadamente 43% (165) são oriundas de anticorpos murinos e 50% (192) humanos. Os anticorpos analisados estão em complexo com antígenos proteicos de 64 diferentes

espécies, sendo o HIV o patógeno mais frequente nesse conjunto de dados, presente em 15% (58) das sequências analisadas. A Figura 5.6 apresenta a frequência de cada subgrupo de gene V das cadeias pesadas analisadas. Os genes das subclasses IGHV1 e IGHV3 são os mais frequentes seguidos pelos da das subclasses IGHV5. Os genes das subclasses IGHV1, IGHV3 e IGHV4 são os mais frequentes em humanos (ARNAOUT et al., 2011; DEWITT et al., 2016), enquanto os genes IGHV1, IGHV5 e IGHV8 são os mais frequentes em camundongos (ARNAOUT et al., 2011). Portanto, como essas são as espécies produtoras de anticorpos mais presentes no Ydb, seria esperado encontrar um maior número de sequências dessas subclasses mais frequentes nessas espécies.

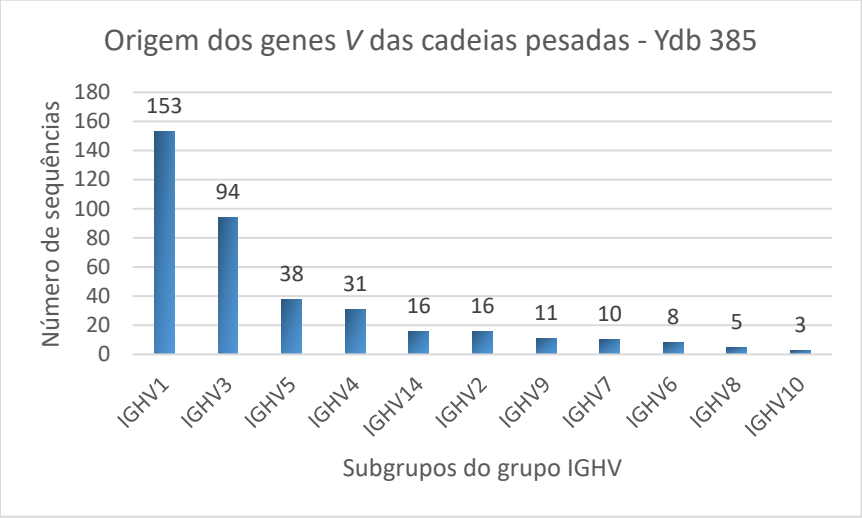


Figura 5.6: Subgrupos dos genes V das cadeias pesadas dos anticorpos analisados.

Como descrito anteriormente, o Ydb armazena os resíduos do paratopo e do epitopo dos complexos antígeno-anticorpo utilizando diferentes critérios de definição de interface. A Figura 5.7 apresenta a contagem dos aminoácidos da interface dos 385 complexos obtidos a partir do filtro descrito anteriormente considerando cada definição. Cada tipo de série representa um dos critérios de definição de interface. A série sólida representa a interface definida pelas interações calculadas entre os resíduos, as séries tracejadas representam as interfaces definidas pela variação na área de acessibilidade ao solvente e a as séries pontilhadas as interfaces definidas por distância. Os aminoácidos estão ordenados pela frequência

na definição de interface baseada em interações calculadas. Independentemente do critério adotado, a tirosina, a arginina e a fenilalanina são os aminoácidos mais frequentes e a leucina e a cisteína são os menos frequentes (NGUYEN et al., 2017; PENG et al., 2014). O número de resíduos presentes na interface, considerando os diversos critérios de definição, é muito variável. Dentre as definições presentes no Ydb, o critério de variação na área de acessibilidade ao solvente de ao menos 50Å^2 gera um conjunto de 1.910 resíduos nas interfaces analisadas enquanto o critério de distância de até 6Å entre átomos gera 13.113 resíduos.

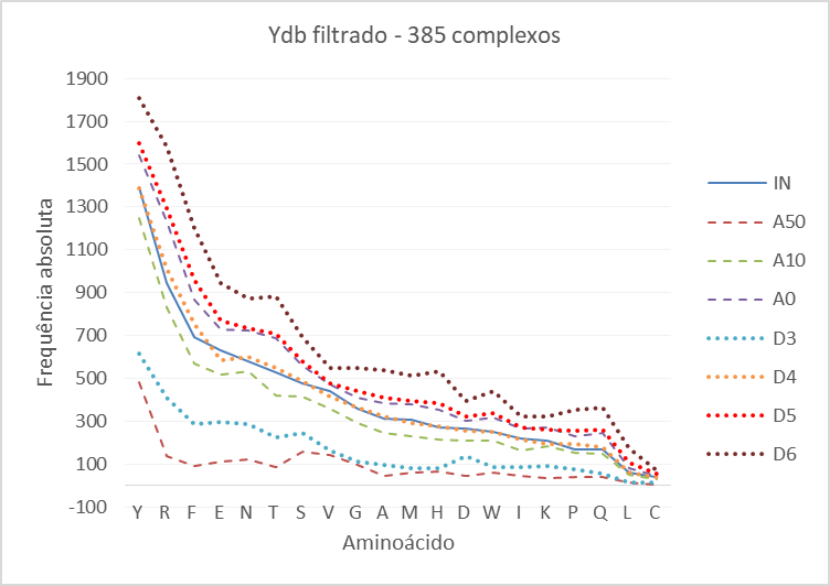


Figura 5.7: Frequência absoluta dos resíduos pertencentes às interfaces de 385 complexos armazenados no Ydb. Cada série representa uma definição diferente de interface: cálculo de interações (IN), variação na área de acessibilidade ao solvente de pelo menos 50Å^2, 10Å^2 ou qualquer variação (A50, A10, A0) e distância entre átomos de 3 a 6Å (D3, D4, D5, D6).

Ao analisar a frequência de cada resíduo relativa ao total de resíduos da interface, como ilustrado nas Figuras 5.8 e 5.9, é possível perceber que, considerando-se um mesmo critério de definição de interface, o valor utilizado como *cutoff* influencia não somente no número de resíduos na interface, mas também, na frequência relativa dos resíduos. Na Figura 5.8, cada série representa uma das definições de interface baseada na variação mínima da área de acessibilidade ao solvente, variando-se esse valor de 0 a 50Å^2 (A0-50) e uma das séries representa a definição por interações calculadas (IN). Nessa figura é

possível perceber que a definição de interface com variação mínima de 50Å^2 é representada por uma série com um comportamento diferente das demais, principalmente para os resíduos tirosina, arginina, fenilalanina, treonina, serina e valina. Na Figura 5.9, cada série representa uma das definições de interface baseada na distância máxima dos átomos, que variam de 3 a 6Å (D3-6) e uma das séries representa a definição por interações calculadas (IN). Nessa figura é possível perceber que a definição de interface com distância máxima de 3Å é representada por uma série com um comportamento diferente das demais definições, principalmente para os resíduos ácido glutâmico, asparagina, serina, glicina, alanina, metionina e histidina.

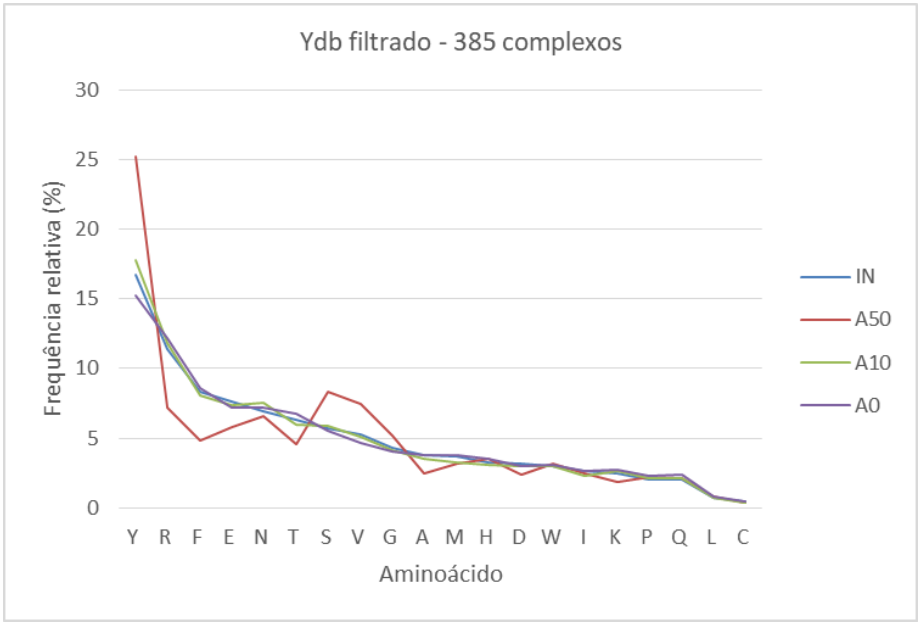


Figura 5.8: Frequência dos resíduos em relação ao total de resíduos da interface para o conjunto de dados do Ydb com 385 complexos. Cada série representa uma das definições de interface baseada na distância máxima dos átomos, variando-se esse valor de 3 a 6Å (D3-6). A série IN representa a definição por interações calculadas.

O gráfico da Figura 5.10 apresenta a frequência de cada resíduo relativa ao total de resíduos da interface de três definições de interface armazenadas no Ydb. Os valores de distância máxima de 4Å na definição de interface baseada em distância dos átomos e de 10Å^2 na variação mínima da área de acessibilidade ao solvente foram escolhidos por apresentarem valores de frequência próximos entre si. Além disso, é apresentada a definição baseada no cálculo das interações e, considerando a frequência

absoluta dos resíduos apresentada na Figura 5.7, estas séries ocupam regiões próximas do gráfico. Essas definições são utilizadas nas análises seguintes.

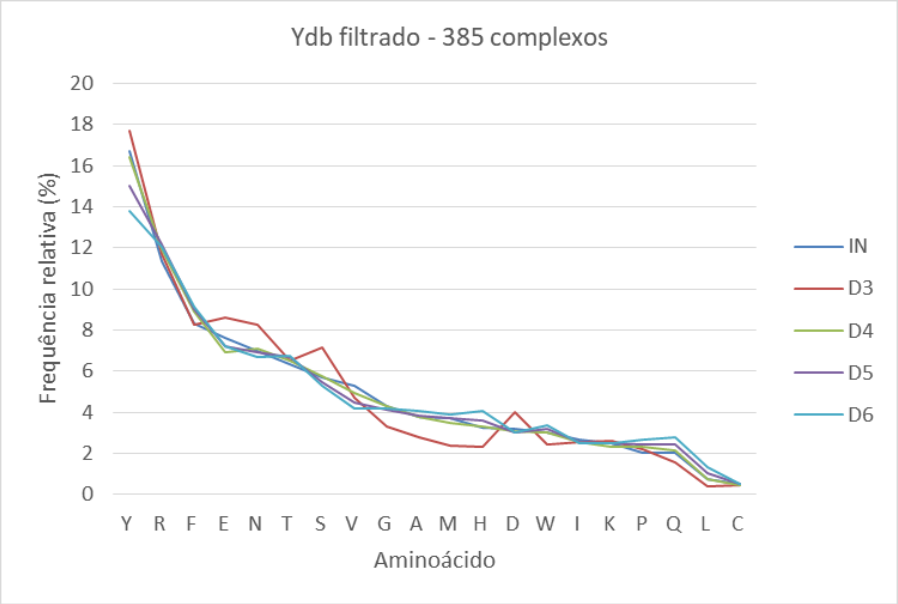


Figura 5.9: Frequência dos resíduos em relação ao total de resíduos da interface para o conjunto de dados do Ydb com 385 complexos. Cada série representa uma das definições de interface baseada na variação mínima da área de acessibilidade ao solvente, variando-se esse valor de 0 a 50Å^2 (A0-50). A série IN representa a definição por interações calculadas.

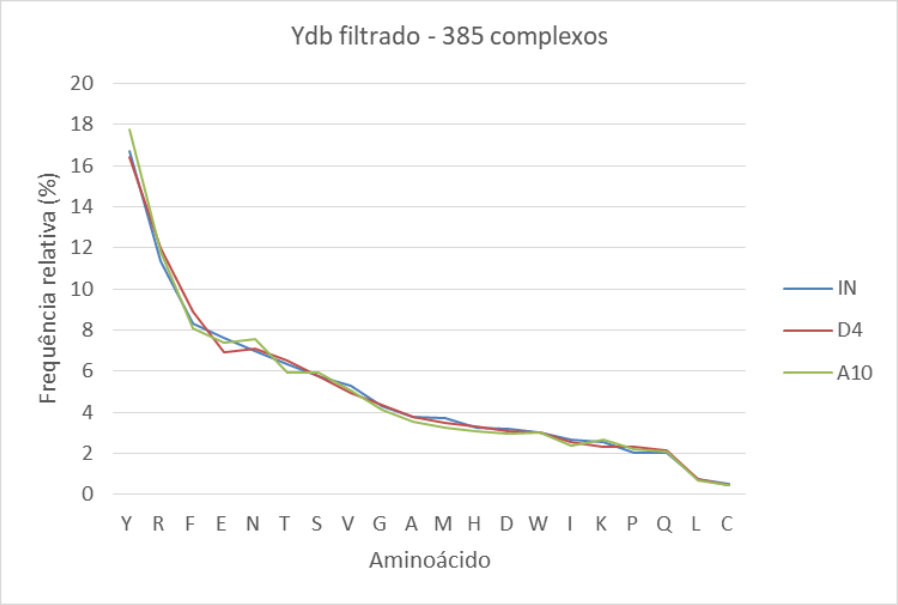


Figura 5.10: Frequência dos resíduos em relação ao total de resíduos da interface para o conjunto de dados do Ydb com 385 complexos. Cada série representa um dos critérios de definição de interface do Ydb: interações calculadas (IN), distância máxima de 4Å entre átomos (D4) e variação mínima de 10Å^2 na área de acessibilidade ao solvente (A10).

Na Figura 5.11 são apresentadas as frequências de cada resíduo relativas ao total de resíduos da interface considerando 4 conjuntos diferentes de estruturas, obtidos utilizando diferentes critérios como apresentado na Tabela 1.1. Observando os 4 gráficos apresentados nessa figura, é possível perceber que os 3 critérios de definição de interface utilizados na análise possuem valores de frequência de cada resíduo próximos. Além disso, é possível constatar que os 4 resíduos mais frequentes e os 4 resíduos menos frequentes em todos os conjuntos analisados, destacados em roxo, são os mesmos, exceto pela fenilalanina no conjunto de Nguyen et al., 2017, destacada em amarelo. Ao comparar esses dados com o gráfico da Figura 5.10, que representa os dados dos complexos do Ydb, é possível perceber o mesmo conjunto de resíduos menos frequentes e pequenas diferenças nos mais frequentes. Além disso, os valores de frequência relativa nos diversos conjuntos são semelhantes. Portanto, ao analisar a frequência relativa dos resíduos como métrica para a comparação dos diferentes conjuntos de dados utilizados nos trabalhos correlatos e os critérios de definição de interface do Ydb, não foi possível encontrar uma grande diferença entre os conjuntos. No entanto, não se pode afirmar que não exista diferença na análise baseada em diferentes conjuntos de dados já que seria necessário avaliar outras características dos mesmos baseadas diferentes critérios, como, por exemplo, as interações realizadas por cada resíduo presente na interface antígeno-anticorpo.

Na Figura 5.12 e no Apêndice 3 são apresentadas as interações realizadas pelas cadeias pesadas dos complexos analisados. Em cada gráfico de setor da Figura 5.12 são apresentas todas as interações realizadas pelos resíduos que estão em uma posição específica do alinhamento, de todas as 385 cadeias pesadas analisadas. Como um mesmo resíduo da cadeia pesada pode realizar mais de uma interação com o antígeno, enquanto outros não realizam interações com ele, o número total de contatos é variável ao longo das posições representadas, mas sempre são apresentados setores com área proporcional a esse total de contatos. Com isso, pode-se dizer que o gráfico de setores apresenta as interações realizadas em cada

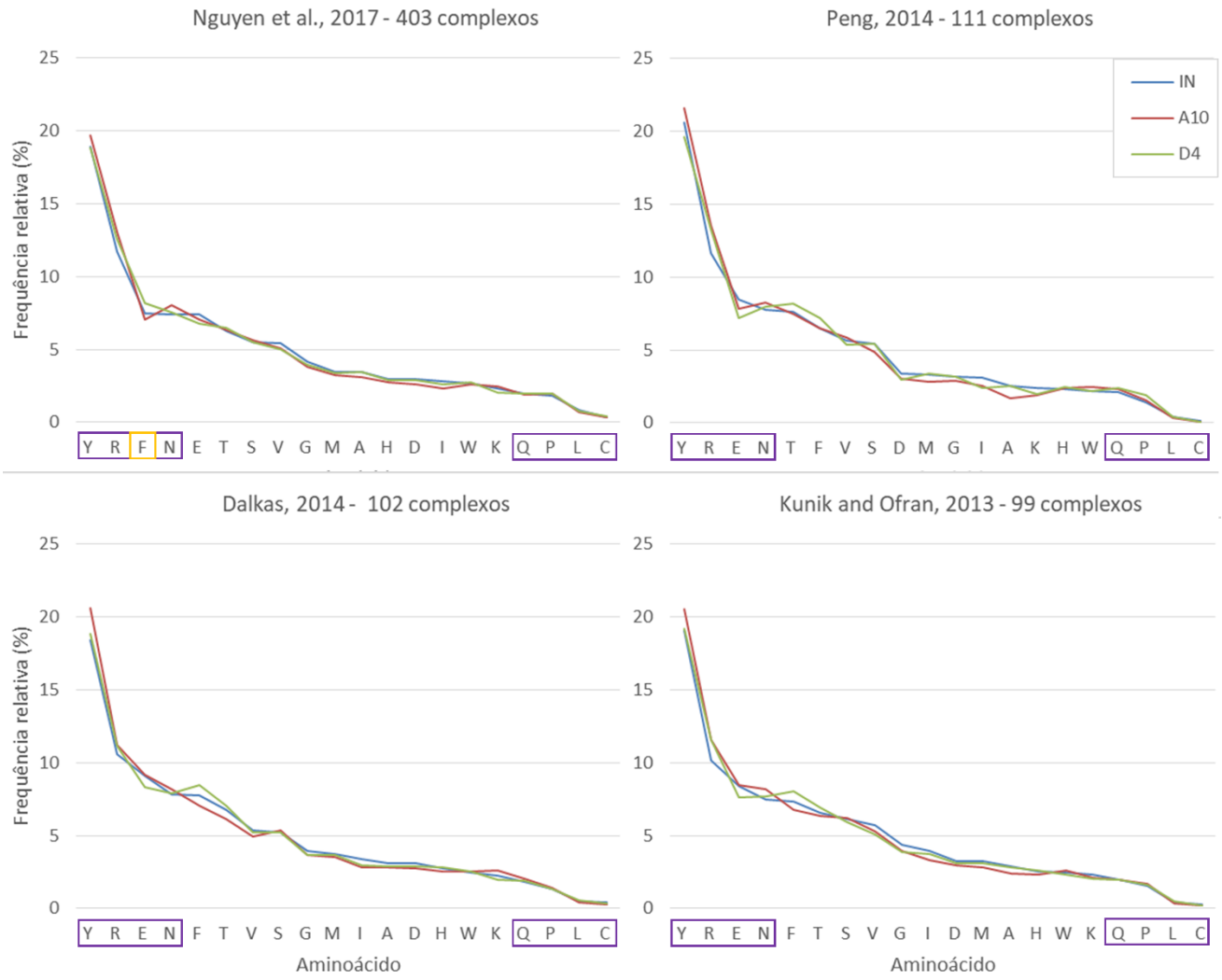


Figura 5.11: Frequência dos resíduos em relação ao total de resíduos da interface para o conjunto de dados de quatro trabalhos correlatos. Cada série representa um dos critérios de definição de interface do Ydb: interações calculadas (IN), distância máxima de 4Å entre átomos (D4) e variação mínima de 10Å^2 na área de acessibilidade ao solvente (A10).

posição normalizadas pelo número de interações realizadas por ela. No entanto, na Tabela 5.4 estão disponíveis o número total de contatos realizados por uma posição e o valor absoluto representado por cada setor. Na Figura 5.12, as posições que não realizam interações são desenhadas em cinza e, para destacar as posições que realizam interações, é desenhado um círculo salmão, em seu redor, que representa o número de cadeias que realizam algum tipo de interação nessa posição.

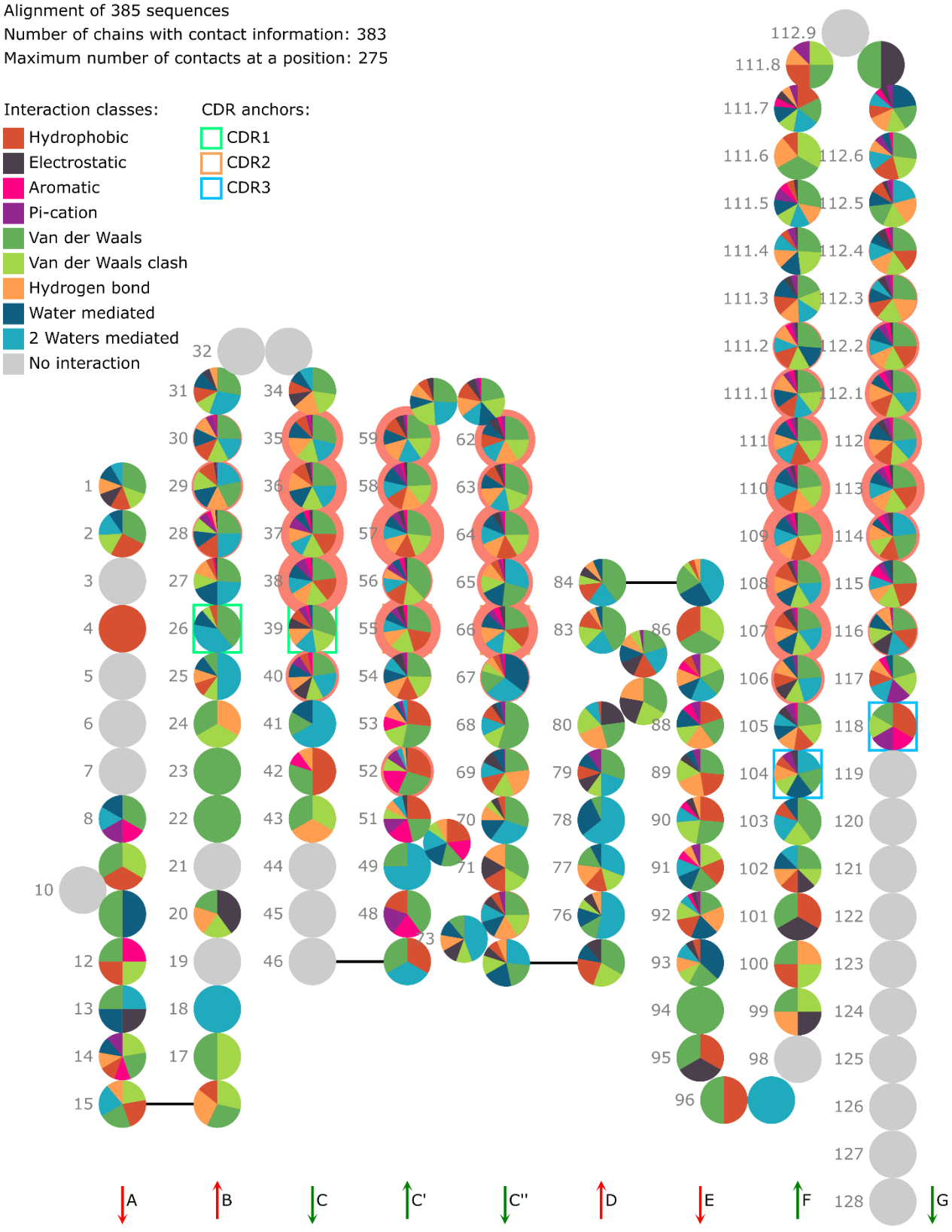


Figura 5.12: Representação *Collier de Diamants* ilustrando as interações realizadas pelas 385 sequências de cadeias pesadas de anticorpos armazenados no Ydb.

A grande maioria das posições destacadas pelos círculos em salmão na Figura 5.12 é localizada nas regiões de CDR. Uma característica comum à maioria dessas posições é ter as interações de Van der Waals, representadas em verde escuro, como interação mais frequentes. Como os CDRs são as regiões normalmente envolvidas nas interações entre antígeno e anticorpo e, na maioria das vezes, são as regiões que ficam próximas do antígeno, essas interações, mesmo sendo de curta distância, são muito frequentes. Às vezes essa distância é tão curta que as interações são consideradas Van der Waals *clash*, representadas em verde claro, quando a distância entre um par de átomos é menor que a soma dos raios de Van der Waals de cada um deles. Além disso, apesar das interações de Van der Waals serem interações normalmente consideradas fracas quando comparadas com as demais, o grande número de interações desse tipo representa uma parte considerável da energia de ligação entre antígeno e anticorpo (KULKARNI-KALE et al., 2014).

Devido à variabilidade do tamanho dos CDRs 3 analisados, como destacado na Figura 5.5, o número de interações realizadas por algumas posições dessa região não aparece com muito destaque na Figura 5.12, devido à ausência dessas posições em diversas cadeias do conjunto analisado. No entanto, como os CDRs 1 e 2 apresentam uma menor variação de tamanho, os círculos que destacam as posições que realizam mais contatos permite concluir que as interações realizadas por posições do CDR1 são menos frequentes que pelas posições dos demais CDRs nas cadeias pesadas (BURKOVITZ; OFRAN, 2016).

Apesar do grande número de interações de Van der Waals realizadas pelas cadeias pesadas dos anticorpos analisados, dois outros tipos de interação também são frequentemente encontrados na interface dos complexos antígeno-anticorpo: interações de hidrogênio, representadas por setores laranja, e interações hidrofóbicas, representadas em vermelho. Como a tirosina é o aminoácido mais frequente nas interfaces e é capaz de realizar esses dois tipos de interação, normalmente, a tirosina é o resíduo do paratopo que mais realiza interações de hidrogênio e hidrofóbicas (DALKAS et al., 2014). Outro tipo de

interação que deve ser destacado são as interações eletrostáticas, representadas em preto na Figura 5.12. Apesar de bem menos frequente que as interações citadas anteriormente, as interações eletrostáticas, individualmente, geram uma contribuição energética maior que as demais. Esse tipo de interação se mostrou mais frequente no CDR3 e em algumas posições do CDR2 (BURKOVITZ; OFRAN, 2016).

Outro tipo de interação que se destaca na Figura 5.12 são as interações de hidrogênio mediadas por água. O papel das interações mediadas por uma molécula de água já foi analisado anteriormente (KULKARNI-KALE et al., 2014; NGUYEN et al., 2017; YOKOTA et al., 2003), no entanto, o número de interações mediadas por duas moléculas de água também é significativo e pode ser explorado em estudos posteriores.

As análises apresentadas nesta seção deram uma visão geral dos dados armazenados no Yvis. No entanto, tanto os dados aqui apresentados como outros que estão armazenados, mas não foram utilizados nessas análises ainda devem ser explorados. Como o Ydb é atualizado frequentemente, as análises aqui apresentadas foram automatizadas, já que elas podem ser repetidas com conjuntos maiores de estrutura que serão depositadas no PDB ao longo do tempo.

6 Conclusões e perspectivas

Os anticorpos são utilizados hoje nas mais diversas áreas como ferramenta para pesquisas e diagnósticos, além de terapia para diversas doenças. O sucesso dessa utilização está relacionado ao conhecimento das estruturas de anticorpos e antígenos e de suas interações. Apesar da ampla utilização dos anticorpos, ainda não são claras as regras que governam o reconhecimento entre antígeno e anticorpo e a especificidade dos anticorpos. Em busca de respostas a essas dúvidas diversos pesquisadores têm trabalhado nas áreas de descoberta e modelagem de estruturas, além de experimentos de sequenciamento de repertório de anticorpos. Devido a esses trabalhos uma grande quantidade de dados relacionados a anticorpos vêm surgindo e existe a carência de ferramentas para análise e visualização de anticorpo e suas interações com os antígenos.

Neste trabalho foram propostas algumas estratégias de análise e visualização em larga escala das cadeias dos anticorpos e suas interações com o antígeno que permitam uma melhor compreensão do processo de reconhecimento e ligação de anticorpos a antígenos. Para a visualização do alinhamento de domínios variáveis de cadeias de anticorpos foi proposta a representação denominada *Collier de Diamants*. Essa representação foi utilizada na implementação da plataforma Yvis que permite não somente a visualização do alinhamento milhares de cadeias de anticorpos, como também a análise de diferentes conjuntos de dados. Dentre esses dados destacam-se as estruturas do PDB pré-processadas e armazenadas no banco de dados do Yvis, atualizados semanalmente, e ferramentas do IMGT de análise de domínios variáveis e de tratamento de dados de sequenciamento de repertório de anticorpos.

A plataforma Yvis pode ser utilizada em diferentes tipos de análises de anticorpos. Por exemplo, a rápida visualização de posições conservadas e divergentes em um conjunto de anticorpos relacionados pode direcionar experimentos de engenharia de anticorpos. Nos estudos de repertórios de anticorpos, a visualização por meio do *Collier de Diamants* e a comparação de uma sequência única com milhares de

sequências podem ressaltar as mutações mais importantes que ocorreram durante o processo de maturação de afinidade. Portanto, a plataforma Yvis apresenta um ambiente para análise de sequências de anticorpos em larga escala e auxilia o pesquisador na formulação de hipóteses referente aos resíduos chave na estrutura e na interação dos anticorpos, auxiliando o entendimento de suas propriedades.

Além da plataforma Yvis, foi implementado o banco de dados Ydb, que armazena dados da composição e das propriedades físico-químicas das cadeias de complexos antígeno-anticorpo, além da caracterização da interface desses complexos utilizando diferentes critérios de definição de interface e de suas interações. A possibilidade de análise conjunta dos dados do Ydb, as diversas definições de interface, juntamente com o cálculo das interações e a atualização semanal fazem com que o Ydb se diferencie dos demais bancos de dados que contêm informações de estruturas de anticorpos. As análises de alguns dos dados presentes no Ydb destacaram as diferenças entre as definições de interface analisadas e as posições mais conservadas e mais divergentes nas sequências de domínios variáveis de cadeias de anticorpos. Além disso, a análise das interações realizadas pelas cadeias pesadas dos anticorpos armazenados no Ydb permitiu a visualização das posições que normalmente realizam interações com o antígeno, sendo que nenhuma delas realiza um tipo exclusivo de interação. Essa análise ainda viabilizou a representação dos tipos de interação geralmente presentes na interface do complexo e deixou em aberto a necessidade de se analisar as interações de hidrogênio entre antígeno e anticorpo mediadas por múltiplas moléculas de água. De qualquer forma, constatou-se que as interações de hidrogênio, sejam mediadas por água ou não, são frequentes na interface antígeno-anticorpo.

Apesar das análises realizadas neste trabalho inúmeras outras podem ser realizadas sobre os dados armazenados no Ydb, seja utilizando dados não analisados neste texto, utilizando técnicas diferentes das utilizadas aqui ou ainda um subconjunto dos dados analisados. Acredita-se que essas análises com dados presentes atualmente no Ydb e os que serão produzidos ao longo do tempo e armazenados nele por meio

de suas atualizações semanais podem gerar um melhor entendimento dos anticorpos e suas interações com os antígenos. Além disso, durante o desenvolvimento do Ydb, percebeu-se que seus dados podem ser de interesse de outros pesquisadores da área e dessa maneira tem-se como mais uma perspectiva deste trabalho o desenvolvido de uma interface *web* e de serviços *web* para acesso a estes dados.

Referências bibliográficas

ABHINANDAN, K. R.; MARTIN, A. C. R. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. **Molecular Immunology**, v. 45, n. 14, p. 3832–3839, 2008.

ALAMYAR, E. et al. IMGT/HighV-QUEST: The IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. **Immunome Research**, v. 8, n. 1, p. 1–15, 2012.

ALLCORN, L. C.; MARTIN, A. C. R. R. SACS--Self-maintaining database of antibody crystal structure information. **Bioinformatics**, v. 18, n. 1, p. 175–181, 1 jan. 2002.

APGAR, J. R. et al. Beyond CDR-grafting: Structure-guided humanization of framework and CDR regions of an anti-myostatin antibody. **mAbs**, v. 8, n. 7, p. 1302–1318, 2016.

ARNAOUT, R. et al. High-resolution description of antibody heavy-chain repertoires in humans. **PLoS ONE**, v. 6, n. 8, 2011.

BENICHOU, J. et al. Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. **Immunology**, v. 135, n. 3, p. 183–191, 2012.

BERMAN, H. M. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 1 jan. 2000.

BHAT, T. N. et al. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. **Proceedings of the National Academy of Sciences of the United States of America**, v. 91, n. 3, p. 1089–1093, fev. 1994.

BICKERTON, G. R.; HIGUERUELO, A. P.; BLUNDELL, T. L. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: The PICCOLO database. **BMC Bioinformatics**, v. 12, n. 1, p. 313, 2011.

BOYD, S. D.; CROWE, J. E. Deep sequencing and human antibody repertoire analysis. **Current Opinion in Immunology**, v. 40, p. 103–109, 2016.

BURKOVITZ, A.; OFRAN, Y. Understanding differences between synthetic and natural antibodies can help improve antibody engineering. **mAbs**, v. 8, n. 2, p. 278–287, 2016.

CARVALHO, M. B.; MOLINA, F.; FELICORI, L. F. Yvis: antibody high-density alignment visualization and analysis platform with an integrated database. **Nucleic Acids Research**, v. 47, n. May, p. 490–495, 2019.

CHEN, S.; VAN REGENMORTEL, M.; PELLEQUER, J. Structure-Activity Relationships in Peptide-Antibody Complexes: Implications for Epitope Prediction and Development of Synthetic Peptide Vaccines. **Current Medicinal Chemistry**, v. 16, n. 8, p. 953–964, 1 mar. 2009.

CHOTHIA, C.; LESK, A. M. Canonical structures for the hypervariable regions of immunoglobulins. **Journal of Molecular Biology**, v. 196, n. 4, p. 901–917, 1987.

COCK, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, jun. 2009.

COLLINS, A. M.; WATSON, C. T. Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. **Frontiers in Immunology**, v. 9, n. OCT, p. 1–12, 2018.

CONROY, P. J. et al. Antibodies: From novel repertoires to defining and refining the structure of biologically important targets. **Methods (San Diego, Calif.)**, v. 116, p. 12–22, mar. 2017.

CORRIE, B. D. et al. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. **Immunological Reviews**, v. 284, n. 1, p. 24–41, 2018.

CROOKS, G. E. et al. WebLogo: a sequence logo generator. **Genome research**, v. 14, n. 6, p. 1188–1190, jun. 2004.

DALKAS, G. A. et al. Cation-π, amino-π, π-π, and H-bond interactions stabilize antigen-antibody interfaces. **Proteins: Structure, Function and Bioinformatics**, v. 82, n. 9, p. 1734–1746, set. 2014.

DAVIES, D. R. et al. Antibody-antigen complexes. **Annual Review of Biochemistry**, v. 59, n. 1, p. 439–473, 1 jun. 1990.

DAVIES, D. R.; COHEN, G. H. Interactions of protein antigens with antibodies. **Proceedings of the National Academy of Sciences**, v. 93, n. 1, p. 7–12, 1996.

DEKOSKY, B. J. et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. **Proceedings of the National Academy of Sciences**, v. 113, n. 19, p. E2636–E2645, 2016.

DEWITT, W. S. et al. A public database of memory and naive B-cell receptor sequences. **PLoS ONE**, v. 11, n. 8, p. 1–18, 2016.

DONDELINGER, M. et al. Understanding the significance and implications of antibody numbering and antigen-binding surface/residue definition. **Frontiers in Immunology**, v. 9, n. OCT, p. 1–15, 2018.

DUNBAR, J. et al. SAbDab: The structural antibody database. **Nucleic Acids Research**, v. 42, n. D1, p. 1140–1146, 2014.

DUNBAR, J.; DEANE, C. M. ANARCI: Antigen receptor numbering and receptor classification. **Bioinformatics**, v. 32, n. 2, p. 298–300, 2015.

EHRENMANN, F. et al. Standardized Sequence and Structure Analysis of Antibody Using IMGT®. In: KONTERMANN, R.; DÜBEL, S. (Eds.). **Antibody Engineering**. Berlin, Heidelberg, Heidelberg: Springer Berlin Heidelberg, 2010. p. 11–31.

EHRENMANN, F.; KAAS, Q.; LEFRANC, M. P. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. **Nucleic Acids Research**, v. 38, n. suppl_1, p. D301–D307, jan. 2010.

FERDOUS, S.; MARTIN, A. C. R. AbDb: antibody structure database—a database of PDB-derived antibody structures. **Database**, v. 2018, p. 1–9, 2018.

FINLAY, W. J. J.; ALMAGRO, J. C. Natural and man-made V-gene repertoires for antibody discovery. **Frontiers in Immunology**, v. 3, n. NOV, p. 1–18, 2012.

FORSSTRÖM, B. et al. Dissecting antibodies with regards to linear and conformational epitopes. **PLoS ONE**, v. 10, n. 3, p. 1–11, 2015.

FRENZEL, A. et al. Designing Human Antibodies by Phage Display. **Transfusion Medicine and Hemotherapy**, v. 44, n. 5, p. 312–318, 2017.

FRIEDENSOHN, S.; KHAN, T. A.; REDDY, S. T. Advanced Methodologies in High-Throughput

Sequencing of Immune Repertoires. **Trends in Biotechnology**, v. 35, n. 3, p. 203–214, 2017.

GEORGIOU, G. et al. The promise and challenge of high-throughput sequencing of the antibody repertoire. **Nature Biotechnology**, v. 32, n. 2, p. 158–168, 2014.

GIUDICELLI, V.; CHAUME, D.; LEFRANC, M. P. IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. **Nucleic Acids Research**, v. 33, n. DATABASE ISS., p. 256–261, 2005.

GONZÁLEZ-MUÑOZ, A. et al. Tailored amino acid diversity for the evolution of antibody affinity. **mAbs**, v. 4, n. 6, p. 664–672, 2012.

GONZALEZ-SAPIENZA, G.; ROSSOTTI, M. A.; TABARES-DA ROSA, S. Single-Domain Antibodies As Versatile Affinity Reagents for Analytical and Diagnostic Applications. **Frontiers in Immunology**, v. 8, p. 977, ago. 2017.

GREIFF, V. et al. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. **Trends in Immunology**, v. 36, n. 11, p. 738–749, nov. 2015.

GRILO, A. L.; MANTALARIS, A. The Increasingly Human and Profitable Monoclonal Antibody Market. **Trends in biotechnology**, v. 37, n. 1, p. 9–16, 2019.

HAMELRYCK, T.; MANDERICK, B. PDB file parser and structure class implemented in Python. **Bioinformatics**, v. 19, n. 17, p. 2308–2310, 2003.

HOEHN, K. B. et al. The Diversity and Molecular Evolution of B-Cell Receptors during Infection. **Molecular Biology and Evolution**, v. 33, n. 5, p. 1147–1157, 2016.

HONEGGER, A.; PLÜCKTHUN, A.; PLUCKTHUN, A. Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis tool. **Journal of Molecular Biology**, v. 309, n. 3, p. 657–670, jun. 2001.

HUBBARD, S.; THORTON, J. M. **NACCESS, Computer Program.**Department of Biochemistry and Molecular Biology, University College of London, UK, , 1996.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science and Engineering**, v. 9, n. 3, p. 99–104, 2007.

ICHINOHE, T. et al. Next-generation immune repertoire sequencing as a clue to elucidate the landscape of immune modulation by host-gut microbiome interactions. **Frontiers in Immunology**, v. 9, n. APR, p. 1–7, 2018.

IMKELLER, K.; WARDEMANN, H. Assessing human B cell repertoire diversity and convergence. **Immunological Reviews**, v. 284, n. 1, p. 51–66, 2018.

JONES, E. et al. **SciPy: Open source scientific tools for Python**, [s.d.]. Disponível em: <http://www.scipy.org/>

JONES, P. T. et al. Replacing the complementarity-determining regions in a human antibody with those from a mouse. **Nature**, v. 321, n. 6069, p. 522–525, 1986.

KAAS, Q. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. **Nucleic Acids Research**, v. 32, n. 90001, p. 208D – 210, 2004.

KAAS, Q.; LEFRANC, M.-P. IMGT Colliers de Perles: Standardized Sequence-Structure

Representations of the IgSF and MhcSF Superfamily Domains. **Current Bioinformatics**, v. 2, n. 1, p. 21–30, 2007.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, v. 22, n. 12, p. 2577–637, 1983.

KAPLON, H.; REICHERT, J. M. Antibodies to watch in 2019. **mAbs**, v. 11, n. 2, p. 219–238, 2019.

KOHLER, G.; MILSTEIN, C. Continuous cultures of fused cells secreting antibody of predefined specificity. **Nature**, v. 256, n. 5517, p. 495–497, 1975.

KOVALTSUK, A. et al. How B-cell receptor repertoire sequencing can be enriched with structural antibody data. **Frontiers in Immunology**, v. 8, 2017.

KOVALTSUK, A. et al. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. **The Journal of Immunology**, v. 201, n. 8, p. 2502–2509, 2018.

KRAWCZYK, K. et al. Structurally mapping antibody repertoires. **Frontiers in Immunology**, v. 9, n. JUL, p. 1–8, 2018.

KRAWCZYK, K.; DUNBAR, J.; DEANE, C. M. Computational Tools for Aiding Rational Antibody Design. **Methods in molecular biology (Clifton, N.J.)**, v. 1529, p. 399–416, 2017.

KRINGELUM, J. V. et al. Structural analysis of B-cell epitopes in antibody:protein complexes. **Molecular Immunology**, v. 53, n. 1–2, p. 24–34, 2013.

KULKARNI-KALE, U. et al. Antigen-Antibody Interaction Database (AgAbDb): a compendium of antigen-antibody interactions. **Methods in molecular biology (Clifton, N.J.)**, v. 1184, p. 149–64, 2014.

KUNIK, V.; OFRAN, Y. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. **Protein Engineering, Design and Selection**, v. 26, n. 10, p. 599–609, 2013.

LASKOWSKI, R. A.; SWINDELLS, M. B. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. **Journal of Chemical Information and Modeling**, v. 51, n. 10, p. 2778–2786, 2011.

LAVINDER, J. J. et al. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. **Current Opinion in Chemical Biology**, v. 24, p. 112–120, 2015.

LEE, J. J. **Mechanize - Automate interaction with HTTP web servers**. Disponível em <https://github.com/python-mechanize/mechanize>.

LEFRANC, M.-P. et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. **Developmental & Comparative Immunology**, v. 27, n. 1, p. 55–77, jan. 2003.

LEFRANC, M.-P. Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of Immunoinformatics. **Frontiers in Immunology**, 2014.

LEINONEN, R. et al. The European Nucleotide Archive. **Nucleic acids research**, v. 39, n. Database issue, p. D28-D31, 2011.

LI, W. et al. Antibody Aggregation: Insights from Sequence and Structure. **Antibodies**, v. 5, n. 3, p. 19, 2016.

LI, W.; GODZIK, A. Cd-hit: A fast program for clustering and comparing large sets of protein or

nucleotide sequences. **Bioinformatics**, v. 22, n. 13, p. 1658–1659, 2006.

LO CONTE, L.; JANIN, J.; CHOTHIA, C. The atomic structure of protein-protein recognition sites. **Journal of Molecular Biology**, v. 285, p. 2177–2198, 1999.

LÓPEZ-SANTIBÁÑEZ-JÁCOME, L.; AVENDAÑO-VÁZQUEZ, S. E.; FLORES-JASSO, C. F. The Pipeline Repertoire for Ig-Seq Analysis. **Frontiers in Immunology**, v. 10, n. April, p. 1–15, 2019.

MACCALLUM, R. M.; MARTIN, A. C.; THORNTON, J. M. Antibody-antigen interactions: contact analysis and binding site topography. **Journal of molecular biology**, v. 262, n. 5, p. 732–45, 1996.

MARILLET, S. et al. Novel Structural Parameters of Ig-Ag Complexes Yield a Quantitative Description of Interaction Specificity and Binding Affinity. **Frontiers in immunology**, v. 8, p. 34, 2017.

MCDONALD, I. K.; THORNTON, J. M. Satisfying hydrogen bonding potential in proteins. **Journal of molecular biology**, v. 238, n. 5, p. 777–93, 1994.

MIAN, I. S.; BRADWELL, A. R.; OLSON, A. J. Structure, function and properties of antibody binding sites. **Journal of molecular biology**, v. 217, n. 1, p. 133–51, 1991.

MIHO, E. et al. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. **Frontiers in Immunology**, v. 9, p. 1–15, 2018.

MORRISON, S. L. et al. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. **Proceedings of the National Academy of Sciences**, v. 81, n. 21, p. 6851–6855, 1984.

NCBI. **PubMed**. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/>.

NGUYEN, M. N. et al. The interfacial character of antibody paratopes: Analysis of antibody-antigen structures. **Bioinformatics**, v. 33, n. 19, p. 2971–2976, 2017.

NGUYEN, M. N.; VERMA, C. S.; ZHONG, P. AppA: a web server for analysis, comparison, and visualization of contact residues and interfacial waters of antibody–antigen structures and models. **Nucleic Acids Research**, n. 8, p. 1–8, 2019.

NIELSEN, S. C. A.; BOYD, S. D. Human adaptive immune receptor repertoire analysis—Past, present, and future. **Immunological Reviews**, v. 284, n. 1, p. 9–23, 2018.

OFRAN, Y.; SCHLESSINGER, A.; ROST, B. Automated Identification of Complementarity Determining Regions (CDRs) Reveals Peculiar Characteristics of CDRs and B Cell Epitopes. **The Journal of Immunology**, v. 181, n. 9, p. 6230–6235, 1 nov. 2008.

PENG, H.-P. et al. Origins of specificity and affinity in antibody-protein interactions. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 26, p. E2656–E2665, 2014.

PONOMARENKO, J. et al. IEDB-3D: structural data within the immune epitope database. **Nucleic acids research**, v. 39, n. Database issue, p. D1164-70, 2011.

PROCTER, J. B. et al. Visualization of multiple alignments, phylogenies and gene family evolution. **Nature Methods**, v. 7, n. 3, p. S25, 2010.

RAMARAJ, T. et al. Antigen-antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures. **Biochimica et Biophysica Acta - Proteins and Proteomics**, v. 1824, n. 3, p. 520–532, 2012.

ROY, A. et al. In silico methods for design of biological therapeutics. **Methods**, v. 131, p. 33–65, 2017.

RUBELT, F. et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. **Nature immunology**, v. 18, n. 12, p. 1274–1278, nov. 2017.

RUIZ, M.; LEFRANC, M.-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. **Immunogenetics**, v. 53, n. 10–11, p. 857–883, 2002.

SELA-CULANG, I.; KUNIK, V.; OFRAN, Y. The structural basis of antibody-antigen recognition. **Frontiers in immunology**, v. 4, n. 302, p. 302, jan. 2013.

SHIRAI, H.; KIDERA, A.; NAKAMURA, H. H3-rules: Identification of CDR-H3 structures in antibodies. **FEBS Letters**, v. 455, n. 1–2, p. 188–197, 1999.

SIRCAR, A.; KIM, E. T.; GRAY, J. J. RosettaAntibody: antibody variable region homology modeling server. **Nucleic acids research**, v. 37, n. Web Server issue, p. W474-9, jul. 2009.

SORMANNI, P.; APRILE, F. A.; VENDRUSCOLO, M. Third generation antibody discovery methods:: In silico rational design. **Chemical Society Reviews**, v. 47, n. 24, p. 9137–9157, 2018.

STAVE, J. W.; LINDPAINTNER, K. Antibody and Antigen Contact Residues Define Epitope and Paratope Size and Structure. **The Journal of Immunology**, v. 191, n. 3, p. 1428–1435, 2013.

STROHL, W. R. Current progress in innovative engineered antibodies. **Protein and Cell**, v. 9, n. 1, p. 86–120, 2018.

SUN, Z. et al. Brief introduction of current technologies in isolation of broadly neutralizing HIV-1 antibodies. **Virus research**, v. 243, p. 75–82, out. 2017.

SWINDELLS, M. B. et al. abYsis: Integrated Antibody Sequence and Structure-Management, Analysis, and Prediction. **Journal of molecular biology**, v. 429, n. 3, p. 356–364, 2017.

TEPLYAKOV, A. et al. Structural diversity in a human antibody germline library. **mAbs**, v. 8, n. 6, p. 1045–1063, 17 ago. 2016.

TILLER, K. E.; TESSIER, P. M. Advances in Antibody Design. **Annual Review of Biomedical Engineering**, v. 17, n. 1, p. 191–216, 2015.

VAN DER WALT, S.; COLBERT, S. C.; VAROQUAUX, G. The NumPy array: A structure for efficient numerical computation. **Computing in Science and Engineering**, v. 13, n. 2, p. 22–30, 2011.

VIART, B. et al. EPI-peptide designer: a tool for designing peptide ligand libraries based on epitope–paratope interactions. **Bioinformatics**, v. 32, n. 10, p. 1462–1470, 2016.

VITA, R. et al. The immune epitope database (IEDB) 3.0. **Nucleic Acids Research**, v. 43, n. D1, p. D405–D412, 2015.

WARDEMANN, H.; BUSSE, C. E. Novel Approaches to Analyze Immunoglobulin Repertoires. **Trends in Immunology**, v. 38, n. 7, p. 471–482, 2017.

WATSON, C. T.; GLANVILLE, J.; MARASCO, W. A. The Individual and Population Genetics of Antibody Immunity. **Trends in Immunology**, v. 38, n. 7, p. 459–470, 2017.

WEBSTER, D. M.; HENRY, A. H.; REES, A. R. Antibody-antigen interactions. **Current Opinion in Structural Biology**, v. 4, n. 1, p. 123–129, 1994.

WEITZNER, B. D. et al. Modeling and docking of antibody structures with Rosetta. **Nature Protocols**,

v. 12, n. 2, p. 401–416, 2017.

WILKINS, M. R. et al. Protein Identification and Analysis Tools in the ExPASy Server. In: **2-D Proteome Analysis Protocols**. New Jersey: Humana Press, 1999. v. 112. p. 531–552.

WILSON, A.; STANFIELD, R. L. Antibody-antigen interactions. **Current Opinion in Structural Biology**, v. 3, n. 1, p. 113–118, 1993.

WU, T. T.; KABAT, E. A. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. **The Journal of experimental medicine**, v. 132, n. 2, p. 211–250, 1970.

WU, X. et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. **Science**, v. 333, n. 6049, p. 1593–1602, 2011.

WU, X. et al. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. **Cell**, v. 161, n. 3, p. 480–485, 2015.

YAARI, G.; KLEINSTEIN, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. **Genome Medicine**, v. 7, n. 1, p. 1–14, 2015.

YOKOTA, A. et al. The role of hydrogen bonding via interfacial water molecules in antigen-antibody complexation: The HyHEL-10-HEL interaction. **Journal of Biological Chemistry**, v. 278, n. 7, p. 5410–5418, 2003.

ZAVRTANIK, U. et al. Structural Basis of Epitope Recognition by Heavy-Chain Camelid Antibodies. **Journal of Molecular Biology**, v. 430, n. 21, p. 4369–4386, 2018.

ZHANG, W. et al. PIRD: Pan immune repertoire database. **bioRxiv**, p. 399493, 2018.

ZHOU, T. et al. Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01. **Science**, v. 329, n. 5993, p. 811–817, 2010.

Apêndices

Apêndice 1 – Nomes padronizados baseados na taxonomia do UniProt padronizado correspondente

Apêndice 2 – Complexos sem redundância intraestrutura armazenados no Ydb

Apêndice 3 – Interações realizadas pelas 385 cadeias pesadas de anticorpo analisadas neste trabalho

Apêndice 1 – Nomes padronizados baseados na taxonomia do UniProt

Nome padronizado do organismo produtor de antígeno	Nome não padronizado presentes no SAbDab ou na estrutura do PDB
Adeno-associated dependoparvovirus A: Adeno-associated virus 2	Adeno-associated virus - 2
Alphacoronavirus: TGEV virulent Purdue	Tgev virulent purdue
Alphapapillomavirus 7	Human papillomavirus type 59
Alphapapillomavirus 9	Human papillomavirus type 16; Human papillomavirus type 58
Androctonus australis: Androctonus australis hector	Androctonus australis hector
Aquifex aeolicus	Aquifex aeolicus vf5; aquifex aeolicus (strainvf5)
Carnivore protoparvovirus 1: Canine parvovirus	Canine parvovirus
Centruroides noxius	Centruroides noxius hoffmann
Cervus canadensis: Cervus elaphus nelsoni	Cervus elaphus nelsoni
Clostridioides difficile: Clostridium difficile	Clostridium difficile
Clostridium botulinum	Clostridium botulinum (strain hall / atcc 3502/ nctc 13319 / type a)
Colobus angolensis	Colobus angolensis palliatus
Deinococcus geothermalis	Deinococcus geothermalis (strain dsm 11300)
Dengue virus: Dengue virus 1	Dengue virus 1; Dengue virus type 1 (strain nauru/westpac/1974)
Dengue virus: Dengue virus 2	Dengue virus 2; Dengue virus 2 puerto rico/pr159-s1/1969; Dengue virus 2 thailand/16681/84
Dengue virus: Dengue virus 3	Dengue virus 3
Dengue virus: Dengue virus 4	Dengue virus 4
Enterovirus A: Coxsackievirus A10	Coxsackievirus a10
Enterovirus A: Coxsackievirus a6	Coxsackievirus a6

Apêndice 1. (Continuação)

Nome padronizado do organismo produtor de antígeno	Nome não padronizado presentes no SAbDab ou na estrutura do PDB
Enterovirus C: Human poliovirus 1	Human poliovirus 1; Human poliovirus 1 mahoney; Poliovirus type 1; Poliovirus type 1 (strain mahoney)
Enterovirus D	Enterovirus d68
Escherichia coli	Escherichia coli (strain k12); Escherichia coli 2-156-04_s3_c3; Escherichia coli dh5[alpha]; Escherichia coli k-12; Escherichia coli o157
Escherichia virus 933W	Enterobacteria phage 933w
Geobacillus thermodenitrificans	Geobacillus thermodenitrificans (strain ng80-2)
Hendra henipavirus	Hendra virus
Hepatitis A virus (Hepatovirus A)	Hepatitis a virus
Hepatitis B virus	Hepatitis b virus; Hepatitis b virus genotype d subtype adw
Hepatitis C virus (Hepacivirus C)	Hepatitis c virus; Hepatitis c virus (isolate glasgow); Hepatitis c virus (isolate h); Hepatitis c virus genotype 1a (isolate h); Hepatitis c virus subtype 2a; Recombinant hepatitis c virus h77(5\'utr-ns2)/jfh1_v787a,q1247l; Hepacivirus c; Recombinant hepatitis c virus hk6a/jfh-1
Hepatitis E virus (Orthohepevirus A)	Hepatitis e virus
Human cytomegalovirus (Human herpesvirus 5)	Human cytomegalovirus; Human cytomegalovirus (strain 5508), Human cytomegalovirus (strain merlin); Human cytomegalovirus (strain merlin), Human cytomegalovirus (strain ad169); Human herpesvirus 5
Human herpesvirus 3 (Varicella-zoster virus)	Human herpesvirus 3 strain oka vaccine
Human herpesvirus 4 (Epstein Barr virus)	Human herpesvirus 4 (strain b95-8); Human herpesvirus 4
Human immunodeficiency virus 1	Hiv-1 m; Human immunodeficiency virus type 1; Human immunodeficiency virus type 1 (isolateyu2); Human immunodeficiency virus type 1 (jrcsfisolate); Human immunodeficiency virus type 1 (malisolate); Human immunodeficiency virus type 1 group msubtype b; Human immunodeficiency virus type 1(clone 12); Human immunodeficiency virus type 3
Human mastadenovirus C: Human adenovirus C serotype 5	Human adenovirus 5; Human adenovirus c serotype 5; human adenovirus5

Apêndice 1. (Continuação)

Nome padronizado do organismo produtor de antígeno	Nome não padronizado presentes no SAbDab ou na estrutura do PDB
Human respiratory syncytial virus	Human respiratory syncytial virus a; Human respiratory syncytial virus a (straina2); Human respiratory syncytial virus a (strainrsb6256)
Human respirovirus 3	Human parainfluenza virus 3
Indiana vesiculovirus: Vesicular stomatitis indiana virus	Vesicular stomatitis indiana virus; Vesicular stomatitis indiana virus (strain sanjuan)
Influenza A virus	Influenza a virus; Influenza a virus (a/anhui/1/2005(h5n1)); Influenza a virus (a/anhui/1-balf_rg45/2013(h7n9)); Influenza a virus (a/brevigmission/1/1918(h1n1)); Influenza a virus (a/california/07/2009(h1n1)); Influenza a virus (a/chicken/thailand/tak-01/2004(h5n1)) ; Influenza a virus (a/hong kong/1-4-ma21-1/1968(h3n2)) ; Influenza a virus (a/puerto rico/8/1934(h1n1)); Influenza a virus (a/solomonislands/3/2006(h1n1)) ; Influenza a virus (a/texas/50/2012(h3n2)); Influenza a virus (a/viet nam/1203/2004(h5n1)); Influenza a virus (a/wsn/1933(h1n1)); Influenza a virus (a/x-31(h3n2)); Influenza a virus (strain a/hong kong/1/1968h3n2) ; Influenza a virus (strain swla/california/04/2009 h1n1) ; Influenza a vírus (a/chicken/henan/109/2013(h7n9)) ; Influenza a vírus (a/turkey/italy/214845/2002(h7n3)); Influenza a virus (a/thailand/1(kan-1)/2004(h5n1)) ; Influenza a virus (straina/duck/alberta/35/1976 h1n1); Influenza a vírus (a/swine/minnesota/a01134337/2010(h3n2)); Influenza a virus (a/hong kong/482/97(h5n1)); Influenza a virus (strain a/aichi/2/1968 h3n2)
Influenza B virus	Influenza b virus; Influenza b virus (b/brisbane/60/2008)
Lactococcus phage p2	Lactococcus lactis phage p2
Lactococcus phage Tuc2009	Lactococcus phage tuc2009
Leishmania donovani	Leishmania donovani (strain bpk282a1)
Mamastrovirus 1: Human astrovirus-2	Human astrovirus-2
Marburg marburgvirus	Lake victoria Marburgvirus; Lake victoria marburgvirus (strain musoke-80)
Metarhizium acridum	Metarhizium acridum (strain cqma 102)
Methanocaldococcus jannaschii	Methanocaldococcus jannaschii (strain atcc43067 / dsm 2661 / jal-1 / jcm 10045 / nbrc 100440)
Methanococcus maripaludis	Methanococcus maripaludis (strain s2 / ll)

Apêndice 1. (Continuação)

Nome padronizado do organismo produtor de antígeno	Nome não padronizado presentes no SAbDab ou na estrutura do PDB
Middle East respiratory syndrome-related coronavirus	Human coronavirus emc; Human coronavirus emc (isolate unitedkingdom/h123990006/2012) ; Middle east respiratory syndrome coronavirus; Middle east respiratory syndrome-relatedcoronavirus ; Betacoronavirus england 1
Neisseria meningitidis	Neisseria meningitidis mc58; Neisseria meningitidis serogroup b (strainmc58)
Norwalk virus	Human calicivirus; Norovirus 13-bh-1/2013/gii.17; Norovirus gii.10; Norovirus hu/gii.4/sydney/nsw0514/2012/au; Norovirus hu/gii-4/saga1/2006/jp; Norwalk virus (strain gi/human/unitedstates/norwalk/1968)
Parechovirus A	Human parechovirus 3
Phage #D	Phage #d
Plasmodium falciparum	Plasmodium falciparum (isolate 3d7); Plasmodium falciparum 3d7; Plasmodium falciparum k1
Plasmodium vivax	Plasmodium vivax (strain salvador i); Plasmodium vivax sai-1
Primate erythroparvovirus 1	Human parvovirus b19
Pseudomonas aeruginosa	Pseudomonas aeruginosa pao1
Rhinovirus A	Human rhinovirus 2
Rhinovirus B	Human rhinovirus 14
Rift Valley fever phlebovirus	Rift valley fever virus
Rotavirus A	Simian rotavirus
Saccharomyces cerevisiae	Saccharomyces cerevisiae (atcc 204508 / s288c); Saccharomyces cerevisiae (strain atcc 204508 /s288c); Saccharomyces cerevisiae s288c
Salmonella choleraesuis	Salmonella enterica subsp. enterica serovartyphimurium
Severe acute respiratory syndrome-related coronavirus	Sars coronavirus
SFTS phlebovirus	Severe fever with thrombocytopenia virus

Apêndice 1. (Continuação)

Nome padronizado do organismo produtor de antígeno	Nome não padronizado presentes no SAbDab ou na estrutura do PDB
Staphylococcus aureus	Staphylococcus aureus (strain bovine rf122 /et3-1) ; Staphylococcus aureus (strain mssa476); Staphylococcus aureus (strain usa300); Staphylococcus aureus subsp. aureus; Staphylococcus aureus subsp. aureus tch60;
Staphylococcus capitis	Staphylococcus capitis vcu116
Thielavia heterothallica	Thermothelomyces thermophila (strain atcc 42464/ bcrc 31852 / dsm 1799)
Tick-borne encephalitis virus	Tick-borne encephalitis virus (strain hypr); Tick-borne encephalitis virus european subtype(strain neudoerfl)
Trypanosoma brucei	Trypanosoma brucei rhodesiense
Trypanosoma congolense	Trypanosoma congolense (strain il3000)
Vibrio cholerae	Vibrio cholerae serotype o1 (strain atcc 39315/ el tor inaba n16961)
Yellow fever virus	Yellow fever virus (strain 17d vaccine)
Zaire ebolavirus	Zaire ebola virus; Zaire ebolavirus (strain mayinga-76); Ebola virus, Zaire ebolavirus
Zika virus	Zika virus (strain mr 766); Zika virus (isolate zikv/human/frenchpolynesia/10087pf/2013)

Apêndice 2 – Complexos sem redundância intraestrutura armazenados no Ydb

Id PDB	Pesada	Leve	Representativo
1A14	H	L	-
1A2Y	B	A	Sim
1A3R	H	L	-
1ACY	H	L	-
1ADQ	H	L	-
1AFV	H	L	-
1AHW	E	D	-
1AI1	H	L	Sim
1AR1	C	D	-
1BGX	H	L	-
1BJ1	H	L	Sim
1BOG	B	A	-
1BQL	H	L	-
1BVK	E	D	-
1BZQ	K	-	-
1C08	B	A	-
1CE1	H	L	Sim
1CFN	B	A	-
1CFS	B	A	-
1CFT	B	A	-
1CU4	H	L	-
1CZ8	Y	X	-
1DEE	F	E	-
1DQJ	B	A	-
1DZB	A	A	Sim
1E4W	H	L	-
1E4X	H	L	Sim
1E6J	H	L	-
1EGJ	H	L	Sim
1EJO	H	L	-
1EO8	H	L	-
1EZV	X	Y	-
1F3R	B	B	Sim
1F58	H	L	Sim
1F90	H	L	-
1FBI	H	L	-
1FDL	H	L	-
1FE8	H	L	-
1FJ1	B	A	-

Id PDB	Pesada	Leve	Representativo
1FNS	H	L	Sim
1FPT	H	L	Sim
1FRG	H	L	-
1FSK	I	H	-
1G6V	K	-	-
1G7H	B	A	-
1G7I	B	A	-
1G7J	B	A	-
1G7L	B	A	-
1G7M	B	A	-
1G9M	H	L	-
1G9N	H	L	-
1GC1	H	L	-
1GGI	H	L	-
1H0D	B	A	Sim
1HEZ	D	C	-
1HH6	B	A	-
1HH9	B	A	-
1HI6	B	A	Sim
1HIM	L	H	-
1HIN	H	L	-
1HYS	D	C	-
1I8I	B	A	-
1I8K	B	A	Sim
1I9R	K	M	-
1IC4	H	L	-
1IC5	H	L	-
1IC7	H	L	-
1IFH	H	L	-
1IKF	H	L	-
1IQD	B	A	-
1J1O	H	L	-
1J1P	H	L	-
1J1X	H	L	-
1J5O	H	L	-
1JHL	H	L	Sim
1JP5	A	A	-
1JPS	H	L	Sim
1JRH	H	L	-

1K4C	A	B	-
1K4D	A	B	-
1KB5	H	L	Sim
1KB9	J	K	-
1KC5	H	L	Sim
1KCR	H	L	-
1KCS	H	L	-
1KEN	T	U	-
1KIP	B	A	-
1KIQ	B	A	-
1KIR	B	A	-
1KTR	H	L	-
1KXQ	G	-	-
1KXT	B	-	-
1KXV	C	-	-
1KYO	J	K	-
1LK3	I	M	Sim
1MCB	-	B	-
1MCC	-	A	-
1MCD	-	A	-
1MCE	-	A	-
1MCF	-	A	-
1MCH	-	A	-
1MCI	-	A	-
1MCJ	-	A	-
1MCK	-	A	-
1MCL	-	B	-
1MCN	-	A	-
1MCQ	-	A	-
1MCR	-	A	-
1MCS	-	A	-
1MHP	H	L	-
1MLC	B	A	-
1MPA	H	L	Sim
1MVF	B	-	-
1MVU	B	A	Sim
1N0X	K	M	Sim
1N5Y	H	L	-
1N64	H	L	Sim
1N6Q	H	L	-
1N8Z	B	A	-
1NAK	H	L	-
1NBY	B	A	-
1NBZ	B	A	-
1NCA	H	L	Sim
1NCB	H	L	-
1NCC	H	L	-
1NCD	H	L	-
1NDG	B	A	-
1NDM	B	A	-
1NFD	H	G	Sim
1NL0	H	L	Sim
1NMA	H	L	-
1NMB	H	L	Sim
1NMC	H	L	-
1NSN	H	L	-
1OAK	H	L	-
1OAZ	H	L	-
1OB1	B	A	-
1OP9	A	-	-
1ORQ	B	A	-
1ORS	B	A	Sim
1OSP	H	L	Sim
1OTS	E	F	-
1OTT	E	F	-
1OTU	E	F	-
1P2C	B	A	Sim
1P4B	H	L	Sim
1P84	J	K	-
1PKQ	B	A	-
1PZ5	B	A	Sim
1Q1J	I	M	-
1QFU	H	L	-
1QFW	I	M	-
1QFW	H	L	-
1QKZ	H	L	-
1QLE	H	L	-
1QNZ	H	L	Sim
1R0A	H	L	Sim
1R3I	H	L	-
1R3J	B	A	-
1R3K	B	A	-
1R3L	B	A	-
1RI8	A	-	-
1RJL	B	A	-
1RVF	H	L	-
1RZJ	H	L	-
1RZK	H	L	-
1S5H	B	A	-
1S78	F	E	-

1SM3	H	L	Sim
1SVZ	B	B	Sim
1SY6	H	L	Sim
1T03	H	L	-
1TET	H	L	-
1TJG	H	L	-
1TJH	H	L	-
1TJI	H	L	-
1TPX	B	C	-
1TQB	B	C	-
1TQC	B	C	-
1TZG	I	M	-
1TZH	B	A	-
1TZI	B	A	-
1U8H	B	A	-
1U8I	B	A	-
1U8J	B	A	-
1U8K	B	A	-
1U8L	B	A	-
1U8M	B	A	-
1U8N	B	A	-
1U8O	B	A	-
1U8P	B	A	-
1U8Q	B	A	-
1U91	B	A	-
1U92	B	A	-
1U93	B	A	-
1U95	B	A	-
1UA6	H	L	-
1UAC	H	L	-
1UJ3	B	A	-
1UWX	H	L	-
1V7M	I	M	Sim
1V7N	J	N	-
1VFB	B	A	-
1W72	I	M	Sim
1WEJ	H	L	Sim
1XCQ	D	C	-
1XCT	D	C	-
1XF5	B	A	-
1XGP	B	A	-
1XGQ	B	A	-
1XGR	B	A	-
1XGT	B	A	-
1XGU	B	A	-
1XGY	H	L	Sim
1XIW	H	G	-
1YJD	H	L	-
1YMH	D	C	-
1YNT	D	C	-
1YQV	H	L	Sim
1YY9	D	C	Sim
1YYL	H	L	-
1YYM	H	L	-
1Z3G	I	M	-
1ZA3	H	L	-
1ZEA	H	L	Sim
1ZMY	A	-	-
1ZTX	H	L	-
1ZV5	A	-	-
1ZVY	A	-	-
1ZWI	A	B	-
2A0L	F	E	-
2A6D	B	A	-
2A6I	B	A	-
2A6K	H	L	-
2ADF	H	L	-
2AEP	H	L	Sim
2AEQ	H	L	-
2AP2	D	C	-
2ARJ	H	L	-
2ATK	A	B	-
2B0S	H	L	-
2B1A	H	L	-
2B1H	H	L	Sim
2B2X	H	L	Sim
2B4C	H	L	-
2BDN	H	L	-
2BOB	A	B	-
2BOC	A	B	-
2BRR	H	L	Sim
2BSE	F	-	-
2CK0	H	L	Sim
2CMR	H	L	-
2DD8	H	L	-
2DQC	H	L	-
2DQD	H	L	-
2DQE	H	L	-
2DQF	E	D	-
2DQG	H	L	-

2DQH	H	L	-
2DQI	H	L	-
2DQJ	H	L	-
2DWD	A	B	-
2DWE	A	B	-
2EH8	H	L	-
2EIZ	B	A	Sim
2EKS	B	A	-
2EXW	C	D	-
2EXY	E	F	-
2EZ0	C	D	-
2F58	H	L	-
2F5B	H	L	-
2FD6	H	L	Sim
2FEC	I	L	-
2FED	C	D	-
2FEE	J	O	-
2FJG	H	L	-
2FJH	B	A	-
2FX7	H	L	-
2FX8	J	N	-
2FX9	H	L	-
2G5B	D	C	Sim
2GHW	B	B	Sim
2GSI	F	E	Sim
2H1P	H	L	-
2H2P	E	F	-
2H2S	E	F	-
2H8P	A	B	-
2H9G	H	L	-
2HFE	A	B	-
2HFG	H	L	-
2HG5	A	B	-
2HH0	H	L	-
2HJF	A	B	-
2HKF	H	L	-
2HLF	C	D	-
2HMI	D	C	-
2HRP	H	L	-
2HT2	C	D	-
2HT3	C	D	-
2HT4	E	F	-
2HTK	C	D	-
2HTL	C	D	-
2HVJ	A	B	-
2HVK	A	B	-
2I5Y	R	Q	-
2I60	H	L	-
2I9L	F	E	-
2IBZ	X	Y	-
2IFF	H	L	-
2IGF	H	L	-
2IH1	A	B	-
2IH3	A	B	Sim
2IPU	G	K	-
2ITC	A	B	-
2ITD	A	B	-
2J4W	H	L	Sim
2J5L	C	B	-
2J6E	I	M	-
2J88	H	L	-
2JEL	H	L	-
2JIX	H	L	-
2JK5	A	B	-
2KH2	B	B	Sim
2LTQ	F	E	Sim
2MPA	H	L	-
2NLJ	B	A	-
2NR6	F	E	-
2NXY	D	C	-
2NXZ	D	C	-
2NY0	D	C	-
2NY1	D	C	Sim
2NY2	D	C	-
2NY3	D	C	-
2NY4	D	C	-
2NY5	H	L	-
2NY6	D	C	-
2NY7	H	L	-
2NYY	D	C	Sim
2NZ9	F	E	-
2OQJ	B	A	-
2OR9	I	M	-
2OSL	H	L	-
2OTU	D	C	Sim
2OTW	B	A	-
2OZ4	H	L	-
2P42	D	-	-
2P43	B	-	-
2P44	B	-	-

2P45	B	-	-
2P46	B	-	-
2P47	B	-	-
2P48	B	-	-
2P49	B	-	-
2P4A	B	-	-
2P7T	A	B	-
2P8L	B	A	-
2P8M	B	A	-
2P8P	B	A	-
2PW1	B	A	-
2PW2	B	A	-
2Q8A	H	L	-
2Q8B	H	L	Sim
2QAD	D	C	-
2QHR	H	L	-
2QQK	H	L	-
2QQL	H	L	-
2QQN	H	L	-
2QR0	R	Q	-
2QSC	H	L	-
2R0K	H	L	-
2R0L	H	L	-
2R0W	H	L	-
2R0Z	H	L	-
2R29	H	L	-
2R4R	H	L	Sim
2R4S	H	L	-
2R56	I	M	-
2R69	H	L	-
2R6P	D	E	-
2R9H	E	F	-
2UUD	J	K	-
2UZI	H	L	Sim
2V17	H	L	Sim
2VC2	H	L	-
2VDK	H	L	-
2VDL	H	L	-
2VDM	H	L	-
2VDN	H	L	-
2VDO	H	L	-
2VDP	H	L	-
2VDQ	H	L	-
2VDR	H	L	-
2VH5	H	L	-
2VIR	B	A	-
2VIS	B	A	-
2VIT	B	A	-
2VWE	E	C	-
2VXQ	H	L	-
2VXS	J	N	-
2VXT	H	L	Sim
2VYR	E	-	-
2W0F	A	B	-
2W65	C	D	Sim
2W9E	H	L	-
2WUB	R	Q	-
2WUC	H	L	-
2WZP	J	-	-
2X7L	G	I	-
2X89	A	-	-
2XQB	H	L	Sim
2XQY	G	L	-
2XRA	H	L	Sim
2XT1	B	-	-
2XTJ	D	B	-
2XV6	D	-	-
2XWT	A	B	Sim
2XXM	B	-	-
2XZQ	H	L	Sim
2Y06	H	L	-
2Y07	H	L	-
2Y36	H	L	-
2Y5T	A	B	-
2Y6S	H	L	-
2YBR	D	E	-
2YC1	A	B	Sim
2YPV	H	L	Sim
2YSS	B	A	-
2ZCH	H	L	Sim
2ZCK	H	L	-
2ZCL	H	L	-
2ZJS	H	L	-
2ZPK	H	L	-
2ZUQ	F	E	-
3A67	H	L	-
3A6B	H	L	-
3A6C	H	L	-
3AB0	B	C	-
3B2U	F	G	Sim

3B2V	H	L	-
3B9K	H	L	Sim
3BAE	H	L	-
3BDY	H	L	-
3BE1	H	L	-
3BGF	H	L	-
3BKJ	H	L	-
3BKY	H	L	-
3BN9	D	C	-
3BSZ	N	M	-
3BT2	H	L	-
3C09	H	L	-
3C2A	H	L	-
3CFI	F	-	-
3CK0	H	L	-
3CSY	E	F	-
3CVH	H	L	Sim
3CX5	U	V	Sim
3CXD	H	L	-
3CXH	U	V	-
3D0L	B	A	-
3D0V	B	A	-
3D85	B	A	Sim
3D9A	H	L	Sim
3DET	C	D	-
3DRO	B	A	-
3DRQ	B	A	-
3DRT	B	A	-
3DSF	H	L	-
3DVG	B	A	-
3DVN	B	A	-
3E8U	H	L	Sim
3EBA	A	-	-
3EFD	H	L	-
3EFF	B	A	-
3EGS	B	A	-
3EHB	C	D	-
3EJY	C	D	-
3EJZ	C	D	-
3EO1	H	G	-
3EOA	H	L	-
3EOB	H	L	-
3ETB	F	F	-
3EYF	B	A	-
3EYS	H	L	-
3EYU	H	L	-
3EZJ	B	-	-
3F58	H	L	-
3F5W	A	B	-
3F7V	A	B	-
3F7Y	A	B	-
3FB5	A	B	-
3FB6	A	B	-
3FB7	A	B	-
3FB8	A	B	-
3FFD	A	B	-
3FKU	U	U	Sim
3FMG	H	L	-
3FN0	H	L	Sim
3G04	B	A	-
3G5V	B	A	-
3G5Y	B	A	Sim
3G6D	H	L	-
3G6J	H	G	-
3G9A	B	-	-
3GB7	A	B	-
3GBM	H	L	-
3GBN	H	L	Sim
3GGW	B	A	Sim
3GHB	H	L	-
3GHE	H	L	-
3GI8	H	L	-
3GI9	H	L	Sim
3GJF	H	L	-
3GO1	H	L	Sim
3GRW	H	L	-
3H0T	B	A	Sim
3H3B	D	D	Sim
3H3P	I	M	-
3H42	H	L	Sim
3HAE	H	L	-
3HB3	C	D	Sim
3HFM	H	L	-
3HI1	H	L	-
3HI6	H	L	-
3HMX	H	L	-
3HPL	A	B	-
3HR5	B	A	-
3I50	H	L	Sim
3IDG	B	A	Sim

3IDI	B	A	-
3IDJ	B	A	-
3IDM	B	A	-
3IDN	B	A	-
3IDX	H	L	Sim
3IDY	B	C	-
3IET	D	C	-
3IFL	H	L	Sim
3IFN	H	L	-
3IFO	A	B	-
3IFP	E	F	-
3IGA	A	B	-
3IU3	A	B	-
3IXT	H	L	-
3IXX	G	H	-
3IXY	I	J	-
3IYW	H	L	-
3J1S	H	L	-
3J3O	H	L	-
3J3P	H	L	-
3J42	K	L	-
3J5M	H	G	-
3J6U	H	L	-
3J70	M	N	Sim
3J8D	D	D	-
3JAB	H	L	-
3JAU	H	L	-
3JBA	H	L	-
3JBC	7	-	-
3JBD	7	-	-
3JBE	7	-	-
3JBF	7	-	-
3JBQ	H	L	-
3JCB	C	B	Sim
3JCC	C	B	-
3JCX	H	L	Sim
3JWD	P	O	-
3JWO	H	L	-
3K1K	D	-	-
3K2U	H	L	-
3K74	B	-	-
3K7U	A	-	-
3K80	B	-	-
3K81	B	-	-
3KJ4	H	L	-
3KJ6	H	L	-
3KLH	D	C	-
3KR3	H	L	-
3KS0	K	J	-
3L5W	B	A	-
3L5X	H	L	Sim
3L95	B	A	-
3LD8	C	B	Sim
3LDB	C	B	-
3LEV	H	L	-
3LEX	H	L	Sim
3LEY	H	L	-
3LH2	K	O	-
3LHP	H	L	-
3LIZ	H	L	Sim
3LQA	H	L	-
3LRH	-	G	-
3LZF	H	L	Sim
3MA9	H	L	Sim
3MAC	H	L	-
3MJ9	H	L	Sim
3MLR	H	L	Sim
3MLS	K	O	-
3MLT	H	L	-
3MLU	H	L	-
3MLV	N	M	-
3MLW	H	L	-
3MLX	I	M	-
3MLY	H	L	Sim
3MLZ	H	L	-
3MNW	B	A	-
3MNZ	B	A	Sim
3MOA	H	L	-
3MOB	H	L	-
3MOD	H	L	-
3MXW	H	L	-
3N85	H	L	-
3NCY	P	S	-
3NFP	H	L	-
3NGB	B	C	-
3NH7	K	O	-
3NID	H	L	-
3NIF	E	F	-
3NIG	E	F	-
3NPS	B	C	Sim

3O0R	H	L	-
3O2D	H	L	Sim
3O41	H	L	Sim
3O45	A	B	-
3O6L	H	L	-
3O6M	H	L	-
3OGC	A	B	-
3OGO	G	-	-
3OPZ	I	M	-
3OR6	A	B	-
3OR7	A	B	-
3P0G	B	-	-
3P0Y	H	L	Sim
3P11	H	L	-
3P30	H	L	-
3P9W	F	-	-
3PGF	H	L	-
3PJS	B	A	-
3PNW	B	A	-
3PP4	H	L	Sim
3Q1S	H	L	-
3Q3G	K	J	-
3QA3	B	A	-
3QG6	H	L	Sim
3QNZ	B	A	Sim
3QO0	B	A	-
3QSK	B	-	-
3QUM	B	A	-
3QUM	K	M	-
3QWO	H	L	Sim
3R08	H	L	-
3R1G	H	L	-
3RAJ	H	L	-
3RHW	F	N	-
3RI5	I	O	-
3RIA	I	O	-
3RIF	H	L	-
3RJQ	B	-	-
3RKD	D	C	Sim
3RU8	H	L	-
3RVV	D	C	Sim
3RVW	D	C	-
3RVX	D	C	-
3S35	H	L	-
3S36	H	L	-
3S37	H	L	-
3S88	H	L	-
3SDY	H	L	-
3SE8	H	L	Sim
3SE9	H	L	Sim
3SGE	H	L	-
3SKJ	H	L	-
3SN6	N	-	-
3SO3	C	B	-
3SOB	H	L	-
3SQO	H	L	-
3STB	B	-	-
3STL	A	B	-
3STZ	A	B	-
3T2N	H	L	Sim
3T3M	E	F	-
3T3P	H	L	Sim
3TT1	I	M	-
3TT3	H	L	-
3U0T	B	A	Sim
3U2S	H	L	Sim
3U30	C	B	-
3U4E	A	B	-
3U7Y	H	L	-
3U9P	H	L	-
3U9U	A	B	-
3UAJ	C	D	-
3UBX	H	L	-
3UC0	I	M	Sim
3UJI	H	L	Sim
3UJJ	H	L	-
3ULU	D	C	-
3ULU	H	L	-
3ULU	F	E	-
3ULV	H	L	-
3ULV	D	C	-
3ULV	F	E	-
3UO1	H	L	-
3UX9	B	B	-
3UYP	A	A	-
3UYR	H	L	-
3UZE	B	B	-
3UZQ	A	A	Sim
3UZV	B	B	-
3V0A	C	-	-

3V4P	H	L	-
3V4U	H	L	Sim
3V4V	H	L	Sim
3V52	H	L	-
3V6O	D	F	Sim
3V6Z	A	B	-
3V7A	E	H	-
3VE0	A	B	-
3VG9	C	B	Sim
3VGA	C	B	-
3VI3	F	E	-
3VI4	H	L	-
3VRL	H	L	-
3W11	C	D	-
3W12	C	D	-
3W13	C	D	-
3W2D	H	L	-
3W9E	A	B	-
3WD5	H	L	-
3WFB	H	L	-
3WFC	H	L	-
3WFD	H	L	Sim
3WFE	H	L	-
3WHE	O	P	-
3WIH	H	L	Sim
3WKM	H	L	Sim
3WLW	C	D	-
3WSQ	H	L	Sim
3WXV	H	L	-
3WXW	H	L	Sim
3X3F	H	L	Sim
3ZDX	H	L	-
3ZDY	H	L	-
3ZDZ	E	F	-
3ZE0	H	L	-
3ZE1	E	F	-
3ZE2	H	L	-
3ZKM	H	L	-
3ZKN	H	L	-
3ZKQ	D	-	-
3ZKS	D	-	-
3ZKX	B	-	-
3ZKX	C	-	-
3ZLQ	C	-	-
3ZTJ	G	H	-
3ZTN	H	L	-
4AEI	H	L	-
4AG4	H	L	-
4AL8	H	L	Sim
4ALA	H	L	-
4AM0	C	D	-
4BEL	D	-	-
4BFB	E	-	-
4BH7	B	A	-
4BH8	B	A	-
4BKL	A	B	-
4BZ1	H	L	-
4BZ2	H	L	-
4C2I	H	L	-
4C57	C	-	-
4C58	B	-	-
4C59	B	-	-
4CAD	E	D	-
4CAU	E	E	-
4CDG	D	-	-
4CKD	H	L	-
4CMH	B	C	-
4CNI	H	L	-
4D3C	H	L	-
4D9Q	E	D	Sim
4D9R	H	L	-
4DAG	H	L	-
4DGI	H	L	-
4DGV	H	L	-
4DGY	H	L	Sim
4DK3	A	-	-
4DK6	B	-	-
4DKA	A	-	-
4DKE	I	M	-
4DKF	H	L	-
4DN4	H	L	-
4DQO	H	L	-
4DTG	H	L	Sim
4DVR	H	L	-
4DW2	H	L	-
4EDW	H	L	-
4EDX	H	L	-
4EIG	B	-	-
4EIZ	C	-	-
4EJ1	C	-	-

4ENE	E	F	Sim
4ERS	H	L	Sim
4ETQ	H	L	-
4F15	K	L	-
4F2M	C	D	-
4F37	H	L	-
4F3F	B	A	-
4F9L	D	D	Sim
4F9P	D	D	-
4FFV	H	L	Sim
4FFW	D	C	-
4FFY	H	L	Sim
4FFZ	Z	Y	-
4FG6	E	F	-
4FHB	D	-	-
4FP8	H	L	-
4FQI	H	L	Sim
4FQJ	H	L	Sim
4FQK	E	F	-
4FQR	a	b	-
4FQV	J	N	-
4FQY	H	L	-
4G3Y	H	L	Sim
4G6A	C	D	Sim
4G6F	H	L	-
4G6J	H	L	Sim
4G6M	H	L	Sim
4G7V	H	L	-
4G7Y	H	L	-
4G80	G	H	-
4GAG	H	L	-
4GAJ	H	L	-
4GFT	B	-	-
4GMS	J	N	-
4GRW	F	-	-
4GRW	E	-	-
4GRW	H	-	-
4GXU	M	N	-
4H0H	B	B	-
4H88	H	L	Sim
4H8W	H	L	Sim
4HC1	M	N	-
4HCR	M	N	-
4HF5	H	L	-
4HFU	H	L	-
4HG4	P	Q	-
4HHA	B	A	Sim
4HIX	H	L	-
4HJ0	C	D	-
4HJG	B	A	-
4HJJ	H	-	-
4HKX	A	B	-
4HKZ	B	A	-
4HLZ	I	J	Sim
4HPO	H	L	Sim
4HPY	H	L	Sim
4HS6	B	A	-
4HS8	H	L	-
4HT1	H	L	Sim
4HWB	H	L	-
4HZL	H	L	-
4I13	B	-	-
4I18	H	L	-
4I1N	B	-	-
4I2X	D	C	-
4I3R	H	L	-
4I3S	H	L	-
4I77	H	L	Sim
4I9W	E	D	-
4IDJ	H	L	-
4IJ3	C	B	-
4IOF	E	F	-
4IOI	B	A	-
4IRZ	H	L	-
4J4P	H	L	Sim
4J6R	H	L	Sim
4J8R	B	A	-
04/jan	A	B	Sim
4JB9	H	L	-
4JDT	H	L	-
4JFX	H	L	-
4JFZ	H	L	-
4JG0	H	L	-
4JG1	H	L	Sim
4JHW	H	L	-
4JKP	H	L	-
4JLR	H	L	-
4JM2	D	C	-
4JM2	A	B	-
4JO1	H	L	-

4JO2	H	L	-
4JO3	I	M	-
4JPK	H	L	-
4JPV	H	L	Sim
4JPW	H	L	Sim
4JQI	H	L	-
4JR9	H	L	Sim
4JRE	H	L	-
4JZJ	H	L	-
4JZN	I	P	-
4JZO	G	H	-
4K24	H	L	-
4K2U	H	L	-
4K3J	H	L	-
4K8R	D	C	-
4K8R	H	L	-
4K94	H	L	-
4K9E	H	L	-
4KDT	A	-	-
4KFZ	D	-	-
4KHT	H	L	-
4KHX	H	L	-
4KI5	E	F	Sim
4KI5	C	D	Sim
4KJP	C	D	-
4KJQ	E	F	-
4KJW	C	D	-
4KK5	E	F	-
4KK6	E	F	-
4KK8	E	F	-
4KK9	E	F	-
4KKA	C	D	-
4KKB	E	F	-
4KKC	C	D	-
4KKL	C	D	-
4KML	B	-	-
4KRL	B	-	-
4KRM	F	-	-
4KRO	B	-	-
4KRO	D	C	-
4KRP	B	-	-
4KRP	D	C	-
4KSD	B	-	-
4KUC	H	D	-
4KV5	J	I	-
4KVN	H	L	-
4KXZ	N	M	-
4L5F	H	L	Sim
4LAJ	M	-	-
4LBE	A	B	-
4LCU	A	B	-
4LDE	B	-	-
4LDL	B	-	-
4LDO	B	-	-
4LEO	A	B	-
4LF3	E	D	-
4LIQ	H	L	-
4LKX	A	B	Sim
4LMQ	E	I	-
4LOU	C	D	-
4LQF	H	L	-
4LSP	H	L	Sim
4LSQ	H	L	-
4LSR	H	L	-
4LSS	H	L	-
4LST	H	L	-
4LSU	H	L	Sim
4LSV	H	L	-
4LU5	I	M	-
4LVH	H	I	Sim
4LVN	C	B	Sim
4LVO	C	B	-
4M1C	E	F	-
4M1D	I	M	Sim
4M1G	H	L	Sim
4M3K	B	-	-
4M48	H	L	-
4M5Z	H	L	Sim
4M62	I	M	Sim
4M7L	H	L	-
4M8Q	A	B	-
4MA7	H	L	-
4MA8	H	L	-
4MHH	H	L	-
4MHJ	K	J	-
4MQX	E	F	-
4MSW	A	B	-
4MWF	A	B	Sim
4MXV	H	L	-
4MXW	H	L	-

4N0Y	H	L	Sim
4N1C	-	B	-
4N1E	-	B	-
4N1H	B	-	-
4N8C	H	L	Sim
4N9G	M	N	Sim
4N9O	B	-	-
4NBX	B	-	-
4NBY	B	-	-
4NBZ	D	-	-
4NC0	B	-	-
4NC1	D	-	-
4NC1	E	-	-
4NC2	B	-	-
4NCO	L	K	-
4NGH	H	L	-
4NHC	H	L	-
4NHH	O	Q	-
4NIK	B	B	-
4NM8	H	L	-
4NNP	H	L	-
4NP4	H	L	-
4NRX	H	L	-
4NZR	H	L	Sim
4NZT	H	L	-
4O02	H	L	-
4O4Y	H	L	Sim
4O51	B	A	-
4O58	H	L	Sim
4O5I	W	X	-
4OD2	B	A	-
4ODX	H	B	-
4OGA	C	D	-
4OGX	H	L	-
4OGY	H	L	-
4OII	I	M	-
4OJF	H	L	Sim
4OKV	A	B	-
4OLU	H	L	-
4OLV	H	L	-
4OLW	H	L	-
4OLX	H	L	-
4OLY	H	L	-
4OLZ	H	L	Sim
4OM0	H	L	-
4OM1	H	L	-
4ONF	H	L	-
4ONG	H	L	-
4OQT	H	L	-
4ORZ	C	-	-
4OT1	H	L	Sim
4P3C	H	L	Sim
4P3D	A	B	-
4P59	H	L	-
4PD4	J	K	-
4PGJ	A	-	-
4PIR	H	-	-
4PLJ	H	L	Sim
4PLK	I	G	-
4POU	B	-	-
4PP1	D	C	-
4PP2	D	C	-
4PS4	H	L	-
4PY8	I	J	-
4Q0X	H	L	-
4Q5Z	O	P	-
4Q6I	H	L	-
4QCI	H	L	-
4QEX	I	M	-
4QHU	B	A	-
4QKX	B	-	-
4QO1	A	-	-
4QTI	H	L	-
4QWW	D	C	-
4QXT	A	B	-
4QXU	H	L	-
4QY8	A	B	-
4QYO	A	B	Sim
4R0L	H	L	-
4R2G	N	M	Sim
4R3S	A	B	-
4R8W	H	L	Sim
4RAU	B	A	-
4RAV	A	B	-
4RDQ	G	F	-
4RFN	B	C	-
4RFO	H	L	-
4RGM	C	B	Sim
4RGN	B	C	-
4RGN	D	E	-

4RGO	H	L	Sim
4RIS	H	L	-
4RQS	D	C	-
4RRP	I	C	-
4RWY	H	L	Sim
4RX4	H	L	-
4S10	B	-	-
4S1Q	H	L	Sim
4S1R	H	L	Sim
4S1S	H	L	-
4TNV	W	f	Sim
4TNW	V	Z	Sim
4TQE	H	L	Sim
4TSA	H	L	-
4TSB	H	L	-
4TSC	H	L	Sim
4TTD	C	D	-
4TUJ	A	B	-
4TUK	H	L	-
4TUL	H	L	Sim
4TVP	H	L	-
4TVP	D	E	-
4TVS	b	-	-
4U0R	B	C	Sim
4U1G	E	F	-
4U3X	A	-	-
4U6G	A	B	-
4U6H	A	B	-
4U6V	K	M	Sim
4UAO	C	B	Sim
4UIF	K	L	-
4UIH	F	G	-
4UT6	I	M	-
4UT9	J	N	Sim
4UTA	H	L	-
4UTB	I	M	-
4UU9	A	B	-
4UUJ	A	B	-
4UV7	H	L	-
4V1D	A	B	-
4V1D	D	E	-
4W2O	A	-	-
4W2Q	G	-	-
4W6W	B	-	-
4W6Y	B	-	-
4WEB	H	L	-
4WEM	B	-	-
4WEN	B	-	-
4WEU	E	-	-
4WFE	G	F	Sim
4WFF	E	D	-
4WFG	G	F	-
4WFH	G	F	-
4WGV	B	-	-
4WGW	B	-	-
4WHT	A	B	-
4WHY	M	N	-
4WUU	E	D	-
4WV1	B	A	-
4WY7	H	L	-
4X7C	C	-	-
4X7D	C	-	-
4X7E	C	-	-
4X7F	D	-	-
4XAK	D	E	-
4XAW	H	L	-
4XBE	H	L	-
4XC1	H	L	-
4XC3	H	L	-
4XCF	H	L	Sim
4XGZ	C	D	-
4XH2	G	I	-
4XI5	D	C	-
4XMK	H	L	-
4XMM	H	L	-
4XMN	H	L	-
4XMP	H	L	Sim
4XNM	B	A	-
4XNQ	B	A	Sim
4XNU	H	L	-
4XNX	H	L	-
4XNY	H	L	-
4XNZ	H	L	-
4XP1	H	L	-
4XP4	H	L	Sim
4XP5	H	L	-
4XP6	H	L	-
4XP9	H	L	-
4XPA	H	L	-
4XPF	H	L	-

4XPH	H	L	-
4XRC	B	A	-
4XT1	C	-	-
4XTR	E	F	-
4XVJ	H	L	Sim
4XVS	H	L	-
4XVT	H	L	Sim
4XVU	C	D	-
4XWG	H	L	-
4XWO	U	V	-
4XX1	H	L	-
4XXD	B	A	-
4XZU	E	F	-
4Y5V	G	H	-
4Y5X	G	H	-
4Y5Y	D	E	-
4Y7M	A	-	-
4Y8D	D	-	-
4YBL	B	C	-
4YBQ	D	C	-
4YC2	H	L	-
4YDI	H	L	Sim
4YDJ	A	B	Sim
4YDK	H	L	Sim
4YDL	B	C	Sim
4YDV	H	L	Sim
4YE4	H	L	Sim
4YFL	H	L	Sim
4YGA	B	-	-
4YHP	H	L	-
4YHZ	H	L	-
4YJZ	L	L	-
4YK4	C	B	-
4YO0	C	D	Sim
4YPG	H	L	-
4YR6	A	B	-
4YWG	I	M	Sim
4YX2	H	L	-
4YXH	H	L	-
4YXK	H	L	-
4YXL	H	L	-
4YZF	E	F	-
4Z0X	B	A	Sim
4Z5R	W	V	-
4Z9K	B	-	-
4ZFF	A	B	-
4ZFG	H	L	-
4ZFO	A	B	Sim
4ZPT	H	L	-
4ZPV	H	L	-
4ZS6	C	D	Sim
4ZS7	H	L	Sim
4ZSO	D	C	-
4ZTO	H	L	Sim
4ZXB	C	D	-
4ZXB	A	B	-
4ZYP	N	O	-
4ZYP	D	E	-
5A1Z	K	L	-
5A2I	H	H	-
5A2J	H	H	-
5A2K	H	H	-
5A2L	H	H	-
5A7X	H	G	-
5A7X	N	M	-
5AAM	B	B	-
5AAW	K	K	-
5ACO	G	J	Sim
5ANM	D	C	-
5AUM	A	B	Sim
5B3J	H	L	-
5B71	D	C	Sim
5B8C	B	A	-
5BJZ	D	H	-
5BK0	B	A	Sim
5BK1	H	L	-
5BK2	C	D	-
5BO1	I	M	-
5BOP	A	-	-
5BOZ	I	-	-
5BV7	C	B	Sim
5BV7	H	L	Sim
5BVP	H	L	-
5C0N	C	D	-
5C0R	H	L	-
5C0S	H	L	-
5C1M	B	-	-
5C2U	B	-	-
5C3L	D	-	-
5C6T	H	L	-

5C7X	M	N	-
5C8J	E	F	-
5CBA	A	B	-
5CBE	C	D	Sim
5CD5	C	D	Sim
5CEZ	H	L	-
5CEZ	D	E	-
5CIL	H	L	-
5CIN	H	L	-
5CJO	H	L	-
5CJQ	H	L	-
5CSZ	H	L	Sim
5CUS	J	N	-
5CZV	H	L	-
5CZX	H	L	-
5D1Q	B	A	-
5D1Q	C	D	-
5D1X	B	A	-
5D1X	C	D	Sim
5D1Z	D	C	Sim
5D1Z	G	H	-
5D70	H	L	Sim
5D71	H	L	-
5D72	M	N	-
5D8J	H	L	-
5D93	C	B	-
5D96	J	I	Sim
5D9Q	D	D	Sim
5D9Q	F	E	-
5DA0	B	-	-
5DD0	H	L	Sim
5DFV	C	D	-
5DFW	H	H	-
5DFZ	E	-	-
5DHV	H	L	Sim
5DHX	B	B	-
5DHY	H	L	-
5DHZ	H	L	-
5DLM	H	L	-
5DMG	E	F	-
5DMI	H	L	-
5DMJ	E	-	-
5DO2	C	D	Sim
5DRZ	H	L	Sim
5DS8	H	L	-
5DSC	E	F	Sim
5DTF	H	L	Sim
5DUB	H	L	-
5DUM	H	L	-
5DUP	H	L	-
5DUR	H	L	-
5DWU	H	L	-
5E0Q	A	-	-
5E1A	A	B	-
5E1H	B	-	-
5E2V	H	L	-
5E2W	H	L	Sim
5E5M	D	-	-
5E7F	B	-	-
5E8D	H	L	-
5E8E	B	A	Sim
5,00E+94	B	A	Sim
5EA0	H	L	-
5EBL	A	B	-
5EBM	A	B	-
5EBW	A	B	-
5EC1	A	B	-
5EC2	A	B	-
5EII	A	B	-
5EN2	A	B	-
5EOC	J	M	Sim
5EOQ	H	L	-
5EOR	H	L	Sim
5EPM	A	B	-
5ERW	A	B	-
5ESV	H	L	Sim
5ESZ	A	B	-
5EU7	F	D	-
5EUL	V	-	-
5EZO	H	L	-
5F1K	D	-	-
5F1O	B	-	-
5F21	B	-	-
5F3B	E	F	Sim
5F3H	E	F	-
5F3J	D	D	-
5F6J	H	F	-
5F72	T	T	-
5F7K	C	-	-
5F7L	B	-	-

5F7M	C	-	-
5F7N	D	-	-
5F7W	D	-	-
5F7Y	D	-	-
5F8R	C	-	-
5F93	H	-	-
5F96	H	L	-
5F97	F	-	-
5F9A	C	-	-
5F9D	C	-	-
5F9O	H	L	Sim
5F9W	B	C	Sim
5FB8	B	A	-
5FCU	H	L	Sim
5FEC	H	L	-
5FGB	E	B	Sim
5FGC	E	B	-
5FHC	A	B	-
5FHC	H	L	-
5FHX	H	-	-
5FOJ	A	-	-
5FV1	-	L	-
5FV2	A	-	-
5FXC	H	H	-
5FYL	H	L	-
5FYL	D	E	-
5G5R	B	-	-
5G5X	B	-	-
5G64	I	M	-
5GGR	H	L	-
5GGS	C	D	Sim
5GGT	H	L	-
5GGV	H	L	Sim
5GHW	H	L	-
5GIR	A	B	-
5GIS	H	L	-
5GJS	H	L	-
5GJT	H	L	-
5GMQ	B	C	-
5GRJ	H	L	-
5GRU	L	L	-
5GRU	H	H	Sim
5GS0	D	C	Sim
5GS2	H	H	Sim
5GS2	D	D	-
5GUX	H	L	-
5GXB	B	-	-
5GZO	C	D	-
5H30	I	M	-
5H32	I	M	-
5H35	H	I	-
5H37	G	H	-
5H8O	A	-	-
5HBT	D	C	Sim
5HBV	D	C	-
5HD8	E	F	-
5HDQ	H	L	-
5HGG	T	-	-
5HHV	H	L	-
5HHX	H	L	-
5HI3	C	D	-
5HI4	C	D	Sim
5HI5	H	L	-
5HJ3	M	N	-
5HM1	A	-	-
5HVF	B	-	-
5HVG	B	-	-
5HVH	B	-	-
5HVH	C	-	-
5HYS	A	B	Sim
5I5K	X	Y	-
5I6X	B	C	-
5I6Z	B	C	-
5I71	B	C	-
5I73	B	C	-
5I74	B	C	-
5I75	B	C	-
5I8C	A	B	Sim
5I8H	G	H	-
5I8H	I	J	-
5I9Q	B	C	Sim
5IES	H	L	-
5IF0	A	B	-
5IFJ	J	K	-
5IG7	G	H	Sim
5IGX	H	L	-
5IHL	E	-	-
5IJK	A	C	-
5IKC	B	A	Sim
5IMK	B	-	-

5IML	B	-	-
5IMM	B	-	-
5IMO	B	-	-
5IP4	A	-	-
5IQ7	H	L	-
5IQ9	A	B	-
5IVN	A	-	-
5IWL	A	A	-
5J13	C	B	-
5J1S	C	-	-
5J1T	C	-	-
5J3H	C	D	-
5J56	B	-	-
5J57	B	-	-
5J9P	A	B	-
5JA8	D	-	-
5JA9	B	-	-
5JDS	B	-	-
5JHL	H	L	Sim
5JMO	C	-	-
5JQ6	H	L	-
5JQH	D	-	-
5JW3	H	L	-
5JXE	G	F	-
5JYL	D	D	-
5JYM	D	D	Sim
5JZ7	H	L	-
5K59	E	F	-
5K9K	H	L	Sim
5K9O	H	L	-
5KAQ	F	G	-
5KEL	Q	U	-
5KEL	J	N	Sim
5KEM	D	E	-
5KEM	B	C	-
5KEN	Q	P	-
5KEN	G	H	-
5KJR	H	L	-
5KN5	D	E	-
5KOV	C	C	-
5KQV	P	Q	-
5KTE	H	L	Sim
5KTZ	7	-	-
5KU0	7	-	-
5KU2	7	-	-
5KVD	H	L	Sim
5KVE	H	L	-
5KVF	H	L	Sim
5KVG	H	L	Sim
5KW9	H	L	-
5KWL	7	-	-
5KZC	E	G	-
5KZP	F	J	-
5L0Q	C	B	-
5L21	B	-	-
5L6Y	H	L	Sim
5LBS	H	L	-
5LCV	H	L	Sim
5LDN	H	L	-
5LHN	B	-	-
5LHP	B	-	-
5LHQ	B	-	-
5LHR	B	-	-
5LQB	H	L	-
5LSP	H	L	-
5LWF	D	-	-
5LWY	H	L	-
5LX9	H	H	-
5LXA	H	H	-
5LXG	H	L	-
5M14	C	-	-
5M15	C	-	-
5M2I	L	-	-
5M2J	D	-	-
5M2M	J	-	-
5M30	D	-	-
5M94	B	-	-
5M95	B	-	-
5MES	H	L	-
5MEV	H	L	-
5MHR	P	O	-
5MI0	B	C	-
5MJE	B	-	-
5MO3	H	L	Sim
5MO9	H	L	-
5MP1	J	K	-
5MP2	D	-	-
5MP3	A	B	-
5MP5	A	B	-
5MU0	E	F	Sim

5MU2	A	B	-
5MUB	I	J	-
5MV3	C	D	-
5MV4	C	D	-
5MVZ	H	L	-
5MWN	N	-	-
5MWN	E	-	-
5MY4	B	A	-
5MY6	B	-	-
5MYK	B	A	Sim
5MYO	B	A	-
5MYX	B	A	Sim
5MZV	D	-	-
5N09	I	M	-
5N0A	I	M	-
5N7B	H	H	-
5N7W	H	L	-
5N88	A	-	-
5NBD	C	-	-
5NBL	E	-	-
5NBM	E	-	-
5NGV	H	L	Sim
5NH3	I	M	-
5NHR	A	B	-
5NJ3	C	D	-
5NJ6	H	L	-
5NJG	E	F	-
5NMV	H	L	Sim
5NPH	H	L	Sim
5NPI	A	A	-
5NPJ	A	A	Sim
5NQW	C	-	-
5NUZ	H	L	Sim
5O02	C	-	-
5O03	C	-	-
5O04	D	-	-
5O04	F	-	-
5O05	C	-	-
5O0W	H	-	-
5O14	C	D	Sim
5O1R	H	L	-
5O2U	D	-	-
5O4G	B	A	-
5O4O	B	A	-
5O6V	H	L	-
5O7P	B	A	-
5OB5	H	L	Sim
5OCA	H	L	-
5OCC	H	L	-
5OCK	H	L	Sim
5OCX	H	L	-
5OCY	H	L	-
5OGI	B	B	-
5OJM	M	-	-
5OMM	C	-	-
5OMN	C	-	-
5OTJ	H	L	Sim
5OVW	I	-	-
5OWP	H	H	-
5SV3	C	-	-
5SX4	H	L	-
5SX5	J	K	-
5SY8	H	L	Sim
5T29	H	L	-
5T33	H	L	-
5T3S	D	E	-
5T3S	H	L	-
5T3X	H	L	-
5T3X	D	E	-
5T3Z	H	L	-
5T3Z	D	E	Sim
5T5B	A	B	-
5T5F	H	L	-
5T5N	M	L	-
5T6L	A	B	-
5T6P	D	C	-
5T78	B	A	-
5T80	H	L	-
5T85	H	L	-
5TBD	E	F	-
5TD8	E	-	-
5TE4	H	L	-
5TE6	H	L	-
5TE7	H	L	Sim
5TFW	H	L	-
5TH9	I	M	-
5THR	T	U	-
5THR	N	K	-
5TIH	H	L	-
5TJW	K	-	-

5TKJ	J	K	-
5TKK	H	L	Sim
5TL5	H	L	-
5TLJ	D	C	-
5TLJ	B	A	-
5TLK	F	E	-
5TLK	H	G	-
5TOJ	E	-	-
5TOK	F	-	-
5TPN	H	L	Sim
5TPW	H	L	-
5TQ0	H	L	Sim
5TQ2	H	L	-
5TQQ	H	L	-
5TR1	I	M	-
5TRU	H	L	-
5TUD	F	E	-
5TZ2	H	L	-
5TZT	H	L	-
5TZU	H	L	-
5U1F	H	L	-
5U3D	B	A	-
5U3J	H	L	Sim
5U3K	A	B	-
5U3L	H	L	-
5U3M	H	L	-
5U3N	H	L	Sim
5U3O	H	L	Sim
5U4L	B	-	-
5U4M	B	-	-
5U5F	B	A	-
5U5M	B	A	-
5U68	F	F	-
5U6A	B	A	Sim
5U7M	H	L	-
5U7M	D	E	-
5U7O	H	L	-
5U7O	D	E	-
5U8Q	H	L	-
5U8R	H	L	-
5UCB	H	L	Sim
5UDC	B	C	Sim
5UEA	H	L	-
5UEK	H	L	-
5UEM	H	L	Sim
5UG0	D	C	-
5UGY	I	M	-
5UJZ	I	I	Sim
5UK0	I	I	-
5UK1	I	I	-
5UK2	H	H	-
5UK4	j	-	-
5UKB	e	-	-
5UKR	H	L	Sim
5UM8	D	E	-
5UM8	H	L	-
5UMN	E	F	Sim
5UOE	M	N	-
5USF	D	-	-
5USH	D	E	Sim
5USL	H	L	-
5UTF	D	E	-
5UTF	H	L	-
5UTY	D	E	-
5UTY	H	L	-
5UTZ	F	G	-
5UWE	H	L	-
5UZ7	N	-	-
5V2A	H	L	-
5V6L	I	M	-
5V6M	H	L	Sim
5V7J	D	E	-
5V7J	H	L	-
5V8L	J	N	Sim
5V8L	I	M	-
5V8M	S	U	-
5VAG	C	B	-
5VAI	N	-	-
5VAK	B	-	-
5VAN	B	-	-
5VAQ	B	-	-
5VCN	F	E	-
5VCO	B	A	-
5VEB	H	L	-
5VGJ	H	L	Sim
5VIC	H	L	-
5VIG	H	L	-
5VJO	C	D	-
5VJQ	C	D	-
5VK6	A	B	-

5VKD	H	L	Sim
5VKE	A	B	-
5VKH	A	B	-
5VL3	H	L	-
5VL7	H	L	-
5VLP	H	L	-
5VN3	M	O	-
5VN8	F	J	-
5VNW	C	-	-
5VOB	H	L	Sim
5VOC	H	L	-
5VOD	H	L	-
5VPG	D	C	-
5VPH	D	C	-
5VPL	D	C	-
5VTA	K	J	-
5VXK	B	-	-
5VXL	B	-	-
5VXM	B	-	-
5VXR	H	L	Sim
5VYF	B	A	-
5VZY	H	L	-
5W06	H	L	-
5W08	K	L	-
5W0D	B	C	Sim
5W0K	H	L	-
5W1K	L	K	-
5W1K	D	C	-
5W1M	D	C	Sim
5W23	H	L	-
5W2B	H	L	-
5W3E	E	G	-
5W3L	E	G	-
5W3M	E	G	Sim
5W3O	D	E	-
5W3P	H	L	-
5W42	H	L	-
5W5X	H	L	-
5W5Z	H	L	Sim
5W6D	H	L	-
5W6D	D	E	-
5W6G	H	L	-
5W9H	H	I	-
5W9I	K	L	Sim
5W9J	B	C	-
5W9K	H	I	-
5W9L	H	I	-
5W9M	B	C	-
5W9N	E	F	-
5W9O	H	I	-
5W9P	D	E	-
5WB9	H	L	Sim
5WDF	A	B	-
5WDU	D	E	-
5WDU	T	S	-
5WDU	M	N	-
5WHK	H	L	-
5WI9	H	L	-
5WK3	W	V	Sim
5WN9	H	H	Sim
5WNA	H	L	Sim
5WNB	H	L	-
5WOB	U	V	-
5WT9	H	L	-
5WTH	E	D	-
5WTT	A	B	-
5WUX	C	D	-
5X08	H	L	-
5X0T	A	B	Sim
5X8L	G	L	-
5X8M	B	C	-
5XBM	E	D	-
5XCQ	A	B	-
5XCR	A	B	-
5XCS	A	B	Sim
5XCT	A	B	Sim
5XCU	D	E	-
5XCV	A	B	-
5XEZ	H	L	-
5XF1	H	L	-
5XHV	H	L	-
5XJ3	J	K	-
5XJ4	H	L	Sim
5XJM	H	L	-
5XKU	C	B	Sim
5XMH	H	L	-
5XS7	H	L	-
5XWD	H	D	-
5XXY	H	L	-
5Y0A	D	E	-

5Y11	A	B	Sim
5Y2L	I	J	-
5Y7Z	D	-	-
5Y80	B	-	-
5Y9J	H	L	Sim
5YAX	B	B	Sim
5YD3	G	G	Sim
5YD4	A	A	-
5YD5	A	A	-
5YE3	B	A	Sim
5YE4	A	B	-
5YHL	H	L	-
5YOY	P	M	-
5YWY	H	L	-
5YY4	A	A	-
5YY5	H	L	Sim
5YY5	C	D	-
5ZIA	G	L	-
5ZV3	H	L	-
5ZXV	C	D	-
6A0Z	H	L	Sim
6A3W	J	K	-
6A4K	I	M	Sim
6A67	C	D	Sim
6A77	H	L	Sim
6A78	I	M	-
6A79	H	L	-
6A96	L	-	-
6AD0	H	L	-
6AJ7	H	L	-
6AL0	H	L	-
6AL1	H	L	-
6AL5	H	L	-
6AOD	B	A	Sim
6APB	H	L	-
6APD	J	L	-
6APD	F	G	Sim
6APP	A	-	-
6AQ7	H	L	Sim
6ARU	C	B	-
6ATT	H	L	-
6AVQ	H	L	Sim
6AVR	H	L	-
6AVU	H	L	-
6AXK	A	B	Sim
6AXL	A	B	-
6AYZ	B	D	-
6AZ2	C	E	-
6AZM	D	C	Sim
6AZZ	F	E	-
6B08	C	B	-
6B0A	H	L	-
6B0E	B	A	-
6B0G	D	C	Sim
6B0H	D	C	-
6B0N	D	E	-
6B0N	H	L	-
6B0S	H	L	Sim
6B3J	N	-	-
6B3S	C	D	-
6B5L	H	L	-
6B5M	H	L	Sim
6B5N	H	L	-
6B5O	H	L	-
6B5P	H	L	-
6B5R	H	L	Sim
6B5S	H	L	-
6B5T	H	L	-
6B70	E	F	-
6B73	C	-	-
6B7Z	E	F	-
6B9J	A	B	-
6B9Y	B	A	-
6B9Z	B	A	-
6BAE	B	A	-
6BAH	B	A	-
6BB4	I	M	-
6BCK	H	L	Sim
6BDZ	H	L	-
6BF4	B	C	-
6BF7	C	D	-
6BF9	C	D	-
6BFQ	H	L	Sim
6BFS	H	L	Sim
6BFT	A	B	-
6BGT	B	A	-
6BIT	J	L	-
6BKB	H	L	Sim
6BKC	H	L	-
6BKD	H	L	-

6BLH	H	L	-
6BLI	D	E	-
6BP2	H	L	-
6BPA	E	F	-
6BPC	B	C	Sim
6BPE	E	F	-
6BQB	H	L	-
6BSP	A	B	-
6BY2	A	B	-
6BY3	A	B	-
6BZU	G	H	-
6BZV	C	D	-
6BZW	C	D	-
6BZY	H	L	Sim
6C5V	H	L	-
6C5W	E	-	-
6C6Y	A	B	-
6C6Z	H	L	-
6C9U	H	L	-
6C9W	B	-	-
6CBP	A	A	Sim
6CBV	H	L	-
6CCB	D	E	-
6CDE	H	L	-
6CDE	M	N	-
6CDE	Q	R	-
6CDI	H	L	-
6CDI	m	n	-
6CDI	Q	R	-
6CDM	A	B	-
6CDO	A	B	Sim
6CDP	A	B	-
6CE0	H	L	-
6CE0	D	E	-
6CEZ	H	L	Sim
6CF2	A	B	-
6CH7	D	E	-
6CH7	Q	R	Sim
6CH8	D	E	-
6CH8	Q	R	-
6CH9	Q	R	-
6CH9	D	E	-
6CHB	J	K	-
6CHB	M	N	-
6CK9	H	L	-
6CK9	D	E	-
6CM3	O	N	-
6CM3	R	S	Sim
6CMG	C	B	-
6CMI	D	C	-
6CMO	H	L	-
6CNV	C	-	-
6CNV	H	L	-
6CO3	H	L	-
6CRK	H	H	-
6CRQ	J	L	-
6CSE	H	L	Sim
6CSF	H	L	-
6CT7	H	L	Sim
6CUE	H	L	-
6CUE	M	N	-
6CUE	7	8	-
6CUF	h	I	-
6CUF	m	n	-
6CUF	8	7	-
6CVK	A	A	Sim
6CW2	A	B	-
6CW3	C	D	-
6CWD	B	B	-
6CWG	D	-	-
6CWT	C	D	-
6CXG	H	L	Sim
6CXY	H	L	Sim
6CYF	F	E	-
6D01	G	H	-
6D0U	H	I	-
6D0X	A	B	Sim
6D11	C	D	-
6D2P	H	L	-
6D6T	J	I	Sim
6D6U	J	I	-
6D9W	H	L	-
6DB5	H	L	-
6DB6	H	L	Sim
6DB7	I	M	-
6DBF	B	-	-
6DBG	C	-	-
6DC8	H	L	-
6DC9	H	L	-
6DCA	J	N	-

6DCW	H	L	Sim
6DDE	E	E	-
6DDM	B	A	Sim
6DDR	B	A	-
6DDV	B	A	-
6DE7	D	E	-
6DE7	H	L	-
6DF1	H	L	Sim
6DF2	H	L	-
6DFI	H	L	-
6DFJ	H	L	Sim
6DID	K	E	-
6DJP	C	D	-
6DJP	E	F	-
6DO1	E	-	-
6DZL	H	K	Sim
6DZM	J	L	-
6E1K	D	C	-
6E3H	H	L	Sim
6E3Y	N	-	-
6E4Y	H	L	-
6E4Z	H	L	-
6E5P	3	4	-
6E5P	O	P	-
6,00E+62	H	L	-
6,00E+63	H	L	-
6EA5	M	N	-
6EA7	H	L	-
6EAY	H	L	Sim
6EDU	L	M	-
6EDU	P	Q	-
6EHG	C	-	-
6EJG	D	D	Sim
6EJM	I	I	Sim
6EK2	H	H	-
6ELU	K	L	-
6EQC	F	F	-
6EQI	B	-	-
6ERX	F	G	-
6ETI	F	E	Sim
6EWB	J	K	-
6EY0	F	-	-
6EY6	K	-	-
6EYO	H	L	-
6F5G	B	-	-
6F7T	B	A	-
6FAX	H	L	-
6FE4	F	-	-
6FEQ	D	C	-
6FGB	H	L	-
6FLA	H	L	-
6FLB	H	L	Sim
6FLC	B	A	Sim
6FN1	C	B	Sim
6FN4	C	B	-
6FRJ	H	H	Sim
6FUZ	N	-	-
6FV0	F	-	-
6FXN	D	E	-
6FY0	I	M	-
6FY1	H	L	-
6FY2	X	Y	Sim
6FY3	H	L	-
6FYT	I	-	-
6FYW	C	-	-
6FZQ	H	H	-
6FZR	H	H	-
6GCI	B	-	-
6GDG	E	-	-
6GG0	H	L	-
6GK7	H	L	-
6GK8	H	L	-
6GS1	H	-	-
6GS4	H	-	-
6GS7	H	-	-
6GV1	H	L	-
6GV4	H	L	-
6H02	B	-	-
6H06	E	F	-
6H0E	C	D	Sim
6H15	D	-	-
6H16	B	-	-
6H1F	A	-	-
6H3T	I	M	-
6H3U	I	M	-
6H5N	C	B	-
6H6Y	H	-	-
6H70	C	-	-
6H71	D	-	-
6H72	D	-	-

6H7J	C	-	-
6H7K	D	-	-
6H7L	D	-	-
6H7M	D	-	-
6H7N	D	-	-
6H7O	D	-	-
6HCO	D	C	-
6HF1	C	B	Sim
6HJP	I	J	-
6HJQ	K	L	-
6HUG	G	-	-
6HUJ	G	-	-
6HUK	G	-	-
6HUO	G	-	-
6HUP	G	-	-
6I04	H	L	-
6I07	B	B	-
6I53	G	-	-
6I6J	C	-	-
6I8S	G	K	-
6I9I	H	L	-
6IBL	D	-	-
6ICC	H	L	-
6ICF	H	L	-
6IDI	M	N	Sim
6IDK	H	L	-
6IDL	I	J	-
6IEA	H	L	Sim
6IEB	H	L	Sim
6IEC	J	K	-
6IEK	B	C	-
6II4	H	L	-
6II8	J	K	-
6II9	P	Q	-
6IUT	H	L	Sim
6IUV	C	D	Sim
6IVZ	H	L	Sim
6IW0	H	L	-
6IW2	K	L	-
6J5D	H	L	Sim
6J5F	H	L	-
6J5G	H	L	-
6J71	B	B	-
6J7W	B	-	-
6JFH	K	I	-
6JFI	K	I	Sim
06/mar	M	N	Sim
6MB3	H	L	-
6MCO	D	E	-
6MCO	H	L	-
6MDT	D	E	-
6MDT	H	L	-
6ME1	A	B	-
6MEH	H	L	-
6MEI	H	L	-
6MEJ	H	L	-
6MEJ	A	B	-
6MEK	F	E	-
6MEK	B	D	-
6MF7	H	L	-
6MFT	A	B	Sim
6MHG	G	S	-
6MHR	D	E	-
6MI2	A	B	-
6MID	H	L	-
6MJZ	H	L	-
6ML8	H	L	-
6MLK	H	L	-
6MLM	E	I	-
6MN7	G	G	Sim
6MTJ	H	L	Sim
6MTJ	D	E	Sim
6MTN	H	L	-
6MTN	D	E	-
6MTO	H	L	-
6MTP	H	L	Sim
6MTQ	H	L	-
6MTT	H	L	Sim
6MU6	H	L	-
6MU6	D	E	-
6MU7	H	L	-
6MU7	D	E	-
6MU8	H	L	-
6MU8	D	E	-
6MUF	D	E	-
6MUF	H	L	-
6MUG	H	L	-
6MUG	D	E	-
6MUI	K	L	-
6MV5	H	L	-

6MW9	O	P	Sim
6MWC	C	D	-
6MWV	K	L	-
6MWX	O	P	-
6MXT	N	-	-
6MYY	G	F	-
6MZJ	H	L	-
6N4B	S	S	-
6N4Y	F	-	-
6N50	E	-	-
6N51	D	-	-
6N7J	H	L	Sim
6NB3	D	E	Sim
6NB4	H	L	-
6NB7	H	L	Sim
6NBF	N	-	-
6NBH	N	-	-
6NBI	N	-	-
6NFJ	B	-	-
6NIY	N	-	-
6NM6	D	E	-
6NM6	H	L	-
6NM6	U	V	-
6NN3	H	L	-
6NNF	H	L	-
6NNF	D	E	-
6NNF	U	V	-
6NNJ	U	V	-
6NNJ	H	L	-
6NNJ	D	E	-
6O39	B	A	-
6O3A	B	A	-
6O3B	G	E	-
6OE4	B	C	Sim
6OE5	B	C	-
6OE5	H	L	-
6QEE	C	B	-
6QEX	C	B	-
6R2S	B	A	-
6R8X	C	B	-

Apêndice 3 – Interações realizadas pelas 385 cadeias pesadas de anticorpo analisadas neste trabalho

Posição	Número de cadeias com interação	Van der Waals	Van der Waals *clash*	Interação de hidrogênio	Mediada por 1 molécula de água	Mediada por 2 moléculas de água	Aromática*	Pi-cátion*	Hidrofóbica	Eletrostática	Total de interações
1	16	11	5	7	7	7	0	0	5	4	62
2	11	10	5	5	0	13	0	0	8	0	52
3	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	1	0	2
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	2	2	0	1	0	1	1	1	0	0	8
9	1	1	1	0	0	0	0	0	1	0	4
10	0	0	0	0	0	0	0	0	0	0	0
11	1	1	0	1	0	0	0	0	0	0	3
12	1	1	1	0	0	0	1	0	1	0	5
13	2	1	0	1	0	1	0	0	0	1	6
14	2	2	2	1	1	0	1	1	1	0	11
15	3	2	2	0	1	3	0	0	2	0	13
16	2	2	2	0	3	0	0	0	1	0	10
17	1	1	1	0	0	0	0	0	0	0	3
18	1	0	0	0	0	3	0	0	0	0	4
19	0	0	0	0	0	0	0	0	0	0	0
20	2	1	1	0	1	0	0	0	0	2	7
21	0	0	0	0	0	0	0	0	0	0	0
22	1	1	0	0	0	0	0	0	0	0	2
23	1	1	0	0	0	0	0	0	0	0	2
24	1	1	1	0	1	0	0	0	0	0	4
25	4	1	1	1	1	13	0	0	1	0	22
26	10	7	1	3	0	18	0	0	1	0	40
27	23	16	7	69	18	39	1	0	3	1	177

Posição	Número de cadeias com interação	Van der Waals	Van der Waals *clash*	Interação de hidrogênio	Mediada por 1 molécula de água	Mediada por 2 moléculas de água	Aromática*	Pi-cátion*	Hidrofóbica	Eletrostática	Total de interações
28	48	32	12	68	7	165	4	8	20	2	366
29	65	44	29	143	162	206	0	2	25	3	679
30	41	28	16	73	37	83	2	6	14	1	301
31	6	5	2	3	1	13	0	0	2	1	33
32	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0
34	5	3	2	1	3	1	0	0	1	1	17
35	160	130	69	240	388	369	6	10	37	4	1413
36	231	197	117	487	500	639	5	5	81	30	2292
37	171	148	89	180	348	308	23	46	95	13	1421
38	238	214	139	419	544	479	41	71	155	7	2307
39	15	12	7	1	8	14	1	2	3	3	66
40	82	42	22	53	55	88	6	13	11	19	391
41	3	1	0	1	0	6	0	0	0	0	11
42	5	3	0	0	1	0	1	0	5	0	15
43	1	1	1	0	1	0	0	0	0	0	4
44	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0
47	1	1	0	0	0	1	0	0	1	0	4
48	2	2	0	0	0	0	1	1	1	0	7
49	2	1	0	0	0	7	0	0	0	0	10
50	6	2	1	3	1	3	2	0	3	0	21
51	12	6	2	0	3	3	5	3	7	1	42
52	78	43	14	3	11	31	30	13	47	0	270
53	10	8	5	1	5	1	3	1	8	0	42
54	31	28	19	27	49	25	1	3	16	8	207
55	205	168	99	98	435	132	20	28	106	26	1317
56	54	51	20	11	41	65	5	7	19	3	276
57	275	249	163	374	684	422	26	48	134	35	2410

Posição	Número de cadeias com interação	Van der Waals	Van der Waals *clash*	Interação de hidrogênio	Mediada por 1 molécula de água*	Mediada por 2 moléculas de água*	Aromática*	Pi-cátion*	Hidrofóbica	Eletrostática	Total de interações
58	147	120	77	215	344	254	13	29	62	13	1274
59	191	157	101	311	474	463	6	14	65	34	1816
60	13	11	9	7	11	42	0	0	3	2	98
61	20	15	11	55	24	55	1	2	8	4	195
62	151	128	84	213	415	293	15	13	50	34	1396
63	97	80	43	117	99	146	3	4	22	12	623
64	219	193	130	372	573	458	13	17	111	39	2125
65	107	78	42	274	173	468	6	7	19	4	1178
66	199	168	109	190	231	230	33	34	110	29	1333
67	47	20	8	112	29	175	2	3	6	0	402
68	20	17	4	1	0	15	0	1	3	1	62
69	35	26	17	47	127	45	0	1	14	10	322
70	10	6	1	5	2	14	0	0	1	1	40
71	2	2	1	0	1	0	0	0	1	1	8
72	36	29	20	43	36	66	0	5	7	8	250
73	1	1	1	1	1	15	0	0	0	1	21
74	7	3	2	5	1	11	0	0	1	1	31
75	3	3	2	1	0	0	0	0	2	1	12
76	9	2	1	3	1	22	0	0	0	1	39
77	4	2	2	1	2	6	0	0	2	0	19
78	9	1	0	6	0	31	0	0	0	0	47
79	5	3	1	1	0	3	0	1	1	1	16
80	25	19	16	0	69	30	0	0	2	19	180
81	4	3	2	0	2	0	0	0	0	2	13
82	27	15	5	47	21	66	0	2	12	9	204
83	19	17	6	3	2	39	0	0	5	0	91
84	8	6	1	1	1	7	0	0	2	1	27
85	13	5	1	19	1	25	0	0	1	0	65
86	3	2	2	0	0	0	0	0	2	0	9
87	4	3	3	2	3	7	1	1	1	0	25

Posição	Número de cadeias com interação	Van der Waals	Van der Waals *clash*	Interação de hidrogênio	Mediada por 1 molécula de água	Mediada por 2 moléculas de água	Aromática*	Pi-cátion*	Hidrofóbica	Eletrostática	Total de interações
88	4	3	2	2	7	0	1	1	3	0	23
89	5	5	3	1	11	0	0	1	4	1	31
90	5	5	4	0	1	3	1	0	5	1	25
91	3	3	3	2	1	3	1	1	3	0	20
92	6	6	4	27	9	3	1	1	5	1	63
93	5	4	2	39	3	3	0	0	1	1	58
94	1	1	0	0	0	0	0	0	0	0	2
95	2	1	0	0	0	0	0	0	1	1	5
96	1	1	0	0	0	0	0	0	1	0	3
97	1	0	0	0	0	3	0	0	0	0	4
98	0	0	0	0	0	0	0	0	0	0	0
99	1	1	1	0	1	0	0	0	0	1	5
100	1	1	1	0	1	0	0	0	1	0	5
101	2	1	0	0	0	0	0	0	1	1	5
102	3	2	1	1	1	1	0	0	1	1	11
103	4	4	2	1	0	3	0	0	1	0	15
104	13	8	5	37	23	41	0	3	3	2	135
105	26	18	12	25	63	18	2	5	11	5	185
106	82	56	31	46	135	216	1	8	24	24	623
107	208	154	96	388	450	571	14	17	73	44	2015
108	190	153	87	364	385	491	10	17	76	22	1795
109	248	225	163	371	571	574	32	40	137	23	2384
110	189	172	122	351	449	435	17	27	103	23	1888
111	142	128	86	265	276	334	18	17	76	11	1353
111.1	88	78	53	197	158	193	12	14	42	9	844
111.2	52	47	26	142	79	93	7	5	25	4	480
111.3	27	23	18	80	81	91	1	5	16	6	348
111.4	16	14	12	39	18	9	0	3	3	1	115
111.5	11	10	4	7	8	13	2	3	2	1	61
111.6	3	3	3	0	2	0	0	0	1	0	12

Posição	Número de cadeias com interação	Van der Waals	Van der Waals *clash*	Interação de hidrogênio	Mediada por 1 molécula de água	Mediada por 2 moléculas de água	Aromática*	Pi-cátion*	Hidrofóbica	Eletrostática	Total de interações
111.7	3	3	2	3	1	7	1	1	3	1	25
111.8	2	2	2	0	1	0	0	1	2	0	10
112.9	0	0	0	0	0	0	0	0	0	0	0
112.8	1	1	0	0	0	0	0	0	0	1	3
112.7	4	4	3	13	5	2	1	1	2	1	36
112.6	10	10	7	3	7	13	1	2	7	0	60
112.5	12	11	11	35	63	57	1	2	6	2	200
112.4	29	27	16	30	74	66	3	3	17	7	272
112.3	44	41	27	82	128	53	3	6	17	8	409
112.2	77	63	37	113	136	142	8	14	42	5	637
112.1	105	89	57	123	217	283	17	16	54	6	967
112	126	107	60	218	277	305	12	20	58	22	1205
113	180	150	101	241	410	311	27	34	109	17	1580
114	84	54	29	62	145	266	10	9	32	8	699
115	33	22	14	11	47	41	7	4	18	2	199
116	51	33	17	91	45	107	2	1	22	18	387
117	21	12	10	5	8	21	3	10	7	0	97
118	4	1	1	0	0	0	1	1	2	0	10
119	0	0	0	0	0	0	0	0	0	0	0
120	0	0	0	0	0	0	0	0	0	0	0
121	0	0	0	0	0	0	0	0	0	0	0
122	0	0	0	0	0	0	0	0	0	0	0
123	0	0	0	0	0	0	0	0	0	0	0
124	0	0	0	0	0	0	0	0	0	0	0
125	0	0	0	0	0	0	0	0	0	0	0
126	0	0	0	0	0	0	0	0	0	0	0
127	0	0	0	0	0	0	0	0	0	0	0
128	0	0	0	0	0	0	0	0	0	0	0
Total	5393	4373	2695	7742	10257	10917	496	696	2346	674	45589

* O número de interações aromáticas e Pi-cátion podem ser menores do que as exibidas porque foram contabilizadas para cada carbono pertencente aos anéis aromáticos. O número de interações mediados por água pode estar superestimado devido a moléculas de água aprisionadas na interface.

Anexo

CARVALHO, M. B.; MOLINA, F.; FELICORI, L. F. Yvis: antibody high-density alignment visualization and analysis platform with an integrated database. **Nucleic Acids Research**, v. 47, n. May, p. 490–495, 2019.

W490–W495 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* *Published online 20 May 2019*
doi: 10.1093/nar/gkz387

Yvis: antibody high-density alignment visualization and analysis platform with an integrated database

Milene B. Carvalho [1,2,*], Franck Molina[3] and Liza F. Felicori[1,*]

[1]Laboratory of Synthetic Biology and Biomimetics, Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil, [2]Departamento de Ciência da Computação, Universidade Federal de São João del Rei, São João del Rei, Minas Gerais, 36301-360, Brazil and [3]Sys2diag, UMR9005 CNRS Alcediag, Montpellier 34184, France

Received March 01, 2019; Revised April 20, 2019; Editorial Decision April 30, 2019; Accepted May 02, 2019

ABSTRACT

As antibodies are a very important tool for diagnosis, therapy, and experimental biology, a large number of antibody structures and sequences have become available in recent years. Therefore, tools that allow the analysis, comparison, and visualization of this large amount of antibody data are crucially needed. We developed the antibody high-density alignment visualization and analysis (Yvis) platform to provide an innovative, robust and high-density data visualization of antibody sequence alignments, called *Collier de Diamants*. The Yvis platform also provides an integrated structural database, which is updated weekly, and many different search and filter options. This platform can help to formulate hypotheses concerning the key residues in antibody structures or interactions to improve the understanding of antibody properties. The Yvis platform is available at http://bioinfo.icb.ufmg.br/yvis/.

INTRODUCTION

Antibodies or immunoglobulins are vertebrate immune system proteins that are produced by B cells and can bind to antigens with high specificity and affinity. For this reason, antibodies are an important tool in diagnosis, therapy and experimental biology (1). To elucidate the antibody characteristics, large numbers of antibody structures and sequences have been generated in the last years. The number of antibodies or antibody fragment structures deposited in Protein Data Bank (PDB) (2) has increased exponentially (3), leading to the development of databases of antibody structures (4–7). Moreover, many antibody sequences have been obtained by high-throughput sequencing of the B-cell receptor repertoire (8,9). This extraordinary and still increasing number of antibody structures and sequences demands integrative data organization and tools for their analysis, comparison and visualization. One of the major bottlenecks in this field is the concomitant visualization of a large amount of antibody data. AbYsis (4) and IMGT/3Dstructure-DB (5) allow antibody visualization, but only a limited number of sequences can be analysed at a time. Indeed, abYsis presents a classical multiple sequence alignment (MSA) that displays a limited number of sequences and positions each time. IMGT/3Dstructure-DB display only one antibody sequence using the IMGT/*Collier des Perles* representation (10) that allows sequence analysis related to the antibody structure. To fill this gap, we developed the antibodY high-density alignment visualization and analysis (Yvis) platform that includes: (i) an updated weekly and curated antibody structure database (Yvis database) and (ii) integrated antibody analysis resources, such as an antibody high-density alignment visualization called *Collier de Diamants,* and multiple filter options to analyse data from user files or from the Yvis database.

MATERIALS AND METHODS

Yvis database: an updated weekly and curated repository of data on antibody PDB structures

The Yvis database is an updated weekly collection of data on antibody PDB structures (in complex with an antigen or not), such as PDB and chain identification, antibody and protein antigen-producing organisms or type (hapten, carbohydrate or nucleic acid), gapped sequences of antibody chains, germline information (assigned V and J genes with their identity values), and antigen-antibody putative contacts.

The Yvis script, developed in Python, extracts a list of antibody PDB structures from SAbDab (7) that is updated weekly. The following data are extracted from this list, processed, and stored in the Yvis database: (i) PDB and chain identifications, (ii) names of the organisms producing antibody and antigen (when applicable) and (iii) antigen molecule description. When the SAbDab list does

*To whom correspondence should be addressed. Tel: +55 32 3379 4935; Email: milenebc@ufsj.edu.br
Correspondence may also be addressed to Liza F. Felicori. Tel: +55 3409 2981; Fax: +55 31 3409 2614; Email: liza@icb.ufmg.br

not contain the antibody- or antigen-producing organism name, Yvis script extracts this information from the corresponding PDB structure file, acquiring the ORGANISM_SCIENTIFIC value from SOURCE record after retrieving the molecule Id from COMPND record. After data extraction, Yvis script checks whether the organism names match the UniProt Taxonomy (11), and correct them if required (Supplementary Table S1). Data are manually curated, if the standard name is not found automatically. These standard names facilitate the Yvis database search, reducing the diversity of organism names, for instance by eliminating all synonyms.

The Yvis script submits antibody chain sequences to IMGT/DomainGapAlign (5) to obtain gapped sequences and germline information. Then, it processes the result page and extracts the gapped sequence of the variable domain of each chain, following the IMGT numbering (12). Moreover, the script extracts and stores the V and J germline genes assigned to the chain sequence, and their identity values.

Finally, to obtain information on the putative antibody–antigen contacts, the Yvis script downloads the PDB structure files and extracts the antibody chain amino acids that potentially interact with a peptide or protein antigen using the Biopython PDB module (13). Then, the distance between each α-carbon of the antibody and antigen amino acids is calculated. If the distance between two α-carbons is not higher than 8 Å, the position that contains the amino acid is marked as making a putative contact. This distance is used because it allows including putative direct interactions between antigen and antibody and also water-mediated interactions (14).

Yvis resources: integrated tools for high-density antibody data visualization and analysis

The Yvis platform integrates resources that allow the analysis of antibody variable domains that have been uploaded as user sequences or selected from the Yvis database. This platform is a web-based application that process sequences in a server or in a user's internet browser, depending on the analysed data. The server-side application was developed using PHP and Mysql, and the client-side using the JavaScript and D3.js framework.

The Yvis Platform offers input and search versatility. With the Yvis platform, users can analyse antibody structures stored in the Yvis database or uploaded by them. Different search options (Figure 1A) are available to select, from the database, a set of antibody structures to be analysed. It is possible to show all antibody chains stored in the Yvis database, or to specify a list of PDB identifiers, or a pair of PDB:chain identifiers. Moreover, users can choose to show free or complexed antibodies, and in the latter case, they can indicate the antigen type (hapten, carbohydrate, nucleic acid or protein). For protein antigens, they can indicate the producing organism. Users can also select antibodies with assigned germline V or J genes, or produced by user-selected organisms. In addition, users can search antibodies by using keywords contained in the literature related to PDB structures. After defining the PDB structure search criteria, the user can apply additional filters to avoid sequence redundancy, such as: (i) to choose only one representative chain of each type (heavy or light) in each PDB structure; (ii) to specify an identity threshold that ensures that none of the filtered sequences has an identity value higher than the user-specified value. This approach was based on Cd-hit (15). Because of the time requested to analyse and group all sequences, the identity filter is not used by default. All these filters can be combined.

Users can also analyse antibody sequences obtained from an IMGT/DomainGapAlign (5) results file, an IMGT/HighV-QUEST (16) gapped amino acid results file, or a FASTA file containing gapped, or ungapped chain sequences or even CDR sequences (Figure 1A, User Input file). When a user submits an IMGT results file, the Yvis platform will process it in the user's browser. Moreover, when a user submits ungapped sequences in a FASTA file, the Yvis platform will number them using ANARCI (17) in the Yvis server.

Independently of the chosen input data (Yvis database or user's data), the Yvis platform will generate an initial set of antibody sequences that can be visualized with the *Collier de Diamants* representation.

Collier de Diamants: a new visualization of high-density multiple sequence alignment of antibody variable domains. After the user's input choice, Yvis presents a first visualization of these data (middle panel in Figure 1B) as a multiple sequence alignment of antibody chains, based on the IMGT/*Collier de Perles* (Pearl Necklace) (10). In Yvis visualization, each sequence is numbered according to the IMGT unique numbering (12), and each position corresponds to a column in a traditional MSA. For each position, a pie chart indicates its amino acid composition. Each pie slice (sector) represents the number of sequences with an amino acid of a specific class in that position. The amino acid class is identified by a specific color, as defined in WebLogo (18). The positions are shown as in the *Collier de Perles*, linking sequences to their 3D structure. A square highlights the CDR anchors, one position before the CDR start and one after the CDR end (i.e. green for CDR1, orange for CDR2, and blue for CDR3), and allows the quick visualization of the residues that compose each CDR. As in Yvis each pearl of the necklace was replaced by a new representation with multiple 'facets', this new visualization was called *Collier de Diamants* (Diamond Necklace).

Like the *Collier de Perles*, the *Collier de Diamants* can be displayed on one (middle panel of Figure 1B) or on two layers (Figure 1B (I)). The two-layer representation shows the variable domain strands in a position closer to their 3D structure, while the one-layer version is closer to the variable domain sequence. As the *Collier de Diamants* uses a pie chart to describe each position of the alignment, it is possible to show the data of countless sequences in the same visualization. Moreover, positions with a conserved amino acid class are easily detected because they are represented by a pie chart with a dominant sector, while a pie chart with many sectors represents a position that is more variable.

Beside the visualization of an MSA, the *Collier de Diamants* displays a quantitative attribute for each position that is represented by a circle around the pie chart (Figure 1B (III)). In the Yvis platform, this attribute represents the

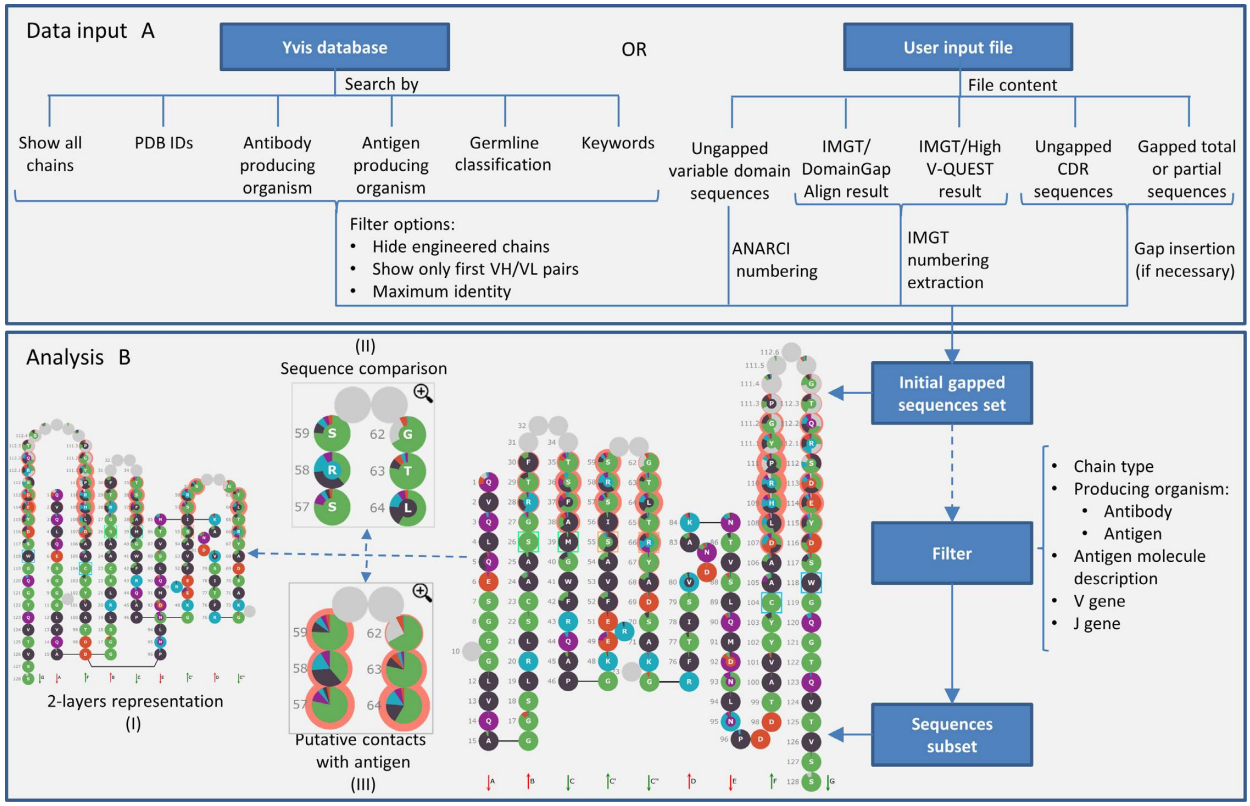


Figure 1. Yvis platform overview. (**A**) The data input box presents two possibilities to input sequences to be analysed in Yvis (user input files and selection from the Yvis database). It presents filter options for sequences from the Yvis database (redundancy and engineered chains) and actions taken by the platform to process the user input files. (**B**) The analysis box presents the options to visualize a multiple sequence alignment of antibody variable domains and the filter possibilities. The user can generate a new subset of sequences to be analysed, by selecting specific filters. The analysis can be displayed by the *Collier de Diamants* on one or two layers (I). Additionally, the user can compare the multiple sequence alignment with a reference sequence (II), and visualize data on putative contacts with the antigen (III).

number of chains that have a putative contact on each position. The radius length of the circle around a pie chart is proportional to the number of sequences with a putative contact at that position. Positions that have many gaps can have a small contact circle, even if all sequences in these positions have an amino acid that makes contact with the antigen. In this manner, positions that are involved in antibody–antigen interactions in the majority of the analysed sequences can be easily detected. However, it is only possible to show this attribute for chains from the Yvis database obtained from an antibody in complex with a protein or peptide antigen in the corresponding PDB structure.

The Yvis platform also allows comparing the MSA with a user-defined sequence (Figure 1B (II)): a target sequence, a germline gene sequence, or a consensus sequence. When the user inserts a sequence to be compared with the MSA, the Yvis displays, in the centre of each pie chart that represents a position, a small inner circle coloured according to the amino acid class of the input sequence that corresponds to that position (same colour code used for the pie chart). This representation facilitates the comparison of the user's sequence with the MSA by just checking whether the colour of the small circle is the same as the one of the larger pie chart sector for that position. In addition to the pie chart representation, Yvis makes available a detailed bar chart for each position, showing the number of each amino acid in the corresponding position.

Additional filter options can be used to select antibody subsets. After having selected the initial set of antibodies from the Yvis database or after uploading antibody sequences, the user can use filters to select sequence subsets and to restrict the analysis to specific chain types, antibody-producing organisms, antigen type or producing organisms, antigen molecule description, and germline information (right side of Figure 1B). As the molecule description is a text field of PDB files extracted from SAbDab, many different descriptions of the same molecule are often available. Therefore, to use this filter, the user must select all descriptions that match the desired molecule. The filter options are restricted to the uploaded or selected sequences. In addition, these filters can be combined to generate antibody subsets. The filters can be used, for example, to analyse only heavy chains or only human antibodies against a specific antigen, among others.

Below the *Collier de Diamants* representation of an initial set of antibody sequences, there is a table containing the available information of all chains presented in the MSA. The colours in each gapped sequence highlight the CDR amino acids (same colour code as for the CDR anchors in the *Collier de Diamants*) and the positions that make the pu-

tative contacts. This table can be exported to be used with other tools.

RESULTS AND DISCUSSION

A continuously updated source of antibody structure information

The Yvis database is updated weekly and currently includes data on 3423 antibody structures (from 22 different antibody-producing organisms) in complex with protein antigens (from 155 different antigen-producing organisms) or non-protein antigens (184 structures with haptens, 22 with nucleic acids, and 106 with carbohydrates), as well as 1,136 structures of antibodies in free form. Most of the antibodies in the database are of human (1444) and mouse (1362) origin.

The organism-producing names are in the standard format defined by UniProt Taxonomy (11), allowing the user to search or filter the database content easily. To achieve that, we automatically modified the antigen-producing organism name of 389 structures, and manually changed the antibody-producing organism name of 58 structures and the antigen-producing organism name of 77 structures. In this database, users can select a set of antibodies, searching mainly by antibody-producing organisms, antigen types, antigen producing-organisms, or assigned germline genes. Moreover, users can restrict the set to be analysed by excluding engineered chains, multiple antibodies in the same structure, and sequences with identity above the user-defined cut-off. These combinations of different search criteria and constraints are not available in other antibody structure databases, such as abYsis (4), IMGT/3Dstructure-DB (5) and SAbDab (7).

Case studies demonstrate the Yvis platform versatility and easy high-density data visualization to help hypothesis formulation

To test the Yvis platform, we carried out three case studies using anti-HIV antibodies. In the first study, we selected anti-HIV gp120 antibodies available in the PDB, through a search of the Yvis database after choosing the HIV antigen-producing organism and the option that forces the result to contain only one chain of each type from each PDB file. After this search, to restrict the analysis to only anti-gp120 antibodies, we used the antigen molecule description filter and limited the analysis to the heavy chains. The alignment presented by *Collier de Diamants* of the 124 identified sequences (Supplementary Figure S1) allowed the easy visualization of known anti-HIV neutralizing antibody characteristics, such as a long CDRH3 (19), and of some conserved positions (e.g. 8, 22, 119 and 121). Moreover, it highlighted that many of the analysed chains could make contact with the antigen via CDR2 and framework region 3 (positions from 57 to 69). This is a known characteristic of some anti-HIV CD4-binding-site antibodies (20). Nevertheless, this is not a very common characteristic of antibody heavy chains.

In the second case study, we downloaded a FASTA file of transcript sequences of an HIV-infected donor (19) from the Sequence Read Archive (21) using the SRR1767440 access number. Then, we submitted the FASTA sequence to IMGT/HighV-QUEST (16), and then uploaded the results file containing the amino acid information for 478 047 heavy chain sequences to the Yvis platform. The platform excluded sequences with ambiguous amino acids, leaving 330 800 sequences (Supplementary Figure S2). We applied the germline filter to restrict the visualization to sequences assigned to the VH1-2*02 allele, because, at least three neutralizing antibodies isolated from this donor were derived from this allele. This resulted in the alignment of 97 751 sequences that were compared with the VH1-2*02 allele by the Yvis comparison sequence feature (Figure 2A). The parts of the sequence corresponding to the D and J genes were filled with gaps. As in the first case study, it was easy to visualize the CDRH3 length (most of the analysed sequences had two insertions in CDR3, and less than 10% of sequences had more than four insertions). For some positions (e.g. 36, 66, 92, 93 and 95), the amino acid class of the uploaded sequences was different from the germline amino acid class, as easily noticeable on the basis of the colour difference between the inner circle and the pie chart sectors. The detailed bar chart for position 36 (Box in Figure 2A) indicated that the glycine (G) residue of the germline gene was mutated into aspartic acid (D) in approximately half of the sequences. As aspartic acid is a charged residue and this position is located in a CDR, probably this modification is selected during antibody expansion to increase antigen binding.

In the third case study, we selected 21 heavy chain sequences from HIV neutralizing antibodies derived from the VH1-2*02 allele (19), and submitted them to IMGT/DomainGapAlign. Then, we uploaded the results file into Yvis and compared the alignment against the VH1-2*02 allele sequence, with gaps in the part corresponding to the D and J genes (Figure 2B). Some positions, mainly in CDR2 (57, 59 and 69) and framework region 3 (82 and 83), had only one sector in the pie chart and its colour was different from the color of the inner circle. This indicates that in all neutralizing antibody sequences, this amino acid is different from the one in the germline gene (new amino acid class). One could hypothesize that positions in which the amino acid class changed might carry important characteristics for the neutralizing activity (for instance, structural or binding features that allow antibody binding to HIV). Comparison of the positions with a change of amino acid class in neutralizing antibodies (Figure 2B) and the same positions in anti-gp120 antibodies with putative contacts (first case study) suggests that in some antibodies, the amino acid class change might have brought a new characteristic that allows some interaction with HIV (mainly positions in CDR2 and framework region).

The case studies presented here demonstrate the usefulness of some of the Yvis resources. Moreover, Yvis facilitates the formulation of hypotheses concerning the subset of analysed sequences. This was achieved by analysing conserved or divergent properties in specific positions or by comparing a set of sequences with their germline gene sequence or another reference sequence.

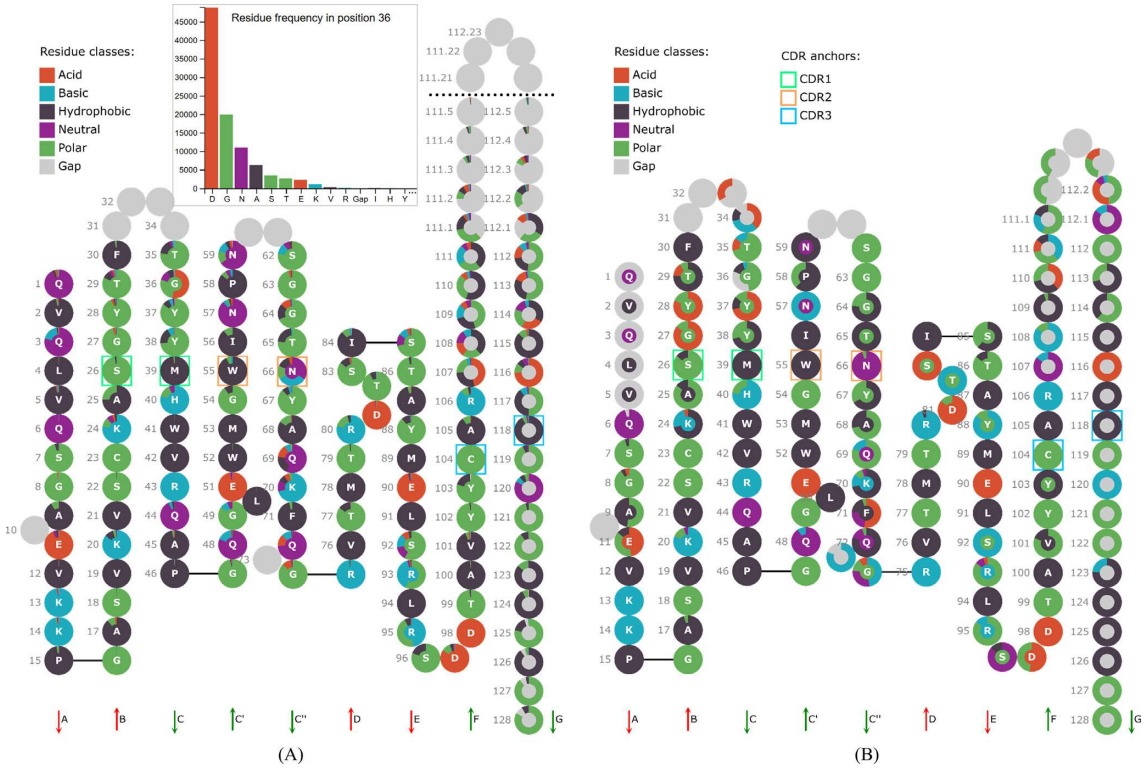


Figure 2. *Collier de Diamants* representation of antibody sequences obtained from HIV-infected patients. (**A**) Visualization of the alignment of 97 751 heavy chain sequences derived from the IGHV1-2*02 allele. Positions between 111.6 and 111.20 and positions between 112.20 and 112.6 were omitted. The box presents part of the detailed bar chart for position 36 showing the residue frequency in this position. (**B**) Visualization of the alignment of 21 heavy chain sequences of HIV neutralizing antibodies. In both visualizations, the IGHV1-2*02 germline sequence was used as comparison sequence. As this sequence includes only the V gene portion of the variable domain, gaps were inserted in the comparison sequence.

CONCLUDING REMARKS

In this article, we presented the Yvis platform that was developed to facilitate the analysis of large numbers of antibody structures and sequences. The Yvis platform includes a very convenient, innovative and robust high-density visualization (*Collier de Diamants*) of antibody sequence alignment that allows the analysis of hundreds of thousands of sequences in a single representation. Moreover, the Yvis integrated database is updated weekly and stores data on antibody PDB structures obtained from SAbDab and PDB files (e.g. PDB and chain identifications, antibody/antigen producing-organism names, molecule description), processed data from IMGT/DomainGapAlign (e.g. gapped sequences, V and J germline allele assignment and the corresponding identity values), and antibody-antigen putative contacts obtained by processing the structure coordinates. The producing-organism names are stored following Uniprot Taxonomy that facilitates database searches based on these names. In addition, there are various database search and filter options. The Yvis platform can also process user files that contain sequences obtained by different methods (e.g. FASTA, IMGT/DomainGapAlign and IMGT/HighV-QUEST files).

The Yvis platform can be used in different types of antibody analysis. For example, the quick visualization of the most conserved or divergent positions in a set of related antibodies can guide antibody engineering and mutagenesis experiments. In antibody repertoire studies, the *Collier de Diamants* visualization, coupled with the sequence comparison feature, can be used to compare thousands of antibody sequences with a specific germline sequence. This can give to researchers some insights into the most important mutations that occurred during the antibody affinity maturation process. Therefore, the Yvis platform offers an environment for antibody sequence analysis that helps to formulate hypotheses concerning the key residues in the antibody structure or interactions and improves the understanding of the antibody properties.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank José Miguel Ortega for providing computational resources.

FUNDING

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (Capes) [PVE 88887.125025/2014-00, Bio-Computacional 51/2013, COFECUB 935/19]; FAPEMIG [APQ-01437-16]. Funding for open access charge: Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais, Programa Interunidades de Pós-graduação em

Bioinformática da Universidade Federal de Minas Gerais and Sys2Diag.
Conflict of interest statement. None declared.

REFERENCES

1. Sela-Culang,I., Kunik,V. and Ofran,Y. (2013) The structural basis of antibody-antigen recognition. *Front. Immunol.*, **4**, 302.
2. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
3. Ferdous,S. and Martin,A.C.R. (2018) AbDb: antibody structure database-a database of PDB-derived antibody structures. *Database*, **2018**, bay040.
4. Swindells,M.B., Porter,C.T., Couch,M., Hurst,J., Abhinandan,K.R., Nielsen,J.H., Macindoe,G., Hetherington,J. and Martin,A.C. (2017) abYsis: integrated antibody sequence and structure-management, analysis, and prediction. *J. Mol. Biol.*, **429**, 356–364.
5. Ehrenmann,F., Kaas,Q. and Lefranc,M.P. (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.*, **38**, D301–D307.
6. Allcorn,L.C. and Martin,A.C. (2002) SACS—self-maintaining database of antibody crystal structure information. *Bioinformatics*, **18**, 175–181.
7. Dunbar,J., Krawczyk,K., Leem,J., Baker,T., Fuchs,A., Georges,G., Shi,J. and Deane,C.M. (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146.
8. Briney,B., Inderbitzin,A., Joyce,C. and Burton,D.R. (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, **566**, 393–397.
9. Kovaltsuk,A., Leem,J., Kelm,S., Snowden,J., Deane,C.M. and Krawczyk,K. (2018) Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *J. Immunol.*, **201**, 2502–2509.
10. Vlachakis,D., Feidakis,C., Megalooikonomou,V. and Kossida,S. (2013) IMGT/Collier-de-Perles: a two-dimensional visualization tool for amino acid domain sequences. *Theoret. Biol. Med. Model.*, **10**, 14.
11. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
12. Lefranc,M.P., Pommie,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
13. Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
14. Viart,B., Dias-Lopes,C., Kozlova,E., Oliveira,C.F., Nguyen,C., Neshich,G., Chavez-Olortegui,C., Molina,F. and Felicori,L.F. (2016) EPI-peptide designer: a tool for designing peptide ligand libraries based on epitope-paratope interactions. *Bioinformatics*, **32**, 1462–1470.
15. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
16. Alamyar,E., Duroux,P., Lefranc,M.P. and Giudicelli,V. (2012) IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol. Biol.*, **882**, 569–604.
17. Dunbar,J. and Deane,C.M. (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**, 298–300.
18. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
19. Wu,X., Zhang,Z., Schramm,C.A., Joyce,M.G., Kwon,Y.D., Zhou,T., Sheng,Z., Zhang,B., O'Dell,S., McKee,K. *et al.* (2015) Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell*, **161**, 470–485.
20. Zhou,T., Lynch,R.M., Chen,L., Acharya,P., Wu,X., Doria-Rose,N.A., Joyce,M.G., Lingwood,D., Soto,C., Bailer,R.T. *et al.* (2015) Structural repertoire of HIV-1-Neutralizing antibodies targeting the CD4 supersite in 14 donors. *Cell*, **161**, 1280–1292.
21. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.