Computational and Structural Biotechnology Journal 18 (2020) 2000–2012

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Methods for sequence and structural analysis of B and T cell receptor repertoires

Shunsuke Teraguchi [a,b], Dianita S. Saputri [b], Mara Anais Llamas-Covarrubias [b,c], Ana Davila [b], Diego Diez [a], Sedat Aybars Nazlica [a], John Rozewicki [a,b], Hendra S. Ismanto [b], Jan Wilamowski [b], Jiaqi Xie [b], Zichang Xu [b], Martin de Jesus Loza-Lopez [a], Floris J. van Eerden [a], Songling Li [b], Daron M. Standley [a,b,*]

[a] *Immunology Frontier Research Center, Osaka University, 3-1 Yamadaoka, Suita, Japan*

[b] *Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Japan*

[c] *Departamento de Biología Molecular y Genómica, Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara, Mexico*

ARTICLE INFO

Article history:
Received 30 April 2020
Received in revised form 8 July 2020
Accepted 8 July 2020
Available online 17 July 2020

Keywords:
B cell receptor
T cell receptor
Immune system
Adaptive immunology
Antigen
Single cell sequencing
Clustering
Machine-learning

ABSTRACT

B cell receptors (BCRs) and T cell receptors (TCRs) make up an essential network of defense molecules that, collectively, can distinguish self from non-self and facilitate destruction of antigen-bearing cells such as pathogens or tumors. The analysis of BCR and TCR repertoires plays an important role in both basic immunology as well as in biotechnology. Because the repertoires are highly diverse, specialized software methods are needed to extract meaningful information from BCR and TCR sequence data. Here, we review recent developments in bioinformatics tools for analysis of BCR and TCR repertoires, with an emphasis on those that incorporate structural features. After describing the recent sequencing technologies for immune receptor repertoires, we survey structural modeling methods for BCR and TCRs, along with methods for clustering such models. We review downstream analyses, including BCR and TCR epitope prediction, antibody-antigen docking and TCR-peptide-MHC Modeling. We also briefly discuss molecular dynamics in this context.

Contents

* Corresponding author.
E-mail address: standley@biken.osaka-u.ac.jp (D.M. Standley).

https://doi.org/10.1016/j.csbj.2020.07.008
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

1. Introduction

B cell receptors (BCRs) and T cell receptors (TCRs) are key molecules in adaptive immune response that provide protection to perturbations, both from the outside (e.g. pathogens) and from within (e.g. mutated or misfolded proteins). Together, BCRs and TCRs constitute a unique class of proteins whose coding sequences are arranged combinatorically in a cell-autonomous manner known as V(D)J recombination. In V(D)J recombination within a given cell, variable (V), diversity (D), and joining (J) segments are selected randomly from among many variants, and joined to make the V (variable) region of a full-length receptor. In addition to V(D)J recombination, BCRs can also undergo subsequent somatic hypermutation (SHM) and clonal selection upon antigen encounter, collectively referred to as "affinity maturation". On a cell population level, these processes create a functionally diverse and dynamic set (repertoire) of B and T cells. The number of possible different BCR or TCR sequence combinations is extremely high, with theoretical estimates in the 10^{12}–10^{18} range [1]. However, the observed populations of receptor sequences in a given individual follow a power law, where most sequences appear only at very low frequency and a minority of sequences appear at higher frequencies (see for example [2] for a recent discussion).

For both BCRs and TCRs, V regions consist of two polypeptide chains, referred to as "light" (BCRs) or "alpha" (TCRs) and "heavy" (BCRs) or "beta" (TCRs). TCRs are composed of a single pair of alpha and beta chains while BCRs contain two pairs of light and heavy chains [1]. For simplicity, in this review, we focus on a single pair of (light-heavy or alpha–beta) chains.

Both BCRs and TCRs belong to the immunoglobulin-like fold in which the canonical antigen binding site is composed of three loops called "complementarity-determining regions" (CDRs), in each receptor chain. The V(D)J recombination junction, in which random nucleotides may be inserted during the recombination, is located in the third CDR (CDR3). As a result, CDR3 is the most diverse among the three CDRs [1]. Much effort has been spent on CDR3 modeling, in particular for soluble BCRs (antibodies).

BCRs interact directly with antigens, and we refer to interface residues as "paratope" on the BCR side and "epitope" on the antigen side (Fig. 1A). TCRs, on the other hand, interact with antigen-derived peptide fragments, which are presented by major histocompatibility complex (MHC) proteins (Fig. 1B). Here, generally "epitope" refers to the antigen-derived peptide and not the MHC contacting residues.

Each human carries up to six class I MHC molecules and up to eight class II molecules. There are thousands of MHC variants (alleles) in the human population, which can differ in their peptide specificity [1]. Peptide-MHC binding affinity shapes the TCR repertoire, and the particular set of MHC alleles carried by an individual become a source of TCR repertoire diversity, affecting the susceptibility to particular diseases (reviewed in [3]). Since BCR maturation requires a co-stimulation from activated helper T cells [4], the BCR and TCR repertoires are not completely independent.

Both BCR and TCR sequences can be captured by current sequencing technologies. Moreover, molecule and cell barcoding technologies are an area of intense research and development. Emerging sequencing and barcoding methods are thus expected to revolutionize our understanding of immune repertoires. As just one example, the number of paired (alpha–beta) TCR sequences for which the peptide-MHC is known has grown by two orders of magnitude in the last two years [5], indicating a need for computational tools that can keep pace with this growth.

In this review, after briefly reviewing recent technologies for repertoire sequencing, we explore tools for interpreting BCR and TCR sequences in terms of their structures and targeted antigens. In this context, we cover structural modeling, epitope prediction, molecular docking, and molecular dynamics. Integration of such tools, along with growth in sequence and associated experimental data, will allow us to more fully describe the immune status of an individual in health and disease.

2. Repertoire sequence analysis

Very early approaches to characterize immune repertoires were limited to estimating the length of the CDR3 loops [6]. Current methods, relying on high-throughput sequencing (HTS) technology, can be used for comprehensive quantification of full-length TCR and BCR V region sequences [7,8]. Though a comprehensive review on the existing technologies for repertoire sequencing analysis is beyond the scope of this review, HTS is the main source of data for subsequent structural analysis. Therefore, we briefly describe the basic information contained in bulk and single-cell RNA-based repertoire sequencing (Fig. 2).

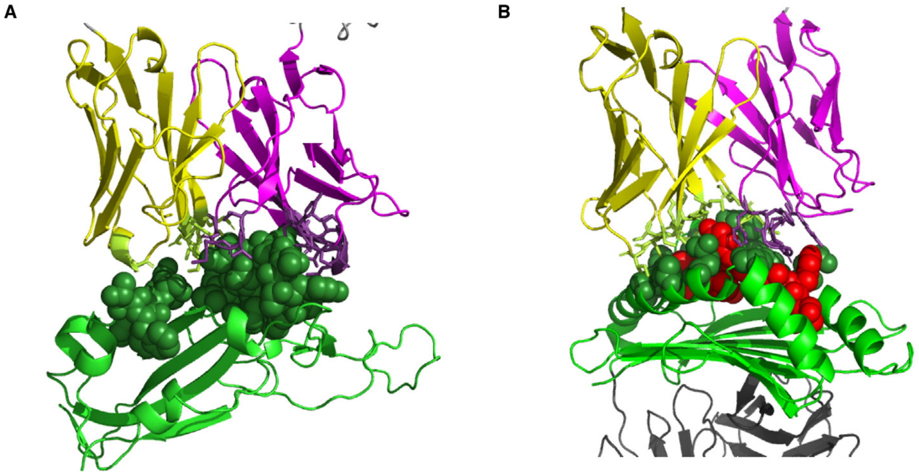

Fig. 1. Paratope and epitope in BCRs and TCRs. A, A crystal structure of SARS-CoV S protein receptor binding domain (green) bound by a neutralizing antibody (PDB identifier: 2DD8); heavy and light chains are colored (magenta and yellow, respectively). Epitope residues are shown as dark green spheres. TCR contacting residues are shown as sticks. B, TCR-peptide-MHC complex for a viral peptide TAX and class I HLA A-0201 (PDB identifier 1BD2). The epitope is shown as red spheres and contacting MHC residues are shown as green spheres, while paratope residues are shown as sticks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

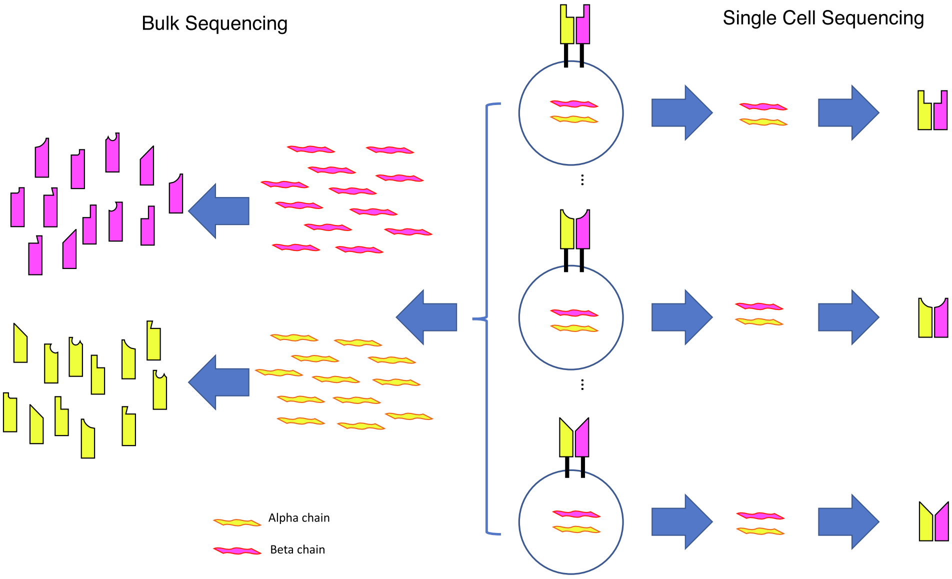

Fig. 2. Conceptual difference of bulk and single cell repertoire sequencing. In bulk sequencing, the information of receptor pairs will be lost while higher coverage tends to be achieved. In single cell sequencing, the pairing information is preserved while currently sample preparation and sequencing costs tend to be higher than in bulk sequencing.

2.1. Bulk sequencing

Early development of HTS repertoire analysis was based on bulk sequencing (i.e. sequencing many cells without preserving their identities). In this approach, the information of light/heavy or alpha/beta pairs is lost. Thus, bulk sequence analysis tends to focus on a single (typically the heavy/beta) chain.

Repertoire sequencing typically uses TCR/BCR enrichment followed by PCR amplification to increase sensitivity and reduce sequencing cost. Since a 100 bp fragment is enough to resolve the CDR3 fragment, short read sequencing is often used. The choice of sequencing technology can have an important impact on quality, since the types and rates of errors can be different. Among preferred platforms are Illumina MiSeq (long reads) and HiSeq (short reads targeting CDR3).

One of the sources of low-quality repertoire data is a biproduct of PCR amplification. Without other information, we cannot distinguish between true nucleotide sequence differences and PCR errors. As a result, PCR errors cause the appearance of spurious sequences, in particular from dominant, highly abundant sequences/clonotypes. Use of Unique Molecular Identifier (UMI) sequences enables correction of PCR amplification biases and quantification of the number of receptors expressed. Thus, the use of technologies with UMI have a distinct advantage.

To date, several pipelines can be used to extract repertoire information from bulk HTS data. These tools generally map sequencing reads to TCR/BCR reference sequences. Then, contigs, the continuous sequences assembled from the mapped reads, can subsequently be annotated by V(D)J gene usage and CDR (1,2,)3 amino acid sequences [9,10]. IMGT/HighV-QUEST (International Immunogenetics Information System V-Query and Standardization) [11,12] uses pairwise alignment and sequence comparison to experimental data to align sequencing reads. IgBLAST [13] utilizes the BLAST algorithm [14] for its search engine. MiXCR [15] is an efficient pipeline equipped with a fast aligner. It can be used for reconstructing TCR/BCR sequences from generic RNA-seq data without PCR amplification of TCRs/BCRs [16]. A detailed assessment on those three tools can be found in [17]. The Immcantation framework [18,19] and TRUST (TCR repertoire utilities for solid tissue) [20] can be also used for the same purpose among many other available tools not covered here

Though single chain information alone is usually not enough to explain the binding of the receptor to the target epitope, there are several methods applicable to bulk sequencing data. For example, diversity analysis of the repertoire sequences can be used for estimating the clonal diversity of an immune repertoire of each individual, as well as repertoire overlap among repertoires of several individuals. This can currently be performed using conventional ecology measures [21–23], or repertoire-designed estimators [24–26]. Also, by analyzing repertoire data from many individuals with additional information like Human Leukocyte Antigen (HLA) allele profiles or disease status, one can associate each TCR with particular labels with the help of statistical hypothesis testing [27,28]. Repertoire information also carries the information of underlying V(D)J recombination. Thus, from repertoire sequences, generative models of V(D)J recombination were developed; and, in turn, these models were used to analyze repertoire sequence data [29–34]. We have collected some of (but not all of) tools used for those sequence analysis as in Table 1.

2.2. Single cell sequencing

The most important limitation of bulk sequencing approaches is the loss of pairing between receptor chains. This limitation is addressed by single cell repertoire profiling methods. These methods use a number of cell barcoding strategies to add a unique barcode to each cDNA in a given cell. New approaches are dramatically improving the ability to measure full length paired receptors at the single-cell level. For example, RAGE-seq (Repertoire and Gene Expression by Sequencing) combines long reads from Oxford Nanopore sequencing with short reads from Illumina sequencers [35]. When combined with droplet based single cell RNA-seq approaches, we can characterize the full-length paired repertoires of thousands of single cells. In addition, off-the-shelf single cell repertoire sequencing platforms are currently available from various companies including 10x Genomics and Takara Bio.

In the case of single-cell gene expression data, TRAPeS (TCR Reconstruction Algorithm for Paired-End Single Cell) [36], TraCeR (Reconstruction of T cell receptor sequences from single cell RNA-seq data) [37] and VDJPuzzle [38] are often used for analysis of TCRs. Meanwhile, BASIC (BCR assembly from single cells) [39], BraCeR (B-cell-receptor reconstruction and clonality inference

Table 1
Repertoire sequence analysis tools.

Tools	Purpose	URL	References
IgBLAST	Bulk Sequence reconstruction	https://www.ncbi.nlm.nih.gov/igblast/	[13]
IMGT/HighV-QUEST		http://www.imgt.org/IMGTindex/IMGTHighV-QUEST.php	[11,12]
MiXCR		https://mixcr.readthedocs.io/en/master/index.html	[15]
TRUST		https://bitbucket.org/liulab/trust/src/master/	[20]
TRAPeS	Single cell Sequence reconstruction	https://github.com/YosefLab/TRAPeS	[36]
TraCeR		https://github.com/teichlab/tracer	[37]
VDJPuzzle		https://github.com/simone-rizzetto/VDJPuzzle	[38]
BASIC		http://ttic.uchicago.edu/~aakhan/BASIC/	[39]
BraCeR		https://github.com/teichlab/bracer/	[40]
VDJtools	General repertoire analysis	https://github.com/mikessh/vdjtools	[21]
Immcantation		https://immcantation.readthedocs.io/en/stable	[19]
Vidjil		http://www.vidjil.org http://bioinfo.lille.inria.fr/vidjil	[22]
ASAP		https://asap.tau.ac.il	[119]
ARGalaxy		https://bioinf-galaxian.erasmusmc.nl/argalaxy/	[120]
bcRep		https://cran.r-project.org/web/packages/bcRep/vignettes/vignette.html	[121]
Immunarch		https://immunarch.com	[23]
Sumrep		https://github.com/matsengrp/sumrep	[122]
DiVE	Specialized in diversity analysis	http://cran.r-project.org/web/packages/DivE/index.html	[24]
RDI		https://rdi.readthedocs.io/en/1.0.0/	[25]
RECOLD		https://github.com/Q-bio-at-IIS/RECOLD/tree/master/codes	[26]
OLGA	Generative model of VDJ recombination	https://github.com/statbiophys/OLGA	[29]
IgoR		https://github.com/qmarcou/IGoR	[34]
SONIA		https://github.com/statbiophys/SONIA	[30]
vampire		https://github.com/matsengrp/vampire/	[26]

from single-cell RNA-seq) [40] and an extension of VDJPuzzle [41] are often used for BCR analysis. These tools mainly differ on the way they assemble the missing information after mapping to reference sequences, and the final results are generally consistent. Since structural modeling has yet to be effectively used for predicting chain pairing, single cell sequencing technologies are critical for TCR or BCR structural modeling. Moreover, expression of TCRs or BCRs requires such pairing and so single cell sequencing is important for most downstream analyses of T or B cells and their cognate antigens.

2.3. Extensions of repertoire sequencing

There have also been exciting developments in the application of HTS technology for experimental discovery of epitopes. In Libra-seq (Linking B cell receptor to antigen specificity through sequencing) [42], the 10x Genomics platform was used to barcode not only BCR sequences but also antigen proteins. By sorting the antigen-bound B cells and then performing single cell sequencing, antigen specific BCRs can be identified from the antigen barcodes. Similarly, by using barcoded peptide-MHC complexes, HTS allow us to generate a large reference dataset of TCR-epitope pairs [43]. Kula et al. [44] developed T-Scan, a high-throughput method that identifies functional antigen targets of CD8 T cells. They started from bulk memory T cells and made antigen libraries such that target cells could present the antigens on MHC molecules. Recognition of target cells by T cells and subsequent next-generation sequencing enabled T-scan to discover CMV antigens as well as the targets of self-reactive TCRs. Gee MH et al. [45] used yeast-display libraries of pMHCs and screened for antigens of orphan T cell receptors on tumor-infiltrating lymphocytes. Kobayashi et al. [46] have developed a cloning and expression system called hTEC10 (human TCR efficient cloning system within 10 days) that can be used to rapidly determine the antigen specificity of TCRs. They applied their system successfully to peptide specificity and cytotoxic activity of TCRs from EBV infection and cancer.

3. TCR and BCR 3D structural modeling

In spite of advances in experiential determination of receptor-antigen interactions, most high-throughput experiments lack residue-level resolution. X-ray crystallography and single-particle electron microscopy (cryo-EM), on the other hand, provide such high-resolution information, but are not suitable for high-throughput analysis. Computational modeling of TCRs and BCRs is now routine and can be performed in a high-throughput manner. Building 3D models of receptors is also the first step in structure-based analysis of receptor antigen interactions. For 3D structural modeling, TCR or BCR V regions are generally divided into "frameworks" and the three CDRs (Fig. 3). Each framework is a double layer of beta sheets that contain the beginning and ending of each CDR loop. There are other loops in V regions, but the CDRs are important because of their high sequence diversity and because they form a continuous surface that constitutes the main antigen binding interface. Of the CDRs, CDR3 is the most diverse in terms of both sequence and structure. CDR3 modeling has been tackled by a wide range of approaches [47]. Software for CDR3 modeling (Table 2) spans the range from simple sequence alignment methods [48], to fragment assembly [49], molecular dynamics (MD) [50] and robotics-based loop closure algorithms [51]. In the most recent antibody modeling assessment (AMA-II) [52], the lowest heavy-chain CDR3 (CDRH3) errors were obtained by our own group using a combination of MD, fragment assembly and manual selection [53]. Based on an internal assessment of our AMA-II results, we developed a purely fragment assembly-based tool, Kotai Antibody Builder [54]. We more recently introduced Repertoire Builder, which exceeded Kotai Antibody Builder in terms of accuracy, with a factor of 100 improvement in speed [55]. In the same time frame, several new tools, including ABodyBuilder [56], TCRModel [57], and PigsPro (Prediction of immunoglobulin structure v2) [58] have been introduced, which show advancement over previously published methods. Because of its high accuracy and ability to scale with the number of input sequences, we will briefly outline the Repertoire Builder approach.

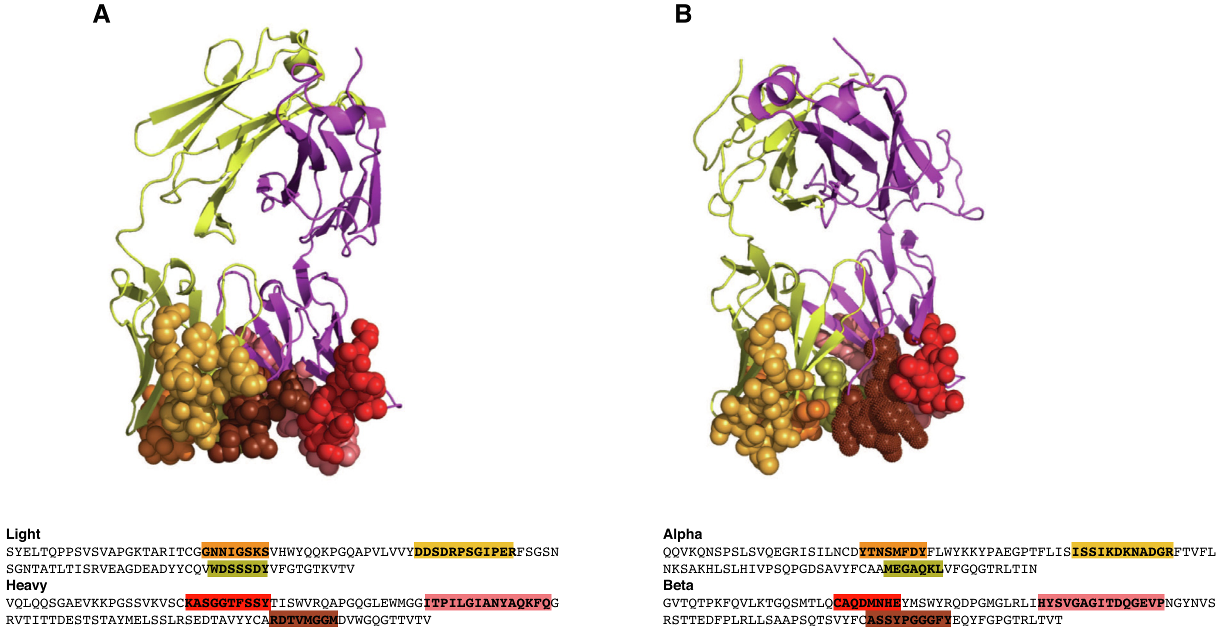
A
B
Light
SYELTQPPSVSVAPGKTARITCGGNNIGSKSVHWYQQKPGQAPVLVVYDDSDRPSGIPERFSGSN
SGNTATLTISRVEAGDEADYYCQVWDSSSDYVFGTGTKVTV
Heavy
VQLQQSGAEVKKPGSSVKVSCKASGGTFSSYTISWVRQAPGQGLEWMGGITPILGIANYAQKFQG
RVTITTDESTSTAYMELSSLRSEDTAVYYCARDTVMGGMDVWGQGTTVTV
Alpha
QQVKQNSPSLSVQEGRISILNCDYTNSMFDYFLWYKKYPAEGPTFLISISSIKDKNADGRFTVFL
NKSAKHLSLHIVPSQPGDSAVYFCAAMEGAQKLVFGQGTRLTIN
Beta
GVTQTPKFQVLKTGQSMTLQCAQDMNHEYMSWYRQDPGMGLRLIHYSVGAGITDQGEVPNGYNVS
RSTTEDFPLRLLSAAPSQTSVYFCASSYPGGGFYEQYFGPGTRLTVT

Fig. 3. BCR and TCR structure. Representative BCR and TCR structures. The location in structure and sequence of the three CDRs are shown for a representative BCR (A) and TCR (B) using the same PDB entries as in Fig. 1.

Table 2
BCR or TCR 3D modeling tools.

Tools	BCR	TCR	URL	References
Repertoire Builder	Yes	Yes	https://sysimm.org/rep_builder/	[55]
PigsPro	Yes	No	http://biocomputing.it/pigspro	[58]
Rosetta Antibody	Yes	No	https://rosie.graylab.jhu.edu/snug_dock	[123]
ABodyBuilder	Yes	No	http://frodock.chaconlab.org/	[56]
LYRA	Yes	Yes	http://www.cbs.dtu.dk/services/LYRA/	[82]
TCRpMHCmodels	No	Yes	http://www.cbs.dtu.dk/services/TCRpMHCmodels/	[83]

In order to improve speed and reduce noise, one aim of Repertoire Builder was to remove 3D structure from the key decision-making steps: sampling and scoring. Working in three dimensions is computationally expensive and also messy, as protein structure files can contain a plethora of sources of noise. As an alternative, we derived feature vectors from pairwise query-template alignments and trained a machine learning model to recognize the good alignments. Feature vectors currently consist of BLOSUM62 matrix elements or gaps for each aligned residue pair and cover the entire V region. The inclusion of residues outside of the CDR region was intended to take the environment of the CDR into account in the choice of template. We note that scoring at the alignment level is not unique to Repertoire Builder; all of the methods do this. What is novel here is the alignment-derived feature vectors. Another trick used by Repertoire Builder was to store templates in the form of structure-aware multiple sequence alignments (MSAs), which can be readily computed using our MAFFT-DASH (Multiple Alignment using Fast Fourier Transform-Database of Aligned Structural Homologs) pipeline and which have been shown to be significantly more accurate than sequence-based MSAs [59]. The query sequence can be added to a stored template MSA efficiently using MAFFT's fragment-adding option, which preserves the relationships between the templates in the stored MSAs [60]. Templates in MSAs are grouped by their CDR lengths. Thus, there is a different template MSA stored for each CDR-length combination. The advantage of using MAFFT-DASH in this manner is primarily a combination of speed and MSA accuracy. We have not assessed whether use of alternative alignment strategies results in a degradation of model quality. The current Repertoire Builder can model 10^4 paired or unpaired sequences in approximately 30 min, which makes it practically useful for high-throughput sequencing discussed above. To our knowledge, Repertoire Builder is the only server that allows multiple BCR or TCR sequences to be input at one time.

4. TCR and BCR clustering

As genomic data continues to grow, methods for clustering nucleotide or amino acid sequences will play major role in sequence and structural analysis. Since generic sequence clustering methods (e.g. [61,62]) are beyond the scope of this review, here we focus on methods specific to immune receptors. A common goal when studying immune repertoires is to understand common features of receptors that are shared by a group of donors of interest (Fig. 4). The implication here is that receptors target the same antigen and epitope will be more common in the donors of interest than in a control group. This is a very general notion that can be applied to either BCRs or TCRs and approached in a variety of ways. Given the broad diversity of immune repertoires, their uneven population distributions, and the relatively low overlap of exact matching sequences among subjects, this task is a significant challenge. To address these issues, several clustering strategies have been developed recently. Below, we review some representative examples, including our own efforts.

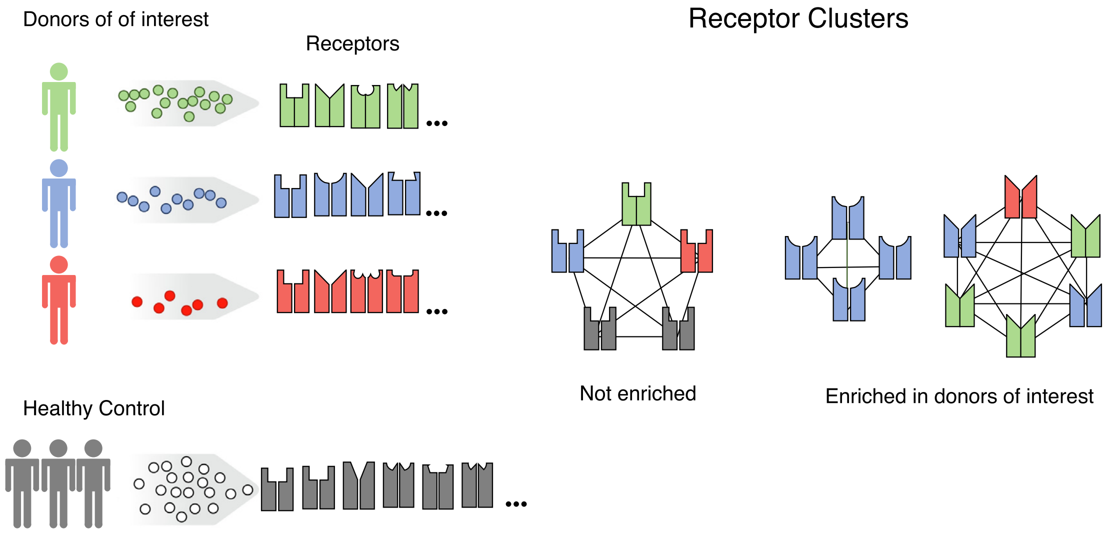

Fig. 4. Receptor clustering. B or T cells of interest are acquired from donors of interest, receptors are sequences and clustered based on sequence features, structure features, or both. Clusters that are enriched in receptors from donors of interest are identified.

4.1. TCR clustering

Based on the observation that there are specific positions in TCR CDR3 regions that contact antigen peptides and that the presence of particular sequence motifs can define TCR clusters, Glanville et al., developed the GLIPH (grouping of lymphocyte interactions by paratope hotspots) algorithm [63,64]. This algorithm clusters TCRs based on local sequence motifs, as well as on other parameters such as global CDR3 similarity, V gene usage, CDR3 length, MHC profile of donor(s) and clone size. GLIPH identifies motifs that are enriched in a given dataset relative to a control group, with the goal of producing groups of TCRs targeting the same peptide-MHC (pMHC). By using this approach, the authors were able to design synthetic antigen-specific TCRs to groups, and confirm their specificity experimentally.

In a similar study, Dash et al. [65] developed TCRdist; a tool that estimates the similarity of two TCR sequences by computing a weighted Hamming distance among the concatenated amino acid sequences of the CDR loops of each TCR. TCRdist assumes a higher weight (3x) for the CDR3 regions. Clusters of highly similar antigen-specific TCRs can be built, and new TCRs of unknown specificity can be assigned to an antigen-specific cluster based on similarity, allowing for the prediction of antigen specificity. Additionally, a diversity score (TCRdiv) that robustly calculates the diversity of epitope-specific repertoires by considering both TCR similarity and exact identity in a generalized Simpson's diversity index, was developed. TCRdist has recently been used to identify clonal expansion of *M. tuberculosis* specific TCRs in a South African cohort where it was able to accurately classify active tuberculosis patients [66].

Though they share the same goal, the focus of those two tools are slightly different. The GLIPH algorithm assumes that the input data is enriched in TCRs targeting a restricted set of epitopes, and tries to cluster these enriched TCRs using common motifs in the dataset. With this approach, they are also able to avoid direct comparison of all pairs of sequences, which is computationally expensive. Thus, GLIPH is suitable for large repertoire analyses of particular disease cohorts. On the other hand, TCRdist is based on direct comparison of each TCRs using a "universal" measure of TCR similarity, and it is thus currently difficult to apply the method to datasets greater than approximately 10^4. However, an advantage of TCRdist is that the calculated distance between a pair of TCRs are always the same, regardless of other factors. Such "universal" definition of TCR similarity/difference is of use when assumptions about shared antigen/epitope cannot be made.

4.2. BCR clustering

Structural studies of antibodies targeting antigens specific to HIV [67], influenza [68] and more recently SARS-CoV-2 [69] have demonstrated that antibodies produced in unrelated donors targeting common antigens and epitopes can share sequence and structural features. We note here that, since B cells can undergo affinity-driven maturation, such receptors need not derive from a similar common clone. Recently, the SAAB + tool was developed to characterize structural properties of CDRs from differentiated B cells [70]. It is likely that more tools trained to identify "convergence" of functionally related antibodies will appear in the future as more sequence data from donors with shared BCR epitopes become available.

To this end, we recently developed InterClone, a method to cluster BCR sequences which are likely to share epitopes [71]. InterClone is based on a comparison of sequence and structural features of pairs of BCRs using a machine learning-based classifier that was trained on known antigen-BCR structures. Like TCRdist, InterClone assigns a "universal" similarity score to each BCR pair. Hierarchical clustering is then used to group sequences of high similarity. As such, InterClone can be used without requiring sequences to be enriched in a particular BCR motif. A sensitivity of 61.9% and specificity of 99.7% were obtained when InterClone was applied to an independent set of anti-HIV antibody sequences [71]. A more robust and computationally efficient version of InterClone that works for both BCRs and TCRs and can perform high-throughput analysis of up to 10^5 sequences is currently being developed.

In addition to the above clustering methods, networks that describe antibody repertoire architecture can be used to compare repertoires. Miho and colleagues [72] developed a platform that builds similarity networks of hundreds of thousands of antibody sequences from both humans and mice. Using this approach, the authors detected global patterns in antibody repertoire architectures that were highly reproducible in different subjects, and tended to converge despite independent VDJ recombination. Furthermore, these repertoire architectures were robust to clonal deletion of private clones.

5. Epitope specificity

5.1. Predicting TCR epitopes

TCRs recognize short peptides presented on class I or II MHC complexes. The ability to predict epitope(s) from TCR sequence and MHC allele would be highly valuable in elucidating disease etiology, monitoring the immune system, developing diagnostic assays and designing vaccines. Traditionally, identifying epitopes is carried out experimentally [73], and is both costly and time-consuming. There is necessarily great interest in methods that can accelerate this process computationally.

To this end, Fischer et al. [74] developed a deep learning approach on TCR CDR3 regions to predict the antigen-specificity of single T cells. Jokinen et al., [75] developed TCRGP to predict whether TCRs recognize certain epitopes using a novel Gaussian process (GP). Their method uses CDR sequences from TCR alpha and beta and learns which CDR recognizes different epitopes. The tool was applied to identify T cells specific to HBV. NetTCR by Jurtz VI et al. [43] utilized convolutional networks for sequence-based prediction of TCR-pMHC specificity. NetTCR uses the recent explosion of next-generation sequencing data to train a sequence based-predictor. Ogishi et al. [76] computationally defined immunogenicity scores through sequence-level simulation of interaction between pMHC complexes and public TCR repertoires. Though their focus is more on immunogenicity of peptides presented to MHC molecules, they also observed correlation between individual TCR-pMHC affinities and the features important for immunogenicity score. Gielis et al. [77] applied random forest-based classifiers for epitope specific TCRs to repertoire level analysis. Their models successfully detected the increase of epitope specific TCRs upon vaccination in two Yellow Fever vaccination studies. The works by Chain and co-workers [78,79] also addressed related questions. In [78], the authors have constructed a classifier to distinguish the TCR beta sequences in expanded repertoires of ovalbumin-stimulated mice from control. Their classifier was based on the frequencies of amino acid triplets in CDR3 and their choice of machine learning algorithm called LPBoost (linear programming boosting) allowed them to identify the responsible motifs in CDR3.

5.2. TCR-pMHC 3D modeling

Unlike BCRs, which can be expressed as soluble antibodies, TCRs remain attached to the cell surface. This, along with their weaker binding affinities to pMHC complexes, has made experimental structural analysis more difficult than for BCRs. Nevertheless, from the known crystal structures of TCR-pMHC complexes, we can see that the range of docking modes is highly restricted, as expected by the similarity of MHCs within a given class (Fig. 5). As a result of this restriction, we and others [80] have approached the problem using structural templates for TCR-pMHC docking.

There are currently few methods for modeling TCR-pMHC complexes. To our knowledge, there are two public servers for this purpose: our own ImmuneScape [81] and the Lymphocyte Receptor Automated Modeling or LYRA-based [82] TCRpMHCmodels [83]. Both of these approaches are ''template-based'' in the sense that existing structures instead of stochastic conformational sampling are used as templates for each of the key modeling steps: TCR, pMHC and TCR-pMHC orientation. They are also both ''bottom-up'' in the sense that models for TCR and pMHC are built and then combined to form the TCR-pMHC complex. One possible conceptual difference is that, in ImmuneScape, CDRs are modeled after the TCR and pMHC templates are combined in order to take the pMHC into account. It will be interesting to compare the two

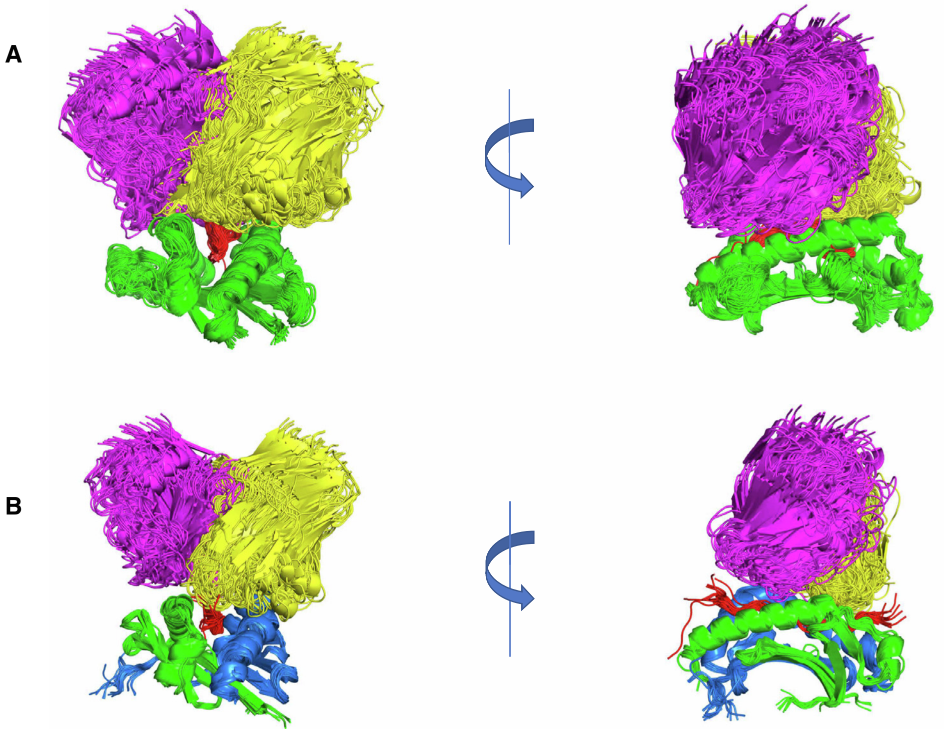

Fig. 5. Restricted docking of TCR-peptide-MHC complexes. A representative set of MHC class I (A) and class-II (B) complexes from the PDB were superimposed using conserved residue positions in the MHC. TCR alpha (yellow) and beta (magenta) chains are contained within a narrow ensemble of binding modes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

approaches in more detail. TCRpMHCmodels compared favorably to an earlier rigid docking-based approach, TCRFlexDock, which suggests that care must be taken in sampling TCR-pMHC orientations beyond that which is observed in typical crystal structures.

5.3. Predicting BCR epitopes

Several computational methods are available to predict BCR epitopes and paratopes. Of the two problems, paratope prediction is much easier, as paratopes tend to correspond to CDR residues, while epitopes can be anywhere on an antigen. This is illustrated in the case of anti-influenza hemagglutinin (HA) antibodies (Fig. 6); a superimposition of all known anti-HA antibodies leaves very little un-targeted surface area.

Paratope prediction methods include the Paratome algorithm [84], which is based on structural consensus between BCRs and uses features from sequence or structure; Prediction of Antibody Contacts or ProABC [85], which applies a random forest learning technique and is based on sequence; Parapred [86], which uses a deep learning architecture to extract patterns from variable regions in sequence; AntibodyInterfacePrediction [87], which uses a support vector machine method (SVM) to classify antibody surface patches based on 3D Zernike descriptors; ProABC-2 [131], which is an upgrade of the original algorithm [85] with convolutional neural networks, and improves performance over existing methods. Additionally, paratope predictors have evolved to be specific to cognate antigen. The antibody i-Patch [89] algorithm introduces a likelihood score for residue contact as a constraint on local docking to generate predicted paratope residues, and thus requires the structure of the antigen-antibody complex. AG-Fast-Parapred [88], which is based on deep neural networks, utilizes antigen sequence information to predict paratope.

With regard to epitope prediction there are many tools available. Previously, methods were built to predict linear epitopes that are contiguous polypeptide chains, an example of which is LBtope (Linear B-Cell epitope prediction server) [90], which discriminated experimentally verified B-cell epitopes from background using SVM. However, the majority of epitopes are non-continuous surface residues characterized by structure as well as sequence. Several methods are available to treat such conformational epitopes. SEPIa [91] uses a combination of two classifiers (naive Bayesian and random forest) from antigen sequence. BepiPred-2.0 [92] uses random forest algorithms to predict epitopes from primary sequence only. Glep [93], is a recent method based on subgraph clustering for the prediction of separated and overlapping epitopes.

Recently, there has been a realization that epitope prediction without reference to a particular antibody is an ill-formed problem, and methods for "antibody-specific epitope prediction" have been introduced [94]. There are currently few options for antibody-specific epitope prediction. The PEASE (Predicting Epitopes using Antibody Sequence) [95] method applies machine learning to predict true contacts of antibody-antigen residue pairs, providing candidates epitope patches. EpiPred [96] identifies the epitope region by rescoring antibody-antigen global docking based on geometric matching of antigen–antibody interfaces and asymmetric potentials. MAbTope [97] predicts epitope residues based on consensus epitopes shared by top-ranked poses; the success of this approach depends on the quality of the docking. PECAN [132] predicts binding interfaces on both antibodies and antigens by learning context-aware structural representations; it applies a unified deep learning framework that consists of a combination of graph convolutional networks, attention and transfer learning. Although there is a clear awareness of the importance of antibody information in epitope prediction, the traditional antigen-centric methods cannot easily be extended to include such information. This is partially because of the increase in the number of degrees of freedom when antibody-antigen interactions are considered.

5.4. BCR-antigen docking

The most direct means of tackling antibody-antigen interactions is through protein docking, a technique that requires structure information of antibody and antigen. This introduces 6 additional degrees of freedom for rigid docking and a host of other issues due to the complexity and inherent uncertainty of protein structural information. Nevertheless, protein docking is a mature field and steady progress has been made in this area. Generally speaking, docking methods can be classified into four categories: Fast Fourier transform (FFT) correlation; Monte-Carlo (MC) simulated annealing; Geometric hashing; and flexible docking [98]. In Table 3, we give a representative list of molecular docking tools or web servers that can be applied to antibody-antigen docking.

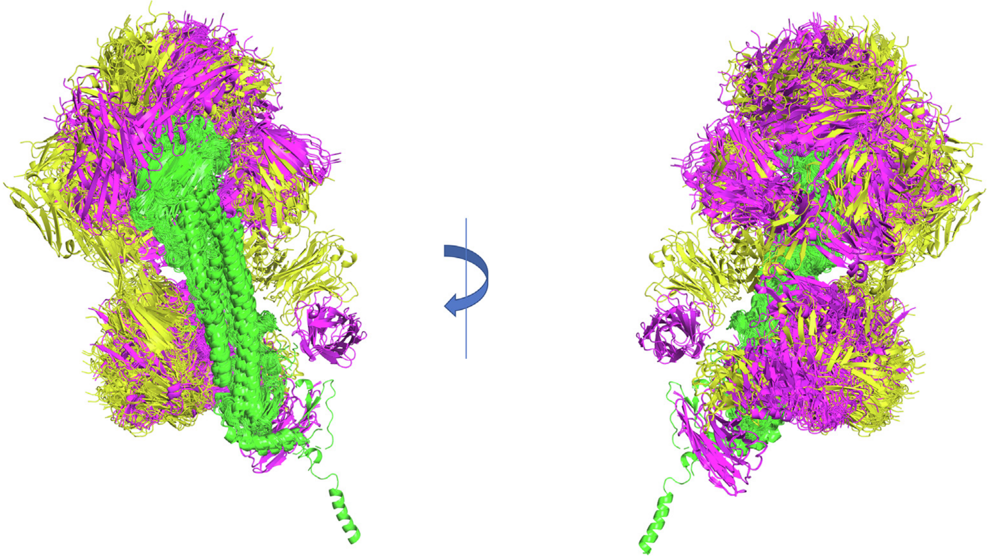

Fig. 6. BCR epitopes on influenza hemagglutinin. A representative set of anti-HA antibodies bound to HA from the PDB were superimposed using conserved residues in HA. HA is a symmetric trimer and antibodies are only shown bound to the chain facing toward the back for simplicity.

Table 3
Antibody docking methods.

Tools	Docking mode	URL	Algorithm	References
ClusPro	Have Ab specific mode	https://cluspro.bu.edu/login.php	FFT based	[99]
SnugDock/ Rosseta	Have Ab specific mode	https://rosie.graylab.jhu.edu/snug_dock	Semi flexible docking with energy minimization	[49,102,123]
FRODOCK2.0	Have Ab specific mode	http://frodock.chaconlab.org/	FFT based	[101]
PatchDock/ FireDock	Have Ab specific mode	https://bioinfo3d.cs.tau.ac.il/PatchDock/, http://bioinfo3d.cs.tau.ac.il/FireDock/	Geometric hashing based	[100,124]
HADDOCK2.2	Not Ab specific mode	https://haddock.science.uu.nl/services/HADDOCK2.2/	MC simulated annealing based	[103]
ZDOCK	Not Ab specific mode	http://zdock.umassmed.edu/	FFT based	[105]
SwarmDock	Not Ab specific mode	https://bmm.crick.ac.uk/~svc-bmm-swarmdock/	Flexible docking with Particle Swarm Optimization (PSO)	[125]
LightDock	Not Ab specific mode	https://lightdock.org/	Flexible docking with Glowworm Swarm Optimization (GSO)	[104]
pyDockWeb/ pyDock	Not Ab specific mode	https://life.bsc.es/pid/pydockweb	FFT based	[126]
HDOCK	Not Ab specific mode	http://hdock.phys.hust.edu.cn/	FFT based	[127]
HexServer	Not Ab specific mode	http://hexserver.loria.fr/	FFT based	[128]
ATTRACT	Not Ab specific mode	http://www.attract.ph.tum.de/services/ATTRACT/	Energy minimization	[129]
GRAMM-X	Not Ab specific mode	http://vakser.compbio.ku.edu/resources/gramm/grammx/	FFT based	[130]

Of these, Cluspro [99], PatchDock [100], FRODOCK [101] and SnugDock [102] provide Antibody-Antigen specific modes and are capable of automatically masking non-CDR regions. Among the four, ClusPro, FRODOCK and PatchDock implement rigid-body or soft docking which do not consider the large conformational changes in the Antibody or Antigen. Although we are not aware of a flexible docking methodology tailored for antibody-antigen interactions [102], SnugDock takes molecular flexibility into account by optimizing the antibody-antigen rigid-body positions, orientation of the H/L chains and conformations of the six CDR loops.

Recently, Vreven et al. used a well-established flexible docking program, HADDOCK [103] and another three representative tools (ClusPro, LightDock [104] and ZDOCK [105]) to systematically analyze 16 antibody-antigen complexes from the well-studied ZDOCK protein–protein interaction benchmark (version 5.0) [106]. The results were evaluated using criteria established by the Critical Assessment of PRedicted Interactions (CAPRI) community where models are classified into the four categories: Incorrect, Acceptable, Medium, or High quality [107]. It was demonstrated that information-driven docking, even using noisy predictions of epitope and paratope, could significantly improve performance over all four algorithms [108]. Notably, HADDOCK was capable of providing high quality models for all 16 entries based on CAPRI criteria in this test. However, this study did not evaluate the tolerance of the docking methods to typical BCR modeling errors.

As with all protein docking from homology models, the success of docking antibody models depends heavily on the quality of the starting structures [109]. Structural uncertainties in the binding regions can occur either from flexibility or modeling errors. Moreover, the regions of greatest uncertainty tend to be the CDRs (especially CDRH3), which is highly likely to form part of the paratope [110]. These issues can be addressed to some extent by use of epitope and paratope predictions. However, few antibody docking methods have been rigorously tested using a large benchmark of realistic models. The bottom line is that structure-based prediction of antibody-antigen interactions from sequence involves a number of interrelated tasks: receptor and antigen model building, initial epitope and paratope prediction, docking, scoring and refinement. The combination of so many critical steps results in complexity, both in terms of software integration and in parameter optimization. Fortunately, the emergence of larger and better BCR sequence datasets will be a motivation to develop well-integrated structure prediction pipelines.

6. Molecular dynamics

In this review, we have focused primarily on high-throughput structure-based methods that can be applied to BCR or TCR repertoires. As is clear from the previous section, combining software methods that work well in isolation introduces complexity. Such complexity arises from conceptual considerations (e.g. parameter optimization) and technical issues (code interoperability). In this regard, MD is conceptually simple: it applies Newtonian mechanics to molecular systems. The force fields describing the interatomic interactions can be taken as given and generally do not have to be optimized. Therefore, even though MD is not a high-throughput method, it can be used to independently confirm BCR- or TCR specific calculations.

As with all proteins, the dynamics of BCRs and TCRs is intimately tied to their functions. Protein dynamics are governed by interactions at the level of individual atoms. The time and length scales involved are, however, difficult to observe experimentally. Molecular dynamics offers the possibility to observe the behavior of proteins and lipids at atomistic resolution, and can therefore contribute to a better understanding of the immune system. The challenges facing such studies are illustrated by recent work by the Deane group, who used a large number of molecular dynamics studies to investigate the influence of point mutations on the structure and dynamics of an epitope derived from the Epstein Barr virus [111]. In their simulations they did not observe a strong relation between the structural and dynamical features of the epitope and its immunogenicity. It is not clear if this is due to limitations in their modelling, or due to the complexity of the immune system. Reboul et al. investigated the immunogenicity of a specific epitope when presented by two structurally highly similar MHC complexes, HLA-B*3508 and HLA-B*3501. Only when the epitope is bound to HLA-B*3508 is a strong interaction with the T cell receptor formed. Simulations showed that the epitope exhibits a much higher flexibility in HLA-B*3501, thereby apparently hindering the formation of a strong interaction by the T cell receptor [112].

Most studies focusing on the T cell receptor only study the dynamics of T cell receptors when bound to a pMHC. In contrast, Dominguez and Knapp compared the dynamics of T cell receptors bound to pMHC and free T cell receptors. In their study they found, apart from expected results as an increased flexibility and increased solvent accessible surface of the CDRs in the free T cell receptor, also differences in the hydrogen bond network of the CDR3α chain in the free TCR versus the pMHC bound TCR [113]. A study combining steered molecular dynamics and single-molecule biophysical experiments [114] studied the formation of catch bonds between the pMHC and the TCR. Catch bonds are a special type of bond in which the lifetime increases when more force is applied. This study suggests that catch bond formation is influenced by conformational changes in the pMHC. A downside of molecular dynamics simulations are the high computational requirements. Fodor et al. were able to distill conformational data from pMHC class I x-ray structures using ensemble refinement, which is a refinement technique to obtain dynamic data without the need of more computationally intensive molecular dynamics simulations [115]. Another way to reduce the computational requirements is by using coarse grained simulations, in which atoms are grouped together into beads. Coarse graining allows for the study of much larger systems on longer time scales. Friess et al. modeled the transmembrane domains of the immunoglobulin M (IgM) B cell receptor, which have been unresolved so far, and subsequently used coarse grained simulations to study their aggregation behavior and association with lipid rafts [116].

7. Conclusions

Recent advances in sequencing technology enable the study of immune responses in unprecedented breadth and depth. As discussed above, the emerging data has spawned the development of a wide range of modeling methods that are applicable to B cells, T cells or both. Current challenges include the integration of data and methodologies. For example, sequence and structural information can, in principle, be combined to yield more accurate descriptions of receptors sharing antigen and epitope specificity. Structural modeling is still not in the mainstream of repertoire analysis; nevertheless, 3D modeling methods present a straightforward direction to encompass "shared features" of functionally related receptors in different donors.

In the context of repertoire analysis, we are often interested in the target antigens and epitopes; however, the scale of publicly available data on targeted antigens and epitopes is currently smaller than that of BCR/TCR sequences, and vastly smaller the actual BCR-antigen or TCR-peptide-MHC interactome. As barcoding methods evolve to include antigens themselves [42], there may soon be new and valuable data available to train methods for functional classification of BCRs and TCRs.

At the point where we are asking not only *what* is targeted but also *why* or *why not*, the use of structural modeling is likely to play a critical role in our understanding of BCR and TCR molecular recognition. As a case in point, at the time of this writing, we are in the midst of the COVID-19 pandemic. This is an example where the target antigens, along with their structures, are largely known, and understanding host immune responses to these antigens is of vital importance in the development of diagnostics, biomarkers, vaccines and therapeutics [117]. Structural similarity among neutralizing antibodies targeting SARS-CoV-2 [69] or between SARS-CoV-1 and SARS-CoV-2 [118] have been noted. With such high stakes driving research and development, integration of emerging technologies in the repertoire analysis domain, including structural analysis, is expected. As the saying goes, "necessity is the mother of invention," and the need for understanding human immune repertoires has never been greater.

Author statement

Shunsuke Teraguchi, Diego Diez, Hendra S. Ismanto and Mara Anais Llamas Covarrubias contributed significantly to writing the sequencing sections. Dianita Saputri, Sedat Aybars Nazlica, Jiaqi Xie and Martin de Jesus Loza Lopez contributed to the experimental epitope determination sections. John Rozewicki, Ana Davila and Jan Wilamowski wrote most of the in-house software sections. Floris J. van Eerden wrote the MD section. Zichang Xu wrote the BCR docking section. Daron M. Standley wrote the overall manuscript and coordinated the efforts of the other members.

Acknowledgements

We would like to thank all members of the Systems Immunology Lab for helpful discussions. This research was supported by JSPS KAKENHI Grant Numbers 18H02430, 20K06610, 20K16286, 20K07538 and by the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number 17am0101108j0001.

References

[1] Murphy K et al. Janeway's immunobiology. New York: Garland Science; 2008.
[2] Mora T, Walczak AM. How many different clonotypes do immune repertoires contain?. Curr Opin Syst Biol 2019;18:104–10.
[3] Turner SJ et al. Structural determinants of T-cell receptor bias in immunity. Nat Rev Immunol 2006;6(12):883–94.
[4] Reinhardt RL, Liang HE, Locksley RM. Cytokine-secreting follicular T cells shape the antibody repertoire. Nat Immunol 2009;10(4):385–93.
[5] Bagaev DV et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Res 2020;48(D1):D1057–62.
[6] Miqueu P et al. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. Mol Immunol 2007;44 (6):1057–64.
[7] Calis JJ, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol 2014;35 (12):581–90.
[8] Hou XL et al. Current status and recent advances of next generation sequencing techniques in immunological repertoire. Genes Immun 2016;17 (3):153–64.
[9] Brochet X, Lefranc MP, Giudicelli V, IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res;2008:36(Web Server issue): p. W503–8.
[10] Ralph DK, Matsen FAT, Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. PLoS Comput Biol;2019:15(7): e1007133.
[11] Alamyar E et al. IMGT((R)) 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 2012;882:569–604.
[12] Li S et al. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun 2013;4:2333.
[13] Ye J, et al., IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res;2013: 41(Web Server issue): W34–40.
[14] Altschul SF et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–402.
[15] Bolotin DA et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods 2015;12(5):380–1.
[16] Bolotin DA et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol 2017;35(10):908–11.
[17] Smakaj E et al. Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. Bioinformatics 2020;36(6):1731–9.
[18] Vander Heiden JA et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 2014;30(13):1930–2.
[19] Gupta NT et al. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 2015;31 (20):3356–8.

[20] Li B et al. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA-seq data. Nat Genet 2017;49(4):482–3.
[21] Shugay M et al. VDJtools: unifying post-analysis of T cell receptor repertoires. PLoS Comput Biol 2015;11(11):e1004503.
[22] Duez M et al. Vidjil: a web platform for analysis of high-throughput repertoire sequencing. PLoS ONE 2016;11(11):e0166126.
[23] Nazarov VI et al. tcR: an R package for T cell receptor repertoire advanced data analysis. BMC Bioinf 2015;16:175.
[24] Laydon DJ et al. Quantification of HTLV-1 clonality and TCR diversity. PLoS Comput Biol 2014;10(6):e1003646.
[25] Bolen CR et al. The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. BMC Bioinf 2017;18(1):155.
[26] Yokota R, Kaminaga Y, Kobayashi TJ. Quantification of inter-sample differences in T-cell receptor repertoires using sequence-based information. Front Immunol 2017;8:1500.
[27] Emerson RO et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet 2017;49(5):659–65.
[28] DeWitt WS, 3rd, et al., Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. Elife;2018:7.
[29] Sethna Z et al. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. Bioinformatics 2019;35(17):2974–81.
[30] Sethna Z, et al., Population variability in the generation and thymic selection of T-cell repertoires. bioRxiv, 2020: p. 2020.01.08.899682.
[31] Davidsen K et al. Deep generative models for T cell receptor protein sequences. Elife 2019:8.
[32] Pogorelyy MV et al. Detecting T cell receptors involved in immune responses from single repertoire snapshots. PLoS Biol 2019;17(6):e3000314.
[33] Murugan A et al. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. Proc Natl Acad Sci U S A 2012;109(40):16161–6.
[34] Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. Nat Commun 2018;9(1):561.
[35] Singh M et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun 2019;10(1):3120.
[36] Afik S et al. Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. Nucleic Acids Res 2017;45(16):e148.
[37] Stubbington MJT et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Methods 2016;13(4):329–32.
[38] Eltahla AA et al. Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. Immunol Cell Biol 2016;94(6):604–11.
[39] Canzar S et al. BASIC: BCR assembly from single cells. Bioinformatics 2017;33(3):425–7.
[40] Lindeman I et al. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat Methods 2018;15(8):563–5.
[41] Rizzetto S et al. B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. Bioinformatics 2018;34(16):2846–7.
[42] Setliff I, et al., High-throughput mapping of B cell receptor sequences to antigen specificity. Cell;2019:179(7):1636–1646 e15.
[43] Jurtz VI, et al., NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. bioRxiv;2018:433706.
[44] Kula T, et al., T-scan: a genome-wide method for the systematic discovery of T cell epitopes. Cell;2019:178(4):1016–1028 e13.
[45] Gee MH, et al., Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes. Cell;2018:172(3): p. 549–563 e16.
[46] Kobayashi E et al. A new cloning and expression system yields and validates TCRs from blood lymphocytes of patients with cancer within 10 days. Nat Med 2013;19(11):1542–6.
[47] Marks C, Deane CM. Antibody H3 Structure Prediction. Comput Struct Biotechnol J 2017;15:222–31.
[48] Marcatili P, Rosi A, Tramontano A. PIGS: automatic prediction of antibody structures. Bioinformatics 2008;24(17):1953–4.
[49] Sircar A, Kim ET, Gray JJ, RosettaAntibody: antibody variable region homology modeling server. Nucleic Acids Res, 2009. 37(Web Server issue): p. W474-9.
[50] Nishigami H, Kamiya N, Nakamura H. Revisiting antibody modeling assessment for CDR-H3 loop. Protein Eng Des Sel 2016;29(11):477–84.
[51] Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods 2009;6(8):551–2.
[52] Almagro JC et al. Second antibody modeling assessment (AMA-II). Proteins 2014;82(8):1553–62.
[53] Shirai H et al. High-resolution modeling of antibody structures by a combination of bioinformatics, expert knowledge, and molecular simulations. Proteins 2014;82(8):1624–35.
[54] Yamashita K et al. Kotai Antibody Builder: automated high-resolution structural modeling of antibodies. Bioinformatics 2014;30(22):3279–80.
[55] Schritt D et al. Repertoire Builder: High-throughput structural modeling of B and T cell receptors. Mol Syst Des Eng 2019;4:761–8.
[56] Leem J et al. ABodyBuilder: Automated antibody structure prediction with data-driven accuracy estimation. MAbs 2016;8(7):1259–68.
[57] Gowthaman R, Pierce BG. TCRmodel: high resolution modeling of T cell receptors from sequence. Nucleic Acids Res 2018;46(W1):W396–401.
[58] Lepore R et al. PIGSPro: prediction of immunoGlobulin structures v2. Nucleic Acids Res 2017;45(W1):W17–23.
[59] Rozewicki J et al. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res 2019;47(W1):W5–W10.
[60] Katoh K, Frith MC. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics 2012;28(23):3144–6.
[61] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26(19):2460–1.
[62] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9.
[63] Glanville J et al. Identifying specificity groups in the T cell receptor repertoire. Nature 2017;547(7661):94–8.
[64] Huang H et al. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol 2020.
[65] Dash P et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 2017;547(7661):89–93.
[66] DeWitt WS et al. A diverse lipid antigen-specific TCR repertoire is clonally expanded during active tuberculosis. J Immunol 2018;201(3):888–96.
[67] Scheid JF et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science 2011;333(6049):1633–7.
[68] Joyce MG et al. Vaccine-induced antibodies that Neutralize Group 1 and Group 2 influenza A viruses. Cell 2016;166(3):609–23.
[69] Robbiani DF et al. Convergent antibody responses to SARS-CoV-2 in convalescent individuals. Nature 2020.
[70] Kovaltsuk A et al. Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. PLoS Comput Biol 2020;16(2):e1007636.
[71] Xu Z, et al., Functional clustering of B cell receptors using sequence and structural features. Mol Syst Des Eng, 2019. in press.
[72] Miho E et al. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. Nat Commun 2019;10(1):1321.
[73] Joglekar AV, Li G. T cell antigen discovery. Nat Methods 2020.
[74] Fischer DS, et al., Predicting antigen-specificity of single T-cells based on TCR CDR3 regions. bioRxiv, 2019: p. 734053.
[75] Jokinen E, et al., TCRGP: Determining epitope specificity of T cell receptors. bioRxiv, 2019: p. 542332.
[76] Ogishi M, Yotsuyanagi H. Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. Front Immunol 2019;10:827.
[77] Gielis S et al. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. Front Immunol 2019;10:2820.
[78] Sun Y et al. Specificity, privacy, and degeneracy in the CD4 T Cell receptor repertoire following immunization. Front Immunol 2017;8:430.
[79] Thomas N et al. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. Bioinformatics 2014;30(22):3181–8.
[80] Lanzarotti E, Marcatili P, Nielsen M. T-cell receptor cognate target prediction based on paired alpha and beta chain sequence and structural CDR loop similarities. Front Immunol 2019;10:2080.
[81] Li S, et al., Structural modeling of lymphocyte receptors and their antigens. Meth Mol Biol, 2019. in press.
[82] Klausen MS et al. LYRA, a webserver for lymphocyte receptor structural modeling. Nucleic Acids Res 2015;43(W1):W349–55.
[83] Jensen KK et al. TCRpMHCmodels: structural modelling of TCR-pMHC class I complexes. Sci Rep 2019;9(1):14530.
[84] Kunik V, Ashkenazi S, Ofran Y, Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. Nucleic Acids Res, 2012. 40(Web Server issue): p. W521-4.
[85] Olimpieri PP et al. Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. Bioinformatics 2013;29(18):2285–91.
[86] Liberis E et al. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics 2018;34(17):2944–50.
[87] Daberdaku S, Ferrari C. Antibody interface prediction with 3D Zernike descriptors and SVM. Bioinformatics 2019;35(11):1870–6.
[88] Deac A, VeliCkovic P, Sormanni P. Attentive cross-modal paratope prediction. J Comput Biol 2019;26(6):536–45.
[89] Krawczyk K et al. Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. Protein Eng Des Sel 2013;26(10):621–9.
[90] Singh H, Ansari HR, Raghava GP. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS ONE 2013;8(5):e62216.
[91] Dalkas GA, Rooman M. SEPIa, a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence. BMC Bioinf 2017;18(1):95.
[92] Jespersen MC et al. BepiPred- 2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 2017;45(W1):W24–9.
[93] Zhao L et al. Novel overlapping subgraph clustering for the detection of antigen epitopes. Bioinformatics 2018;34(12):2061–8.
[94] Jespersen MC et al. Antibody specific B-cell epitope predictions: leveraging information from antibody-antigen protein complexes. Front Immunol 2019;10:298.
[95] Sela-Culang I et al. PEASE: predicting B-cell epitopes utilizing antibody sequence. Bioinformatics 2015;31(8):1313–5.

[96] Krawczyk K et al. Improving B-cell epitope prediction and its application to global antibody-antigen docking. Bioinformatics 2014;30(16):2288–94.

[97] Bourquard T et al. MAbTope: a method for improved epitope mapping. J Immunol 2018;201(10):3096–105.

[98] Janin J. Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 2010;6(12):2351–62.

[99] Kozakov D et al. The ClusPro web server for protein-protein docking. Nat Protoc 2017;12(2):255–78.

[100] Schneidman-Duhovny D, et al., PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res, 2005. 33(Web Server issue): p. W363-7.

[101] Ramirez-Aportela E, Lopez-Blanco JR, Chacon P, FRODOCK 2.0: fast protein-protein docking server. Bioinformatics, 2016. 32(15): p. 2386-8.

[102] Sircar A, Gray JJ, SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. PLoS Comput Biol, 2010. 6(1): p. e1000644.

[103] van Zundert GCP et al. The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. J Mol Biol 2016;428(4):720–5.

[104] Roel-Touris J, Bonvin A, Jimenez-Garcia B. LightDock goes information-driven. Bioinformatics 2020;36(3):950–2.

[105] Pierce BG et al. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 2014;30(12):1771–3.

[106] Vreven T et al. Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 2015;427(19):3031–41.

[107] Lensink MF, Velankar S, Wodak SJ, Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. Proteins, 2017. 85(3): p. 359-377.

[108] Ambrosetti F, et al., Modeling antibody-antigen complexes by information-driven docking. Structure 2020. 28(1): p. 119-129 e2.

[109] Anishchenko I, Kundrotas PJ, Vakser IA. Modeling complexes of modeled proteins. Proteins 2017;85(3):470–8.

[110] Norman RA et al. Computational approaches to therapeutic antibody design: established methods and emerging trends. Brief Bioinform 2019.

[111] Knapp B, Dunbar J, Deane CM. Large scale characterization of the LC13 TCR and HLA-B8 structural landscape in reaction to 172 altered peptide ligands: a molecular dynamics simulation study. PLoS Comput Biol 2014;10(8):e1003748.

[112] Reboul CF et al. Epitope flexibility and dynamic footprint revealed by molecular dynamics of a pMHC-TCR complex. PLoS Comput Biol 2012;8(3):e1002404.

[113] Dominguez JL, Knapp B. How peptide/MHC presence affects the dynamics of the LC13 T-cell receptor. Sci Rep 2019;9(1):2638.

[114] Wu P, et al., Mechano-regulation of peptide-MHC Class I conformations determines TCR antigen recognition. Mol Cell 2019;73(5): p. 1015-1027 e7.

[115] Fodor J et al. Previously hidden dynamics at the TCR-peptide-MHC interface revealed. J Immunol 2018;200(12):4134–45.

[116] Friess MD, Pluhackova K, Bockmann RA. Structural model of the mIgM B-cell receptor transmembrane domain from self-association molecular dynamics simulations. Front Immunol 2018;9:2947.

[117] Tay MZ et al. The trinity of COVID-19: immunity, inflammation and intervention. Nat Rev Immunol 2020.

[118] Cao Y et al. Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. Cell 2020.

[119] Avram O et al. ASAP – A webserver for immunoglobulin-sequencing analysis pipeline. Front Immunol 2018;9:1686.

[120] H, IJ, et al., Antigen receptor galaxy: A user-friendly, web-based tool for analysis and visualization of T and B cell receptor repertoire data. J Immunol 2017;198(10): p. 4156–4165.

[121] Bischof J, Ibrahim SM. bcRep: R package for comprehensive analysis of B cell receptor repertoire data. PLoS ONE 2016;11(8):e0161569.

[122] Olson BJ et al. sumrep: a summary statistic framework for immune receptor repertoire comparison and model validation. Front Immunol 2019;10:2533.

[123] Weitzner BD et al. Modeling and docking of antibody structures with Rosetta. Nat Protoc 2017;12(2):401–16.

[124] Mashiach E, et al., FireDock: a web server for fast interaction refinement in molecular docking. Nucleic Acids Res 2008;36(Web Server issue): p. W229–32.

[125] Torchala M et al. SwarmDock: a server for flexible protein-protein docking. Bioinformatics 2013;29(6):807–9.

[126] Jiménez-García B, Pons C, Fernández-Recio J. pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics 2013;29(13):1698–9.

[127] Yan Y et al. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res 2017;45(W1):W365–73.

[128] Macindoe G, et al., HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 2010;38(Web Server issue): p. W445–9.

[129] de Vries SJ et al. A web interface for easy flexible protein-protein docking with ATTRACT. Biophys J 2015;108(3):462–5.

[130] Tovchigrechko A, Vakser IA, GRAMM-X public web server for protein-protein docking. Nucleic Acids Res, 2006. 34(Web Server issue): p. W310–4.

[131] Ambrosetti F et al. *proABC-2: PRediction Of AntiBody Contacts v2 and its application to information-driven docking*. bioRxiv 2020. https://doi.org/10.1101/2020.03.18.967828.

[132] Pittala S, Bailey-Kellogg C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. Bioinformatics 2020;36(13):3996–4003.