Ab-Ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope

Wing Ki Wong[1], Sarah A. Robinson[1], Alexander Bujotzek[2], Guy Georges[2], Alan P. Lewis[3], Jiye Shi[4], James Snowden[4], Bruck Taddese[5], and Charlotte M. Deane[1,*]

[1]Department of Statistics, University of Oxford, Oxford, UK
[2]Roche Pharma Research and Early Development, Large Molecule Research, Roche Innovation Center Munich, Penzberg, Germany
[3]Data and Computational Sciences, GlaxoSmithKline Research and Development, Gunnels Wood Road, Stevenage, UK
[4]UCB Pharma, 216 Bath Road, Slough, UK
[5]Discovery Sciences, BioPharmaceuticals, R&D, AstraZeneca, Cambridge, UK
[*]To whom correspondence should be addressed.

Abstract

Motivation: Solving the structure of an antibody-antigen complex gives atomic level information of the interactions between an antibody and its antigen, but such structures are expensive and hard to obtain. Alternative experimental sources include epitope mapping and binning experiments which can be used as a surrogate to identify key interacting residues. However, their resolution is usually not sufficient to identify if two antibodies have identical interactions. Computational approaches to this problem have so far been based on the premise that antibodies with similar sequences behave similarly. Such approaches will fail to identify sequence-distant antibodies that target the same epitope.

Results: We present Ab-Ligity, a structure-based similarity measure tailored to antibody-antigen interfaces. Using predicted paratopes on model antibody structures, we assessed our ability to identify those antibodies that target highly similar epitopes. Most antibodies adopting similar binding modes can be identified from sequence similarity alone, using methods such as clonotyping. In the challenging subset where the antibody sequences differ significantly, Ab-Ligity is still able to predict antibodies that would bind to highly similar epitopes (area under the precision-recall curve of 0.86). We compared Ab-Ligity's performance to an existing tool InterComp, and showed improved performance alongside a significant speed-up. These results suggest that Ab-Ligity will allow the identification of diverse (sequence-dissimilar) antibodies that bind to the same epitopes from large datasets such as immune repertoires.

Availability: http://opig.stats.ox.ac.uk/resources

Contact: deane@stats.ox.ac.uk

Supplementary information: Supplementary Materials are available at *Biorxiv*.

1 Introduction

Antibodies are immune proteins that have high specificity and affinity against their target antigens. Their target specificity is mainly determined by the intermolecular interactions at the

antibody-antigen interface. The types of interactions used in antibody-antigen binding are known to be distinct from those observed in general protein-protein interactions [1].

The highest resolution method for studying antibody-antigen binding configurations is co-crystal complex structures. These give atomic level information but are expensive and difficult to obtain [2]. Experimental mapping is often used as a surrogate as it is able to identify the binding regions of the antigen ("epitopes") and antibody ("paratopes"; [3]). Competition assays exploit the cross-blocking effect of antibodies that displace one another if they bind to similar or neighbouring epitopes [4, 5]. This method gives a coarse representation of which binders may share similar target sites, as minimal epitope overlap is sufficient for a pair of antibodies to compete with each other [5]. A more refined approach is hydrogen deuterium exchange (HDX). HDX assesses the solvent accessibility of the bound and unbound forms of the partner proteins, and highlights regions with the maximum changes upon binding (*e.g.* [6, 7]). The resolution is typically up to the range of peptides. To achieve residue-level resolution, point mutations of the interacting proteins can be used to indicate the key binding residues. Mutagenesis studies measure the binding kinetics upon mutation of specific residues, but structural integrity may be compromised by the mutations, leading to spurious results [3]. All three of these experimental techniques provide an approximation of the binding regions, but are usually unable to provide a fine mapping of exact epitopes and paratopes.

Computational techniques have also been developed to identify antibodies that bind in similar ways. These have generally been exploited on large immunoglobulin sequencing datasets (*e.g.* [8]). Some of these techniques require a large number of known binders to a given epitope, to be able to identify further binders [9]. The informatics approaches employed to analyse datasets, when none or only a few binders to a given epitope are known, are mainly dependent on sequence similarity, based on the concept of "clonotype" analysis [8], which considers the genotype and the sequence identity of CDRH3 (the third complementarity determining region on the heavy chain) [8, 10, 11]. Clonotyping exploits the evolutionary origins of antibodies, using both the concept that antibodies from the same clone and lineage tend to bind similarly [12], and that the CDRH3 region, the most sequence-variable region, is often responsible for much of the binding [13]. Whilst clonotype analysis can identify antibodies with similar binding modes, there are many cases where binding remains the same even when the CDRH3 sequences and/or the genotypes are different (*e.g.* in anti-lysozyme antibodies).

To capture potential chemical interactions and the antibody-antigen binding configurations, macromolecular docking has been used [14]. Some studies have also presented the possibility of enriching computational modelling with experimental information (*e.g.* [15]). However, this process is slow and not scalable to the large datasets of antibody sequences available in the early discovery stage [16].

In small molecule discovery, comparing the spatial arrangements of pharmacophores (the atom features involved in interactions) has been proposed as an alternative to docking when searching for similar binders. Pharmacophoric points are encoded using geometric hashing algorithms (*e.g.* [17, 18, 19]). This approach is much faster whilst giving comparable results to docking [19]. Some of the descriptors for ligand-binding pockets have been adapted for protein-protein interactions. I2ISiteEngine uses atom type triangulation to describe protein-protein interfaces and successfully identifies similar interfaces across different protein families in a number of case studies [20]. However, the all-atom description means it is far slower than its counterpart developed for the comparison of small molecules [17]. Recently, a number of tools have been proposed for protein-protein interface comparisons. MaSIF compares molecular surfaces of protein-protein interfaces using geometric deep learning [21]. Whilst this technique has shown success using bound structures to search for similar binders, the performance did not

hold on the more challenging unbound set. To apply this tool in early-stage antibody discovery, MaSIF's high sensitivity to the precise location and chemical properties of the atoms suggests that it would depend heavily on antibody model accuracy. InterComp has been proposed for protein-protein interface comparison. It aligns residues in the surfaces for comparison and scores their similarity using a combination of the Cα atom distances and their BLOSUM substitution score [22]. The approach successfully identified similar surface patches but the algorithm is computationally intensive.

In this manuscript, we describe Ab-Ligity, an antibody version of the small molecule method Ligity [19]. Ab-Ligity uses residue points tokenised by their physicochemical properties to allow rapid operation. To simulate a real-life scenario, we applied Ab-Ligity to antibody models and predicted paratopes, and showed that it can accurately predict antibodies that share similar target epitopes. As the majority of the similar binders share similar sequence composition and lengths, and can be identified by sequence-based metrics, we considered only cases that could not be identified by such sequence-based metrics. In these more challenging cases of sequence-dissimilar antibodies, Ab-Ligity still accurately predicts antibodies that have similar binding modes. Ab-Ligity also shows better performance than InterComp, in a fraction of the time. Finally we highlight two case studies where Ab-Ligity predicts sequence-dissimilar and length-mismatched antibodies that share highly similar epitopes.

2 Methods

2.1 Antibody-antigen co-crystal datasets

We selected all paired antibody-antigen complexes from SAbDab [2] as of 27th January, 2020, that were solved by X-ray crystallography, had no missing residues in all six CDRs (using the North definition; [13]), and were co-crystallised with a protein antigen of more than 50 residues. To avoid redundancy, we only retain one copy of each antibody. In the case of multiple copies, the complex with the best resolution is selected. Nine hundred and twenty two antibody-antigen complexes were identified.

We defined epitopes as residues on the antigen with any atoms within 4.5Å of its cognate antibody. We extracted these residues using Biopython [23].

2.2 Antibody modelling and paratope prediction

We numbered antibody sequences using the IMGT scheme [24] and defined the CDR regions by the North scheme [13], with ANARCI [25]. We then modelled the full set using ABody-Builder [26], barring it from using sequence-identical structures as templates. The ABody-Builder template library was built using all structures available on the 27th January, 2020, in SAbDab [2].

We used Parapred to predict the paratope residues [27]. Parapred gives a score to each residue in the CDRs, and two residues before and two after, to indicate how likely it will participate in binding. As suggested in the original paper, we selected residues with a score of ≥ 0.67 to be the predicted paratope residues on the models (see Supplementary Materials for Parapred prediction performance). The coordinates of these predicted paratope residues were obtained from the corresponding model. Models and predicted paratopes were used throughout the manuscript for calculating the structural similarity between antibodies.

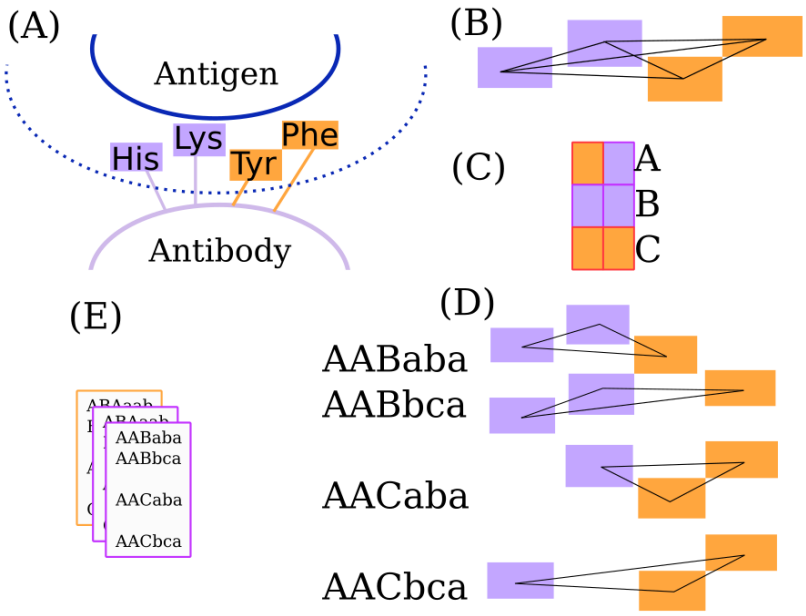

Figure 1: The Ab-Ligity workflow. (A) Binding site residues within 4.5Å of the binding partner are tokenised as stated in Table 1. (B) All distances between tokenised points are calculated and hashed into 1.5Å-wide distance bins for the paratopes and 1.0Å-wide distance bins for epitopes. (C) Each pair of tokens is given a unique hash code. (D) A six-character hash code is generated for each triplet. (E) The hashes of a binding site are stored in a frequency table.

Table 1: Residue groupings for tokenisation.

Group	Residues
Small	Glycine (G), Alanine (A)
Nucleophilic	Serine (S), Threonine (T), Cysteine (C)
Hydrophobic	Valine (V), Leucine (L), Isoleucine (I), Methionine (M), Proline (P)
Aromatic	Phenylalanine (F), Tyrosine (Y), Tryptophan (W)
Acidic	Aspartic acid (D), Glutamic acid (E)
Amine	Asparagine (N), Glutamine (Q)
Basic	Histidine (H), Lysine (K), Arginine (R)

2.3 Ab-Ligity calculations

The workflow of Ab-Ligity is illustrated in Figure 1. The binding residues are tokenised according to Table 1. This tokenisation scheme was chosen as it is simple and intuitive, and gave similar results to several other more complex choices (results not shown). For each binding site residue, a point is placed representing the residue group on the C-α atom of the residue. These points collectively represent a paratope on an antibody or an epitope on an antigen.

Ab-Ligity uses the hashing function outlined in [19]. In short, it considers all combinations of triplets formed from a set of tokenised residues in a binding site. In a given triplet, each edge is represented by its vertices' tokens and its length. Each combination of tokens has a unique hash

code. Edge lengths are discretised into bins of 1.5Å in paratopes and 1.0Å in epitopes to reduce computational complexity. The final hash code for a given triplet is determined by the three vertex hash codes sorted alphabetically, followed by the hash codes of the three corresponding length bins. The resulting hash table for the target binding site stores the frequency of these hash codes from all triplets.

Binding site similarity is calculated by the Tversky index of the pairs of hash tables:

$$S(X,Y) = \frac{|X \cap Y|}{|X \cup Y| + \alpha|X - Y| + \beta|Y - X|} \tag{1}$$

where X and Y are the two hash tables, $|X \cap Y|$ and $|X \cup Y|$ are the intersection and union, and $|X - Y|$ and $|Y - X|$ are the differences between the two tables respectively. For this study we used $\alpha = \beta = 0.5$.

2.4 Selecting an epitope similarity threshold

For each pair of antibody-antigen complexes, we calculated the epitope similarity based on the crystal information. We selected the epitope similarity threshold by evaluating pairs of proposed crystal paratope and epitope similarity thresholds (see Supplementary Materials) and chose the epitope similarity score that gives the best classification performance, *i.e.* the highest Matthews correlation coefficient (MCC; see Supplementary Materials). This becomes the ground truth for evaluation. For Ab-Ligity, we selected the crystal epitope similarity threshold to be 0.2.

2.5 Performance evaluation settings

We calculated the Ab-Ligity paratope similarity based on antibody models and their predicted paratopes. The majority of the antibodies that engage the same epitope have highly similar CDRH3 sequences and would be predicted by to bind in the same manner by sequence comparison (clonotype). To test if Ab-Ligity makes accurate predictions for those cases without similar CDRH3s, we assessed the performance on both the "full set", and a subset where the CDRH3 identity is ≤ 0.8 ("CDRH3$\leq$0.8 set").

We evaluated the performance of Ab-Ligity on these two sets. The number of positive and negative cases for these combinations are listed in Table 2.

The classification performance is reported as the area under the precision-recall curve (AUPRC). An AUPRC of 1.0 indicates a perfect classifier; whereas a random classifier would have an AUPRC of: (Number of positives)/(Total number of comparisons).

2.6 Benchmark

We compared our prediction performance and computational time with InterComp, a surface comparison tool for protein-protein interfaces [22]. We selected a qualitatively equivalent epitope similarity threshold using the same method outlined earlier, and evaluated the performance of both methods based on Ab-Ligity's definitions of similar crystal epitopes. (See Supplementary Materials for the corresponding analysis using InterComp's definition.) The algorithm run-time was measured on a single 3.40GHz i7-6700 CPU core.

Table 2: Number of positive and negative comparisons in the datasets, based on Ab-Ligity's definition of similar epitopes.

Set	Positive	Negative
Full	572	29,839
CDRH3≤0.8	190	29,730

Table 3: Area under the precision-recall curve (AUPRC) on the full and CDRH3≤0.8 sets, using Ab-Ligity and InterComp.

Set	Ab-Ligity	InterComp
Full	0.84	0.85
CDRH3≤0.8	0.86	0.81

3 Results

To test the power of Ab-Ligity to identify antibodies that bind to the same epitope, we built models of 922 unique protein-binding antibodies and predicted their paratope residues with Parapred [27]. Using these models and predicted paratopes we tested if Ab-Ligity was able to identify antibodies that bind to the same epitopes (as defined by the crystal complexes). Since the majority of the similar binders have the same CDRH3 length and highly identical CDRH3 sequences, we also tested Ab-Ligity performance when we removed these "easy" comparisons (CDRH3 identity >0.8). The number of comparisons for each of these tests are summarised in Table 2.

3.1 Using Ab-Ligity to predict antibodies that bind to highly similar epitopes

We tested whether Ab-Ligity could recover antibody pairs that target highly similar epitopes from the full set (Table 3). We found that Ab-Ligity is able to identify the majority of the similar binders with good accuracy (AUPRC of 0.84).

Table 2 shows that 33.2% (190/572) of the pairs that bind to similar epitopes have CDRH3 sequence identities ≤0.8, and would not be identified by sequence-based methods such as clonotype. On this subset of cases, Ab-Ligity's performance remained high (AUPRC of 0.86; Table 3). These results demonstrate that Ab-Ligity is able to accurately identify sequence-dissimilar antibodies that have similar binding modes.

3.2 Sensitivity analysis on Parapred predictions

As Ab-Ligity uses predicted paratopes, we examined how their accuracy influences Ab-Ligity performance.

The Parapred prediction threshold of 0.67 was selected in the original paper such that the size of the predicted paratopes replicates that of the actual paratopes. At this threshold, the precision and recall were 0.67 and 0.70. We increased the threshold to 0.80 (increasing precision and reducing the number of residues predicted) to investigate the effect on Ab-Ligity. The

performance of Ab-Ligity was found to remain similar (AUPRCs of 0.85 in the full set and 0.83 in the CDRH3≤0.8 set; see Supplementary Materials).

3.3 Selecting similarity thresholds

At a model paratope similarity threshold of 0.2, Ab-Ligity gives a precision of 0.73 and a recall of 0.81 (see Supplementary Materials).

3.4 Comparing Ab-Ligity to InterComp

We compared Ab-Ligity to an existing protein-protein interface comparison tool, InterComp. We ran the same evaluation as for Ab-Ligity. In Table 3, we show the results of Ab-Ligity and InterComp using our definition of similar crystal epitopes. In the Supplementary Materials, we give the corresponding results for the methods using InterComp's definition of similar epitopes. Table 3 shows that the two methods have comparable performance with Ab-Ligity showing improved performance on the more challenging CDRH3≤0.8 set.

As well as outperforming InterComp, Ab-Ligity is also significantly faster. On the full set of 922×922 pairwise comparisons, Ab-Ligity takes 0.5 CPU-minute to generate the 922 hash tables for all paratopes, and 18.5 minutes for an optimised all-against-all similarity calculation. InterComp does not require pre-processing but the all-against-all query takes 65.5 minutes.

It is now possible to model large portion of next generation sequencing datasets [28]. The speed-up by Ab-Ligity would allow rapid comparison of binding sites. For example, Ab-Ligity would take one day to process on 150 cores a next generation dataset of 100,000 antibodies, compared to five days for InterComp.

3.5 Anti-lysozyme antibodies with dissimilar CDRH3 sequences against highly similar epitopes

To show the power of Ab-Ligity to predict similar binding of antibodies with dissimilar CDRH3 sequences, we examined three anti-lysosyzme antibodies, HyHEL-26, HyHEL-10 L-Y50F mutant and HyHEL-63 (annotated as 1NDM_BA_C, 1J1O_HL_Y, and 1NBY_BA_C respectively in Figure 2). These antibodies are known to all adopt the same binding mode against lysozyme (Figure 2A). Their CDRH3 loops are seven-residue long, but differ by at least three residues between any pair, so would not be considered to share the same clonotype (Figure 2B).

We used Ab-Ligity to calculate the similarity of their binding using models of predicted paratopes and find that it would predict all three to bind to the same epitope (Figure 2B).

3.6 CDRH3 sequences with different lengths engage the same epitope in HIV core gp120

Ab-Ligity may also be useful in cases where antibodies have different CDRH3 lengths but similar binding modes. Sequence-based metrics are not applied to such cases. One example of mismatched CDRH3 lengths binding the same epitope are two HIV-1 neutralizing antibodies, VRC03 (PDB code and chain ID 3SE8_HL) and VRC06 (4JB9_HL). They target the gp120 core with highly similar binding modes (Figure 3). Their heavy and light chain germlines are the

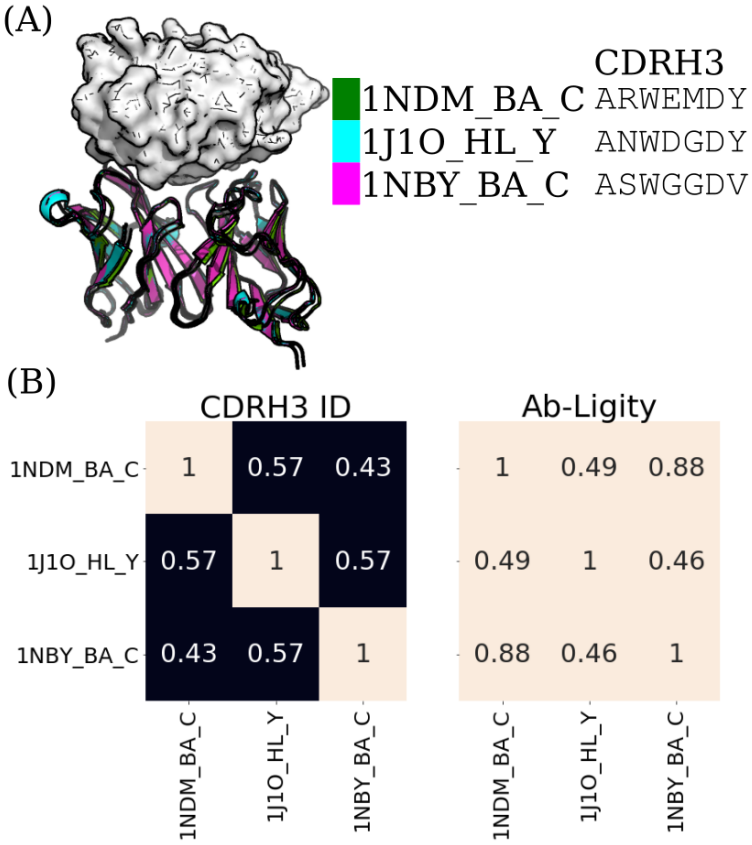

Figure 2: Analysis of anti-lysozyme antibodies with dissimilar CDRH3 sequences and highly similar epitopes. (A) Structural superposition of three anti-lysozyme antibodies co-crystallised with their antigens in white. The antigens from the three crystal structures are aligned. The legend shows the colours of the antibodies with their PDB codes followed by the heavy-light chain and antigen chain identifiers. The CDRH3 sequences are displayed next to the respective antibody identifiers. (B) Heatmaps of CDRH3 sequence identity and Ab-Ligity paratope similarity. The row and column labels correspond to the structures shown in (A). Ab-Ligity paratope similarity is calculated on the antibody model and predicted paratope as outlined in the manuscript. Pairs of antibodies with CDRH3 identity of >0.80 would have been considered similar by clonotype. For Ab-Ligity, a similarity score of >0.20 suggests that the antibodies bind to highly similar epitopes.

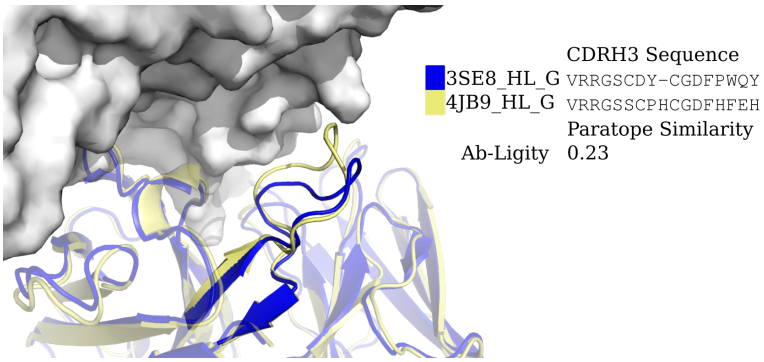

Figure 3: Analysis of two anti-gp120 antibodies. VRC03 antibody (PDB code and chain IDs: 3SE8_HL_G) is coloured in blue; VRC06 (PDB code and chain IDs: 4JB9_HL_G) is coloured in yellow. The gp120 core antigen is displayed as white surfaces and the superimposed antibodies are in cartoons. The CDRH3 loops are in solid shades of the cartoon representation. The PDB code, heavy and light chain ID, and antigen chain ID are separated by ('_') and listed in the legend with the corresponding CDRH3 sequences. CDRH3 sequences are shown by aligning their IMGT positions and '-' indicates a gap in the alignment according to the IMGT numbering scheme [24]. The Ab-Ligity paratope similarity score for the pair is listed.

same but their CDRH3 sequences differ in length and composition. Aligning by the IMGT-numbered positions [24], only 9 out of the 17 residues are aligned and identical. These two antibodies are known to engage a similar epitope [29] and Ab-Ligity based on models of these antibodies and predicted paratopes correctly predicts this (Ab-Ligity score of 0.23).

4 Discussion

We have presented Ab-Ligity, an antibody-protein binding site structural similarity metric that can identify sequence-dissimilar antibodies that engage the same epitope. Our results show that Ab-Ligity is able to identify antibodies that bind to the same epitopes using model structures and predicted paratopes, even for pairs of sequence-dissimilar antibodies. We further show that Ab-Ligity is able to find antibodies that bind to the same epitope with dissimilar CDRH3 lengths, beyond the constraints of most sequence-based metrics. We also compared Ab-Ligity to InterComp, an existing protein surface similarity metric and found improved performance for the harder cases with dissimilar sequences with a far faster run-time.

Binding site comparison with tools such as Ab-Ligity, opens up an alternative way to search for binders with similar binding modes. Typically in antibody discovery, multiple diverse hits are desired to avoid developability issues in the downstream optimisation process [30]. Ab-Ligity has the ability to predict if antibodies share similar target epitopes, without being sequence-similar. This ability coupled with the fast run-time of Ab-Ligity makes it suitable for searching through large datasets of antibody sequences to find sets of sequence-diverse binders to the same epitope.

Acknowledgements

The authors thank Dr. Jean-Paul Ebejer for providing the original implementation of Ligity, and members at Roche and GlaxoSmithKline for helpful discussion.

Funding

This work was supported by funding from the Engineering and Physical Sciences Research Council and Medical Research Council [grant number EP/L016044/1].

Conflict of Interest: none declared.

References

[1] M. Wang, D. Zhu, J. Zhu, R. Nussinov, and B. Ma, "Local and global anatomy of antibody-protein antigen recognition," *Journal of Molecular Recognition*, vol. 31, no. 5, p. e2693, 2018.

[2] J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane, "SAbDab: the structural antibody database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1140–D1146, 2013.

[3] W. M. Abbott, M. M. Damschroder, and D. C. Lowe, "Current approaches to fine mapping of antigen–antibody interactions," *Immunology*, vol. 142, no. 4, pp. 526–535, 2014.

[4] J.-W. Kwak and C.-S. Yoon, "A convenient method for epitope competition analysis of two monoclonal antibodies for their antigen binding," *Journal of Immunological Methods*, vol. 191, no. 1, pp. 49–54, 1996.

[5] Y. N. Abdiche, A. Y. Yeung, I. Ni, D. Stone, A. Miles, W. Morishige, A. Rossi, and P. Strop, "Antibodies targeting closely adjacent or minimally overlapping epitopes can displace one another," *PloS One*, vol. 12, no. 1, p. e0169535, 2017.

[6] Q. Zhang, J. Yang, J. Bautista, A. Badithe, W. Olson, and Y. Liu, "Epitope mapping by HDX-MS elucidates the surface coverage of antigens associated with high blocking efficiency of antibodies to birch pollen allergen," *Analytical Chemistry*, vol. 90, no. 19, pp. 11315–11323, 2018.

[7] C. Puchades, B. Kűkrer, O. Diefenbach, E. Sneekes-Vriese, J. Juraszek, W. Koudstaal, and A. Apetri, "Epitope mapping of diverse influenza Hemagglutinin drug candidates using HDX-MS," *Scientific Reports*, vol. 9, no. 1, p. 4735, 2019.

[8] J. D. Galson, J. Trück, A. Fowler, M. Münz, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly, "In-depth assessment of within-individual and inter-individual variation in the b cell receptor repertoire," *Frontiers in Immunology*, vol. 6, p. 531, 2015.

[9] D. M. Mason, S. Friedensohn, C. R. Weber, C. Jordi, B. Wagner, S. Meng, P. Gainza, B. Correia, and S. T. Reddy, "Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space," *bioRxiv*, p. 617860, 2019.

[10] J. Trück, M. N. Ramasamy, J. D. Galson, R. Rance, J. Parkhill, G. Lunter, A. J. Pollard, and D. F. Kelly, "Identification of antigen-specific B cell receptor sequences using public repertoire analysis," *The Journal of Immunology*, vol. 194, no. 1, pp. 252–261, 2015.

[11] C. Soto, R. G. Bombardi, A. Branchizio, N. Kose, P. Matta, A. M. Sevy, R. S. Sinkovits, P. Gilchuk, J. A. Finn, and J. E. Crowe, "High frequency of shared clonotypes in human B cell receptor repertoires," *Nature*, vol. 566, no. 7744, p. 398, 2019.

[12] Y.-C. Hsiao, Y.-J. J. Chen, L. D. Goldstein, J. Wu, Z. Lin, K. Schneider, S. Chaudhuri, A. Antony, K. B. Pahuja, Z. Modrusan, D. Seshasayee, S. Seshagiri, and I. Htzel, "Restricted epitope specificity determined by variable region germline segment pairing in rodent antibody repertoires," *mAbs*, vol. 12, no. 1, p. 1722541, 2020.

[13] B. North, A. Lehmann, and R. L. Dunbrack Jr, "A new clustering of antibody CDR loop conformations," *Journal of Molecular Biology*, vol. 406, no. 2, pp. 228–256, 2011.

[14] M. F. Lensink, N. Nadzirin, S. Velankar, and S. J. Wodak, "Modeling protein-protein, protein-peptide and protein-oligosaccharide complexes: CAPRI 7th edition," *Proteins: Structure, Function, and Bioinformatics*, 2019.

[15] D. A. Cannon, L. Shan, Q. Du, L. Shirinian, K. W. Rickert, K. L. Rosenthal, M. Korade III, L. E. van Vlerken-Ysla, A. Buchanan, T. J. Vaughan, *et al.*, "Experimentally guided computational antibody affinity maturation with *de novo* docking modelling and rational design," *PLoS Computational Biology*, vol. 15, no. 5, p. e1006980, 2019.

[16] M. I. J. Raybould, W. K. Wong, and C. M. Deane, "Antibody-antigen complex modelling in the era of immunoglobulin repertoire sequencing," *Molecular Systems Design & Engineering*, vol. 4, no. 4, pp. 679–688, 2019.

[17] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, "SiteEngines: recognition and comparison of binding sites and protein–protein interfaces," *Nucleic Acids Research*, vol. 33, no. Suppl_2, pp. W337–W341, 2005.

[18] D. J. Wood, J. d. Vlieg, M. Wagener, and T. Ritschel, "Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement," *Journal of Chemical Information and Modeling*, vol. 52, no. 8, pp. 2031–2043, 2012.

[19] J.-P. Ebejer, P. W. Finn, W. K. Wong, C. M. Deane, and G. M. Morris, "Ligity: A non-superpositional, knowledge-based approach to virtual screening," *Journal of Chemical Information and Modeling*, vol. 59, no. 6, pp. 2600–2616, 2019.

[20] A. Shulman-Peleg, S. Mintz, R. Nussinov, and H. J. Wolfson, "Protein-protein interfaces: Recognition of similar spatial and chemical organizations," in *International Workshop on Algorithms in Bioinformatics*, pp. 194–205, Springer, 2004.

[21] P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia, "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, no. 2, pp. 184–192, 2020.

[22] C. Mirabello and B. Wallner, "Topology independent structural matching discovers novel templates for protein interfaces," *Bioinformatics*, vol. 34, no. 17, pp. i787–i794, 2018.

[23] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[24] M.-P. Lefranc, V. Giudicelli, P. Duroux, J. Jabado-Michaloud, G. Folch, S. Aouinti, E. Carillon, H. Duvergey, A. Houles, T. Paysan-Lafosse, S. Hadi-Saljoqi, S. Sasorith, G. Lefranc, and S. Kossida, "IMGT®, the international ImMunoGeneTics information system® 25 years on," *Nucleic Acids Research*, vol. 43, no. D1, pp. D413–D422, 2014.

[25] J. Dunbar and C. M. Deane, "ANARCI: antigen receptor numbering and receptor classification," *Bioinformatics*, vol. 32, no. 2, pp. 298–300, 2016.

[26] J. Leem, J. Dunbar, G. Georges, J. Shi, and C. M. Deane, "ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation," *mAbs*, vol. 8, no. 7, pp. 1259–1268, 2016.

[27] E. Liberis, P. Veličković, P. Sormanni, M. Vendruscolo, and P. Liò, "Parapred: antibody paratope prediction using convolutional and recurrent neural networks," *Bioinformatics*, vol. 34, no. 17, pp. 2944–2950, 2018.

[28] M. I. Raybould, C. Marks, A. Kovaltsuk, A. P. Lewis, J. Shi, and C. M. Deane, "Evidence of antibody repertoire functional convergence through public baseline and shared response structures," *BioRxiv*, 2020.

[29] Y. Li, S. O'Dell, R. Wilson, X. Wu, S. D. Schmidt, C.-M. Hogerkorp, M. K. Louder, N. S. Longo, C. Poulsen, J. Guenaga, B. K. Chakrabarti, N. Doria-Rose, M. Roederer, M. Connors, J. R. Mascola, and R. T. Wyatt, "HIV-1 neutralizing antibodies display dual recognition of the primary and coreceptor binding sites and preferential binding to fully cleaved envelope glycoproteins," *Journal of Virology*, vol. 86, no. 20, pp. 11231–11241, 2012.

[30] M. I. J. Raybould, C. Marks, K. Krawczyk, B. Taddese, J. Nowak, A. P. Lewis, A. Bujotzek, J. Shi, and C. M. Deane, "Five computational developability guidelines for therapeutic antibody profiling," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4025–4030, 2019.