XXIX Congresso de Iniciação Científica
UNICAMP
2021

An Exploratory Analysis of Linear Epitopes and Associated Antibody Complementarity-Determining Regions in the Immune Epitope Database

Henrique da Fonseca Simões[1, *] and João Meidanis[1, †]

[1]*Institute of Computing (IC) - University of Campinas (Unicamp)*

The Immune Epitope Database (IEDB) is a freely accessible database containing a significant part of the immune epitope assays described in relevant studies published over the last decades. Each epitope is classified as linear, discontinuous (possibly in multi-chain) or non-peptidic. In this work, we analyzed the subset of linear epitopes, including information about the antibody binding regions, the so-called complementarity-determining regions (CDRs). We found out that only a small percentage of assays contribute nonredundant linear epitopes with CDR3 chains, summing up to 485 entries. Furthermore, we explored some properties of these data, concluding that: (1) human and mouse hosts make up the vast majority of assays; (2) most epitopes range between 5 and 25 amino acids, whereas CDR light and heavy chain expected sizes are around 9–10 and 13–14 amino acids, respectively; (c) amino acid composition is far from uniform, with rare amino acids such as tryptophan being overrepresented, while more common ones such as lysine appearing underrepresented. Studies on other epitope-CDR datasets in the literature show similar conclusions.

Keywords: IEDB; antibody; machine learning

I. INTRODUCTION

Recently, increasing numbers of machine learning-based approaches are being used in attempts to solving different problems in bioinformatics, including immunology problems [1, 5, 8]. A machine learning algorithm is basically a system that reliably improves its performance at a particular task based on experience (*i.e.* data) [9]. Therefore, a key aspect to the success of these methods is the quality and representativeness of the data employed by them to recognize patterns and generate results.

With this in mind, we explored the subset of linear epitopes from the Immune Epitope Database (IEDB) [16], in preparatory analysis prior to using them in machine learning systems.

II. BACKGROUND

The mammalian immune system is vastly complex and comprises several defense mechanisms against potentially dangerous agents (**antigens**), which can be carbohydrates, proteins, and other molecules. The identification of these antigens usually occurs when an immune cell **receptor** protein binds to the antigen ligands, also called **epitopes**. When epitopes are proteins, they can be a continuous subsequence from the antigen's amino acid sequence, in which case they are refered as **linear epitopes**, or a discontinuous subsequence.

The first antigen-specific receptor discovered was the **antibody** (Ab), a Y-shaped proteic molecule from the immunoglobulin superfamily synthesized by **B-cells** consisting of two identical **light** (L) chains and two identical **heavy** (H) chains. Antibodies, as part of our adaptative immune system, have highly variable regions on the top of the Y, called **complementarity-determining regions** (CDRs), where the binding happens [11]. Among these, it is known that CDR3 is where most conformational variations happen [4].

Since the discovery of these mechanisms, a significant number of published studies have sequenced amino acids both from epitopes and antibodies (and their CDR3). In 2004, the National Institute of Allergy and Infections Diseases (NIAID) established the IEDB, with the goal of making all these experimentally determined immune epitopes freely available to the public [14].

By 2012, more than 95% of the publications in PubMed [10] were already curated by PhD specialists working on IEDB [15]. An overview of the number of new assays over the years can be seen in Figure 2. IEDB assay curation is not restricted by species, which means that any species might be represented the dataset as a host. In fact, more than a hundred species (strains) have at least one entry (Figure 1).

[*] henrique.simoes@students.ic.unicamp.br
[†] meidanis@unicamp.br

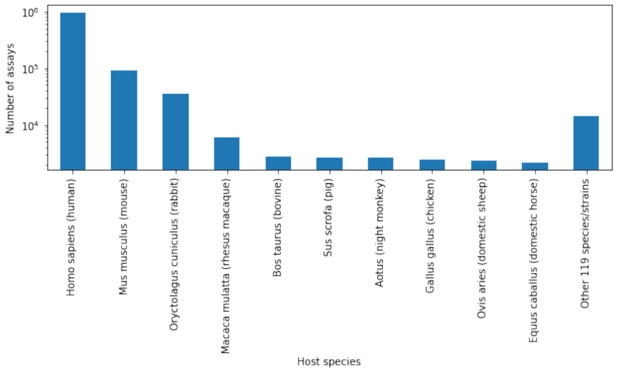


Figure 1: Most frequent host species in IEDB B-cell assays. The first 10 species had their strains grouped, while the others had not.

III. METHODOLOGY

In this analysis, we used the data avalaible in IEDB[1] on August 15, 2021. The complete dataset for B-cell experiments, which includes assay information, was obtained from the "database export" option in "more IEDB" menu. To retrieve linear epitope sequence data, we performed a search on the homepage by selecting *linear* epitopes from *B-cell* assays with *positive* outcome. Positivity is an important property, since negative results associates epitopes and CDRs that probably do not bind. An important distinction to be made is that a positive assays does not imply all amino acids in the sequence are interacting residues. As mentioned by Vita et al. [15], IEDB epitope sequences in many cases are not minimal, but match the exact sequences that were actually tested in the assays.

In a data review, we identified that linear epitopes retrieved by IEDB also contain discontinuous and even non-peptidic epitopes, making up about 3.1% and 4.0%, respectively, of the 1040 non-redundant epitopes. Therefore, we filtered downloaded linear epitopes even further, removing these cases. In the process, we removed blank spaces from the beginning and end of epitope sequences, and selected those containing only capital letters, using the regular expression `^[A-Z]$`. This excludes discontinuous epitopes, which are encoded as sequences of comma-separated codes composed of amino acids and their positions (*e.g.* `W126`), as well as non-peptidic sequences, since these usually contain lowercase letters[2] and numbers (*e.g.* 2,4-dinitrophenyl group). Due to modified epitopes being represented using a plus sign (`+`) in their sequences, they are also correctly disconsidered with the aforementioned approach.

Among entries encoded as discontinuous, we also found fully linear sequences. However, these had one or two amino acids only, and were not included in the linear subset.

On comparative analysis with discontinuous epitopes, we use a similar filtering process on all B-cell assay epitopes, but searching for sequences that match the amino-acid-and-position format, using the regular expression `^(?:[A-Z][0-9]+,\s*)*[A-Z][0-9]+$`.

In order to analyze amino acid composition, we followed Ofran et al. [12]. For each analyzed sequence type, we computed the amino acid distribution with respect to each sequence, and then averaged the results. The natural logarithm was applied to the ratio of the averaged values over Swiss-Prot amino acid frequency in order to obtain the representativeness. Swiss-Prot [2] amino acid statistics released[3] on June 21, 2021 were used for this normalization. Aiming at assessing the effect of sequence frequency in amino acid distribution, we evaluated both redundant and non-redundant sequences.

To process the data, we used `Python` 3.8.10, `pandas` [13] and `matplotlib` [3]. All scripts used to generate results shown herein are available at `https://github.com/henriquesimoes/xxix-pibic`.

IV. RESULTS

After filtering the data, we found 1128 entries for linear epitopes (560 of them unique), 22439 entries for discontinuous epitopes (3519 unique when not considering amino acid relative positions) and 613 triples (linear epitope, CDRL3, CDRH3), with 485 of them being unique. Figure 5 shows amino acid representativeness for epitopes and CDR3 with respect to proteins in general. The host species and sequence length distribution from assays of linear epitope with binding CDR3 are shown respectively in Figures 3 and 4.

V. DISCUSSION

Despite having a significant amount of data from the literature, and in a growing rate, IEDB non-redudant linear epitope data with information of binding antibodies is still scarce, corresponding to around 0.043% of all 1,125,362 assays. This will likely bring challenges to machine learning-based approaches, such as generalization issues due to low representativeness of possible binding patterns.

[1] `https://www.iedb.org`

[2] Since we are applying this filter only on data labeled as linear epitope, if there is any non-peptidic epitope described with uppercase letters only, there is no other way of knowing whether it is mislabeled other than reviewing its publication paper.

[3] `https://web.expasy.org/docs/relnotes/relstat.html`

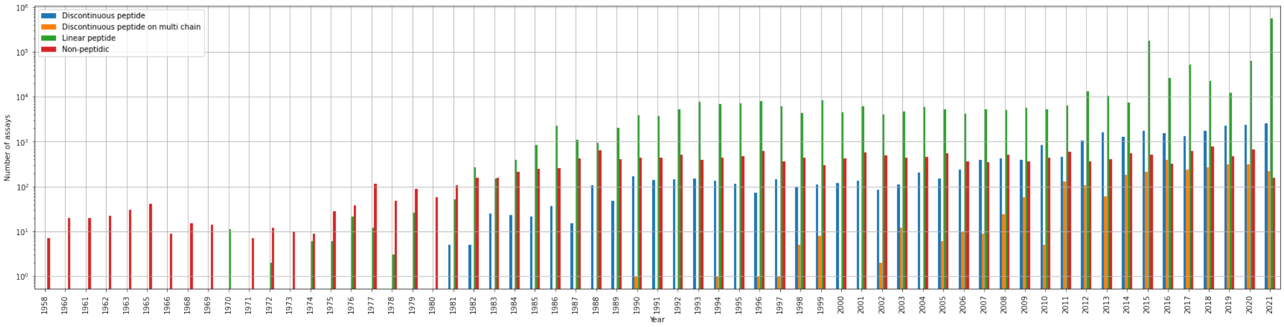


Figure 2: Number of new curated assays based on the epitope type included in IEDB over the years until August 15, 2021. A more precise description of each epitope type can be found in the IEDB Curation Manual, available at `http://curationwiki.iedb.org/wiki/index.php/Curation_Manual2.0#Epitope_Objects`.

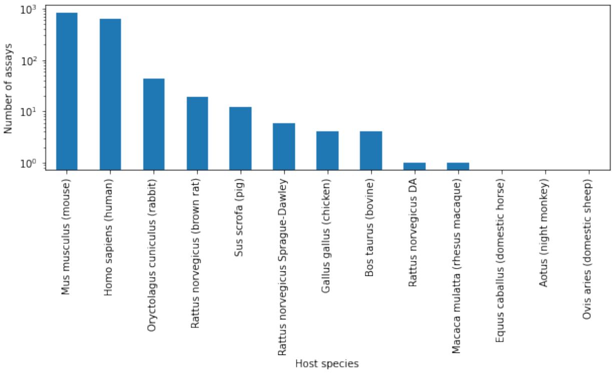


Figure 3: Host species from assays of linear epitope with associated CDR3 chains. Same species from Figure 1 have been grouped here.

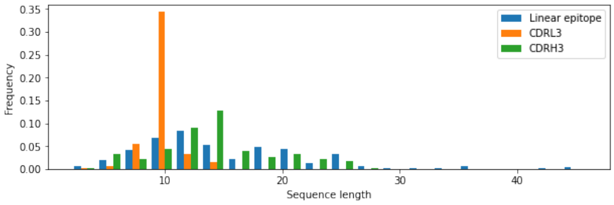


Figure 4: Length distribution for non-redundant linear epitopes and their associated antibody CDR3.

Furthermore, as it can be noticed from Figure 3, in this subset, there are remarkably more antibodies from humans and mice. This is expected, since clinical trials are usually conducted with these species. However, it also limits capturing possible patterns that may occur in other organisms and could be important, for instance, to unravel which mammalian antibody may bind to a given linear epitope.

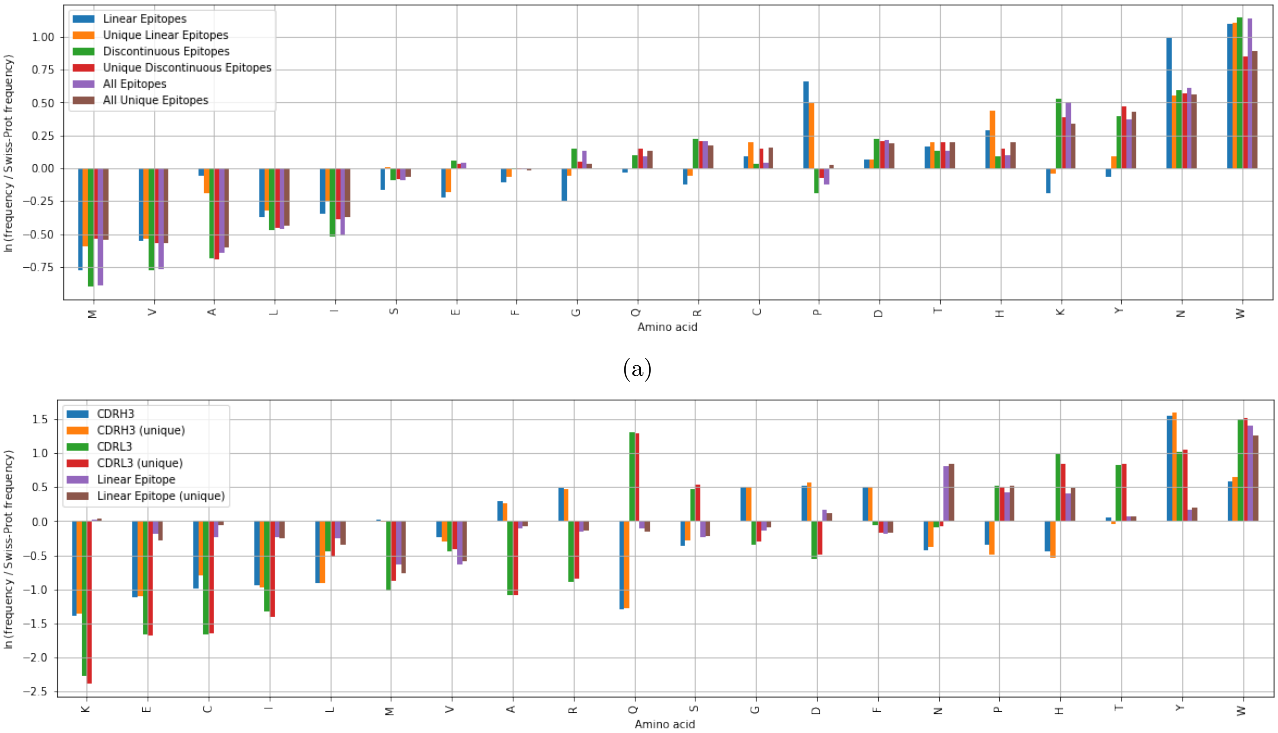
Similarly to Ofran et al. [12], we found that CDR3 amino acid representativeness varies (Figure 5b), with lysine (`K`) and glutamate (`E`) underrepresented while tryptophan (`W`) and tyrosine (`Y`) are overrepresented. However, glutamine (`Q`) has opposite representativeness in our data for light and heavy chains, while both are underrepresented in their work. This suggests that the selected antibody sequences have indeed these uncovered properties. Therefore, if an algorithm uses this subset, it might be able to correctly capture propensity of amino acids that have been also identified by Kringelum et al. [6].

Another interesting thing to note with respect to amino acid preference is that epitopes seem distinguishable from general proteins (Figure 5a). Nevertheless, Swiss-Prot naturally includes non-surface residues, which might introduce a bias, as discussed by Kringelum et al. [6]. To remove this bias, they evaluated sequence distribution by sampling from 12 bins based on surface exposure. However, this characteristic is relative to each sequence and is, as such, not available in Swiss-Prot statistics. Consequently, we did not take this into account.

Finally, CDR3 light and heavy chains have different expected length, with CDRL3 of 9–10 amino acids and CDRH3 13–14 being more prevalent. This is similar to properties identified by Kunik and Ofran [7] with respect to length distribution diversity and chain relative sizes.

VI. CONCLUSION

We explored the Immune Epitope Database and found 485 non-redundant sequences of linear epitopes and associated antibody CDR3. Still, they are mainly from two species: *Mus musculus* and *Homo sapiens*. This may be a bias to take into account when designing new systems. Moreover, CDR3 sequence properties seem to correspond to ones previously found, including sequence length and amino acid composition. This suggests that machine learning systems trained with the analyzed subset are likely to face important challenges, but may capture intrinsic properties from antibody complementarity-determining regions.



(a)

(b)

Figure 5: Amino acid representativeness based on Swiss-Prot amino acid distribution for (a) epitopes and (b) epitope-CDRH3-CDRL3 triples. Negative values mean underrepresentation and positive values mean overrepresentation with respect to the propensity in proteins in general. Uniqueness has been evaluated isolated for each epitope type (a), while a non-redundant entry for receptor triples (b) took into account epitope, CDRL3 and CDRH3 together.

ACKNOWLEDGMENTS

This work was supported[4] by the São Paulo Research Foundation (FAPESP) under grant no. 2020/11194-4. We also thank Unicamp's Student Support Service (SAE) for financially supporting H. F. Simões.

[1] R. Akbar, P. A. Robert, M. Pavlović, J. R. Jeliazkov, I. Snapkov, A. Slabodkin, C. R. Weber, L. Scheffer, E. Miho, I. H. Haff, D. T. T. Haug, F. Lund-Johansen, Y. Safonova, G. K. Sandve, and V. Greiff. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, 34(11):108856, 2021.

[2] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. M. Bairoch. UniProtKB/Swiss-Prot. *Methods in Molecular Biology*, 406:89–112, 2007.

[3] T. A. Caswell, M. Droettboom, A. Lee, E. S. de Andrade, J. Hunter, T. Hoffmann, E. Firing, J. Klymak, D. Stansby, N. Varoquaux, J. H. Nielsen, B. Root, R. May, P. Elson, J. K. Seppänen, D. Dale, J.-J. Lee, D. McDougall, A. Straw, P. Hobson, C. Gohlke, hannah, T. S. Yu, E. Ma, A. F. Vincent, S. Silvester, C. Moad, N. Kniazev, E. Ernest, and P. Ivanov. matplotlib/matplotlib: Rel: v3.4.2, May 2021. URL `https://doi.org/10.5281/zenodo.4743323`.

[4] C. Chothia and A. M. Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology*, 196(4):901–917, 1987. ISSN 0022-2836. doi:https://doi.org/10.1016/0022-2836(87)90412-8.

[4] Opinions, hypothesis and conclusions, or recommendations expressed in this material are responsibility of the authors, and do not necessarily reflect FAPESP's point of view.

[5] J. Graves, J. Byerly, E. Priego, N. Makkapati, S. V. Parish, B. Medellin, and M. Berrondo. A review of deep learning methods for antibodies. *Antibodies*, 9 (2):12, 2020. doi:10.3390/antib9020012.

[6] J. V. Kringelum, M. Nielsen, S. B. Padkjær, and O. Lund. Structural analysis of B-cell epitopes in antibody:protein complexes. *Molecular Immunology*, 53(1):24–34, 2013. ISSN 0161-5890. doi: https://doi.org/10.1016/j.molimm.2012.06.001.

[7] V. Kunik and Y. Ofran. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Engineering, Design and Selection*, 26(10):599–609, 06 2013. ISSN 1741-0126. doi: 10.1093/protein/gzt027.

[8] G. Liu, H. Zeng, J. Mueller, B. Carter, Z. Wang, J. Schilz, G. Horny, M. E. Birnbaum, S. Ewert, and D. K. Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36(7):2126–2133, 11 2019. ISSN 1367-4803. doi:10.1093/bioinformatics/btz895.

[9] T. M. Mitchell. Machine Learning. 1997.

[10] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1):D8–D13, 11 2017. ISSN 0305-1048. doi:10.1093/nar/gkx1095.

[11] L. B. Nicholson. The immune system. *Essays in biochemistry*, 60(3):275–301, 2016. doi: 10.1042/EBC20160017.

[12] Y. Ofran, A. Schlessinger, and B. Rost. Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *The Journal of Immunology*, 181(9): 6230–6235, 2008. doi:10.4049/jimmunol.181.9.6230.

[13] The pandas development team. pandas-dev/pandas: Pandas 1.2.4, Apr. 2021. URL `https://doi.org/10.5281/zenodo.4681666`.

[14] R. Vita, L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette, and B. Peters. The Immune Epitope Database 2.0. *Nucleic Acids Research*, 38(suppl_1):D854–D862, 2010.

[15] R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. The Immune Epitope Database (IEDB) 3.0. *Nucleic Acids Research*, 43(D1):D405–D412, 10 2014. doi: 10.1093/nar/gku938.

[16] R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47 (D1):D339–D343, 2019.