Patterns

CellPress
OPEN ACCESS

Preview

Artificial intelligence for antibody reading comprehension: AntiBERTa

Yoonjoo Choi[1,*]
[1]Combinatorial Tumor Immunotherapy MRC, Chonnam National University Medical School, Hwasun-gun, Jeollanam-do 58128, South Korea
*Correspondence: kalicuta@gmail.com
https://doi.org/10.1016/j.patter.2022.100535

Utilizing publicly available antibody big data resources, Leem et al. (2022) developed an antibody-specific language model, AntiBERTa, to understand the "language" of antibodies. Case studies reveal that AntiBERTa can be an extremely useful tool for antibody engineering.

There are a large number of players in the immune system to protect biological individuals from harmful foreign substances. Among those, the B cell is a main player in the adaptive immune system. B cells are equipped with receptor molecules (B cell receptor) that can be secreted upon activation. The secreted molecules, antibody, are known to be astronomically diverse (estimated 10^{13}–10^{15}).

The high diversity of the antibody is a two-faced Janus. The immune system can respond to nearly any type of antigen, mainly due to the large diversity of antibodies. According to Antibodypedia,[1] 4.6 million monoclonal antibodies are currently available for 19,000 genes. The diversity also enables antibodies to be highly successful as biotherapeutics. In 2021, FDA approved the 100th therapeutic antibody.[2] The coronavirus pandemic has been currently boosting the development of therapeutic antibodies for COVID-19, and several new antibodies are waiting to treat SARS-CoV-2-infected patients.

On the other hand, such rich diversity may not be always advantageous. Despite the fact that antibodies have been (perhaps the most) extensively studied and the antibody-related biopharmaceutical industry continues to mature, there seem to be a lot of things to learn about antibodies, as evidenced in the increasing growth of papers with the publication keyword, "antibody." It is simply practically impossible to experimentally explore the entire antibody repertoire. Thus, computational approaches using artificial intelligence (AI) techniques have become essential for antibody research.

The advancement of AI and big data are not separable. Recent advances in next-generation sequencing technology now enable the construction of a large volume of antibody repertoires. The observed antibody space (OAS) database[3,4] is a compilation of known repertoire studies and databases. Since the release of OAS, many practical applications have been developed including computational antibody humanization using AI.[5,6]

The antibody repertoire big data resources also provide an in-depth view and biological insights into antibodies.[7] Here, Leem et al. present an antibody-specific language model in a timely manner. AntiBERTa (antibody-specific bidirectional encoder representation from transformers) is a 12-layer transformer model pre-trained using the OAS database.[8] In fact, there has been a language model for general proteins[9] (ProtBERT), and there have been other antibody-specific language models, such as DeepAb[10] and Sapiens.[6] Comparing with those existing methods, however, AntiBERTa is more versatile and specific with deeper layers.

It is remarkable that AntiBERTa nicely partitions memory and naive B cells, whereas other models showed relatively less distinct results; i.e. the antibody-specific deep-layered model indeed learns the language of antibodies and finds the origin of B cell. One of the direct applications can be the estimation of antibody humanness and immunogenicity for the development of safer therapeutic antibodies. It is well known that antibodies with high human content tend to be less immunogenic. As demonstrated in the separation of memory and naive B cells, AntiBERTa is shown to be successful in classifying their species origin (murine, chimeric, humanized, and human).

The antibody-specific model generally provides better descriptions of antibodies than the general protein model. The authors found that residue pairs with high self-attention scores give structural insights into long-range interactions, which were not identified by the general protein model. The insight naturally leads to the prediction of paratopes, antigen binding sites. From several case studies, the authors showed that AntiBERTa successfully identifies paratope residues that are not in complementarity determining regions (CDR).

While the authors demonstrated that AntiBERTa outperforms other methods and provided convincing rationales, they also leave something to be desired. As the authors stated in the main manuscript, AntiBERTa can be directly applicable to antibody-structure prediction and humanization (or both at the same time), but the authors left it as potential applications. In the near future, we hope to meet practical application tools based on the AntiBERTa model.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (NRF-2020R1A5A2031185 and NRF-2020M3A9G3080281).

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

1. Björling, E., and Uhlén, M. (2008). Antibodypedia, a portal for sharing antibody and antigen validation data. Mol. Cell. Proteomics *7*, 2028–2037. https://doi.org/10.1074/mcp.m800264-mcp200.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2. Mullard, A. (2021). FDA approves 100th monoclonal antibody product. Nat. Rev. Drug Discov. *20*, 491–495. https://doi.org/10.1038/d41573-021-00079-7.

3. Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C.M., and Krawczyk, K. (2018). Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J. Immunol. *201*, 2502–2509. https://doi.org/10.4049/jimmunol.1800708.

4. Olsen, T.H., Boyles, F., and Deane, C.M. (2022). Observed Antibody Space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. *31*, 141–146. https://doi.org/10.1002/pro.4205.

5. Marks, C., Hummer, A.M., Chin, M., and Deane, C.M. (2021). Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics *37*, 4041–4047. https://doi.org/10.1093/bioinformatics/btab434.

6. Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., and Bitton, D.A. (2022). BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. mAbs *14*, 2020203. https://doi.org/10.1080/19420862.2021.2020203.

7. Marks, C., and Deane, C.M. (2020). How repertoire data are changing antibody science. J. Biol. Chem. *295*, 9823–9837. https://doi.org/10.1074/jbc.rev120.010181.

8. Leem, J., Mitchell, L.S., Farmery, J.H., Barton, J., and Galson, J.D. (2022). Deciphering the Language of Antibodies Using Self-Supervised Learning. Patterns *3*, 100513.

9. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. IEEE Trans. Pattern Anal. Mach. Intell. 1. https://doi.org/10.1109/TPAMI.2021.3095381.

10. Ruffolo, J.A., Sulam, J., and Gray, J.J. (2022). Antibody structure prediction using interpretable deep learning. Patterns *3*, 100406. https://doi.org/10.1016/j.patter.2021.100406.