AbRSA: A robust tool for antibody numbering

Lei Li[1§], Shuang Chen[2§], Zhichao Miao[3,4§], Yang Liu[1], Xu Liu[1], ZhiXiong Xiao[1] and Yang Cao[1,5*]

[1] Center of Growth, Metabolism and Aging, Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, PR China.

[2] Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, 610041, PR China.

[3] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[4] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[5] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109-2218, USA

[§]These authors contributed equally in this study.

*To whom correspondence should be addressed.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pro.3633

ABSTRACT

The remarkable progress in cancer immunotherapy in recent years has led to the heat of great development for therapeutic antibodies. Antibody numbering, which standardizes a residue index at each position of an antibody variable domain, is an important step in immunoinformatic analysis. It provides an equivalent index for the comparison of sequences or structures, which is particularly valuable for antibody modeling and engineering. However, due to the extremely high diversity of antibody sequences, antibody numbering tools cannot work in all cases. This article introduces a new antibody numbering tool named AbRSA, which integrates heuristic knowledge of region-specific features into sequence mapping to enhance the robustness. The benchmarks demonstrate that, AbRSA exhibits robust performance in numbering sequences with diverse lengths and patterns compared with the state-of-the-art tools. AbRSA offers a user-friendly interface for antibody numbering, complementarity-determining region (CDR) delimitation and 3D structure rendering. It is freely available at http://cao.labshare.cn/AbRSA/.

Keywords:

Immunoinformatic; Antibody; Antibody numbering; Complementarity-determining region;

Abbreviations

CDR: complementarity determining region

FR: framework region

HMM: Hidden Markov Model

V gene: variable gene

D gene: diversity gene

J gene: joining gene

Introduction

Antibodies, or immunoglobulins, are extremely important proteins in the immune system that can identify and neutralize antigens due to their unique sequence composition of two identical pairs of heavy and light chains. Each pair contains a constant domain and a variable domain. The latter forms antigen-binding sites and determines binding specificity and affinity. The sequence of the variable domain is highly diverse, resulting from the rearrangement of a variable (V) gene, a diversity (D) gene and a joining (J) gene (or a J gene and a V gene). The variable domain can be further divided into two types of regions: complementarity-determining regions (CDRs) and framework regions (FRs). CDRs exhibit a direct relationship with antigen binding, while FRs provide support to the CDR conformation.

The general 3D structure of all variable domains is very similar, folding into a group of beta strands linked by loops.[1] Six of the loops at the top are the CDRs (CDR H1/2/3 in the heavy chain and CDR L1/2/3 in the light chain). Because of the regularity of the structure, it is highly beneficial to number each residue using a standard scheme for sequence comparisons and engineering, for example, to facilitate the description of critical residues and to delimit CDRs for humanization.[2-12] Antibody numbering has become a fundamental technique used in immunoinformatic analyses. The first numbering scheme, known as the Kabat scheme, was introduced in 1970.[13] In subsequent years, the Chothia scheme[14-16] and its variants,[17] such as the IMGT[18] and AHo[19] schemes, were introduced. These schemes are

similar but vary in the numbering of some insertion and deletion positions. The Kabat scheme is based on sequence patterns, while in the Chothia scheme, the definition is modified by incorporating structural information. IMGT and AHo unify the numbering of antibodies and T-cell receptors.

Antibody numbering is generally achieved by mapping a query sequence onto known sequences that have been numbered in advance. The state-of-the-art software used for antibody numbering introduces complicated strategies. For example, AbNum[17] built the profiles of six residues at the start and end of FRs as anchor regions. For a query sequence, AbNum searches for the anchor regions and aligns the other regions to known sequences, which finally determine the numbering. IgBlast and the germline knowledge-based approach are used to perform sequence alignment against pre-annotated databases of germline genes and map the numberings to the query sequence.[20,21] Other software options, such as ANARCI,[22] PyIgClassify,[23] ProABC[24] and DIGIT,[25] pre-annotated a large set of antibody sequences, and made sequence alignment to build hidden Markov models (HMMs). Then query sequences were aligned to the HMMs and annotated by transferring the numbering of equivalent position from known sequence to the query. These software tools have successfully numbered lots of antibody sequences. However, antibody sequences are highly diverse. If the query sequence has no known anchor residues or similar patterns they cannot be numbered with these software. This has been observed in many benchmark tests, for

example, 1%~2% antibody sequences of Kabat databases cannot be numbered by AbNum.[17] 9,560 in 1,936,119 VH sequences cannot be numbered by ANARCI.[22] Although the error rates (2%~0.5%) are small, the absolute value of errors is large because the cardinality of antibody sequences is huge. Besides, the latest developed antibody technology obtained sequences not only in vivo as before but also in vitro, which further creates the divergence.[26] Therefore, no matter for high throughput analysis or individual inspection, it is valuable to develop more robust antibody numbering tools. In this work, we introduce heuristic knowledge into sequence mapping and propose a new method, referred to as the Antibody Region-Specific Alignment (AbRSA), which improves robustness when numbering antibodies with diverse patterns accord to our benchmark tests. In order to serve as an easy-to-use tool, AbRSA offers a user-friendly interface that can identify whether a query sequence is an antibody heavy or light chain, how the sequence is numbered and which region is the CDR. AbRSA provides a new option for immunoinformatic analyses.

Results

The popular antibody numbering methods are mainly classified into sequence-alignment based or HMM based numbering. Both of them require a large collection of pre-annotated antibody sequences to build the library or model for sequence mapping. The dependence on

pre-annotated sequences achieves success in known patterns but fails in irregular ones. AbRSA belongs to the sequence-alignment based method. In order to enhance the robustness, it divides the sequences into segments, which allows the lower weights in more diverse regions, particularity CDRs, and higher weights in the conserved regions (see Materials and Methods). In this section, we will describe the results of AbRSA on the benchmark set and simulated set, and compared the results with three state-of-the-art tools. Also we will introduce its online usage to facilitate users which are not experts in programming.

Testing AbRSA on the benchmark set

To assess the overall performance, AbRSA is applied to the benchmark set, which includes 1816 pre-annotated antibody sequences collected from Protein Data Bank and has no overlaps with the training sequences (see Materials and Methods). The results show that AbRSA numbered all the sequences, while the AbRSA without region-specific deviation only numbered 83.6% sequences. By comparing the pre-annotated numbering with the results of the two programs, AbRSA shows identical numberings except one mismatch of a heavy chain (PDB code: 4I3R), while the AbRSA without region-specific deviation failed in 13.0% numbered sequences. This result indicates that the region-specific alignment is superior to the pure sequence-alignment method.

For comparison, three state-of-the-art programs (AbNum, ANARCI and an integrated tool in proABC) are applied to the same benchmark set. AbNum employs a sequence-profile matching method, while ANARCI and proABC use HMMs for sequence numbering. proABC, AbNum and ANARCI were unable to number 22, 32, and 8 heavy chains and 8, 5 and 0 light chains, respectively (see Table 1). By comparing the results with pre-annotated numberings, proABC and ANARCI show 16 and 4 mismatches, respectively. These results imply that it is a particularly difficult task to number every antibody sequence using program tools.

Detailed analysis shows that those unsuccessful numbering sequences exhibit two types of features in CDRs: pseudo-conserved patterns and unusual lengths. For example, the WGTE segment in CDR H3 appears similar to the conserved pattern WGXG (where X represents any type of amino acid), and the SGGG segment in CDR L3 is similar to conserved pattern FGXG in FR, which tend to mislead the pattern recognition (Fig. 1). Another types are the ultra-long CDR H3 (>25 residues) and the ultra-short CDR H1 (<4 residues), which are quite different from the regular patterns. These unusual sequences cannot be recognized if there is no additional information to guide the program. AbRSA classifies the sequence into regions and attempts to number them using the general pattern recognition in CDRs and the precise pattern recognition in FRs. This is the reason that AbRSA can number these sequences, even though their pseudo-conserved patterns and ultra-lengths of CDRs are not included in the consensus sequences or training sequence set.

Testing robustness: increasing diversities of antibody sequences

The benchmark test shows that AbRSA can number antibodies with unusual patterns. However, its maximum performance for diverse sequences remains unknown. Hence, we employed a well-established antibody sequence simulation model AB[33] to generate a huge simulated antibody set (187,200 heavy chains and 70,200 light chains) in order to expand the diversity of benchmark sequences (see Materials and Methods). Briefly, random mutations

are introduced in antibody sequences using a transition probability matrix to mimick somatic hypermutations. The number of mutation is gradually increased from 10 to 50 with a step of 5. The more mutations imply the larger diversity of the sequence. We applied AbRSA to this simulated set. When the number of mutations in heavy chains is less than 10, the success rate of numbering is 100%. As the number further increases, the success rates decrease from 99.5%, 98.9%, 97.7%, 96.7%, 95.2%, 94.0% and 91.6% to 90.6% in the end, which changes half of the original antibody sequence [Fig. 2(A)]. Similarly, when the numbers of mutations in light chains are less than 30, the success rates keep 100%. As the number further increases, the success rate slightly decreased from 99.2%, 99%, 98.6% and 97.0% to 96.3% [Fig. 2(B)]. For comparison, we also applied ANARCI and AbNum to the simulated set (see Testing methods in Materials and Methods). In case of heavy chains, the success rates shows the similar trend to AbRSA when the number of mutation increases. However, ANARCI and AbNum accumulates more unsuccessful numberings, especially in the end, when introducing more mutations in a sequence, the success rate decreases to 82.7% and 60.7%, respectively, which is 8.1% and 30.0% less than AbRSA. In case of light chains, ANARCI and AbNum also accumulate more unsuccessful numberings than AbRSA. In the end, the success rate is 90.2% and 74.5% while that of AbRSA is 96.3% [Fig. 2(A), 2(B)]. This result shows that AbRSA was rather robust to the simulated somatic hypermutations, due to the prior knowledge of the region-specific feature.

Testing specificity: distinguishing between antibodies and non-antibodies

AbRSA shows highly robust performance for antibody sequence variations in the above test. Then, the question arises of whether AbRSA can distinguish between antibodies and non-antibodies, which is critical for application to unknown sequences. To answer this question, we compared the alignment results for both antibody and non-antibody sequences. The former are obtained from the benchmark data, which included 1,816 heavy or light chains. The latter are from the Swiss-Prot database,[27] including 551,744 sequences, excluding those annotated as antibodies, V genes, immunoglobulins, or heavy or light chains. The alignment results are quantified based on the sequence identity between the query and consensus sequences. As CDRs exhibit diverse lengths, which may affect the results, we quantified the sequence identity in FRs, rather than the whole sequence.

The result is illustrated in Figure 3, which shows the frequency distribution of the sequence identities. There are two peaks in the figure. The black curve indicates non-antibodies, whose average sequence identity is approximately 38%, ranging from 0-66%. The gray curve indicates antibodies, whose average sequence identity is approximately 94%, ranging from 74-100%. The non-overlapping curves demonstrate that AbRSA can clearly distinguish between antibody and non-antibody sequences. We checked the four non-antibodies exhibiting the highest sequence identity (≥65%) and found that they were

immunoglobulin-like proteins, but not real antibodies. For comparison, we also assessed the performance of ANARCI and AbNum using the non-antibody set (see Testing methods in Materials and Methods). The result shows that ANARCI identified all the sequences as non-antibody, while AbNum identified most non-antibodies except a viral T-cell receptor beta chain-like protein (UniProt ID: P11364) that was recognized as the light chain of antibody.

Web service of AbRSA

To make the best use of AbRSA, we set up a user-friendly web service. AbRSA is implemented in C++ and PHP. The AbRSA pipeline for computation is shown in Figure 4. The input could be either the protein sequence or structure. Multiple protein sequences are supported if the sequences are in FASTA format. The input will then be subjected to region-specific alignment with heavy- and light-chain consensus sequences (see supporting information). The program will judge whether the sequence is a heavy chain or light chain, or neither, by comparing sequence identities with consensus sequences. If the identities are lower than the cutoff (70%, minimum value from the training set), the query sequence will not be recognized as an antibody, and the program will loop back to search for more heavy or light chains in the query sequence in the case of fusion proteins, which may contain multiple antibody variable domains. After all possible heavy or light chains are found; the program

will output the numbering results and the location of FRs and CDRs in the sequences. If the input is a protein structure (PDB format file), the web server will extract protein sequences and subject to region-specific alignment with heavy- and light-chain consensus sequences as the above process. In addition, it uses an interactive molecular visualization JavaScript library, called 3Dmol.js,[28] to render three-dimensional (3D) graphics in the web browser. It highlights the residues of CDRs according to the numbering results using different colors in the 3D graphics. As 3Dmol is hardware-accelerated, the 3D view can be rotated, translated, and re-sized by dragging and scrolling the mouse smoothly. This feature could help to determine the location and conformation of CDRs.

Discussion

Antibody sequence numbering is a well-established, but critical topic in the field of antibody research. AbRSA can complement existing methods by focusing on unusual antibody sequences. Traditionally, the performance of antibody numbering is related to the amount of sequences used for method training. Recently developed methods employing a larger training set can number more antibodies than earlier methods. However, these methods may not be applicable to unusual sequences. Integrating additional biological insights, beyond the sequences themselves, could solve the problem. In this report, we show that AbRSA is very robust to unusual antibodies whose patterns are not included in the training data, indicating

that the region-specific features contribute to antibody numbering. The current AbRSA tool supports a Kabat and Chothia numbering scheme. In the future, we will continuously improve our program and provide additional numbering schemes to the user.

Materials and Methods

Training set

The training dataset consisted of antibody sequences downloaded from the UniProt database.[29] Redundant sequences were eliminated using CD-HIT[30] with a sequence identity cutoff of 85%. In total, 503 heavy and 475 light chains were collected. All sequences were numbered and double checked by using two well-established programs, AbNum and ANARCI. The sequences that could not be numbered by the programs were discarded.

Benchmark set

All the antibodies were collected from the Protein Data Bank.[31] The redundant sequences were removed by CD-HIT. To avoid overlap of the benchmark and training datasets, the sequence whose identity was over 85% (the average identity for any pair of antibody sequences that we knew) were also removed. In total, 983 heavy and 833 light chains were obtained. The pre-annotated numbering was obtained using AbNum and ANARCI. 39 chains could not be numbered (no output) by both software, and 13 chains showed conflicting results. Those sequences were manually numbered using 3D structure alignments.[32]

Simulated set

To expand the diversity of benchmark sequences, a well-established antibody sequence simulation model AB[33] is used to generate the simulated set in two steps. Firstly, the benchmark set is clustered using a sequence similarity cutoff of 80% by CD-HIT. One representative sequence in each cluster was randomly selected as the starting sequence of simulation. The sequences that could not be numbered by AbNum and ANARCI are eliminated. In all, 208 heavy and 78 light chains were obtained. Secondly, random mutations are introduced in these sequences using the antibody-specific amino acid substitution model AB, which is guided by 20×20 matrix of replacement rates between amino acids. The number of mutations is gradually increased from 10 to 50 with a step of 5. Each of the steps is repeated 100 times which finally create 900×208=187,200 heavy chains and 900×78=70,200 light chains.

Antibody region-specific alignment method

Antibody sequence diversity is mainly generated through two processes: somatic recombination and hypermutation.[34,35] The former joins the alleles of V, D and J genes or those of only V and J genes randomly. The regions where the genes recombine are located in CDR H3 and CDR L3, resulting in the diverse lengths of these regions. The latter process consists of frequent random mutations, which often occur at CDRs. As a result, CDRs exhibit

frequent substitutions, insertions and deletions, while FRs generally show a small number of substitutions. The origin of diversity tells us the variability of antibody is region specific. Although it has been implicitly incorporate in previous methods, region specific feature has not been taken full advantage yet. Hence we tried to strengthen the feature explicitly by using different alignment strategy in each region of the consensus sequence of antibody. Two consensus sequences (see supporting information) of heavy and light chains came from the antibody database abYsis,[36] where the residue distribution in variable domains is analyzed. The two consensus sequences were numbered by the standard numbering scheme of Kabat and Chothia. The sequences were divided into four types of regions: (1) FRs; (2) CDRs; (3) insertion positions (IPs) following the definition of the numbering scheme, such as H82 and L95 in the Chothia scheme (see supporting information); and (4) conserved positions (CPs), where one or two amino acids show a frequency above 95%, such as cysteine at H22 and aspartic acid at L82 in the Chothia scheme (see supporting information). These four types allow region-dependent features to be incorporated into sequence mapping.

To number an antibody, we employed a modified Needleman-Wunsch dynamic programming algorithm[37] to map the query sequence to the consensus sequence. The score at each position in the alignment is calculated as follows:

$$S(i,j)=\max\begin{cases}S(i-1,j)-w_1\\ S(i-1,j-1)+\max\{Blosum62(i,j)\}\cdot w_2\\ S(i,j-1)-w_1\end{cases} \quad (1)$$

where

$$w_1=\begin{cases}P_{CPs} & if\ \ j\in CPs\\ P_{IPs} & else\ if\ \ j\in IPs\\ P_{FRs} & else\ if\ \ j\in FRs\\ P_{CDRs} & else\ if\ \ j\in CDRs\\ 11 & else\ if\ \ j\in others\end{cases} \quad (2)$$

and

$$w_2=\begin{cases}S_{CPs} & if\ \ j\in CPs\\ 1 & else\ if\ \ j\in others\end{cases} \quad (3)$$

In the above equations, i is the residue index of the query sequence, and j is the residue index of the consensus sequence. The alignment score ($\max\{Blosum62(i,j)\}$) is calculated by enumerating BLOSUM62 matrix for query residue and all possible residues at a given position of the consensus sequence. P_{CPs}, P_{IPs}, P_{FRs} and P_{CDRs} are the gap penalties at CPs, IPs, FRs and CDRs, respectively. S_{CPs} is the weight of the match score at CPs. The total score is the summation of $S(i,j)$ at each position. We explored the combination of S_{CPs}, P_{IPs}, P_{CPs}, P_{FRs} and P_{CDRs} by iterating their values from 0 to 100 using the training set.

We found that the best numbering performance (100% identical to pre-annotated numbering) for the training data was achieved when P_{IPs}=1, P_{CDRs}=11, P_{FRs}=26, P_{CPs}=55 and S_{CPs}=5. These parameters are in accordance with the assumption that the gap penalty

increases in a step-by-step manner from IPs, to CDRs, FRs and CPs. The scoring weight (S_{CPs}) for residues located at CPs was five times greater than that of the other residues.

Testing methods for ANARCI, AbNum and proABC

The latest stand-alone version of ANARCI (v1.3) was used in this work. AbNum and proABC were tested on their web servers (http://bioinf.org.uk/abs/abnum/, http://circe.med.uniroma1.it/proABC/). As using the simulated set (187,200 and 70,200 sequences) and non-antibody set (551,744 sequences), we used their subsets for testing. The subset of simulated set consists of randomly selected 10 out of 100 mutants in each group. It includes 187,200×10%=18,720 heavy chains and 70,200×10%=7,020 light chains. The success rates of antibody numbering using this subset are consistent with the results using 5% sequences. The subset of non-antibody data includes 7,459 sequences that are most similar to antibody consensus sequences (share over 45% identities) in the whole non-antibody set. We did not use proABC in comparison studies on the simulated set and non-antibody set for the reason of time consumption.

Availability of data and materials

The web tool, benchmark set and simulated set are freely available to the public at http://cao.labshare.cn/AbRSA/ .

Acknowledgements

The authors wish to thank Professor Yang Zhang and Chengxin Zhang in the University of Michigan, Professor Zhihua Zhang in Beijing Institute of Genomics, Professor Haiying Hang in Institute of Biophysics, Chinese Academy of Sciences for invaluable discussion. This work was supported by the National Natural Science Foundation of China under Grant (number 31401130 and 81830108), the funding for prevention and control technology of African swine fever (2018NZ0151).

Competing interests

The authors declare that they have no competing interests.

References

1. Krawczyk K, Kelm S, Kovaltsuk A, Galson JD, Kelly D, Truck J, Regep C, Leem J, Wong WK, Nowak J, Snowden J, Wright M, Starkie L, Scott-Tucker A, Shi J, Deane CM (2018) Structurally mapping antibody repertoires. Front Immunol 9:1698.

2. Riechmann L, Clark M, Waldmann H, Winter G (1988) Reshaping human antibodies for therapy. Nature 332:323-327.

3. Verhoeyen ME, Saunders JA, Broderick EL, Eida SJ, Badley RA (1991) Reshaping human monoclonal antibodies for imaging and therapy. Dis Markers 9:197-203.

4. Marcatili P, Rosi A, Tramontano A (2008) PIGS: automatic prediction of antibody structures. Bioinformatics 24:1953-1954.

5. Sircar A, Kim ET, Gray JJ (2009) RosettaAntibody: antibody variable region homology modeling server. Nucleic Acids Res 37:W474-W479.

6. Kuroda D, Shirai H, Jacobson MP, Nakamura H (2012) Computer-aided antibody design. Protein Eng Des Sel 25:507-521.

7. Shirai H, Prades C, Vita R, Marcatili P, Popovic B, Xu J, Overington JP, Hirayama K, Soga S, Tsunoyama K, Clark D, Lefranc MP, Ikeda K (2014) Antibody informatics for drug discovery. Biochim Biophys Acta 1844:2002-2015.

8. Olimpieri PP, Marcatili P, Tramontano A (2015) Tabhu: tools for antibody humanization. Bioinformatics 31:434-435.

9. Jarasch A, Skerra A (2017) Aligning, analyzing, and visualizing sequences for antibody engineering: Automated recognition of immunoglobulin variable region features. Proteins 85:65-71.

10. Clavero-Alvarez A, Di Mambro T, Perez-Gaviro S, Magnani M, Bruscolini P (2018) Humanization of antibodies using a statistical inference approach. Sci Rep 8:14820.

11. Leem J, Georges G, Shi J, Deane CM (2018) Antibody side chain conformations are position-dependent. Proteins 86:383-392.

12. Roy A, Nair S, Sen N, Soni N, Madhusudhan MS (2017) In silico methods for design of biological therapeutics. Methods 131:33-65.

13. Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 132:211-250.

14. Al-Lazikani B, Lesk AM, Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. J Mol Biol 273:927-948.

15. Chothia C, Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 196:901-917.

16. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan

EA, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, Poljak RJ (1989) Conformations of immunoglobulin hypervariable regions. Nature 342:877-883.
17. Abhinandan KR, Martin AC (2008) Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. Mol Immunol 45:3832-3839.
18. Lefranc MP, Pommie C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55-77.
19. Honegger A, Pluckthun A (2001) Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. J Mol Biol 309:657-670.
20. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41:W34-W40.
21. Zhao S, Lu J (2010) A germline knowledge based computational approach for determining antibody complementarity determining regions. Mol Immunol 47:694-700.
22. Dunbar J, Deane CM (2016) ANARCI: antigen receptor numbering and receptor classification. Bioinformatics 32:298-300.
23. North B, Lehmann A, Dunbrack RJ (2011) A new clustering of antibody CDR loop conformations. J Mol Biol 406:228-256.
24. Olimpieri PP, Chailyan A, Tramontano A, Marcatili P (2013) Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. Bioinformatics 29:2285-2291.
25. Chailyan A, Tramontano A, Marcatili P (2012) A database of immunoglobulins with integrated tools: DIGIT. Nucleic Acids Res 40:D1230-D1234.
26. Sormanni P, Aprile FA, Vendruscolo M (2018) Third generation antibody discovery methods: in silico rational design. Chem Soc Rev 47:9137-9157.
27. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45-48.
28. Rego N, Koes D (2015) 3Dmol.js: molecular visualization with WebGL. Bioinformatics 31:1322-1324.
29. Pundir S, Martin MJ, O'Donovan C (2017) UniProt protein knowledgebase. Methods Mol Biol 1558:41-55.
30. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680-682.
31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242.
32. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302-2309.

33. Mirsky A, Kazandjian L, Anisimova M (2015) Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences. Mol Biol Evol 32:806-819.
34. Market E, Papavasiliou FN (2003) V(D)J recombination and the evolution of the adaptive immune system. PLOS Biol 1:e16.
35. Diaz M, Casali P (2002) Somatic immunoglobulin hypermutation. Curr Opin Immunol 14:235-240.
36. Swindells MB, Porter CT, Couch M, Hurst J, Abhinandan KR, Nielsen JH, Macindoe G, Hetherington J, Martin AC (2017) abYsis: Integrated antibody sequence and structure-management, analysis, and prediction. J Mol Biol 429:356-364.

37. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443-453.

Table 1. Results of antibody numbering using the benchmark dataset. The values are numbers of failed cases using four software tools. The result was classified as "No output", which means no light or heavy chain is detected by the tool, and "Mismatch", which indicates the predicted CDRs and FRs are not correct, for heavy and light chains according to Kabat or Chothia schemes.

Tool	Heavy Chain (N=983)				Light Chain (N=833)			
	Kabat		Chothia		Kabat		Chothia	
	No output	Mismatch	No output	Mismatch	No output	Mismatch	No output	Mismatch
proABC	22	12	-[a]	-[a]	8	4	-[a]	-[a]
AbNum	32	0	32	1	5	0	5	0
ANARCI	8	2	8	2	0	2	0	2
AbRSA	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**

[a] Software does not support.

Figure legends

Figure 1. Examples of pseudo-conserved patterns in CDR3. QueryH and QueryL are heavy- and light-chain sequences (PDB code: 4OCR, 3MLT), respectively. Their CDRs are colored in blue, and those of CDR3 are underlined. The sky-blue shaded letters "WGTE" and "SGGG" in CDR3 are pseudo-conserved patterns, which are similar to real conserved patterns (yellow shaded letters "WGQG" and "FGDG"). Pseudo-conserved patterns may mislead numbering tools. Please see the sequences as text in the supplementary file.

Figure 2. Benchmark results on simulated set. It illustrates the success rate of simulated antibody numbering versus the number of mutations in the original antibodies. A: results for heavy chain. B: results for light chain. The black, blue and red lines indicate the results of ANARCI, AbNum and AbRSA respectively.

Figure 3. Frequency distribution of the sequence identities of antibodies and non-antibodies. The red curve indicates the frequency of sequence identities between antibody queries and the consensus sequence. The black curve indicates the frequency of sequence identities between non-antibody queries and the consensus sequence. The two curves exhibit no overlap, which suggests the antibodies and non-antibodies show obvious differences in sequence identity.

Figure 4. A: Pipeline of the AbRSA web service. B: The input of AbRSA. The sequences in the textbox are examples. C: The output numbering results of AbRSA using an antibody structure (PDB code: 3i75). D: The CDRs are highlighted in colors in the 3D viewer (input PDB code: 3i75).

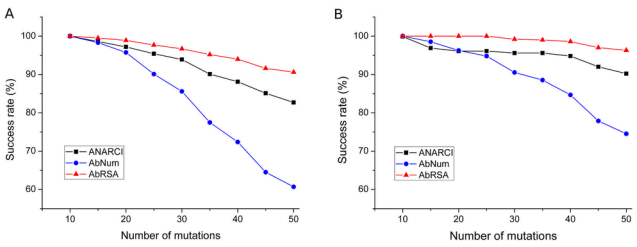
Heavy Chain

```
  1 EVQVVESGGGVVQPGRSLRLSCTASGFTFSNFAMGWVRQAPGKGLEWVAFISSDGSNKNY   60
 61 GDSVKGRFTISRDNSKNTVFLQMNSLRVEDTALYYCAKDVGDYKSDEWGTEYYDISISYP  120
121 IQDPRAMVGAFDLWGQGTMVTVSPAS
```

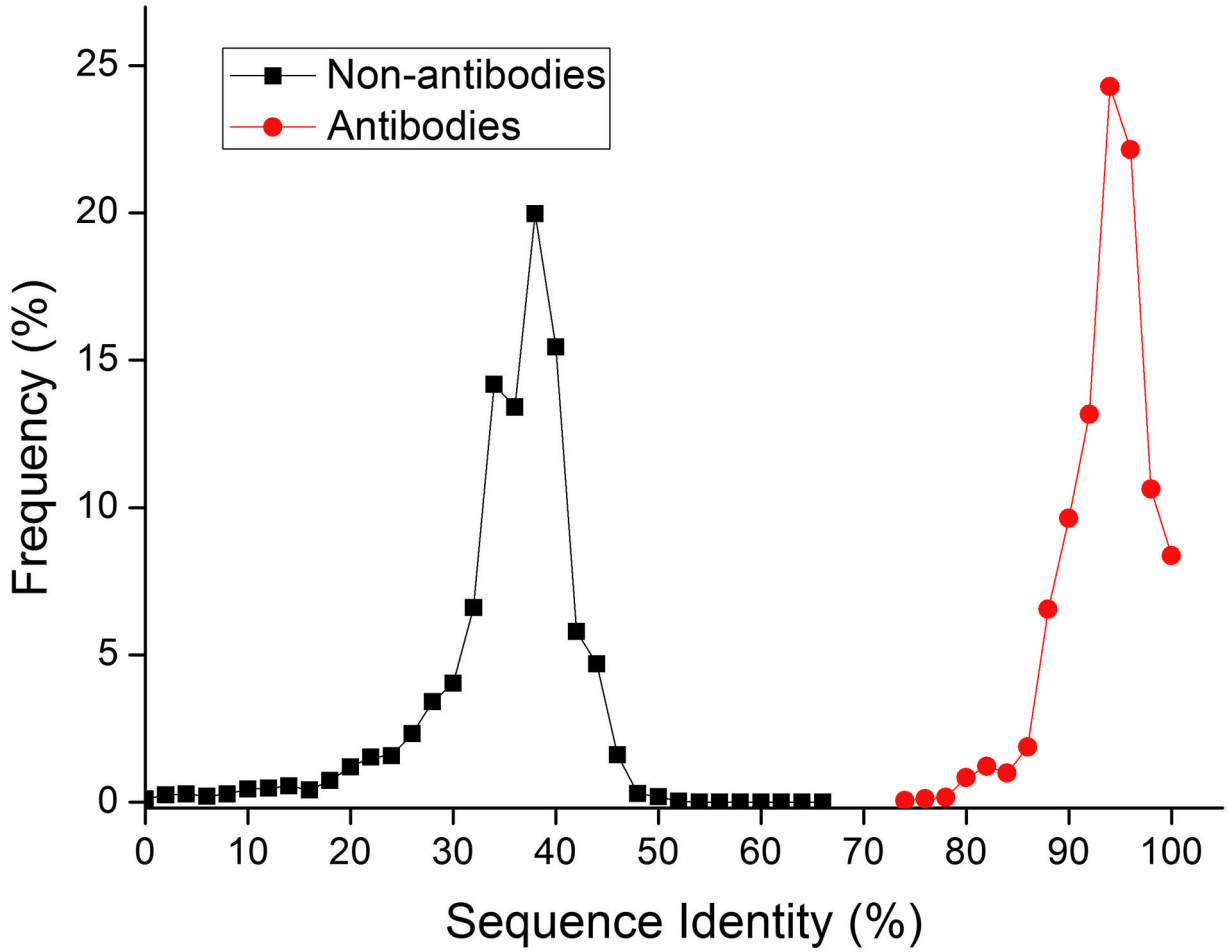
Light Chain

```
  1 SYDLTQPPSVSVSPGQTASISCSGDKLDDKYVSWYYQRPGQSPVLLMYQDFKRPSGIPER   60
 61 LSGSKSGKTATLTISGTQSLDEGDYYCQAWDASTGVSGGGTKLTVLFGDGTRLTVLGQPK  120
```

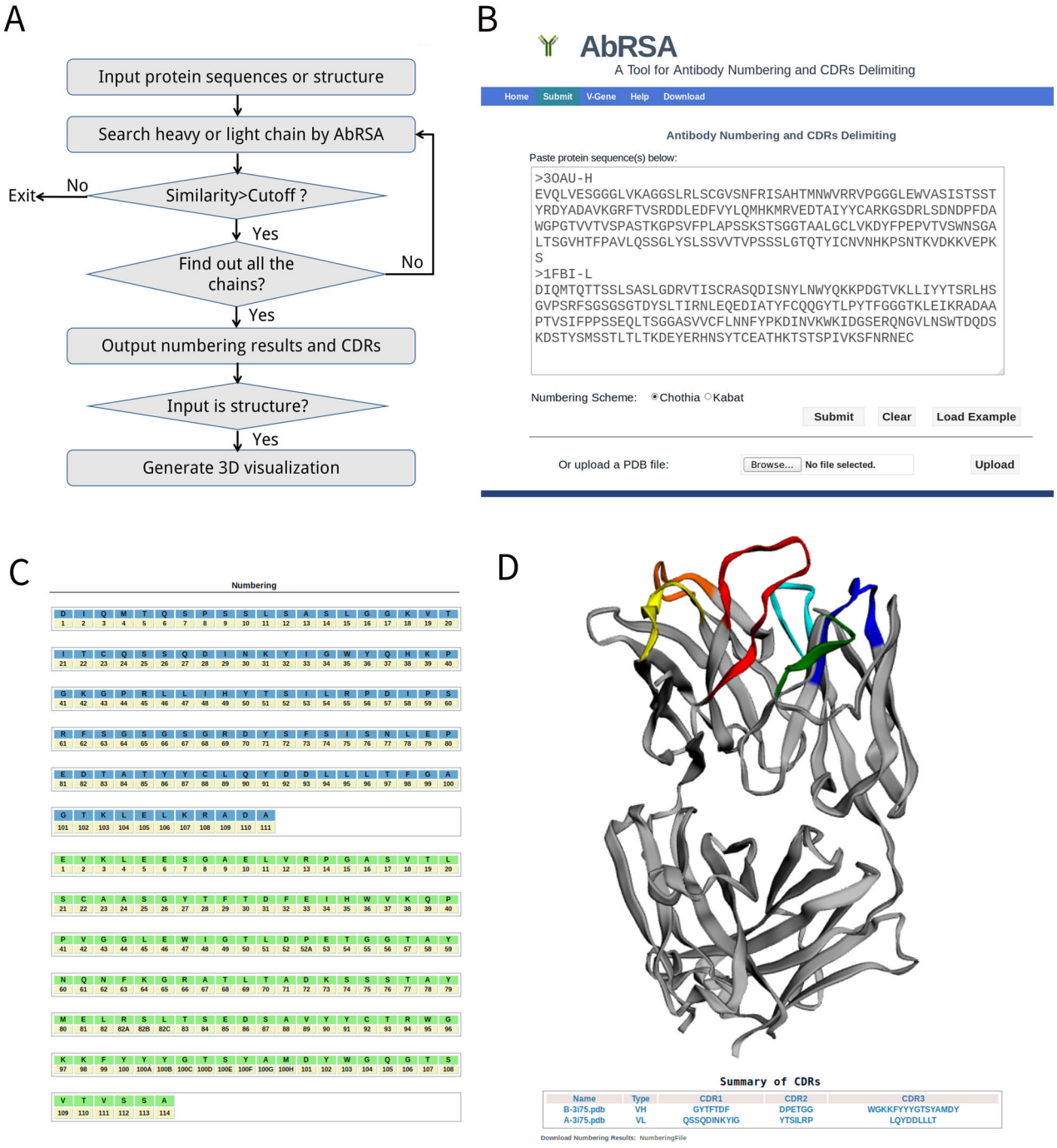
PRO_3633_Figure1.tiff



PRO_3633_Figure2.tiff



PRO_3633_Figure3.tiff



PRO_3633_Figure4.tiff