W490–W495 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* *Published online 20 May 2019*
doi: 10.1093/nar/gkz387

Yvis: antibody high-density alignment visualization and analysis platform with an integrated database

Milene B. Carvalho [1,2,*], Franck Molina[3] and Liza F. Felicori[1,*]

[1]Laboratory of Synthetic Biology and Biomimetics, Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil, [2]Departamento de Ciência da Computação, Universidade Federal de São João del Rei, São João del Rei, Minas Gerais, 36301-360, Brazil and [3]Sys2diag, UMR9005 CNRS Alcediag, Montpellier 34184, France

Received March 01, 2019; Revised April 20, 2019; Editorial Decision April 30, 2019; Accepted May 02, 2019

ABSTRACT

As antibodies are a very important tool for diagnosis, therapy, and experimental biology, a large number of antibody structures and sequences have become available in recent years. Therefore, tools that allow the analysis, comparison, and visualization of this large amount of antibody data are crucially needed. We developed the antibody high-density alignment visualization and analysis (Yvis) platform to provide an innovative, robust and high-density data visualization of antibody sequence alignments, called *Collier de Diamants*. The Yvis platform also provides an integrated structural database, which is updated weekly, and many different search and filter options. This platform can help to formulate hypotheses concerning the key residues in antibody structures or interactions to improve the understanding of antibody properties. The Yvis platform is available at http://bioinfo.icb.ufmg.br/yvis/.

INTRODUCTION

Antibodies or immunoglobulins are vertebrate immune system proteins that are produced by B cells and can bind to antigens with high specificity and affinity. For this reason, antibodies are an important tool in diagnosis, therapy and experimental biology (1). To elucidate the antibody characteristics, large numbers of antibody structures and sequences have been generated in the last years. The number of antibodies or antibody fragment structures deposited in Protein Data Bank (PDB) (2) has increased exponentially (3), leading to the development of databases of antibody structures (4–7). Moreover, many antibody sequences have been obtained by high-throughput sequencing of the B-cell receptor repertoire (8,9). This extraordinary and still increasing number of antibody structures and sequences demands integrative data organization and tools for their analysis, comparison and visualization. One of the major bottlenecks in this field is the concomitant visualization of a large amount of antibody data. AbYsis (4) and IMGT/3Dstructure-DB (5) allow antibody visualization, but only a limited number of sequences can be analysed at a time. Indeed, abYsis presents a classical multiple sequence alignment (MSA) that displays a limited number of sequences and positions each time. IMGT/3Dstructure-DB display only one antibody sequence using the IMGT/*Collier des Perles* representation (10) that allows sequence analysis related to the antibody structure. To fill this gap, we developed the antibodY high-density alignment visualization and analysis (Yvis) platform that includes: (i) an updated weekly and curated antibody structure database (Yvis database) and (ii) integrated antibody analysis resources, such as an antibody high-density alignment visualization called *Collier de Diamants,* and multiple filter options to analyse data from user files or from the Yvis database.

MATERIALS AND METHODS

Yvis database: an updated weekly and curated repository of data on antibody PDB structures

The Yvis database is an updated weekly collection of data on antibody PDB structures (in complex with an antigen or not), such as PDB and chain identification, antibody and protein antigen-producing organisms or type (hapten, carbohydrate or nucleic acid), gapped sequences of antibody chains, germline information (assigned V and J genes with their identity values), and antigen-antibody putative contacts.

The Yvis script, developed in Python, extracts a list of antibody PDB structures from SAbDab (7) that is updated weekly. The following data are extracted from this list, processed, and stored in the Yvis database: (i) PDB and chain identifications, (ii) names of the organisms producing antibody and antigen (when applicable) and (iii) antigen molecule description. When the SAbDab list does

*To whom correspondence should be addressed. Tel: +55 32 3379 4935; Email: milenebc@ufsj.edu.br
Correspondence may also be addressed to Liza F. Felicori. Tel: +55 3409 2981; Fax: +55 31 3409 2614; Email: liza@icb.ufmg.br

not contain the antibody- or antigen-producing organism name, Yvis script extracts this information from the corresponding PDB structure file, acquiring the ORGANISM_SCIENTIFIC value from SOURCE record after retrieving the molecule Id from COMPND record. After data extraction, Yvis script checks whether the organism names match the UniProt Taxonomy (11), and correct them if required (Supplementary Table S1). Data are manually curated, if the standard name is not found automatically. These standard names facilitate the Yvis database search, reducing the diversity of organism names, for instance by eliminating all synonyms.

The Yvis script submits antibody chain sequences to IMGT/DomainGapAlign (5) to obtain gapped sequences and germline information. Then, it processes the result page and extracts the gapped sequence of the variable domain of each chain, following the IMGT numbering (12). Moreover, the script extracts and stores the V and J germline genes assigned to the chain sequence, and their identity values.

Finally, to obtain information on the putative antibody–antigen contacts, the Yvis script downloads the PDB structure files and extracts the antibody chain amino acids that potentially interact with a peptide or protein antigen using the Biopython PDB module (13). Then, the distance between each α-carbon of the antibody and antigen amino acids is calculated. If the distance between two α-carbons is not higher than 8 Å, the position that contains the amino acid is marked as making a putative contact. This distance is used because it allows including putative direct interactions between antigen and antibody and also water-mediated interactions (14).

Yvis resources: integrated tools for high-density antibody data visualization and analysis

The Yvis platform integrates resources that allow the analysis of antibody variable domains that have been uploaded as user sequences or selected from the Yvis database. This platform is a web-based application that process sequences in a server or in a user's internet browser, depending on the analysed data. The server-side application was developed using PHP and Mysql, and the client-side using the JavaScript and D3.js framework.

The Yvis Platform offers input and search versatility. With the Yvis platform, users can analyse antibody structures stored in the Yvis database or uploaded by them. Different search options (Figure 1A) are available to select, from the database, a set of antibody structures to be analysed. It is possible to show all antibody chains stored in the Yvis database, or to specify a list of PDB identifiers, or a pair of PDB:chain identifiers. Moreover, users can choose to show free or complexed antibodies, and in the latter case, they can indicate the antigen type (hapten, carbohydrate, nucleic acid or protein). For protein antigens, they can indicate the producing organism. Users can also select antibodies with assigned germline V or J genes, or produced by user-selected organisms. In addition, users can search antibodies by using keywords contained in the literature related to PDB structures. After defining the PDB structure search criteria, the user can apply additional filters to avoid sequence redundancy, such as: (i) to choose only one representative chain of each type (heavy or light) in each PDB structure; (ii) to specify an identity threshold that ensures that none of the filtered sequences has an identity value higher than the user-specified value. This approach was based on Cd-hit (15). Because of the time requested to analyse and group all sequences, the identity filter is not used by default. All these filters can be combined.

Users can also analyse antibody sequences obtained from an IMGT/DomainGapAlign (5) results file, an IMGT/HighV-QUEST (16) gapped amino acid results file, or a FASTA file containing gapped, or ungapped chain sequences or even CDR sequences (Figure 1A, User Input file). When a user submits an IMGT results file, the Yvis platform will process it in the user's browser. Moreover, when a user submits ungapped sequences in a FASTA file, the Yvis platform will number them using ANARCI (17) in the Yvis server.

Independently of the chosen input data (Yvis database or user's data), the Yvis platform will generate an initial set of antibody sequences that can be visualized with the *Collier de Diamants* representation.

Collier de Diamants: a new visualization of high-density multiple sequence alignment of antibody variable domains. After the user's input choice, Yvis presents a first visualization of these data (middle panel in Figure 1B) as a multiple sequence alignment of antibody chains, based on the IMGT/*Collier de Perles* (Pearl Necklace) (10). In Yvis visualization, each sequence is numbered according to the IMGT unique numbering (12), and each position corresponds to a column in a traditional MSA. For each position, a pie chart indicates its amino acid composition. Each pie slice (sector) represents the number of sequences with an amino acid of a specific class in that position. The amino acid class is identified by a specific color, as defined in WebLogo (18). The positions are shown as in the *Collier de Perles*, linking sequences to their 3D structure. A square highlights the CDR anchors, one position before the CDR start and one after the CDR end (i.e. green for CDR1, orange for CDR2, and blue for CDR3), and allows the quick visualization of the residues that compose each CDR. As in Yvis each pearl of the necklace was replaced by a new representation with multiple 'facets', this new visualization was called *Collier de Diamants* (Diamond Necklace).

Like the *Collier de Perles*, the *Collier de Diamants* can be displayed on one (middle panel of Figure 1B) or on two layers (Figure 1B (I)). The two-layer representation shows the variable domain strands in a position closer to their 3D structure, while the one-layer version is closer to the variable domain sequence. As the *Collier de Diamants* uses a pie chart to describe each position of the alignment, it is possible to show the data of countless sequences in the same visualization. Moreover, positions with a conserved amino acid class are easily detected because they are represented by a pie chart with a dominant sector, while a pie chart with many sectors represents a position that is more variable.

Beside the visualization of an MSA, the *Collier de Diamants* displays a quantitative attribute for each position that is represented by a circle around the pie chart (Figure 1B (III)). In the Yvis platform, this attribute represents the

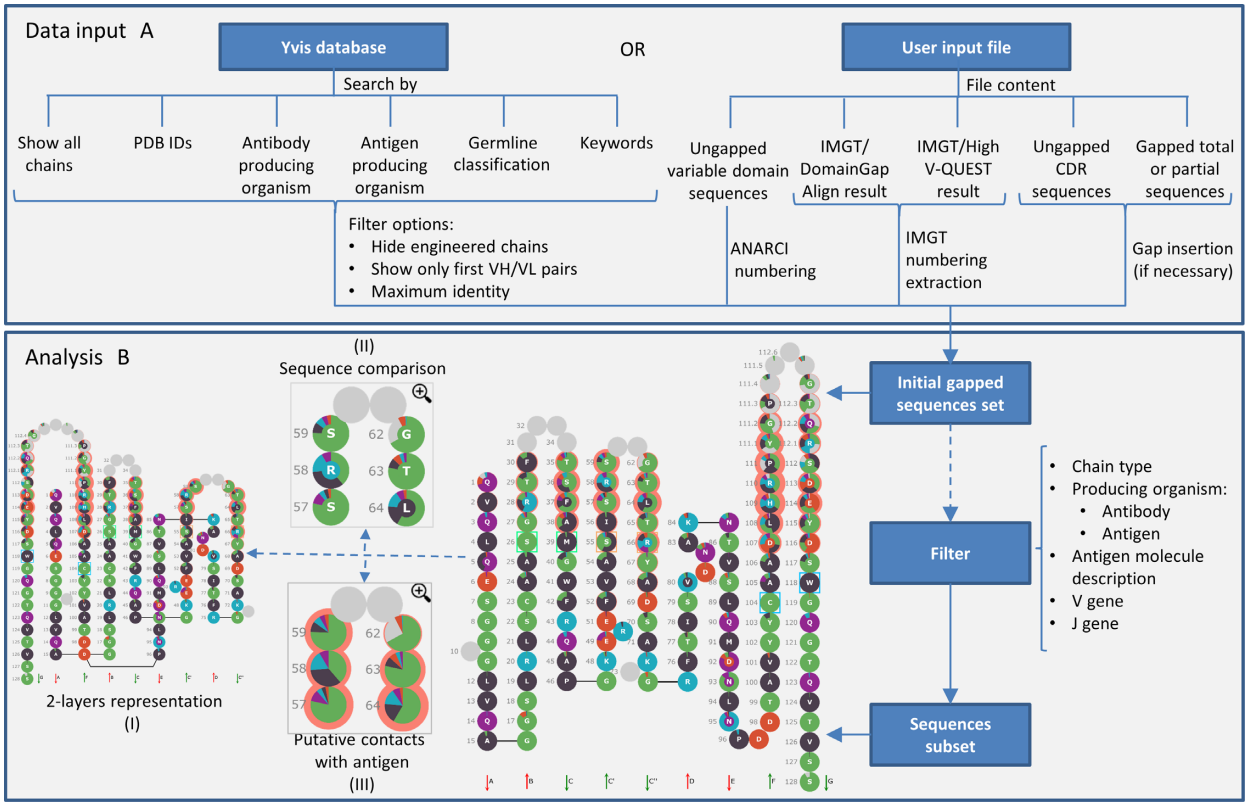

Figure 1. Yvis platform overview. (**A**) The data input box presents two possibilities to input sequences to be analysed in Yvis (user input files and selection from the Yvis database). It presents filter options for sequences from the Yvis database (redundancy and engineered chains) and actions taken by the platform to process the user input files. (**B**) The analysis box presents the options to visualize a multiple sequence alignment of antibody variable domains and the filter possibilities. The user can generate a new subset of sequences to be analysed, by selecting specific filters. The analysis can be displayed by the *Collier de Diamants* on one or two layers (I). Additionally, the user can compare the multiple sequence alignment with a reference sequence (II), and visualize data on putative contacts with the antigen (III).

number of chains that have a putative contact on each position. The radius length of the circle around a pie chart is proportional to the number of sequences with a putative contact at that position. Positions that have many gaps can have a small contact circle, even if all sequences in these positions have an amino acid that makes contact with the antigen. In this manner, positions that are involved in antibody–antigen interactions in the majority of the analysed sequences can be easily detected. However, it is only possible to show this attribute for chains from the Yvis database obtained from an antibody in complex with a protein or peptide antigen in the corresponding PDB structure.

The Yvis platform also allows comparing the MSA with a user-defined sequence (Figure 1B (II)): a target sequence, a germline gene sequence, or a consensus sequence. When the user inserts a sequence to be compared with the MSA, the Yvis displays, in the centre of each pie chart that represents a position, a small inner circle coloured according to the amino acid class of the input sequence that corresponds to that position (same colour code used for the pie chart). This representation facilitates the comparison of the user's sequence with the MSA by just checking whether the colour of the small circle is the same as the one of the larger pie chart sector for that position. In addition to the pie chart representation, Yvis makes available a detailed bar chart for each position, showing the number of each amino acid in the corresponding position.

Additional filter options can be used to select antibody subsets. After having selected the initial set of antibodies from the Yvis database or after uploading antibody sequences, the user can use filters to select sequence subsets and to restrict the analysis to specific chain types, antibody-producing organisms, antigen type or producing organisms, antigen molecule description, and germline information (right side of Figure 1B). As the molecule description is a text field of PDB files extracted from SAbDab, many different descriptions of the same molecule are often available. Therefore, to use this filter, the user must select all descriptions that match the desired molecule. The filter options are restricted to the uploaded or selected sequences. In addition, these filters can be combined to generate antibody subsets. The filters can be used, for example, to analyse only heavy chains or only human antibodies against a specific antigen, among others.

Below the *Collier de Diamants* representation of an initial set of antibody sequences, there is a table containing the available information of all chains presented in the MSA. The colours in each gapped sequence highlight the CDR amino acids (same colour code as for the CDR anchors in the *Collier de Diamants*) and the positions that make the pu-

tative contacts. This table can be exported to be used with other tools.

RESULTS AND DISCUSSION

A continuously updated source of antibody structure information

The Yvis database is updated weekly and currently includes data on 3423 antibody structures (from 22 different antibody-producing organisms) in complex with protein antigens (from 155 different antigen-producing organisms) or non-protein antigens (184 structures with haptens, 22 with nucleic acids, and 106 with carbohydrates), as well as 1,136 structures of antibodies in free form. Most of the antibodies in the database are of human (1444) and mouse (1362) origin.

The organism-producing names are in the standard format defined by UniProt Taxonomy (11), allowing the user to search or filter the database content easily. To achieve that, we automatically modified the antigen-producing organism name of 389 structures, and manually changed the antibody-producing organism name of 58 structures and the antigen-producing organism name of 77 structures. In this database, users can select a set of antibodies, searching mainly by antibody-producing organisms, antigen types, antigen producing-organisms, or assigned germline genes. Moreover, users can restrict the set to be analysed by excluding engineered chains, multiple antibodies in the same structure, and sequences with identity above the user-defined cut-off. These combinations of different search criteria and constraints are not available in other antibody structure databases, such as abYsis (4), IMGT/3Dstructure-DB (5) and SAbDab (7).

Case studies demonstrate the Yvis platform versatility and easy high-density data visualization to help hypothesis formulation

To test the Yvis platform, we carried out three case studies using anti-HIV antibodies. In the first study, we selected anti-HIV gp120 antibodies available in the PDB, through a search of the Yvis database after choosing the HIV antigen-producing organism and the option that forces the result to contain only one chain of each type from each PDB file. After this search, to restrict the analysis to only anti-gp120 antibodies, we used the antigen molecule description filter and limited the analysis to the heavy chains. The alignment presented by *Collier de Diamants* of the 124 identified sequences (Supplementary Figure S1) allowed the easy visualization of known anti-HIV neutralizing antibody characteristics, such as a long CDRH3 (19), and of some conserved positions (e.g. 8, 22, 119 and 121). Moreover, it highlighted that many of the analysed chains could make contact with the antigen via CDR2 and framework region 3 (positions from 57 to 69). This is a known characteristic of some anti-HIV CD4-binding-site antibodies (20). Nevertheless, this is not a very common characteristic of antibody heavy chains.

In the second case study, we downloaded a FASTA file of transcript sequences of an HIV-infected donor (19) from the Sequence Read Archive (21) using the SRR1767440 access number. Then, we submitted the FASTA sequence to IMGT/HighV-QUEST (16), and then uploaded the results file containing the amino acid information for 478 047 heavy chain sequences to the Yvis platform. The platform excluded sequences with ambiguous amino acids, leaving 330 800 sequences (Supplementary Figure S2). We applied the germline filter to restrict the visualization to sequences assigned to the VH1-2*02 allele, because, at least three neutralizing antibodies isolated from this donor were derived from this allele. This resulted in the alignment of 97 751 sequences that were compared with the VH1-2*02 allele by the Yvis comparison sequence feature (Figure 2A). The parts of the sequence corresponding to the D and J genes were filled with gaps. As in the first case study, it was easy to visualize the CDRH3 length (most of the analysed sequences had two insertions in CDR3, and less than 10% of sequences had more than four insertions). For some positions (e.g. 36, 66, 92, 93 and 95), the amino acid class of the uploaded sequences was different from the germline amino acid class, as easily noticeable on the basis of the colour difference between the inner circle and the pie chart sectors. The detailed bar chart for position 36 (Box in Figure 2A) indicated that the glycine (G) residue of the germline gene was mutated into aspartic acid (D) in approximately half of the sequences. As aspartic acid is a charged residue and this position is located in a CDR, probably this modification is selected during antibody expansion to increase antigen binding.

In the third case study, we selected 21 heavy chain sequences from HIV neutralizing antibodies derived from the VH1-2*02 allele (19), and submitted them to IMGT/DomainGapAlign. Then, we uploaded the results file into Yvis and compared the alignment against the VH1-2*02 allele sequence, with gaps in the part corresponding to the D and J genes (Figure 2B). Some positions, mainly in CDR2 (57, 59 and 69) and framework region 3 (82 and 83), had only one sector in the pie chart and its colour was different from the color of the inner circle. This indicates that in all neutralizing antibody sequences, this amino acid is different from the one in the germline gene (new amino acid class). One could hypothesize that positions in which the amino acid class changed might carry important characteristics for the neutralizing activity (for instance, structural or binding features that allow antibody binding to HIV). Comparison of the positions with a change of amino acid class in neutralizing antibodies (Figure 2B) and the same positions in anti-gp120 antibodies with putative contacts (first case study) suggests that in some antibodies, the amino acid class change might have brought a new characteristic that allows some interaction with HIV (mainly positions in CDR2 and framework region).

The case studies presented here demonstrate the usefulness of some of the Yvis resources. Moreover, Yvis facilitates the formulation of hypotheses concerning the subset of analysed sequences. This was achieved by analysing conserved or divergent properties in specific positions or by comparing a set of sequences with their germline gene sequence or another reference sequence.

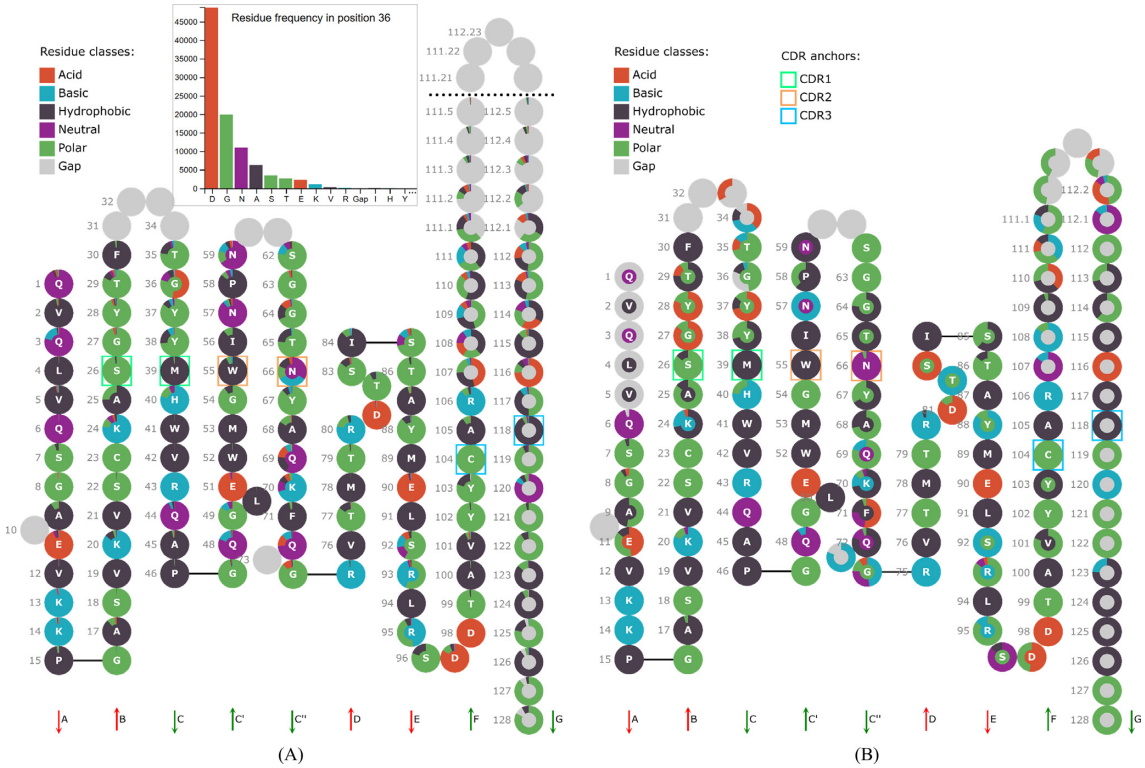

Figure 2. *Collier de Diamants* representation of antibody sequences obtained from HIV-infected patients. (**A**) Visualization of the alignment of 97 751 heavy chain sequences derived from the IGHV1-2*02 allele. Positions between 111.6 and 111.20 and positions between 112.20 and 112.6 were omitted. The box presents part of the detailed bar chart for position 36 showing the residue frequency in this position. (**B**) Visualization of the alignment of 21 heavy chain sequences of HIV neutralizing antibodies. In both visualizations, the IGHV1-2*02 germline sequence was used as comparison sequence. As this sequence includes only the V gene portion of the variable domain, gaps were inserted in the comparison sequence.

CONCLUDING REMARKS

In this article, we presented the Yvis platform that was developed to facilitate the analysis of large numbers of antibody structures and sequences. The Yvis platform includes a very convenient, innovative and robust high-density visualization (*Collier de Diamants*) of antibody sequence alignment that allows the analysis of hundreds of thousands of sequences in a single representation. Moreover, the Yvis integrated database is updated weekly and stores data on antibody PDB structures obtained from SAbDab and PDB files (e.g. PDB and chain identifications, antibody/antigen producing-organism names, molecule description), processed data from IMGT/DomainGapAlign (e.g. gapped sequences, V and J germline allele assignment and the corresponding identity values), and antibody-antigen putative contacts obtained by processing the structure coordinates. The producing-organism names are stored following Uniprot Taxonomy that facilitates database searches based on these names. In addition, there are various database search and filter options. The Yvis platform can also process user files that contain sequences obtained by different methods (e.g. FASTA, IMGT/DomainGapAlign and IMGT/HighV-QUEST files).

The Yvis platform can be used in different types of antibody analysis. For example, the quick visualization of the most conserved or divergent positions in a set of related antibodies can guide antibody engineering and mutagenesis experiments. In antibody repertoire studies, the *Collier de Diamants* visualization, coupled with the sequence comparison feature, can be used to compare thousands of antibody sequences with a specific germline sequence. This can give to researchers some insights into the most important mutations that occurred during the antibody affinity maturation process. Therefore, the Yvis platform offers an environment for antibody sequence analysis that helps to formulate hypotheses concerning the key residues in the antibody structure or interactions and improves the understanding of the antibody properties.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank José Miguel Ortega for providing computational resources.

FUNDING

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (Capes) [PVE 88887.125025/2014-00, Bio-Computacional 51/2013, COFECUB 935/19]; FAPEMIG [APQ-01437-16]. Funding for open access charge: Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais, Programa Interunidades de Pós-graduação em

Bioinformática da Universidade Federal de Minas Gerais and Sys2Diag.
Conflict of interest statement. None declared.

REFERENCES

1. Sela-Culang,I., Kunik,V. and Ofran,Y. (2013) The structural basis of antibody-antigen recognition. *Front. Immunol.*, **4**, 302.
2. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
3. Ferdous,S. and Martin,A.C.R. (2018) AbDb: antibody structure database-a database of PDB-derived antibody structures. *Database*, **2018**, bay040.
4. Swindells,M.B., Porter,C.T., Couch,M., Hurst,J., Abhinandan,K.R., Nielsen,J.H., Macindoe,G., Hetherington,J. and Martin,A.C. (2017) abYsis: integrated antibody sequence and structure-management, analysis, and prediction. *J. Mol. Biol.*, **429**, 356–364.
5. Ehrenmann,F., Kaas,Q. and Lefranc,M.P. (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.*, **38**, D301–D307.
6. Allcorn,L.C. and Martin,A.C. (2002) SACS—self-maintaining database of antibody crystal structure information. *Bioinformatics*, **18**, 175–181.
7. Dunbar,J., Krawczyk,K., Leem,J., Baker,T., Fuchs,A., Georges,G., Shi,J. and Deane,C.M. (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146.
8. Briney,B., Inderbitzin,A., Joyce,C. and Burton,D.R. (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, **566**, 393–397.
9. 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.
10. Vlachakis,D., Feidakis,C., Megalooikonomou,V. and Kossida,S. (2013) IMGT/Collier-de-Perles: a two-dimensional visualization tool for amino acid domain sequences. *Theoret. Biol. Med. Model.*, **10**, 14.
11. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
12. Lefranc,M.P., Pommie,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and 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.
13. Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
14. Viart,B., Dias-Lopes,C., Kozlova,E., Oliveira,C.F., Nguyen,C., Neshich,G., Chavez-Olortegui,C., Molina,F. and Felicori,L.F. (2016) EPI-peptide designer: a tool for designing peptide ligand libraries based on epitope-paratope interactions. *Bioinformatics*, **32**, 1462–1470.
15. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
16. Alamyar,E., Duroux,P., Lefranc,M.P. and Giudicelli,V. (2012) IMGT® 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.*, **882**, 569–604.
17. Dunbar,J. and Deane,C.M. (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**, 298–300.
18. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
19. Wu,X., Zhang,Z., Schramm,C.A., Joyce,M.G., Kwon,Y.D., Zhou,T., Sheng,Z., Zhang,B., O'Dell,S., McKee,K. *et al.* (2015) Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell*, **161**, 470–485.
20. Zhou,T., Lynch,R.M., Chen,L., Acharya,P., Wu,X., Doria-Rose,N.A., Joyce,M.G., Lingwood,D., Soto,C., Bailer,R.T. *et al.* (2015) Structural repertoire of HIV-1-Neutralizing antibodies targeting the CD4 supersite in 14 donors. *Cell*, **161**, 1280–1292.
21. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.