Nucleic Acids Research, 2006, Vol. 34, Database issue ***D777–D780***
doi:10.1093/nar/gkj053

Epitome: database of structure-inferred antigenic epitopes

Avner Schlessinger[1,2,3,*], Yanay Ofran[1,2], Guy Yachdav[1,2,3] and Burkhard Rost[1,2,3]

[1]CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St Nicholas Avenue room 804, New York, NY 10032, USA, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St Nicholas Avenue room 804, New York, NY 10032, USA and [3]NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St. Nicholas Avenue room 804, New York, NY 10032, USA

Received August 15, 2005; Revised September 23, 2005; Accepted October 4, 2005

ABSTRACT

Immunoglobulin molecules specifically recognize particular areas on the surface of proteins. These areas are commonly dubbed B-cell epitopes. The identification of epitopes in proteins is important both for the design of experiments and vaccines. Additionally, the interactions between epitopes and antibodies have often served as a model for protein–protein interactions. One of the main obstacles in creating a database of antigen–antibody interactions is the difficulty in distinguishing between antigenic and non-antigenic interactions. Antigenic interactions involve specific recognition sites on the antibody's surface, while non-antigenic interactions are between a protein and any other site on the antibody. To solve this problem, we performed a comparative analysis of all protein–antibody complexes for which structures have been experimentally determined. Additionally, we developed a semi-automated tool that identified the antigenic interactions within the known antigen–antibody complex structures. We compiled those interactions into Epitome, a database of structure-inferred antigenic residues in proteins. Epitome consists of all known antigen/antibody complex structures, a detailed description of the residues that are involved in the interactions, and their sequence/structure environments. Interactions can be visualized using an interface to Jmol. The database is available at http://www.rostlab.org/services/epitome/.

BACKGROUND

Protein–antigen structures

Antigen–antibody complexes have long been used as a model for understanding the general phenomenon of molecular recognition (1–5). The number of experimental high-resolution 3D structures of antibody–antigen complexes in the PDB (6) has significantly increased over the last years. Several groups have used these data to analyze and characterize antigenic interactions, i.e. interactions between the protein (the antigen) and the Complementarity Determining Regions (CDRs) of the antibody (7,8). An important first step in studying antigenic interactions is the characterization of CDRs. MacCallum *et al*. (8) observed that the hypervariable loops of CDRs adopt only a limited number of backbone conformations that are determined by a few key residues. Two recent studies have suggested that the amino acid composition and the length of CDRs determine the type of antigen that can be bound (9,10). Several studies have attempted to differentiate the residues on the antigen surface that are involved in the antigenic interaction from all others (5,7,11). The results of these studies were rather inconsistent. Differences in the data sets chosen (some of which were very small) and in the methodologies may explain some of those inconsistencies. Most importantly, however, the definitions of the CDRs often differed greatly, i.e. if two studies investigate the same PDB complex and use the same methodology, they might disagree on which of the interactions are antigenic (7). An important ramification of this problem was unveiled by Blythe and Flower (12), who showed that most existing B-cell epitope prediction methods do not work adequately. One explanation for this observation could be that most methods rely on inaccurate identifications of epitopes.

*To whom correspondence should be addressed. Tel: +1 212 851 4669; Fax: +1 212 305 7932; Email: as2067@columbia.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Definition of the CDRs

Antibodies are composed of a skeleton of beta-sheets. Most of the amazing variety of antibodies is realized by differences in six hypervariable loops of the CDRs. Therefore, the CDRs have previously been defined through these six loops. The first definition of CDRs was as regions in the Kabat sequence variability plot (13,14). The residues in these regions are identified through an alignment between the query sequence and a consensus motif for antibodies. Although widely used, the Kabat CDR-definitions can be problematic because CDRs that are in structural loops often have very unusual sequences that are not captured by regular sequence motifs (15). In fact, any method based only on sequence information is prone to mis-aligning and therefore mis-assigning loopy CDRs. Chothia and co-workers (16) therefore based their CDR identification on structural information. Initially, hypervariable loops were defined according to a few structures. Later, the numbering of the residues that was used to locate the CDRs was changed to account for structures that became available subsequently (17). Studies also differ in their definition of secondary structures, thereby increasing the inconsistency in defining hypervariable loops. Additional disadvantages of both the Kabat and Chothia *et al*. method are described elsewhere (http://www.bioinf.org.uk/abs/).

Here, we address these problems through a comprehensive study of all known antigen–antibody complexes in the PDB. Analyzing the structures, we identified the consensus residues on the antibodies and thereby identified the CDRs on all known protein–antibody complexes (details below). This initial set of CDRs facilitated the automatic generation of a database with all known antigenic residues in the PDB; we also included the sequence environment and a detailed description of the CDR with which they interact. Several databases of antibody–antigen complex structures are available (15,18,19). Some of these databases focus on the structural aspects of the interaction (19,20). There are also databases that compile B-cell epitopes without their corresponding antibodies (12,21). However, none of these databases explicitly locates the CDRs or identifies the antigenic residues semi-automatically. In this sense, our resource is more comprehensive and easily adjustable to growing data, as more 3D structures of antigen–antibody complexes become available. Thus, the databases mentioned above, particularly the ones that are not structure based, are complementary to Epitome.

DATABASE

Extraction of 3D structures and identification of CDRs

In order to identify all structures in the PDB that contain at least one antibody–antigen complex, we searched with BLAST (22) for a consensus sequence of an antibody against the PDB. The rationale for using BLAST rather than PSI-BLAST was to avoid capturing molecules such as T-cell receptors which, despite their similarity to antibodies, participate in cell-mediated immune response, and therefore represent a different type of antigenic interaction. We then added PDB structures that contain an immunoglobulin fold from the Structural Classification of Proteins database (SCOP) (23) and PDB entries that are identified as antibody–antigen complexes through keywords (e.g. 'antibody' and 'antigen'). We discarded all complexes with T-cell receptors or MHC molecules, since these are formed during cell-mediated immune response. We labeled residues as interacting if any of their respective atoms were within a sphere of ≤6Å (24). This resulted in our final list of interactions between antibodies and antigens. Thus, we define antibody–antigen interaction as spatial proximity between a residue within the CDRs and a residue on the surface of the antigenic protein.

We located the CDRs in the known protein–antibody complexes through the following knowledge-based approach. We began by creating multiple structure alignments of antibody structures using SKA (25,26). Since the light and heavy chains have different CDRs, two different multiple structure alignments were performed corresponding to each type of antibody chain. Additionally, due to the fact that our database included several redundant sequences, we ran the structural alignment program on a sequence-unique subset of all protein–antibody complexes. As antibody sequences are highly similar to each other, the criteria for the redundancy of the complex set was determined by the antigen sequences; sequence redundancy was reduced at HSSP-values of 0 (corresponding to <33% pairwise sequence identity for long alignments) (27–30). Then, we identified structurally aligned positions that interact with a protein in more than 10% of the complexes of the alignment. We defined the borders of the CDRs through those highly populated positions. Given the CDRs in the aligned antibodies, we transferred their location to the antibody chains of the corresponding sequence–structure family that they represent by structural pairwise alignments using Combinatorial Extension (CE) (31) (Figure 1). Finally, we defined all the residues on the protein surface that are in contact with the residues on the antibody CDRs as antigenic residues.

Content statistics

Epitome currently contains 142 antigens from protein–antibody complex structures with a current total of 10 180 antigenic interactions. A total of 63 of the complexes consist of antigens that are sequence-unique, i.e. 63 are such that no other antigen in the database has a level of sequence similarity

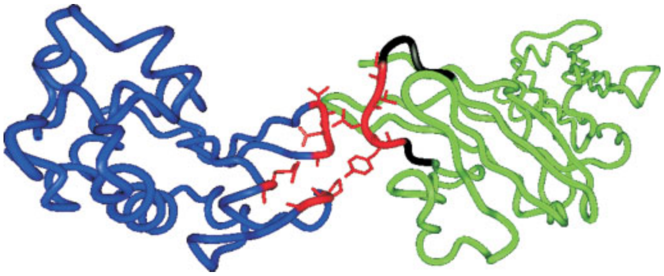
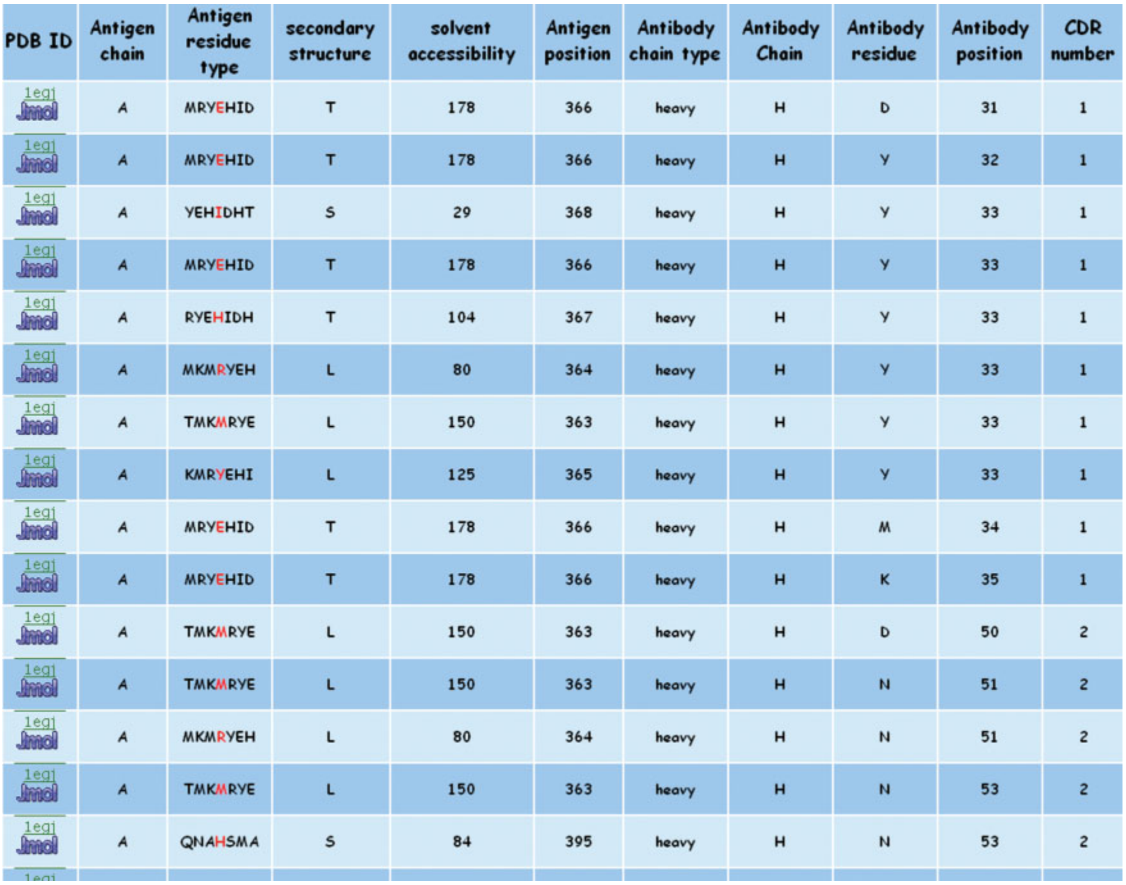


Figure 1. Antigenic residues according to Epitome. Complex structure of quail lysozyme (in blue) and the light chain of an antibody (in green), as taken from PDB ID 1bql (33). The residues that are defined to be in CDR 1 of the light chain according to Kabat definition (13) are colored in black. Residues in red are all the residues that are involved in the interaction according to Epitome. Note that not all of the residues on the antibody surface that are located on 'Kabat' CDR are involved in the antigenic reaction. Additionally, although 1bql antibody chains did not participate in the multiple structure alignment, i.e. the information about the location of the CDR was transferred from a homologous structure, the interaction was correctly identified.



PDB ID	Antigen chain	Antigen residue type	secondary structure	solvent accessibility	Antigen position	Antibody chain type	Antibody Chain	Antibody residue	Antibody position	CDR number
1egj Jmol	A	MRYEHID	T	178	366	heavy	H	D	31	1
1egj Jmol	A	MRYEHID	T	178	366	heavy	H	Y	32	1
1egj Jmol	A	YEHIDHT	S	29	368	heavy	H	Y	33	1
1egj Jmol	A	MRYEHID	T	178	366	heavy	H	Y	33	1
1egj Jmol	A	RYEHIDH	T	104	367	heavy	H	Y	33	1
1egj Jmol	A	MKMRYEH	L	80	364	heavy	H	Y	33	1
1egj Jmol	A	TMKMRYE	L	150	363	heavy	H	Y	33	1
1egj Jmol	A	KMRYEHI	L	125	365	heavy	H	Y	33	1
1egj Jmol	A	MRYEHID	T	178	366	heavy	H	M	34	1
1egj Jmol	A	MRYEHID	T	178	366	heavy	H	K	35	1
1egj Jmol	A	TMKMRYE	L	150	363	heavy	H	D	50	2
1egj Jmol	A	TMKMRYE	L	150	363	heavy	H	N	51	2
1egj Jmol	A	MKMRYEH	L	80	364	heavy	H	N	51	2
1egj Jmol	A	TMKMRYE	L	150	363	heavy	H	N	53	2
1egj Jmol	A	QNAHSMA	S	84	395	heavy	H	N	53	2
1egj										

Figure 2. Screenshot of a database entry. Each line of the table represents different antigenic interaction, i.e. interaction of a protein surface residue with an antibody surface residue that is located on one of the antibody's 6 CDRs. Note that the search could be performed using any of the table fields and that there is additional link to visualize the interaction using Jmol (http://jmol.sourceforge.net/).

to any other of the 63 that would enable coarse-grained homology modeling.

Input and fields

Epitome users can search for epitopes either by querying the database or by entering a sequence and 'BLASTing' for similar sequences that are stored in the database. The fields that can be queried include one or more of the following: PDB identifier (four-letter code used by the PDB, e.g. 1pdb); Antigen chain ID (PDB identifier for the chain of the antigen, e.g. 1pdb_C), antigen residue type (one letter code for amino acids, e.g. Y corresponds to Tyrosine), antigen residue secondary structure state as defined by DSSP (32) (1 letter code; GHI corresponds to helical structures, EB to strands and TSL to other), antigen residue solvent accessibility (the input is the accessible surface in Å^2 as defined by DSSP (32) and the search is on all residues with accessibility values that are bigger or equal to the input value), antigen residue position (the residue number as annotated in the PDB file), heavy/light chain (the interaction involves residues that are located either on the light or the heavy or both chains of the antibody), antibody chain identifier (similar to the antigen chain identifier), antibody residue type (one letter code for amino acids, e.g. C corresponds to Cysteine), antibody residue position in the PDB (the position of the antibody residue that is involved in the interaction as annotated by the PDB) and CDR number (possible values: 1, 2, 3).

Output

Results for database queries are presented as a table that lists all features of the result sets (Figure 2). The antigen results include the residues in the environment of the antigen (highlighted in red). If a user performs a BLAST sequence search against the Epitome database to find PDB structures containing antigens with similar sequences, the output will be all complex structures consisting of proteins with high degree of similarity to the input sequence, the corresponding *E*-value and BLAST score of the pairwise sequence alignments. Additionally, each PSI-BLAST hit contains a link that can trigger another database query.

Updates

Since most Epitome entries were identified using the SCOP database, Epitome updates will follow updates of SCOP, i.e. Epitome will be updated twice a year as soon as SCOP updates its parseable files. Additionally, all the other programs used to create the database are installed locally and can be run automatically.

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu (Columbia) for computer assistance and to Andrew Kernytsky and Henry Bigelow for helpful comments on the manuscript. Thanks also to the anonymous reviewer for an immensely supportive, helpful and enjoyable critique. This work was supported by the grants RO1-GM64633-01 from the National Institutes of Health (NIH), and RO1-LM07329-01 from the National Library of Medicine (NLM). Last, not least, thanks to Helen Berman (Rutgers), Phil Bourne (UCSD) and their crews for maintaining an excellent PDB, and to all experimentalists who enabled this analysis by making their data publicly available. Funding to pay the Open Access publication charges for this article was provided by the National Library of Medicine (NLM).

Conflict of interest statement. None declared.

REFERENCES

1. Jones,S. and Thornton,J.M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.
2. Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.
3. Chen,R., Mintseris,J., Janin,J. and Weng,Z. (2003) A protein–protein docking benchmark. *Proteins*, **52**, 88–91.
4. Jones,S. and Thornton,J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.
5. Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
6. Berman,H.M., Westbrook,J., Feng,Z., Gillliland,G., Bhat,T.N. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
7. Davies,D.R. and Cohen,G.H. (1996) Interactions of protein antigens with antibodies. *Proc. Natl Acad. Sci. USA*, **93**, 7–12.
8. MacCallum,R.M., Martin,A.C. and Thornton,J.M. (1996) Antibody–antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.*, **262**, 732–745.
9. Collis,A.V., Brouwer,A.P. and Martin,A.C. (2003) Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J. Mol. Biol.*, **325**, 337–354.
10. Almagro,J.C. (2004) Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit.*, **17**, 132–143.
11. Van Regenmortel,M.H.V. (1992) *Structure of Antigens*. CRC Press, Inc., 2000 Corporate Blvd, N.W., Boca Raton, Florida 33431.
12. Blythe,M.J. and Flower,D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.*, **14**, 246–248.
13. Wu,T.T. and Kabat,E.A. (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. Johnson,G. and Wu,T.T. (2000) Kabat Database and its applications: 30 years after the first variability plot. *Nucleic Acid Res.*, **28**, 214–218.
15. Allcorn,L.C. and Martin,A.C. (2002) SACS—self-maintaining database of antibody crystal structure information. *Bioinformatics*, **18**, 175–181.
16. Chothia,C., Lesk,A.M., Tramontano,A., Levitt,M., Smith-Gill,S.J., Air,G., Sheriff,S., Padlan,E.A., Davies,D. and Tulip,W.R. (1989) Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883.
17. Al-Lazikani,B., Lesk,A.M. and Chothia,C. (1997) Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.*, **273**, 927–948.
18. Saha,S., Bhasin,M. and Raghava,G.P. (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics*, **6**, 79.
19. Peters,B., Sidney,J., Bourne,P., Bui,H.H., Buus,S., Doh,G., Fleri,W., Kronenberg,M., Kubo,R., Lund,O. *et al.* (2005) The design and implementation of the immune epitope database and analysis resource. *Immunogenetics*, **57**, 326–336.
20. Kaas,Q., Ruiz,M. and Lefranc,M.P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res.*, **32**, D208–D210.
21. McSparron,H., Blythe,M.J., Zygouri,C., Doytchinova,I.A. and Flower,D.R. (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci.*, **43**, 1276–1287.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
24. Ofran,Y. and Rost,B. (2003) Analysing six types of protein–protein interfaces. *J. Mol. Biol.*, **325**, 377–387.
25. Petrey,D., Xiang,Z., Tang,C.L., Xie,L., Gimpelev,M., Mitros,T., Soto,C.S., Goldsmith-Fischman,S., Kernytsky,A., Schlessinger,A. *et al.* (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53** (Suppl. 6), 430–435.
26. Petrey,D. and Honig,B. (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.*, **374**, 492–509.
27. Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acid Res.*, **31**, 3789–3791.
28. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
29. Sander,C.S.R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
30. Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
31. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
32. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **12**, 2577–2637.
33. Chacko,S., Silverton,E.W., Smith-Gill,S.J., Davies,D.R., Shick,K.A., Xavier,K.A., Willson,R.C., Jeffrey,P.D., Chang,C.Y., Sieker,L.C. *et al.* (1996) Refined structures of bobwhite quail lysozyme uncomplexed and complexed with the HyHEL-5 Fab fragment. *Proteins*, **26**, 55–65.