Molecular
BioSystems

Cite this: *Mol. BioSyst.*, 2011, **7**, 3327–3334

www.rsc.org/molecularbiosystems

PAPER

Predicting antibody complementarity determining region structures without classification†

Yoonjoo Choi and Charlotte M. Deane*

Received 6th June 2011, Accepted 21st September 2011
DOI: 10.1039/c1mb05223c

Antibodies are used extensively in medical and biological research. Their complementarity determining regions (CDRs) define the majority of their antigen binding functionality. CDR structures have been intensively studied and classified (canonical structures). Here we show that CDR structure prediction is no different from the standard loop structure prediction problem and predict them without classification. FREAD, a successful database loop prediction technique, is able to produce accurate predictions for all CDR loops (0.81, 0.42, 0.96, 0.98, 0.88 and 2.25 Å RMSD for CDR-L1 to CDR-H3). In order to overcome the relatively poor predictions of CDR-H3, we developed two variants of FREAD, one focused on sequence similarity (FREAD-S) and another which includes contact information (ConFREAD). Both of the methods improve accuracy for CDR-H3 to 1.34 Å and 1.23 Å respectively. The FREAD variants are also tested on homology models and compared to RosettaAntibody (CDR-H3 prediction on models: 1.98 and 2.62 Å for ConFREAD and RosettaAntibody respectively). CDRs are known to change their structural conformations upon binding the antigen. Traditional CDR classifications are based on sequence similarity and do not account for such environment changes. Using a set of antigen-free and antigen-bound structures, we compared our FREAD variants. ConFREAD which includes contact information successfully discriminates the bound and unbound CDR structures and achieves an accuracy of 1.35 Å for bound structures of CDR-H3.

1 Introduction

The Protein Data Bank (PDB)[1] currently contains the structures of about 750 immunoglobulins. This enables good models to be created for the majority of any antibody structure *via* homology modelling (framework region predicted to be ~1.2 Å RMSD on average).[2] Due to the high variability of the CDRs, these regions are predicted far less accurately.[2–4] CDRs, however, are particularly important as they are the major contributors to the binding of the antigen along with the relative orientation of the antibody light and heavy chains.[5–7] Predicting the conformations of CDRs accurately is therefore vital for an understanding of antibody–antigen complexes and has increased in importance with the rise of therapeutic antibodies in healthcare.[8] The treatment of many diseases has benefited from the use of therapeutic antibody drugs[9] and homology modelling has already played an important role in the development of several such drugs.[10]

Despite the high sequence diversity of CDR loops, five of the six CDRs (L1, L2, L3, H1 and H2) are thought to have a set of limited structural conformations (canonical structures).[11] Reasonably accurate predictions can be made for these five non-CDR-H3 loops using a set of sequence based canonical rules.[3,4,12,13] More recently, the canonical structures have been updated and it was shown that non-CDR-H3 loops are largely predictable (estimated 85%) using sequence, gene source and framework regions.[14]

There have been several efforts to identify similar sequence rules for CDR-H3.[15–17] However, no definitive canonical rules for all CDR-H3 loops have been found. Furthermore, there are many examples where different structural conformations and side chain arrangements of CDR-H3 loops occur because of their corresponding antigens.[18–24] Some CDR-H3 loops in fact take on different structural conformations dependent on which antigen or which part of an antigen they bind.[25,26]

Much of the previous work on CDR structure prediction has been dependent upon the canonical rules. Using these rules, candidate CDR loops are selected from a database of CDR loop structures and grafted onto the rest of the antibody structure. If the query loop is not identifiable by the rules then other strategies such as *ab initio* methods take over.[27–30]

Extensive benchmarks of CDR prediction have not been performed. Most methods have been tested on a small number of structures and concentrated on CDR-H3 loops. For example, ABGEN[29] was tested on a set of 15 antibody structures and

Department of Statistics, Oxford University, 1 South Parks Road, Oxford, OX1 3TG, UK. E-mail: deane@stats.ox.ac.uk; Fax: +44 (0)1865 272595; Tel: +44 (0)1865 281301

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c1mb05223c

the CDR-H3 loops were predicted within the range of 0.98–4.06 Å RMSD. WAM[30] gave relatively accurate predictions of short CDR-loops (up to nine residues ≤ 1.7 Å), but inaccurate results for longer loops (19 test structures). Marcatili *et al.* (2008) tested four examples and their results varied from 1.66–3.06 Å for CDR-H3 loops of between 9 and 13 residues in length.[31] A recent method RosettaAntibody was tested on a set of 54 antibody structures.[2] For non-CDR-H3 loops, fragments were selected using BLAST[32] from a database of antibody structures and grafted onto the framework. For CDR-H3 loops, the *ab initio* Rosetta protocol[33,34] was used, but with a database of fragments expanded to contain 230 antibody structures. Kinked CDR-H3 loops were specifically identified with sequence based rules[17] and predicted using a special fragment library containing kinked conformations. The method gave reasonably accurate results for CDR-H3 loops (on average 2.91 Å RMSD for all backbone atoms in model structures and 2.15 Å in native structures).

In early studies, there were attempts to predict CDR structures using non-antibody structures.[35,36] As an extension of the idea, here we assume that the problems of the classification and prediction of antibody CDRs are in principle no different from those of loops whose anchor regions are anti-parallel β-sheets in soluble proteins. Here we demonstrate that CDRs can be predicted without prior classifications or knowledge of the gene source. We use FREAD,[37] a database search protein loop structure prediction method, that was shown to be an effective loop modelling tool outperforming MODELLER,[38] RAPPER[39] and PLOP.[40]

The structural conformations of the CDR loops and the framework regions are known to be affected by their external environment[41,42] even without antigens.[43,44] Although many CDRs may be predictable using sequence rules alone, these may not be sufficient to capture structural conformations in a modelling situation due to the high structural variability of CDR loops and its dependence upon surroundings. As FREAD predicts loops using only local similarities (local sequence and geometrical matches) and does not take into account such environmental effects, we developed a contact profile extension.

FREAD was initially tested on 97 non-redundant test structures. It produced accurate predictions for non-CDR-H3 loops using a database containing only antibody structures (0.81 Å (L1), 0.42 Å (L2), 0.96 Å (L3), 0.98 Å (H1) and 0.88 Å (H2) on average), but relatively less accurate predictions for CDR-H3 loops (2.25 Å on average). In order to overcome the relatively poor predictions of CDR-H3, we adapted FREAD to take greater account of sequence (FREAD-S). This improved our results for CDR-H3 (1.38 Å on average), but not for other CDRs, suggesting that CDR-H3 is the most sequence dependent of the CDRs. ConFREAD (FREAD-S with contact information) showed the best performance among the FREAD variants. However it caused a loss of prediction (for CDR-H3, 1.23 Å on average with 70% coverage).

The FREAD variants were compared to RosettaAntibody on model structures. All the methods showed accurate results in non-CDR-H3 predictions (about 1 Å on average). CDR-H3 prediction is once again less accurate (3.12 and 2.91 Å for FREAD and RosettaAntibody respectively). FREAD-S and ConFREAD both also improve prediction for CDR-H3 loops on models (2.07 and 2.91 Å for FREAD-S and RosettaAntibody).

Finally, in order to test the generic applicability, we modelled antigen-bound antibody structures using their corresponding antigen-free structures, computationally docking the antigen and then predicting the CDRs using FREAD. In this case, sequence only based rules may not discriminate which sampled fragments are best. In this test, for all CDR loops, ConFREAD gives the most accurate results among the FREAD variants (2.56, 2.61, 3.5 and 1.35 Å by FREAD, FREAD-S and ConFREAD, respectively, for native free against native bound structures in the same CDR-H3 subset).

2 Results and discussion

2.1 Using FREAD to predict CDRs on native structures

Initially FREAD was used to predict the CDRs on a large non-redundant set of 97 structures (native set, for more details see Materials and methods) using DB-I (the database including antibody structures).

FREAD was able to predict all the cases in the native set (Fig. 1A and C). As can be seen in Fig. 1A, FREAD produced results of similar accuracy to those previously reported by Sivasubramanian *et al.* ((2009) (0.81 Å (L1), 0.42 Å (L2), 0.96 Å (L3), 0.98 Å (H1), 0.88 Å (H2) and 2.25 Å (H3)). FREAD is not using any antibody specific knowledge such as conserved residues or structural classes to find matched fragments. The environment specific substitution score[37,45,46] (ESSS, the main selection method in FREAD), which was developed for general loop structure prediction appears to be able to accurately identify near native fragments for CDRs.

2.2 Investigating the applicability of the FREAD sequence score

It is noticeable that FREAD, like all methods, predicts CDR-H3 poorly, on average twice as badly as any of the other CDRs. In standard FREAD the first-ranked prediction is taken as the fragment with the lowest anchor RMSD among the predicted fragments which have a high sequence score (ESSS over 25). However, CDR loops are known to have high sequence-specificity and single mutations can cause changes in antigen affinity and structural conformation.[47] A second version of FREAD was therefore tested, FREAD-S, which selects the fragment with the highest ESSS as the first-ranked prediction. Fig. 1A shows that this method improved the accuracy of CDR-H3 prediction (2.25 → 1.38 Å). However, such improvements were not seen for the other CDR loops.

2.3 Contact profile and ConFREAD

As antibody CDRs change their structural conformations upon binding, it may be possible to use contacts between the antigen and the antibody and between different parts of the antibody to improve prediction. The contact profile describes the contacts of a fragment of protein structure with one value for each residue in the fragment. Each residue in the CDR fragment is annotated between 0 and 2 dependent on the types of contacts that a residue has (for full details of the contact profile, see Section 4.3). For example, the different contact

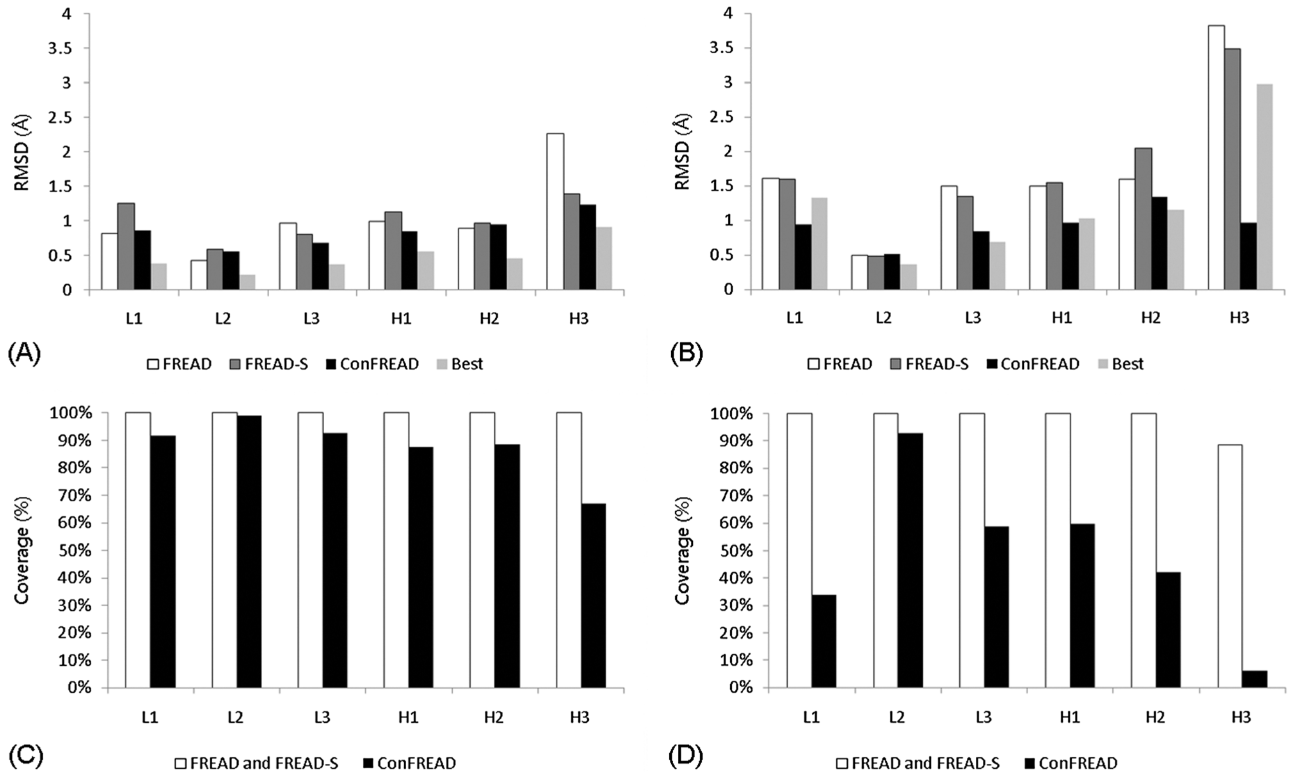

Fig. 1 The results of CDR prediction on the native set. (A) Average global RMSD (DB-I), (B) average global RMSD (DB-E), (C) coverage (DB-I) and (D) coverage (DB-E). Best refers to the average RMSD of the lowest RMSD structures FREAD produced. The detailed results including standard deviations are in Tables S3 and S4 (ESI†).

types can be seen in the CDR-H3 loops of chains B and H of 1XF2, an antibody–DNA complex (Fig. S3, ESI†). Chain H is actively involved in antigen binding while chain B is not. The two CDR-H3s of these chains share 100% sequence identity (VRGGYRPYYAMDY) whereas their contact profiles (2222100111212 and 2222012111212 for chains B and H respectively) share 76.9% contact identity. The structural distortion occurs where the contact information is different.

As shown in Fig. 1A, using ConFREAD slightly improved prediction over FREAD-S and FREAD. However, it gives lower prediction coverage. For example, in the case of CDR-H3 it drops from 100% (FREAD-S) to 70% (ConFREAD).

2.4 Predicting CDRs using non-antibody structures

If CDRs are no different from general loops, it may be possible to predict CDRs using non-antibody protein structures. The database excluding antibody structures (DB-E) was built and used to predict the CDRs of the native set.

FREAD and FREAD-S were able to predict most CDRs to a reasonable accuracy, but both perform worse than using a database containing antibody fragments (Fig. 1B). In the case of CDR-H3, coverage is no longer 100% and accuracy has dropped (2.25 → 3.81 Å for FREAD and 1.38 → 3.48 Å for FREAD-S).

In the case of ConFREAD, with DB-E, accurate results are achieved but coverage drops substantially (Fig. 1B and D). ConFREAD's lack of coverage but highly accurate predictions for CDR-H3 using a database without antibody structures (6.1% coverage, 0.66 Å average RMSD) may indicate that contacts are highly important for CDR-H3 shape and that such contacts are not seen in non-antibody structures.

It should be noted that all the fragments found for the CDR-H3 prediction using DB-E are from antibody-related structures (3CFB: 3EFD, 2W9D: 3H33, 1NLB: 3LS5, 1PKQ: 3GO1, 3EYQ: 3LS5, 2IPU: 3EYU). It is noticeable that non-CDR-H3 predictions by ConFREAD are in the same accuracy range as those using (DB-I : DB-E 0.86 : 0.93 Å, 0.55 : 0.57 Å, 0.68 : 0.83 Å, 0.84 : 0.96 Å, 0.94 : 1.33 Å for CDR-L1, L2, L3, H1 and H2 respectively).

2.5 Predicting CDRs on model structures and comparison to RosettaAntibody

The FREAD variants were also compared to the recently published predictor RosettaAntibody. In order to demonstrate that FREAD is able to make predictions in a near-true modelling situation, CDR-H3 loops were predicted on the model set used in RosettaAntibody (the RA set). In this case, surroundings (such as antigens) were estimated by superimposing the model structure excluding the CDR loops onto the native structures.

FREAD and RosettaAntibody achieved similar accuracy on all CDR loops on the RA-native set (Fig. 2A and C). As seen previously on the native set, FREAD-S did not improve the prediction for non-CDR-H3 loops. However, FREAD-S did lead to significant improvement in CDR-H3 prediction (1.85 → 1.52 Å).

On the RA-model set, for CDR-H3, FREAD-S gave 98% coverage and an RMSD 2.07 Å compared to 2.91 Å for

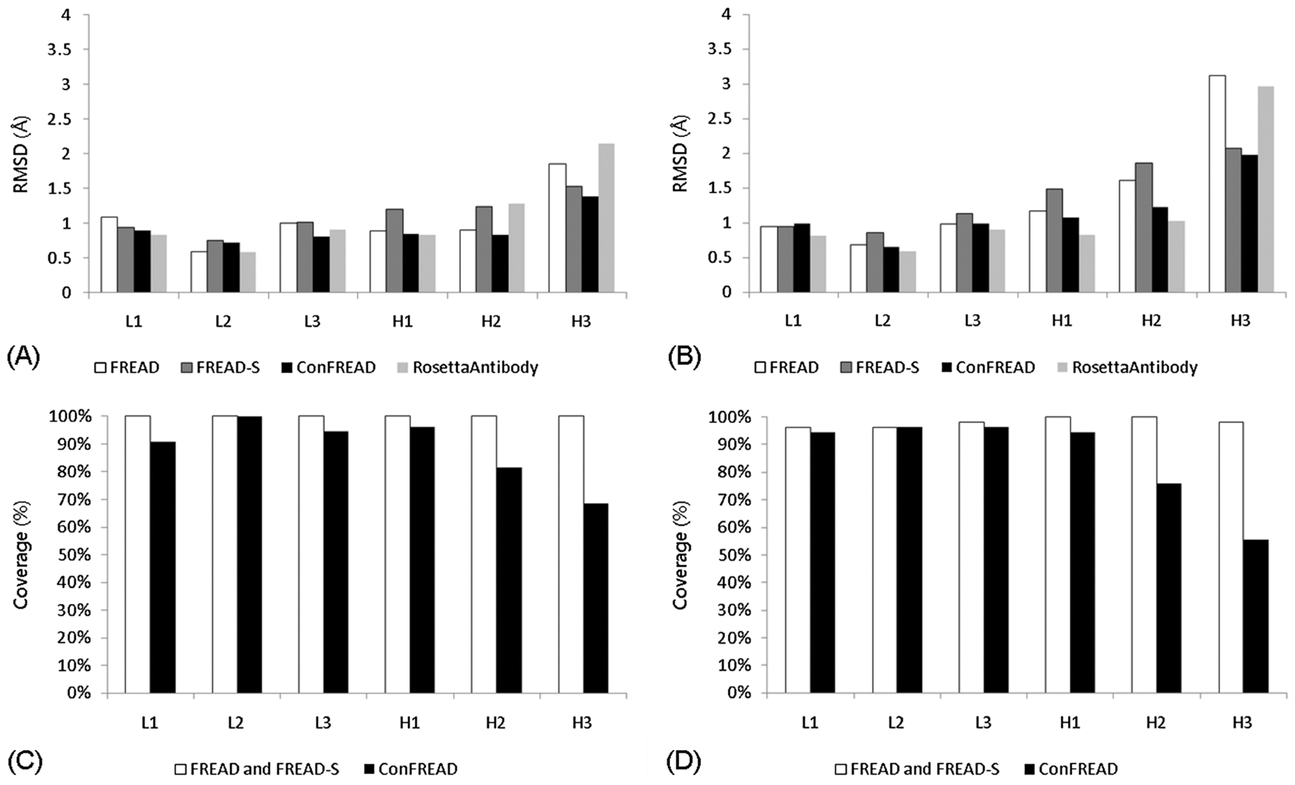

Fig. 2 The results of the FREAD variants compared to RosettaAntibody on the RA sets. (A) Average global RMSD on the RA-native set, (B) average global RMSD on the RA-model set, (C) coverage (RA-native set), (D) coverage (RA-model set). The detailed results including standard deviations are in Tables S5 and S6 (ESI†).

RosettaAntibody in the same subset (Fig. 2B and D). This level of accuracy on modelled structures with an approximate framework is encouraging. ConFREAD once again improves prediction accuracy with a drop in coverage. The improvement in accuracy is however only marginal (Tables S7 and S8 (ESI†) for detailed results).

2.6 Predicting CDRs of antigen-bound structures using antigen-free structures as a template

In the bound–free set, the antigen-free structure was used as a template to predict its CDR loops in a bound form. In each case the two structures share 100% sequence identity but in one no antigen is bound but in the other an antigen is. In this set, predicting CDRs using only sequence based approaches may not be successful. ZDOCK was run on the free structure where all the residues apart from the CDR residues were blocked to bind. Ten thousand antigen positions were generated and the best placed antigen was chosen.

Here we are examining only the current limit of antibody–antigen docking sampling. The purpose of this study is not benchmarking a docking method, but investigating structural transitions of local CDR regions upon environment changes and how one can capture the structural variability. Note that only ConFREAD makes use of the antigen position.

The FREAD variants were run on this free structure with the docked antigen. It was assumed that the bound CDR forms are unknown. Hence, the self-predictions from both free and bound structures were eliminated (otherwise FREAD would give the free structure as the top prediction for any CDR of the bound structure).

For all CDRs, ConFREAD gave the best prediction on average (Fig. S4 and Table S9, ESI†). Again, the most significant improvement can be seen in CDR-H3 prediction. For non-CDR-H3 loops, FREAD showed similar accuracies to native CDR loops of the free structures. However, for CDR-H3, although the coverage decreased, ConFREAD successfully discriminated bad fragments and gave better results than using the native free CDRs.

Although the antigens were not correctly placed, ConF-READ was still able to give relatively accurate predictions (Table 1 and Fig. 3). This is probably due to the roughness of the contact profile method. The contact profile is given only by atomic distances and does not depend on exact antigen positions.

3 Discussion

CDR structure prediction has been reasonably successful for five of the six CDR loops in antibodies, through the use of the canonical rules. The sixth CDR, CDR-H3 has in general proved more challenging and is predicted less accurately. We approach CDR loop structure prediction as a special version of the general loop structure prediction problem using FREAD, a database search method. On a non-redundant set of 97 antibody structures, we show that CDRs can be predicted successfully using FREAD. However, CDR-H3 loops are predicted least accurately. In order to overcome

Table 1 The prediction results of the bound–free set for CDR-H3. Free–bound refers to the difference between the native antigen free CDR-H3 and its native bound counterpart. The second last column is for the best fragments found using FREAD and the last column shows differences between the centroids of the native antigen and the docked antigen. For the results on non-CDR-H3 loops, see Fig. S4 and Table S9 (ESI†)

Free	Bound	Free–bound	FREAD	FREAD-S	ConFREAD	Best	Antigen position
1NGZ	1N7M	2.65	1.20	5.63	1.50	0.71	21.66
1D5I	1D6V	1.95	5.22	1.63	0.73	0.73	28.06
2A6J	2A6I	1.05	4.28	2.18	—	1.39	0.74
1Q9K	1Q9Q	4.09	0.60	4.37	1.21	0.60	25.54
1KCV	1KCS	1.99	1.66	0.63	0.63	0.63	0.89
1CR9	1CU4	2.50	2.50	2.95	—	1.35	0.57
1GGC	1GGI	2.24	2.88	0.60	—	0.60	1.74
1CGS	2CGR	2.79	5.59	3.17	—	2.19	24.93
1NBV	1CBV	1.76	3.28	2.64	—	2.25	24.82
1HIL	1IFH	3.12	3.08	3.08	3.08	2.50	0.64
1OAQ	1OAU	1.59	3.93	5.65	0.92	0.36	0.71
1MNU	1MPA	2.43	8.66	2.39	—	2.39	1.54
Mean		2.35	3.57	2.91	1.35	1.31	
STD		0.79	2.20	1.66	0.91	0.82	

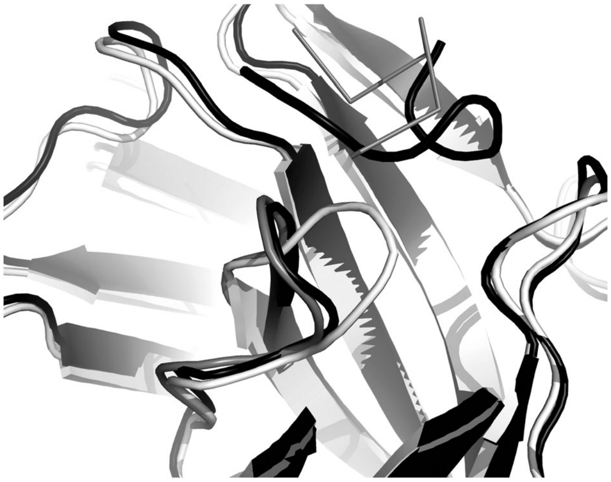

Fig. 3 Predicting CDR-H3 using a docked antigen. The native antigen and bound CDR-H3 of the bound structure (1KCS) are black and its free structure (1KCV) is white (1.99 Å). Their contact profiles are 223311012 and 222211012 respectively. The thin grey stick structure is the docked antigen and the prediction made by ConFREAD is grey (0.63 Å).

this, we modified FREAD (FREAD-S) to focus more on sequence similarity in prediction. FREAD-S improved prediction for CDR-H3, but not for the other CDR loops.

CDR loops are supposed to be actively involved in antigen interactions and sometimes interact with other parts of the antibody structure. CDR-H3 can also take on multiple structural conformations dependent on its antigen. In fact, CDR-H3 and CDR-L3 are both involved more in external contacts than the others. About 40% of those CDR types have external contacts whereas this is only about 10% in the other CDRs (Fig. S5, ESI†). However CDR-L3's contacts do not change much upon antigen binding.

In this paper, we have examined different FREAD variants on different sets and situations under the assumption that CDR prediction is no different from general loop structure prediction. In any test set, non-CDR-H3 loops are accurately predicted using any FREAD variants and the contact filter does not show significant improvements in prediction. This may be due to the contact compositions of non-CDR-H3 loops. CDR-L1, L2, H1 and H2 generally have internal contacts or no contacts and the contact pattern of CDR-L3 is conserved upon antigen binding.

Most CDRs can be predicted using local similarities such as local sequence and geometrical matches. The conservation of contact patterns in terms of antigen binding and the fact that non-CDR-H3 loops have few external contacts mean that most of their structural conformations can be captured using sequence rules. However, in the case of CDR-H3, the changes in contact patterns upon antigen binding and the wide length variation make classification and prediction difficult.

This method, like all modelling procedures, suffers from limitations. In particular, FREAD is unable to offer 100% coverage (something which should improve as the PDB increases in size) and although average accuracies for CDRs are often within the range of 1 Å, the standard deviations are of the same magnitude. These results were also achieved on static framework regions, though in some cases only approximate ones such as the Rosetta models. Overall, however, the results demonstrate how a fragment-based database search method can predict accurately the conformations of CDRs without recourse to the canonical rules.

4 Materials and methods

4.1 Test sets and CDR definition

A total of 2009 antibody and antibody-related structures were collected using the union of the data in IMGT (Ver. 4.3.0),[48] immunoglobulin superfamily in SCOP (Ver. 1.75)[49] and immunoglobulin homologous superfamily (2.60.40.10) in CATH (Ver. 3.3).[50] This union set was used to create a FREAD database excluding potential antibody structures (see section 4.4). Since the union set may contain non-immunoglobulins such as T-cell receptors, for test sets, only the immunoglobulin structures which contain both heavy and light chains and CDRs annotated by IMGT were kept leaving 588 immunoglobulin structures.

- Native set: a non-redundant set created from the above structures. All the antibody structures in the set share less than 80% sequence identity in their variable regions. If any missing

Table 2 The RMSD between CDRs of the native antigen-bound and free structures in the bound–free set after superimposing the framework regions. The most structurally different CDRs in each pair are in bold. The largest structure changes upon antigen binding occur in CDR-H loops. For the contact pattern changes upon binding, see Fig. S5 (ESI†)

Free	Bound	L1	L2	L3	H1	H2	H3
1NGZ	1N7M	2.35	1.28	2.30	1.61	1.77	**2.65**
1D5I	1D6V	0.27	0.39	0.37	0.39	**1.99**	1.95
2A6J	2A6I	1.20	1.05	0.95	0.68	**2.17**	1.05
1Q9K	1Q9Q	1.70	0.77	0.46	0.81	0.42	**4.09**
1KCV	1KCS	0.34	0.26	0.93	0.89	0.83	**1.99**
1CR9	1CU4	0.56	0.43	0.64	1.55	1.09	**2.50**
1GGC	1GGI	1.43	1.48	1.28	2.03	1.18	**2.24**
1CGS	2CGR	5.25	2.53	3.92	**5.63**	4.30	2.79
1NBV	1CBV	1.25	0.73	1.02	**2.34**	1.65	1.76
1HIL	1IFH	1.59	1.34	0.57	**3.27**	2.27	3.12
1OAQ	1OAU	0.35	0.14	0.86	0.18	0.23	**1.59**
1MNU	1MPA	**2.43**	1.28	1.29	0.84	0.93	2.43

residues were found in any of the CDRs in the structure, it was discarded. This left 97 non-redundant antibody structures (56 antigen-bound structures and 41 free structures) which contain CDR-L1 through to H3 (the full list is given in Table S1, ESI†).

- RA set: the CDR loops from the 54 structures used to test RosettaAntibody.[2] This set was used to compare FREAD to RosettaAntibody and test its prediction ability on models. Two sets were prepared; RA-native set (54 native structures) and RA-model set (54 homology modelled structures) (the full list is given in Table S2, ESI†).
- Bound–free set: taken from Babor and Kortemme (2009), 12 pairs of antibody structures. The two sequences in each pair are 100% identical, but in one case, the structure contains an antigen and in the other it does not. The antigen-free structure is used as a model for the antigen-bound structure. The CDR conformations found in the bound structure were predicted using only the coordinates from the antigen-free structure. The antigen is positioned in the antigen-free structure by docking (see Section 4.5).

In the original Babor and Kortemme set, there are 14 pairs (the full list with RMSD differences is given in Table 2).[51] Here we exclude two of them. The 1AJ7–2RCS pair where the framework is very different between the two (C_α RMSD $>$ 3 Å) and the 1FL5–1FL6 pair where residues are missing in the CDRs.

CDRs are defined as given in IMGT for both the native and bound–free set. In the RA sets, the Chothia numbering scheme was used as it was used by RosettaAntibody in the original paper. In this case, the CDR loops of the light chain are composed of residues 24–34 (L1), 50–56 (L2) and 89–97 (L3), and the heavy chain CDR loops are composed of residues 26–35 (H1), 50–56 (H2) and 95–102 (H3).

4.2 Measurement of accuracy

The measure of accuracy used is global loop RMSD, calculated by superimposing entire structures except for the CDR loop regions and then calculating the RMSD of the loop main chain atoms (nitrogen, α-carbon, carbon and oxygen). The predictions are those ranked first by the modelling method. The method was tested using a leave one out cross validation so in all cases results from self-prediction were eliminated. Additionally, any self-predictions from both bound and free structures were removed in the bound–free set.

4.3 Contact profile

A contact residue is defined in terms of the distance between a pair of C_β (C_α for glycine) atoms (including antigens if present). The atoms considered are from the same chain, different chains and the antigen. Within the same chain, ten residues at either side of the residues of interest are excluded from the contact calculation. The antigen can be a protein, peptide or hapten. If the antigen is a hapten, all heavy atoms in the hapten are considered in the contact calculation. This relatively simple definition is easy to calculate and does not depend on the quality of sidechain modelling. There are four possible contact types; (1) non-contact represented as "0", (2) external contact (such as external chains or antigen) only "1", (3) internal chain contact only "2" and (4) both external and internal contacts "3" (see an example in Fig. S1, ESI†). The contact profiles of two loops can therefore be compared independently of the sequence.

Contact profile information is available from three different sources.

(a) The actual contact profile of the correct target fragment can be calculated from the full target structure; we term this the target contact profile (unavailable in a modelling situation).

(b) The contact profile of a database fragment in its original structure termed the database contact profile.

(c) The contact profile calculated when a database fragment is grafted into the target structure termed the predicted contact profile.

Here we use only the database contact profile and the predicted contact profile both of which would be available in a modelling situation. Therefore the prediction made is entirely blind as to the target contact profile.

4.4 FREAD

4.4.1 FREAD algorithm. FREAD[37,52] is a database search method for loop structure prediction which uses four main filters. First, a fast database search is performed using anchor C_α separations (compare the C_α separations of two residues at either side of the loop in the target to the equivalent 4 residues in the database fragments). Second, an environment specific substitution score (ESSS) filters out fragments that have low sequence similarities to the target loop. The score table is a quantitative measure of the probability that an amino acid is substituted for another amino acid in a certain environment.[45,53] Third, a statistical energy function[54] removes physically impossible/implausible loop fragments. Finally the anchor RMSD, the RMSD between the target and the database structures for two residues at either side of the loop, is calculated and used to rank the predictions.

4.4.2 FREAD databases. Two separate databases were built.

- DB-I, which contains all the chains from the 2009 structures (identified as antibody and antibody-related).
- DB-E, which contains proteins with sequence identity $\leq$ 99%, resolution $\leq$ 3 Å and *R*-factor $\leq$ 0.3. All the chains

in DB-I are eliminated leaving 28 099 PDB chains (a structure database excluding antibody and antibody-related structures).

4.4.3 Selection procedure and FREAD variants. In the standard FREAD protocol, the maximum anchor RMSD cut-off value between the target and database anchors (two residues at each loop terminal) is set at 0.7 Å and the ESSS cut-off is fixed at 25. All putative database fragments within these limits are then sorted and the one with the lowest anchor RMSD is selected as the first-ranked prediction. If FREAD is unable to identify a match of the same length within the database, it extends its search space by increasing the length of the loop region by two (one at the N and one at the C termini) until a prediction is made (stopped if length 26 reached). This extension is possible as FREAD has been shown to have relatively length independent accuracy.

Here we test a new variant of FREAD: FREAD-S. As described above, FREAD normally selects its first-ranked prediction according to anchor RMSD; in FREAD-S the list is ordered instead by the ESSS. This puts far more weight on the sequence component of the scoring system.

ConFREAD is an extended version of FREAD-S, which makes use of contact information. It acts as a filter on the database fragments in the prediction list given by FREAD. ConFREAD operates by each FREAD prediction in turn (within the anchor RMSD and ESSS cut-off) and calculating both their database contact profile (the contact profile of the database fragment in the structure it was taken from) and their predicted contact profile (the profile of this database fragment when it is grafted onto the target structure). ConFREAD predicts only fragments which share 100% contact identity.

4.5 Idealised docking

Twelve pairs of structures which share 100% sequence identity were chosen, but in each pair one member is bound to an antigen and the other is not (the bound–free set). The antigen-free structures are taken as initial models and their corresponding antigens (taken from their counterparts) are docked using ZDOCK.[55] All the residues apart from the CDRs were blocked. Ten thousand antigen locations were generated and the best docked antigen, that closest to the native antigen position, was taken. After the docking step, ConFREAD was run on the antigen-free structures with the docked antigen. Self-predictions (fragments from both antigen-free and antigen bound structures) were discarded. The test protocol is outlined in Fig. S2, ESI.†

It should be noted that the aim of this study is not to test the docking method. The best docked antigen positions are taken as limitations of current protein docking programmes. The primary goal is to examine whether ConFREAD can produce accurate results even if the framework is approximate and the antigen is not exactly positioned.

Acknowledgements

The authors would like to thank Professor J. Gray and his colleagues at Johns Hopkins University for their data. The authors also give special thanks to Dr Y. Jang of Boncho Korean medical clinic for insightful discussions. Yoonjoo Choi is funded by Department of Statistics, St. Cross College and Oxford University.

References

1 S. Dutta, K. Burkhardt, J. Young, G. J. Swaminathan, T. Matsuura, K. Henrick, H. Nakamura and H. M. Berman, *Mol. Biotechnol.*, 2009, **42**, 1–13.
2 A. Sivasubramanian, A. Sircar, S. Chaudhury and J. Gray, *Proteins: Struct., Funct., Genet.*, 2009, **74**, 497–514.
3 B. Al-Lazikani, A. M. Lesk and C. Chothia, *J. Mol. Biol.*, 1997, **273**, 927–948.
4 V. Morea, A. M. Lesk and A. Tramontano, *Methods*, 2000, **20**, 267–279.
5 D. R. Davies and H. Metzger, *Annu. Rev. Immunol.*, 1983, **1**, 87–117.
6 A. Narayanan, B. D. Sellers and M. P. Jacobson, *J. Mol. Biol.*, 2009, **388**, 941–953.
7 K. R. Abhinandan and A. C. Martin, *Protein Eng., Des. Sel.*, 2010, **23**, 689–697.
8 G. Walsh, *Nat. Biotechnol.*, 2006, **24**, 769–776.
9 J. Reichert and A. K. Pavlou, *Nat. Rev. Drug Discovery*, 2004, **3**, 383–384.
10 T. Schwede, A. Sali, B. Honig, M. Levitt, H. M. Berman, D. Jones, S. Brenner, S. K. Burley, R. Das, N. V. Dokholyan, R. L. J. Dunbrack, K. Fidelis, A. Fiser, A. Godzik, Y. Huang, C. Humblet, M. P. Jacobson, A. Joachimiak, S. R. J. Krystek, T. Kortemme, A. Kryshtafovych, G. T. Montelione, J. Moult, D. Murray, R. Sanchez, T. R. Sosnick, D. M. Standley, T. Stouch, S. Vajda, M. Vasquez, J. D. Westbrook and I. A. Wilson, *Structure (London)*, 2009, **17**, 151–159.
11 C. Chothia and A. M. Lesk, *J. Mol. Biol.*, 1987, **196**, 901–917.
12 J. Bajorath and S. Sheriff, *Proteins: Struct., Funct., Genet.*, 1996, **24**, 152–157.
13 P. A. Ramsland, L. W. Guddat, A. B. Edmundson and R. L. Raison, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 453–461.
14 B. North, A. Lehmann and R. L. J. Dunbrack, *J. Mol. Biol.*, 2011, **406**, 228–256.
15 A. C. Martin and J. M. Thornton, *J. Mol. Biol.*, 1996, **263**, 800–815.
16 B. Oliva, P. A. Bates, E. Querol, F. X. Aviles and M. J. E. Sternberg, *J. Mol. Biol.*, 1998, **279**, 1193–1210.
17 H. Shirai, A. Kidera and H. Nakamura, *FEBS Lett.*, 1999, **455**, 188–197.
18 V. Manivel, N. C. Sahoo, D. M. Salunke and K. V. Rao, *Immunity*, 2000, **13**, 611–620.
19 E. C. Mundorff, M. A. Hanson, A. Varvak, H. Ulrich, P. G. Schultz and R. C. Stevens, *Biochemistry*, 2000, **39**, 627–632.
20 H. P. Nguyen, N. O. Seto, C. R. MacKenzie, L. Brade, P. Kosma, H. Brade and S. V. Evans, *Nat. Struct. Biol.*, 2003, **10**, 1019–1025.
21 D. K. Sethi, A. Agarwal, V. Manivel, K. V. Rao and D. M. Salunke, *Immunity*, 2006, **24**, 429–438.
22 G. J. Wedemayer, P. A. Patten, L. H. Wang, P. G. Schultz and R. C. Stevens, *Science*, 1997, **276**, 1665–1669.
23 J. Yin, S. E. Andryski, A. E. Beuscher, 4th, R. C. Stevens and P. G. Schultz, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 856–861.
24 J. Yin, E. C. Mundorff, P. L. Yang, K. U. Wendt, D. Hanway, R. C. Stevens and P. G. Schultz, *Biochemistry*, 2001, **40**, 10764–10773.
25 J. Yin, A. E. Beuscher, 4th, S. E. Andryski, R. C. Stevens and P. G. Schultz, *J. Mol. Biol.*, 2003, **330**, 651–656.
26 J. P. Schuermann, S. P. Prewitt, C. Davies, S. L. Deutcher and J. J. Tanner, *J. Mol. Biol.*, 2005, **347**, 965–978.
27 R. E. Bruccoleri, E. Haber and J. Novotny, *Nature*, 1988, **335**, 564–568.
28 R. E. Bruccoleri and M. Karplus, *Biopolymers*, 1987, **26**, 137–168.
29 C. Mandal, B. D. Kingery, J. M. Anchin, S. Subramaniam and D. S. Linthicum, *Nat. Biotechnol.*, 1996, **14**, 323–328.
30 N. R. Whitelegg and A. R. Rees, *Protein Eng., Des. Sel.*, 2000, **13**, 819–824.
31 P. Marcatili, A. Rosi and A. Tramontano, *Bioinformatics*, 2008, **24**, 1953–1954.
32 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.

33 C. A. Rohl, C. E. Strauss, K. M. Misura and D. Baker, *Methods Enzymol.*, 2004, **383**, 66–93.
34 K. T. Simons, C. Kooperberg, E. Huang and D. Baker, *J. Mol. Biol.*, 1997, **268**, 209–225.
35 P. de la Paz, B. J. Sutton, M. J. Darsley and A. R. Rees, *EMBO J.*, 1986, **5**, 415–425.
36 A. C. Martin, J. C. Cheetham and A. R. Rees, *Proc. Natl. Acad. Sci. U. S. A.*, 1989, **86**, 9268–9272.
37 Y. Choi and C. M. Deane, *Proteins: Struct., Funct., Genet.*, 2010, **78**, 1431–1440.
38 A. Fiser, R. K. G. Do and A. Sali, *Protein Sci.*, 2000, **9**, 1753–1773.
39 M. A. DePristo, P. I. W. de Bakker, S. C. Lovell and T. L. Blundell, *Proteins: Struct., Funct., Genet.*, 2003, **51**, 41–55.
40 M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw and R. A. Friesner, *Proteins: Struct., Funct., Genet.*, 2004, **55**, 351–367.
41 B. C. Braden and R. J. Poljak, *FASEB J.*, 1995, **9**, 9–16.
42 I. A. Wilson and R. L. Stanfield, *Curr. Opin. Struct. Biol.*, 1994, **4**, 857–867.
43 X. Y. Pei, P. Holliger, A. Murzin and R. L. Williams, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 9637–9642.
44 J. M. Rini, U. Schulze-Gahmen and I. A. Wilson, *Science*, 1992, **255**, 959–962.
45 S. Lee and T. L. Blundell, *Bioinformatics*, 2009, **25**, 1976–1977.
46 S. Kelm, J. Shi and C. M. Deane, *Bioinformatics*, 2010, **26**, 2833–2840.
47 J. C. Krause, D. C. Ekiert, T. M. Tumpey, P. B. Smith, I. A. Wilson and J. E. J. Crowe, *mBio*, 2011, **2**(1), e00345-10, DOI: 10.1128/mBio.00345-10.
48 M. P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, L. Reginier, F. Ehrenmann, G. Lefranc and P. Duroux, *Nucleic Acids Res.*, 2009, **37**, D1006–D1012.
49 A. Andreeva, D. Howorth, J. M. Chandonia, S. Brenner, T. Hubbard, C. Chothia and A. Murzin, *Nucleic Acids Res.*, 2007, **36**, D419–D425.
50 L. H. Greene, T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thorton and C. A. Orengo, *Nucleic Acids Res.*, 2007, **35**, D291–D297.
51 M. Babor and T. Kortemme, *Proteins: Struct., Funct., Genet.*, 2009, **75**, 846–858.
52 C. M. Deane and T. L. Blundell, *Protein Sci.*, 2001, **10**, 599–612.
53 J. Shi, T. L. Blundell and K. Mizuguchi, *J. Mol. Biol.*, 2001, **310**, 243–257.
54 R. Samudrala and J. Moult, *J. Mol. Biol.*, 1998, **275**, 895–916.
55 R. Chen, L. Li and Z. Weng, *Proteins: Struct., Funct., Genet.*, 2003, **52**, 80–87.