BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 18 2013, pages 2285–2291
doi:10.1093/bioinformatics/btt369

Structural bioinformatics

Advance Access publication June 26, 2013

Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server

Pier Paolo Olimpieri[1], Anna Chailyan[1], Anna Tramontano[1,2,*] and Paolo Marcatili[1,*]

[1]Department of Physics, Sapienza University and [2]Istituto Pasteur – Fondazione Cenci Bolognetti, 00185 Rome, Italy

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Antibodies or immunoglobulins are proteins of paramount importance in the immune system. They are extremely relevant as diagnostic, biotechnological and therapeutic tools. Their modular structure makes it easy to re-engineer them for specific purposes. Short of undergoing a trial and error process, these experiments, as well as others, need to rely on an understanding of the specific determinants of the antibody binding mode.

Results: In this article, we present a method to identify, on the basis of the antibody sequence alone, which residues of an antibody directly interact with its cognate antigen. The method, based on the random forest automatic learning techniques, reaches a recall and specificity as high as 80% and is implemented as a free and easy-to-use server, named prediction of Antibody Contacts. We believe that it can be of great help in re-design experiments as well as a guide for molecular docking experiments. The results that we obtained also allowed us to dissect which features of the antibody sequence contribute most to the involvement of specific residues in binding to the antigen.

Availability: http://www.biocomputing.it/proABC.

Contact: anna.tramontano@uniroma1.it or paolo.marcatili@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 19, 2013; revised on June 13, 2013; accepted on June 18, 2013

1 INTRODUCTION

The past two decades have seen monoclonal antibody (mAb) therapy come to age. With >30 molecules approved for clinical practice and hundreds currently being tested, mAbs are rapidly emerging as one of the most important classes of biological therapeutics. Despite their benefits, mABs obtained from both human and xenogeneic sources have some deficiencies, such as short *in vivo* life, low stability and high chances to raise an immunogenic reaction in patients. To overcome these hurdles, a number of strategies based on genetic recombination have been developed and optimized, which allow the modification and improvement of almost all the clinically relevant aspects of an antibody molecule but require expensive and time-demanding trial-and-error experimental procedures, a process that can be speeded up by the understanding of the structure and binding mode of the specific antibody (Morea *et al.*, 2000). The antibody molecule, with few exceptions, contains one or more tetramers of two identical pairs of polypeptide chains, the heavy and the light chains. Each chain consists of homologous domains, two for the light chain (one variable and one constant domain) and four or more for the heavy chain (one variable and three or more constant domains). All the domains share a similar tertiary structure, the so-called immunoglobulin fold, which is characterized by two anti-parallel beta sheets. The antigen-binding site (ABS) is mainly composed of six loops, three from the light and three from the heavy chain [also known as the hypervariable (HV) loops]. In a seminal study on antibody sequences (Wu and Kabat, 1970), such large variability was exploited to correctly define these HV sequence stretches as the complementarity determining regions (CDRs) in antibody recognition. Later studies (Novotny *et al.*, 1983) confirmed that this definition largely overlaps with the structurally based definition of the ABS. A number of other biological mechanisms are in place to increase the sequence diversity of antibody regions containing the ABS to enlarge the size of the antibody repertoire, and therefore the number of different antigens that can be targeted by the immune system (Di Noia and Neuberger, 2007; Schatz and Swanson, 2011; Teng and Papavasiliou, 2007).

Analyses of the rapidly growing number of antibody crystal structures in complex with their antigens pointed out that, even though almost all the intermolecular interactions are made by residues in the CDR (Kunik *et al.*, 2012; MacCallum *et al.*, 1996), the specific interaction pattern of each antibody depends on a subset of residues within or outside the CDR regions that are important either to maintain the correct three-dimensional (3D) conformation (Narciso *et al.*, 2011) or to specify the physicochemical environment of the ABS.

Knowing the role played by a specific residue is a key aspect in antibody rational design and engineering. This information can be inferred by analyzing the 3D structure of the antibody molecule or, when the latter is not available, by building and analyzing its 3D model. Modeling of antibody structures is a field that has attracted much attention, and available methods can produce models of very good accuracy (Accelrys Software Inc., 2012; Marcatili *et al.*, 2008; Molecular Operating Environment, 2012; Sircar *et al.*, 2009; Whitelegg and Rees, 2000). It should be noted here that antibody structure prediction still has pitfalls, mainly as far as the prediction of the conformation of the third HV loop of the heavy chain is concerned (Kuroda *et al.*, 2012; Ramos, 2012; Sircar, 2012).

Despite the abundance of methods specifically devoted to the prediction and analysis of antibodies (Lefranc *et al.*, 2009), few tools are available to provide information about the paratope, i.e. the subset of residues that contact the antigen. Paratome (Kunik *et al.*, 2012) is a recently published online tool for

*To whom correspondence should be addressed.

identification of antigen-binding region that extends the definition of CDRs to include contacts outside the canonical ABS region. However, Paratome does not provide any information on the specific residues that are directly involved in the binding, on the type of interaction (hydrogen bond, hydrophobic and other non-bonded interactions) or on the atoms involved in the interaction (main chain, side chain or both). This information is of relevance for any type of antibody engineering experiment, such as *in silico* maturation.

To overcome this problem, we developed prediction of Antibody Contacts (proABC), a web server for predicting which residues of an antibody are involved in recognizing its cognate antigen. It is based on a machine-learning method trained on sequence and sequence-derived features. Starting from the antibody sequence alone, proABC estimates, for each residue in its sequence, the probability that it interacts with the cognate antigen. Three different types of interaction are considered and predicted separately (hydrogen bond, hydrophobic and other non-bonded interactions). The results are displayed in an intuitive manner allowing an easy yet comprehensive examination of the residues that could directly interact with the antigen (also known as specificity determining residues). proABC also builds a 3D model of the antibody, in which residues are colored according to their contact probability.

The server is available at http://www.biocomputing.it/proABC.

2 METHODS

2.1 Datasets

The two datasets used for training and testing the predictors contain 313 and 44 antibody-antigen complexes, respectively. We scanned the sequences of all the molecules contained in the Protein Data Bank (PDB) database (October 15, 2012) using isotype-specific Hidden Markov Model (HMM) profiles developed by us (Chailyan *et al.*, 2012) and found 1294 antibody molecules. Using the PISCES web server (Wang and Dunbrack, 2003), we removed all the structures with a resolution worse than 3 Å, ending up with 1139 molecules. Among them, we selected all the immunoglobulins solved in complex with the antigen. This step was performed by examining whether any atom not belonging to the immunoglobulin light or heavy chain falls within a 10 Å radius from the ABS barycenter. When such atoms were found, the antibody was labeled as 'bound', and all the chains to which these atoms belong were considered as 'antigens'. Cases where non-immunoglobulin atoms were present near the binding site but were not labeled with a different chain identifier (usually small molecules) were manually examined to define the antigen/hapten.

Of the initial set of 1139 immunoglobulin structures, 637 were found to be in a bound state. We removed all complexes where the immunoglobulin sequence shared a sequence identity higher than 95% with any other using the cd-hit software (Li and Godzik, 2006), and ended up with 313 antibody structures. This dataset (hereafter called 'RF dataset') was used for the cross-validated training of the predictors.

To make an unbiased comparison of our results with those of Paratome, we also trained a predictor using a dataset obtained from the RF dataset after removing all the complexes in which the antibody sequence shares >95% sequence identity with antibodies deposited in the PDB database after February 2011. The antibodies released between February 2011 and October 2012 were collected in a third dataset (hereafter called 'Validation dataset') and culled to remove redundancy both within the dataset and with the other datasets.

2.2 Interaction identification

We calculated non-bonded contacts, hydrogen bonds and hydrophobic interactions for all the complexes using the software Ligplot (Wallace *et al.*, 1995) with default parameters. We also categorized these three groups of interactions according to which atoms of the residue (side chain and/or main chain) contact the antigen. We obtained 18 different interaction tables containing the number of non-bonded contacts, h-bonds and hydrophobic interactions occurring between an antigen and a residue, its main chain alone and its side chain alone for both the heavy and the light chain. We used each table to train a different predictor for every combination of interaction types (cont, h-bond, hydro), location of interaction (whole, main, side) and Ig chain (H, L).

2.3 Random forest analysis

We aligned, following the Chothia numbering scheme, the heavy and the light chain sequences in the RF dataset using HMM profiles that we developed earlier (Chailyan *et al.*, 2012). For the H3 alignment, we followed the method described in Lefranc *et al.* (2003) and Morea *et al.* (1998). The insertions were introduced at the center of the region comprised between the conserved residue Cys92 and Gly104 (Cys104 and Gly119 according to the international ImMunoGeneTics information system (IMGT) numbering).

Each position of the heavy and the light chain multiple alignment was considered as a variable; therefore, we had 135 variables for the heavy chain and 125 variables for the light chain. In other words, we predicted the binding properties of an amino acid (the target site) taking into account all the amino acids in the chains.

Each position can host one of the 20 amino acids or a gap, resulting in a 21-letter alphabet. We adopted two different encodings for the amino acids. The first strategy used the complete alphabet for all the variables (predictor A). The second strategy (predictors B and C) used the complete alphabet only for the target site and a reduced alphabet for all other positions in the sequence. The B and C strategies differ for the inclusion/exclusion of the antigen volume variable (see later in the text). We chose a reduced alphabet based on the 11 amino acid classes described in Pommie *et al.* (2004) and reported in Table 1. In this 12-letter alphabet (11 amino acid classes and the gap symbol), specifically derived for immunoglobulins, aromatic residues such as tyrosine, tryptophan and phenylalanine as well as glycine and proline are considered as different classes. This allows us to capture the pivotal role played by these amino acids both in antibody antigen recognition (Birtalan *et al.*, 2008; Koide

Table 1. The eleven classes used to encode amino acids in the reduced 12-letter alphabet adopted for models B and C

Cluster	Amino acid
Aliphatuc	Ala, Val, Ile, Leu
Sulfur	Cys, Met
Hydroxyl	Ser, Thr
Acidic	Asp, Glut
Basic	His, Lys, Arg
Amide	Asn
Phenylalanine	Phe
Tryptophan	Trp
Tyrosine	Tyr
Glycine	Gly
Proline	Pro

and Sidhu, 2009) and in maintenance of the ABS correct structural conformation (North *et al*., 2011).

2.4 Antigen volume

Several studies highlighted the correlation between antigen volume and the shape of the ABS (Collis *et al*., 2003; Lee *et al*., 2006). Consequently, we also evaluated the contribution of antigen volume to the overall performance of our predictors. We determined the antigen volumes using the 3 V software (Voss and Gerstein, 2010) and found that the distribution of the volumes of the antigens in our RF dataset is bimodal with a median value of 1538 $Å^3$. Complexes were classified in two categories depending on whether the antigen volume is larger or smaller than the median. This variable was used in training predictors A and B.

2.5 Canonical structures and HV loop length

Hypervariable loops and their main chain conformations (canonical structures) (Chothia and Lesk, 1987) are key parameters for antigen recognition. Five among the six loops (L1, L2, L3, H1, H2) adopt only few canonical structures, while the H3 loop is the most variable in both sequence and structural conformation (Al-Lazikani *et al*., 1997; Morea *et al*., 1998). We used the canonical structures and the HV loop length as variables. These were obtained using the tools provided by the Digit database (Chailyan *et al*., 2012).

We decided to use these variables that implicitly take into account the 3D structure of the antibody, rather than the modeled structure itself to avoid biases arising from incorrect modeling of some parts of the molecule, especially the H3 loop.

2.6 Germline families

Germline heavy and light chain variable regions (VH and VL) gene germline genes have been extensively used to study the biophysical properties of different antibodies (Ewert *et al*., 2003) and for VH/VL packing prediction (Chailyan *et al*., 2011a, b). We determined the source organism and the germline family and included them as variables in our predictors (Chailyan *et al*., 2012).

2.7 Random forest

In this study, we applied the R (v.4.6) implementation of the Random Forest (randomForest package).

We built separate predictors for all the antibody positions that are in contact with the antigen in at least 10 of the 313 complexes in our dataset. For all the other positions, the proABC output is simply the frequency with which the residue at that position interacts with the antigen in our dataset. For each position, we generated 18 different predictors, taking into account the type of interaction (non-bonded contact, h-bond, hydrophobic), the atoms involved in the interaction (whole residue, side chain and main chain) and the Ig heavy and the light chain.

Predictors were built in two-steps. In the first step, we used all the variables reported in Table 2 and fed a forest of 1500 trees. The total number of variables to be tested at each tree node (mtry) was set to its default value (corresponding to the square root of the total number of variables).

Mean decrease Gini (MDG) values resulting from this first step were ranked, and the average MDG was calculated (Chailyan *et al*., 2011a, b). In the second step, we used only the features with an MDG value larger than average MDG and built 1500 trees with default mtry. The entire process was subjected to a 10-fold cross-validation.

Three different models, named A, B and C (Table 2), were developed as described earlier in the text. For comparison, we also built a predictor (naive predictor) trained on each position alone, independently of all the others, using the 21-letter alphabet and building 100 trees. The results of the 10-fold cross-validation were evaluated using the following statistical measures: Precision, Accuracy, Recall, Matthew's correlation coefficient (MCC) and area under the curve (AUC).

Table 2. Random forest models

Model	Sequence	Antigen Volume	CDRs Lengths	Germline	Position
A	20 + gap	Yes	Yes	Yes	20 + gap
B	11 + gap	Yes	Yes	Yes	20 + gap
C	11 + gap	No	Yes	Yes	20 + gap
Naive	No	No	No	No	20 + gap

Note: The different sets of variables were adopted to train models A, B, C and the naïve predictor. All predictors use the complete amino acid alphabet to encode the residue at the specific position for which the interaction is being predicted ('Position' column). The complete alphabet is used in model A to encode the whole sequence, while the reduced 12-letter alphabet described in Table 1 is adopted in models B and C ('Sequence' column). Models A, B and C share the same sequence-derived features (canonical structures, HV loop length and germline family). The 'Antigen Volume' binary variable, which labels antigens with a volume larger or smaller than 1538 $Å^3$, is used only in models A and B.

2.8 Web interface

The web interface to the server is implemented in php and Jquery. Plots and tables are generated with the R package GoogleVis v.0.33 (Gesmann and de Castillo, 2011). Images are generated with Pymol (DeLano, 2002).

3 RESULTS

3.1 Performance evaluation

We compared the results of the three predictors with those obtained using the naive predictor. The corresponding ROC curves are shown in Figure 1, and the MCC and the AUC values for all the models are shown in Table 3. In all cases, our models clearly outperform the naive predictor indicating that information on the whole antibody sequence effectively contributes to the prediction performance. All predictors had similar values of AUC and MCC for non-bonded contacts, whereas model B proved to be better at predicting both hydrogen bonds and hydrophobic interactions (See Supplementary Fig. S1). We obtained a slightly worse performance for main chain hydrophobic interactions and hydrogen bonds, possibly because of the small number of these interactions present in our dataset.

Despite the importance of antigen volume for the prediction of some specific positions (as discussed later), predictor C still has a good overall classification ability. This gave us the possibility to include the antigen volume as an optional feature in the proABC web server, so that we use model B when the antigen volume is known and model C in all other cases.

3.2 Variable importance

As shown in the previous section, the predictors that use all the variables (sequence, antigen volume, canonical structures and germline families) greatly outperform the naïve one that only takes into account the contact frequency and residue identity at the target site.

Table 4 lists the 20 most important variables and their relative contribution to the correct prediction of the target site interactions. Supplementary Table S1 includes the complete list of variables and their relative importance. The analysis of the most important variables for specific target sites provides interesting insights into the antibody ABS and the process of antigen recognition. The antigen volume, for instance, is critical to correctly predict interactions with the N-terminals and C-terminals of the H1 and L1 loops, whereas residues in the center of these loops are influenced by the antigen volume to a significantly lower extent. These findings agree with a recently published analysis of the ABS anatomy (Raghunathan *et al.*, 2012).

It has been noticed (Collis *et al.*, 2003) that the H3 loop packs with L1 in the ABS reducing the accessibility of its C-terminal residues. Our data support this observation, as we observe a correlation between the H3 length and the number of contacts in the C-terminal region of L1, the shorter the loop, the higher the number of contacts ($C_{L1} = -0.06H3_{Len} + 1.61$, $P = 2.16 \times 10^{-5}$, Pearson correlation coefficient = 0.24).

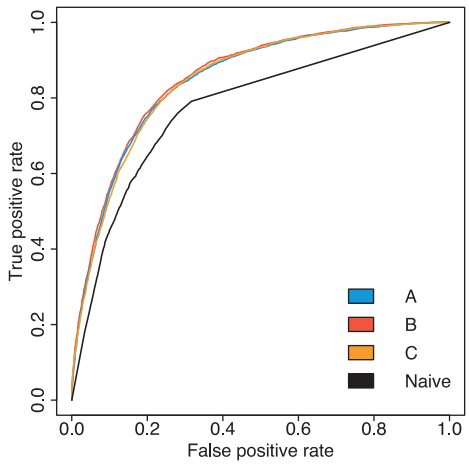


Fig. 1. Non-bonded contact prediction ROC curves for models A,B and C and the naive predictor

3.3 Comparison with other methods

To the best of our knowledge, proABC is unique in its purpose. The most similar tool is Paratome (Kunik *et al.*, 2012), an online tool for antigen-binding region identification. The performance of the two algorithms is evaluated here using the independent validation set described in Methods section. The results are summarized in Table 5. Paratome shows a very high total recall at the expense of low precision and specificity. This can be explained considering the somewhat large threshold that Paratome uses to define contacts (6.0 Å cut-off). This choice is due to the fact that Paratome aims at defining a plausible region of antigen binding rather than at predicting residue-specific interactions. On the other hand, proABC, despite its lower recall, outperforms Paratome in terms of precision and specificity. The MCC of the two methods shows that, overall, proABC performs significantly better.

3.4 Web server interface and 3D model

We integrated our predictors and several other tools, including a routine to build a 3D model of the antibody (Marcatili *et al.*, 2008), in a web server with a user-friendly interface (http://www.biocomputing.it/proABC). The web server includes six different pages: (i) Home; (ii) Plot; (iii) Summary; (iv) High Quality (HQ) figures; (v) Structure; and (vi) Logs.

The main page allows the user to input the light and the heavy chain sequences and, optionally, the antigen volume. The input is pre-processed to obtain all the information that can be derived from the sequence, namely germline families, canonical structures and length of each HV loop, shown in the Summary page, and passed to the relevant predictor. The web server uses predictor B (that has the best overall performance) if the user provides information on the antigen volume (smaller or larger than 1538 $Å^3$), and predictor C if no information on the antigen volume is given. The Plot page reports the results as graphs/tables listing the probability of each residue to interact with the antigen. The Summary page also includes links to download the results and the 3D model. The Image and Structure tabs allow the users to visualize the antibody 3D model and high-quality images of the antibody in which residues are colored according to their contact

Table 3. Matthews correlation coefficient and area under the curve values for each classifier and each type of interaction

	Non-bonded contacts (%)			Hydrogen bonds (%)			Hydrophobic interactions (%)		
	All	Side	Main	All	Side	Main	All	Side	Main
AUC									
A	84.8	83.7	82.2	73.8	73.3	75.9	79.4	79.6	70.8
B	85.1	85.0	82.8	76.3	76.6	76.1	80.7	80.5	72.2
C	84.7	84.5	82.6	75.9	75.9	75.2	80.1	80.4	71.3
Naive	77.7	78.0	69.8	64.8	64.7	58.8	72.0	73.6	59.9
MCC									
A	51.9	48.2	41.5	25.5	26.0	19.8	36.2	36.6	11.2
B	52.2	51.0	40.2	26.9	27.0	22.0	38.5	38.4	14.1
C	51.2	49.8	40.4	26.9	25.3	20.8	37.5	38.0	14.2
Naive	41.4	41.1	25.4	18.5	20.2	12.6	30.0	33.5	0.0

Table 4. The top 20 variables ordered according to their overall importance

Heavy chain		Light chain	
Variable	Importance	Variable	Importance
Germline family VL	208,56	Germline family VH	111,16
Position	140,90	Germline family VL	90,48
Germline family VH	107,98	Position	82,35
H:95 + 1	97,65	L:96	70,88
H:95 + 2	93,78	H3 Length	68,93
H:101 − 3	91,44	L:92	51,37
H:95	87,91	L:50	51,37
L1 Canonical structure	87,62	L:94	51,28
H:95 + 3	86,54	L:91	49,73
H:101 − 4	84,89	L:30	41,40
H:101 − 2	82,69	L:93	39,71
H:50	75,04	L:55	39,52
H:95 + 4	67,92	L:32	38,75
H:33	66,77	L:34	37,53
H:52	62,25	L1 Canonical structure	37,26
H:53	58,77	L3 Canonical structure	30,07
H:56	58,33	H2 Canonical structure	28,34
H:101 − 1	52,05	L:89	26,85
Antigen volume	50,83	Antigen volume	25,70
H:58	49,05	L:30	24,96

Note: The variable importance has been calculated by summing the mean decrease Gini value of the variable for each position. H3 residues are numbered according to their relative position with respect to H:95 and H:101 (i.e. H:95, H:95 + 1, H:95 + 2, . . . , H:101-2, H:101-1, H:101).

Table 5. Comparison between proABC and Paratome

Index	proABC	Paratome
True positives	624	778
False positives	286	1386
True negatives	1264	164
False negatives	156	2
Recall	80%	100%
Precision	69%	36%
Specificity	82%	11%
MCC	60%	19%

Note: Comparison of proABC and Paratome in terms of true positives, true negatives, false positives, false negatives, MCC, recall, precision and specificity of the two methods.

probability (Fig. 2). The Log page shows the current job status, indicating which operations are being performed.

Figure 3 shows an example of application of the server. The selected case is that of the humanized antibody Gevokizumab, the crystal structure of which (both free and in complex with its antigen interleukin-1 beta) has been recently solved

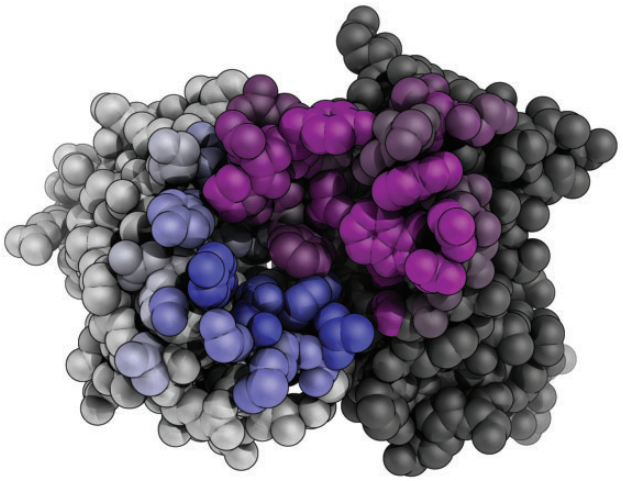


Fig. 2. Three-dimensional model generated by the proABC server for Gevokizumab. Residues are colored according to their predicted contact probabilities (light gray to blue gradient for the light chain, dark gray to purple gradient for the heavy chain)

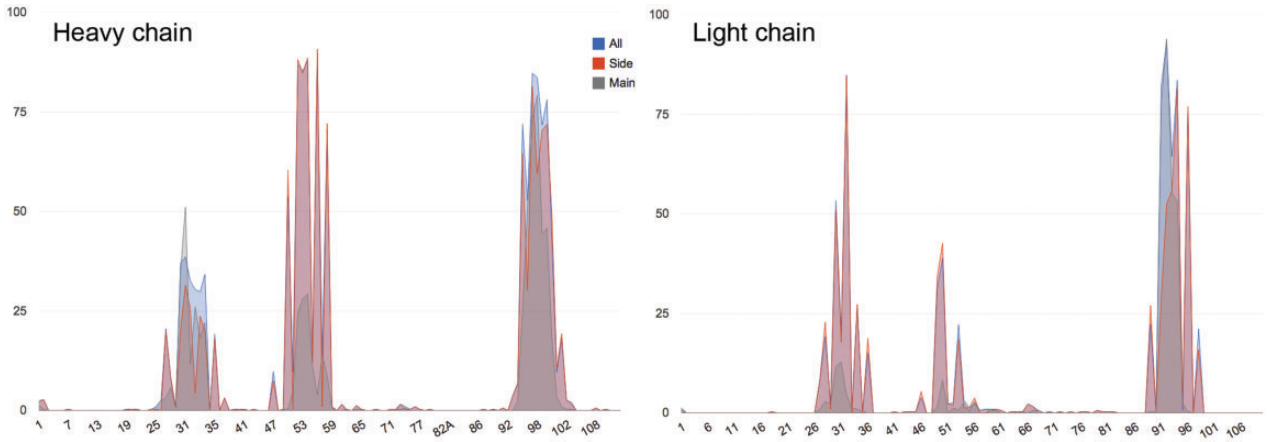
(Blech *et al*., 2013), but was obviously not used for the prediction. None of the antibodies in our dataset shares >75% sequence identity with Gevokizumab. Figure 3 and Supplementary Table S2 show the probabilities assigned by proABC to residues in the Gevokizumab light and heavy chain sequences and highlights the correct and incorrect predictions obtained in this example (considering a residue as predicted to be in contact if the corresponding probability is >0.5). As can be seen, most of the predictions are correct. proABC predicts a marginal role for the H1 loop and a prominent role of H2 in antigen recognition. Interestingly, the L3 loop is predicted to be the most important light chain region for antigen binding, with four contacts and two rarely observed main chain hydrogen bonds. This is what is observed in the crystal structure of the complex.

On the other hand, two contacts in L1 (position 27 and 28) were not predicted with high probability in this case. These two residues are rarely observed to be in contact with the antigen (13 and 18 cases in our dataset, respectively) and therefore are very difficult to classify correctly, highlighting the importance of increasing the number of solved structures of antibody complexes to further improve the performance of our method.

4 CONCLUSIONS

In this article, we described a method for predicting the probability of site-specific interactions between an antibody and its cognate antigen that reaches specificity and recall values of the order of 80%. This implies that the predictions are sufficiently accurate not only for investigating the properties of specific antibodies, but also as input for the design of novel antibodies, for example, in affinity maturation projects, and as a guidance for docking methods if the structure of the antigen is known or can be modeled.

Interestingly, our data show that the prediction accuracy improves when the complete sequence of the antibody rather than



Pos	All	Side	Main	Seq
56	89.7	90.8	3.9	D
54	87.9	88.6	29.3	D
52	87.1	88.2	24.5	W
53	85.3	84.7	28.1	W
97	84.7	81.5	73.1	Y
98	83.7	59.5	79.2	D
100	78.1	71.9	45.8	P
95	72.0	64.6	25.7	N
99	71.7	70.5	44.5	P
58	70.7	72.1	8.9	S
50	53.7	60.4	0.4	H
96	52.8	30.2	50.8	R
...	...	...	...	...
32	29.8	23.8	18.3	M
57	13.9	0.96	13.9	E

Pos	All	Side	Main	Seq
92	93.6	52.5	94.0	K
32	84.9	84.6	4.6	Y
94	83.6	81.7	53.3	L
91	82.1	29.1	77.8	G
96	77.0	77.0	0.32	W
93	64.4	55.9	55.2	M
30	53.4	51.1	11.6	S
32	...	...	...	...
50	38.9	42.7	8.3	Y
53	22.2	18.5	0.6	K
28	19.2	22.8	2.9	D
27	8.2	8.5	1.0	Q

Fig. 3. The results of the proABC server on the sequence of the humanized antibody Gevokizumab are compared with the experimentally observed interactions computed using Ligplot on the solved structure of the antibody in complex with its cognate antigen (PDB code:: 4G6M). The plots report the non-bonded contact probability for each residue and, separately, for its side chain and its main chain. The same information is reported in a tabular form as well. False positive predictions were made for H:100, H:95, H:50 (all and side) and H:96 (all and main), false negative predictions for H:32, L:50, L:53, L:28, L:27 (all, side), H:57 (all, main) and H:58 (main) while the other predictions shown in Fig.3 are all true positives. Each residue for which proABC returns an interaction probability higher than 0.5 is considered as a predicted contact

that of its binding site alone is taken into account. We could also derive information about which position/features of the antibody sequence influence the participation to the binding interface of specific sites, thus permitting a detailed structural analysis of the complex interplay between the framework and ABS in determining the specificity of the interaction, an issue of interest from both a theoretical and practical point of view.

The main advantages of the method presented here are that (i) it relies only on the sequence of the variable fragment of the antibody of interest, (ii) it is able to take into account the size of the antigen when the information is available and (iii) it can be automatically updated as new data on the structure of antibody complexes become available.

The associated server is straightforward to use and provides publication-ready output images as well as information on the properties of the analyzed immunoglobulin, including the determination of the corresponding germline family and the canonical structures and length of each HV loop. It also permits the user to obtain a model of the 3D structure of the antibody quickly and easily by using a well-tested and very accurate method (Marcatili, *et al.*, 2008). The server (http://www.biocomputing.it/proABC) is free and open to all users and there is no login requirement.

ACKNOWLEDGEMENTS

The authors are grateful to all other members of the Biocomputing Unit for useful discussions and for testing the server.

Funding: KAUST Award No. KUK-I1-012-43 made by King Abdullah University of Science and Technology (KAUST), FIRB RBIN06E9Z8_005, PRIN 20108XYHJS and the Epigenomics Flagship Project – EPIGEN.

Conflict of Interest: none declared.

REFERENCES

Accelrys Software Inc. (2012) *D.S.M.E., Release 3.5*. Accelrys Software Inc., San Diego.

Al-Lazikani,B. *et al.* (1997) Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.*, **273**, 927–948.

Birtalan,S. *et al.* (2008) The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J. Mol. Biol.*, **377**, 1518–1528.

Blech,M. *et al.* (2013) One target-two different binding modes: structural insights into gevokizumab and canakinumab interactions to interleukin-1beta. *J. Mol. Biol.*, **425**, 94–111.

Chailyan,A. *et al.* (2011a) Structural repertoire of immunoglobulin lambda light chains. *Proteins*, **79**, 1513–1524.
Chailyan,A. *et al.* (2011b) The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *FEBS J.*, **278**, 2858–2866.
Chailyan,A. *et al.* (2012) A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res.*, **40**, D1230–D1234.
Chothia,C. and Lesk,A.M. (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.*, **196**, 901–917.
Collis,A.V. *et al.* (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.
DeLano,W. (2002) The PyMOL Molecular Graphics System, Version 1.5.0.4, Schrödinger, LLC.
Di Noia,J.M. and Neuberger,M.S. (2007) Molecular mechanisms of antibody somatic hypermutation. *Ann. Rev. Biochem.*, **76**, 1–22.
Ewert,S. *et al.* (2003) Biophysical properties of human antibody variable domains. *J. Mol. Biol.*, **325**, 531–553.
Gesmann,M. and de Castillo,D. (2011) Using the Google visualisation API with R. *R. J.*, **3**, 5.
Koide,S. and Sidhu,S.S. (2009) The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem. Biol.*, **4**, 325–334.
Kunik,V. *et al.* (2012) Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Res.*, **40**, W521–W524.
Kuroda,D. *et al.* (2012) Computer-aided antibody design. *Protein Eng. Des. Sel.*, **25**, 507–521.
Lee,M. *et al.* (2006) Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *J. Org. Chem.*, **71**, 5082–5092.
Lefranc,M.P. *et al.* (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
Lefranc,M.P. *et al.* (2009) IMGT, the International ImMunoGeneTics information system. *Nucleic Acids Res.*, **37**, D1006–D1012.
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.
MacCallum,R.M. *et al.* (1996) Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.*, **262**, 732–745.
Marcatili,P. *et al.* (2008) PIGS: automatic prediction of antibody structures. *Bioinformatics*, **24**, 1953–1954.
Molecular Operating Environment (MOE) 2012.10 (2012) Chemical Computing Group Inc., Montreal, QC, Canada.
Morea,V. *et al.* (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J. Mol. Biol.*, **275**, 269–294.
Morea,V. *et al.* (2000) Antibody modeling: implications for engineering and design. *Methods*, **20**, 267–279.
Narciso,J.E. *et al.* (2011) Analysis of the antibody structure based on high-resolution crystallographic studies. *N. Biotechnol.*, **28**, 435–447.
North,B. *et al.* (2011) A new clustering of antibody CDR loop conformations. *J. Mol. Biol.*, **406**, 228–256.
Novotny,J. *et al.* (1983) Molecular anatomy of the antibody binding site. *J. Biol. Chem.*, **258**, 14433–14437.
Pommie,C. *et al.* (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J. Mol. Recognit.*, **17**, 17–32.
Raghunathan,G. *et al.* (2012) Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J. Mol. Recognit.*, **25**, 103–113.
Ramos,O.H. (2012) Computer-assisted modeling of antibody variable domains. *Methods Mol. Biol.*, **907**, 39–55.
Schatz,D.G. and Swanson,P.C. (2011) V(D)J recombination: mechanisms of initiation. *Ann. Rev. Genet.*, **45**, 167–202.
Sircar,A. (2012) Methods for the homology modeling of antibody variable regions. *Methods Mol. Biol.*, **857**, 301–311.
Sircar,A. *et al.* (2009) RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.*, **37**, W474–W479.
Teng,G. and Papavasiliou,F.N. (2007) Immunoglobulin somatic hypermutation. *Ann. Rev. Genet.*, **41**, 107–120.
Voss,N.R. and Gerstein,M. (2010) 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.*, **38**, W555–W562.
Wallace,A.C. *et al.* (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, **8**, 127–134.
Wang,G. and Dunbrack,R.L. Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
Whitelegg,N.R. and Rees,A.R. (2000) WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Eng.*, **13**, 819–824.
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.