Bioinformatics Advance Access published January 18, 2016

EPI-Peptide Designer : a tool for designing specific peptide ligand libraries based on Epitope-Paratope Interactions

Viart B [1], Gonzalez E [1], Dias-Lopes C [1], Oliveira C F B [1], Nguyen C [3], Neshich G [2] , Chávez-Olórtegui C [1] , Molina F [3] , and Felicori L [1*]

[1] Universidade Federal do Minas Gerais, Brazil
[2] Embrapa Informática Agropecuária, Campinas, SP, Brazil
[3] Sys2Diag, FRE3690-CNRS/ALCEDIAG, Montpellier, France

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Motivation:

Antibodies are an important class of biological drugs, but with limitations, such as inadequate pharmacokinetics, adverse immunogenicity and high production costs. Synthetic peptides with high affinity and specificity for the desired target represent an important alternative to antibodies. However, no computational tool exists to guide the design of these peptides.

Results:

To identify the interacting residues in a given antibody-antigen interface we used Interface Interacting Residue (I2R), a selection method based on computed molecular interactions. The aggregation of all the molecular interactions between epitope and paratope residues allowed us to transform the 3D antibody-antigen complex structures into interface graphs. Based on these data and the probability of molecular interaction we developed EPI-Peptide Designer tool that uses predicted paratope residues for an epitope of interest to generate targeted peptide ligand libraries. EPI-Peptide Designer successfully predicted 301 peptides able to bind to LiD1 target protein (65% of the experimentally tested peptides). This tool should enable the development of a new generation of synthetic interacting peptides that could be very useful in the biosensor, diagnostic and therapeutic fields.

Availability:

All software developed in this work are available at http://www.biocomp.icb.ufmg.br/biocomp/

Contact: liza@icb.ufmg.br

1 INTRODUCTION

Protein-protein interactions are at the heart of biological processes and protein functions are highly related to their binding properties (Chakrabarti and Janin, 2002). For instance, the immune response relies on antigen recognition by a specific antibody and the Antibody-Antigen (Ab-Ag) complex represents a specific type of protein-protein interaction characterized by high affinity and specificity. Identifying the key residues and interaction patterns on the Ab-Ag interface could help improving antibody humanization as well as the design of new antibodies (Morea *et al.*, 2000) and peptide ligands based on the antibody properties.

The use of peptides for therapeutic purpose instead of antibodies has plenty of advantages such as lower manufacturing costs, less immunogenic profile, greater stability and better organ/tumor penetration. Several chemical approaches have been generated to overcome therapeutic peptides limitations such as low oral bioavailability and biodistribution (Vlieghe *et al.*, 2010). Indeed, much research effort is focused on the use of peptide ligands as a viable alternative to antibodies in targeted therapies (Wada, 2013). For instance, mimetic peptides derived from the anti-HER2/ERBB antibody can inhibit the tyrosine kinase activity of this receptor and consequently impair tumour growth (Park *et al.*, 2000; Ponde *et al.*, 2011). Presently, over 50 peptide drugs are approved for clinical use (Reichert J., 2010). To guide the design and increase the affinity and specificity of these peptide drugs, different tools, based on various methodologies (e.g., directed evolution, high-throughput protein screening or rational design based on protein-peptide interactions) have emerged (Pei and Wavreille, 2007; Yin *et al.*, 2007; Vanhee *et al.*, 2011). *In silico* rational design of peptides based on molecular interactions is also a fundamental proof-of-concept for the current understanding of the physical-chemical basis of molecular recognition. Moreover, this approach could become a powerful complement to the current library-based screening methods because it allows targeting specific patches on the surface of a protein (Fleishman *et al.*, 2011). Computational design also gives the opportunity to program protein-protein interactions for specific applications. However, currently no computational methodology to design this kind of peptides is available.

In this work, we propose a computational method to generate libraries of peptide ligands or paratope mimetics based on the Epitope-Paratope Interaction (EPI) patterns and on a target epitope input sequence. This software, called EPI-Peptide Designer, uses a set of Ab-Ag complex structures from the Protein Data Bank (PDB) (Berman *et al.*, 2000) and the Blue Star STING server and STING_DB (Neshich *et al.*, 2006) containing hundreds of interaction descriptors reported in residue by residue fashion

*to whom correspondence should be addressed

to compute the Bayesian probabilities of molecular interactions between epitope and paratope. EPI-Peptide Designer generates peptide binder sequences based on the epitope sequence entered by the user and the patterns extracted from the Ab-Ag interfaces. The method was experimentally validated using as target a dermonecrotic protein LiD1 from the brown spider venom. We have synthesized a library of 460 peptides and 65% of them were able to bind to LiD1. This is, to our knowledge, the first generator of peptide ligand libraries based on EPI.

2 METHODS

Dataset extraction

To extract structures of Ab-Ag complexes from the PDB (Berman *et al.*, 2000), we first used the datasets from Ramaraj *et al.* and Kunik *et al.* to select the antibody light and heavy chains to be used as reference sequences. After redundancy removal using CD-Hit (Fu *et al.*, 2012), we processed the two reference sequence datasets with Interface Research Algorithm (IRA), a BioJava program we developed. IRA automatically computed the Smith and Waterman local alignment (Smith *et al.*, 1981) of each sequence against each chain of all the PDB files that contain at least three protein chains. Using a threshold determined by aligning the reference dataset against itself, IRA labelled each chain as Antibody Light, Antibody Heavy or Antigen. IRA selected structures that contain at least one antigen, one light chain and one heavy chain spatially close (i.e., presenting inter-atomic contacts using the 5 Ångström (Å) distance cut-off). From these, the PDB files with X-ray resolution lower or equal to 2.5Å and present in STING RDB were extracted (Neshich *et al.*, 2006).

Interface selection

To analyse the interface of Ab-Ag complexes, we used three different interface selection methods. First, in the selection based on the distance between atoms of the antigen and the antibody (distance-based selection, DBS) (Chothia and Janin, 1975; Lo Conte *et al.*, 1999), an amino acid of the antigen is considered to be part of the Distance Selected Epitope (DSE), if one or more of its atoms are at a distance below a chosen cut-off (in our study, from 3 to 8 Ångström). The Distance Selected Paratope (DSP) is selected in the same manner. Second, in the approach based on the difference of Solvent Accessible Surface (ΔSAS), interfaces are selected based on the loss of solvent accessibility between the separated and the complexed protein (Lo Conte *et al.*, 1999). Third, we developed a selection method in which the interface computed molecular interactions are extracted from STING RDB (Neshich *et al.*, 2006). In this method, the interface is defined by all the amino acids that are involved in the molecular interactions between the antigen and the antibody chains and that are called, therefore, Interface Interacting Residues (I2R). The selected antibody residues form the I2R Paratope and the selected antigen amino acids constitute the I2R Epitope.

Computation of the interface molecular interactions

Molecular interactions (salt bridges, hydrogen bonds, aromatic stacking and hydrophobic interactions) were taken from STING RDB IFR (Mancini *et al.*, 2004). This tool identifies all potential intra- and inter-protein chain contacts stored in STING RDB (Neshich *et al.*, 2006) by (1) classifying the atoms in groups according to their electrostatic behaviour and position in the amino acid (main or side chain) and (2) by then selecting atoms based on the type of contacts they potentially can make and on the experimentally defined distance restrictions (Harris and Mildvan, 1999; Sobolev *et al.*, 1999; Swindells, 1995).

Redundancy removal

To extract meaningful information from the interface dataset, we removed redundancies by selecting only the DSE and DSP sequences from the complex (with a cut-off of 6Å). Using the CD-Hit global sequence identity score (Fu *et al.*, 2012), we only selected interfaces with a score lower than 0.90 for both interface sides. Global sequence identity score is define as the number of identical amino acids in alignment divided by the length of the shorter sequence. The selected files were manually curated to confirm their quality. This provided us with a non-redundant dataset composed of 101 PDB structures, 21 antibody-peptide complexes (here, peptides are defined as molecules smaller than 30 amino acids) and 80 antibody-protein complex.

Interface statistical analysis

To compute the percentage of occurrence (%Occ) of the epitopes and paratopes selected by I2R we used :

$$\%Occ_n = \frac{Occ_n}{Occ_{total}} \times 100,$$

where n is an amino acids, $\%Occ_n$ is the percentage of occurrence of n, Occ_n is the occurrence of n and Occ_{total} is the occurrence of all the residues. The results were compared to all STING RDB protein-protein interaction (Neshich *et al.*, 2006) occurrence values after exclusion of our 101 PDB Files. The statistical comparison of the amino acids was done using a t-test of differential distribution and was considered significant when the p-value was lower than 0.01.

Comparison of the interface selection methods

To compare the interface residue selection by the three methods we computed the Receiver Operating Prime Curve (ROC') of the performance of the distance-based selection and ΔSAS, using various cut-offs, against I2R. As the aim was the comparison of selected interface residues, the true negatives were not considered. We computed the ROC' curve as follows. The True Positive Rate (TPR), also called recall, was computed as:

$$TPR = \frac{TP}{TP + FN}$$

and the False Discovery Rate (FDR) as:

$$FDR = \frac{FP}{FP + TP}$$

where TP is the True Positive, FP the False Positive and FN the False Negative.

Computation of the most frequent interface partners using graph analysis

To analyse the interface in a multi-level manner, we developed Interface to Graph Generator (IGG). IGG is a BioJava program that takes as input PDB codes and two sets of chains. Molecular interactions between those two sets are recovered from PDB structures using STING RDB (Neshich *et al.*, 2006). The interface is

automatically transformed into a graph, where all I2Rs are vertices and all interactions are edges. The vertex label holds the information concerning the interface side and the amino acid type (Table 1). The edges are labelled according to the type of interaction, such as hydrogen bonds, salt bridges, hydrophobic interactions and aromatic stacking. Using GASTON (Nijssen and Kok, 2004), we extracted the most conserved sub-graphs from the complete set of interfaces containing two and three nodes. Subgraphs plot was done using R (R Development Core Team, 2008) and the "igraph" package (Csardi and Nepusz, 2006).

Table 1. Amino acids group used for graph and subgraphs analysis

Group	Residue
Small	A,G
Charged +	K,R,H
Charged -	D,E
Hydrophobic	V,I,L,C,M,P
Alcohol	S,T
Aromatic	Y,W,F
Polar	Q,N

Assessment of paratope residue prediction

Based on the Bayesian probabilities extracted from the epitope-paratope graphs, we predicted the amino acid sequence and the interaction of a given paratope using a given epitope sequence. To evaluate the prediction of residues and interactions, we used a leave-one-out cross validation of the 21 antibody-peptide PDB interfaces from our dataset. Antigens were considered as peptides if their size was equal or lower than 30 amino acids. The evaluation considered each residue from the input epitope and defined as True Positive (TP) a correct "interaction type and paratope residue" couple, as False Positive (FP) any interaction where the interaction type or the residue group was incorrect, as False Negative (FN) any existing couple not added by the program and as True Negative (TN) any possible not existing and not added interaction type-paratope residue couple.

EPI-Peptide Design tool

Using all the Ab-Ag interaction patterns and the residue occurrence data obtained in this study, we developed EPI-Peptide Designer in BioJava. EPI-Peptide Designer includes the IGG program described above. The program takes as input a real or putative epitope sequence (linear or conformational; gaps in the sequence can be represented by -), a cut-off score representing the importance of the epitope sequence in the design and the number and size of peptides needed by the user. To design peptide ligands, EPI-Peptide Designer uses the Base Residue Library (BRL) composed of all residues from all the paratopes in the input dataset. The computed probabilities include: probability of an epitope residue type to do an interaction and, for each type of interaction, the probability of the target paratope residue type and the influence of the epitope neighbour residues on the interaction. Using these probabilities and the input sequence, EPI-Peptide Designer ranks the predicted paratope residues in decreasing order of likelihood. The paratope residues are then added according to the decreasing order of likelihood to the BRL until the defined cut-off score is reached (i.e., for a BRL of 100 residues and a cut-off score of 10%, EPI-Peptide Designer will add 10 residues to the BRL). The thus obtained biased amino acid library (i.e., modified to become specific for a given epitope sequence) is then used to generate random EPI-peptide sequences of the length and in the number defined by the user.

EPI-peptide design, peptide synthesis on cellulose membranes and binding assay

In order to test the effectiveness of the method, we generated 800 EPI-Peptides using the protein LiD1 (GI: 33348850,Felicori *et al.* (2006)) catalytic sequence epitope ($_{37}$FDDNANPEYTYHGIP$_{51}$) and default parameter of EPI-Peptide Designer (Ab-peptide dataset, length of 15 amino acid and a score of 50). To ensure solubility, only sequences which contained less than 50% hydrophobic residues; at least 25% of charged residues and less than 75% of D, E, H, K, N, Q, R, S, T and Y were selected and synthesized (Following recommendations from Life technologies peptide solubility website, http://www.lifetechnologies.com). Four hundred and sixty peptides were synthesized on a cellulose membrane as previously described by Laune *et al.* The membrane was blocked by incubation with 3% BSA and 5% sacarose at room temperature overnight, and then membranes were probed LiD1 covalently linked to biotin at a concentration of 20μg/ml in blocking buffer at room temperature for 90 min. Biotinalytion of LiD1 was conducted using commercial available Biotinylation kit (Sigma-Aldrich, BK101). Protein binding was revealed by incubation (at room temperature for 90 min) with alkaline phosphatase-conjugated avidin (1:10,000) and 5-bromo-4-chloro-3-indolyl phosphate (BCIP) plus 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) as substrate. To remove molecules and precipitated blue die attached, membranes were sequentially treated with dimethylformamide, 1% SDS, 0.1% 2-mercaptoethanol in 8 M urea, ethanol/water/acetic acid (50:40:10, vol/vol/vol) and, finally, methanol and further employed in other assays. Peptide reactivity was assessed based on manual reading and consensus of triplicate assays. Positive sequences were analysed by GibbsCluster (Andreatta *et al.*, 2013) and Weblogo (Crooks *et al.*, 2004) tools.

3 RESULTS

Analysis of the Interface Interacting Residues (I2R) allows evaluating the distance-based selection and the difference of solvent-accessible surface methods

To compare the three interface residue selection techniques, we selected interfaces from the 101 PDB structures by computing the Euclidean distance DBS, the ΔSAS and the interface molecular interactions (I2R). We then compared the selections made with the DBS and ΔSAS methods against the I2Rs by computing the ROC' curves (Fig.1). Comparison of the selection made based on the Euclidean distance with the extracted I2Rs showed that the maximum precision was obtained with a 3Å distance, while the maximum TPR (also called Recall) was reached with 8Å. The DBS had a higher surface under the curve and the highest value

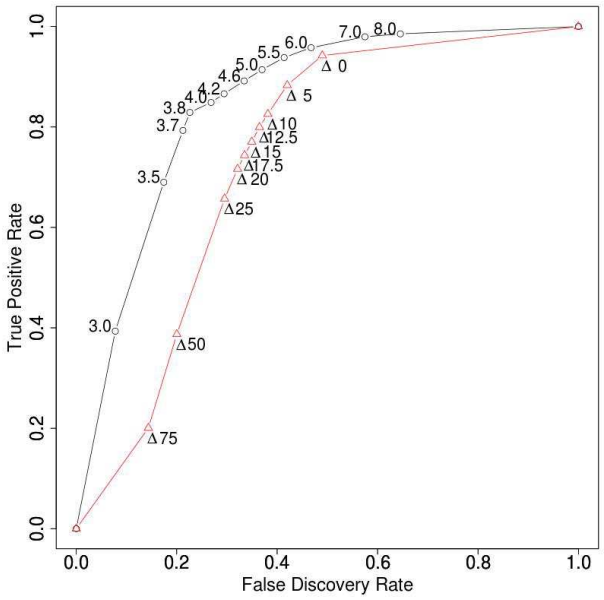

Fig. 1. Comparison of DBS (black circles) and ΔSAS (red triangles) residue selection using different cut-offs relative to the I2R method.

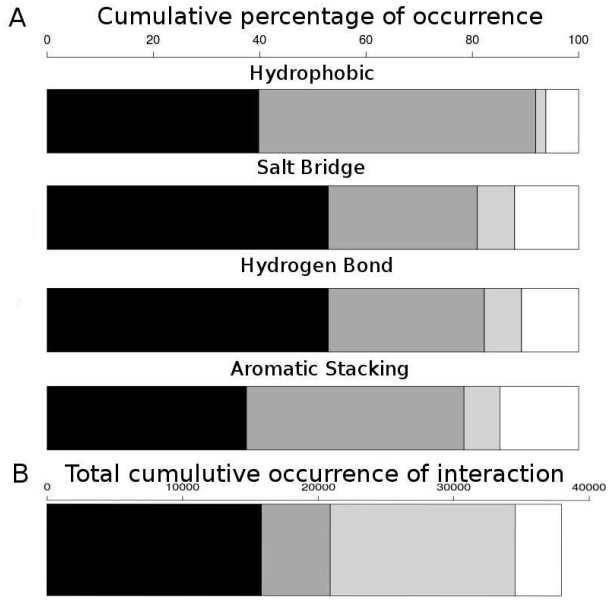

Fig. 2. **A:**Percentage of molecular interactions by type using DSB from 0 to 3Å (black), from 3 to 4Å (dark grey),from 4 to 5Å light grey) and from 5 to 8Å (white). **B:** Cumulative occurrence of hydrophobic interactions (black), salt bridges (dark grey), hydrogen bonds (light grey) and aromatic stacking (white) at the antigen-antibody interface.

of $TPR - FDR$ was reached for a distance of 3.8Å. Most DBS-based Ab-Ag structure studies use a cut-off between 4Å and 6Å. For a distance of 5Å, with this plot, 91.5% (TPR) of interacting residues were selected; however, 32% of the selected residues did not to do any kind of interaction. Surprisingly, to reach the maximum TPR, a distance cut-off of 8Å was needed. As most of the molecular interaction maximum distances are lower than 6Å, we further investigated the interaction repartition.

As all interface interactions are not selected by the 5Å cut-off, we were interested in the interaction repartition in function of the distance. The bar plots (Fig.2A) of the interactions relative to the chosen distance showed that the distance of 5Å, as expected based on the previous results, allowed the selection of most interactions, but still missed 8.5% of them, specifically 2% of all salt bridges, 5.2% of all hydrogen bonds and 6.5% of all aromatic stacking, but none of the hydrophobic interactions. The hydrogen bonds with a distance bigger than 5Å were all water-mediated, thus explaining the unusual long distance. The cumulative bar plot of the interactions (Fig.2B) showed that the hydrophobic interactions were quantitatively the most important, followed closely by hydrogen bonds. Conversely, salt bridges and aromatic stacking were less frequent on the antibody-antigen interface.

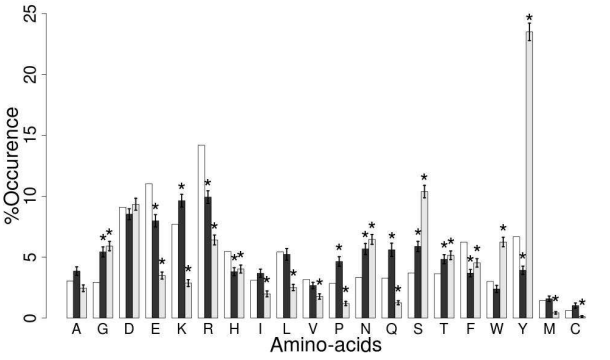

Fig. 3. Comparison of the occurrence (in percentage) of all interacting residues in STING RDB (white), I2R epitopes (black) and I2R paratopes (grey). Error bars are calculated as the standard deviation divided by the root square of the set size. Stars represent statistically significant differences compared to STING RDB, p value <0.01 using a standard t-test.

Amino acid occurrence in epitopes and paratopes selected with the Interface Interacting Residue (I2R) method

Compared to all interacting residues in STING RDB, I2R paratopes (grey columns in Fig.3) were significantly enriched in Tyr, Ser, Trp, Gly, Asn and Thr. I2R paratopes were depleted of most of the other amino acids, but for Ala, Asp and Phe the occurrence of which was not significantly different compared with all STING RDB interacting residues. I2R epitopes (black columns in Fig.3) were enriched in Gly, Pro, Asn, Gln, Ser, Thr and Cys and depleted of Glu, Arg, His Phe and Tyr.

A bipartite graph representation of the paratope-epitope interactions indicated that the interacting residues had a very asymmetric distribution (Fig.4). In the paratope, Tyr, the most frequent residue, interacted with almost all the epitopic amino acids via different types of interactions. Tyr interacted most frequently with hydrophobic amino acids, particularly Pro, Gln, Gly, Phe, and

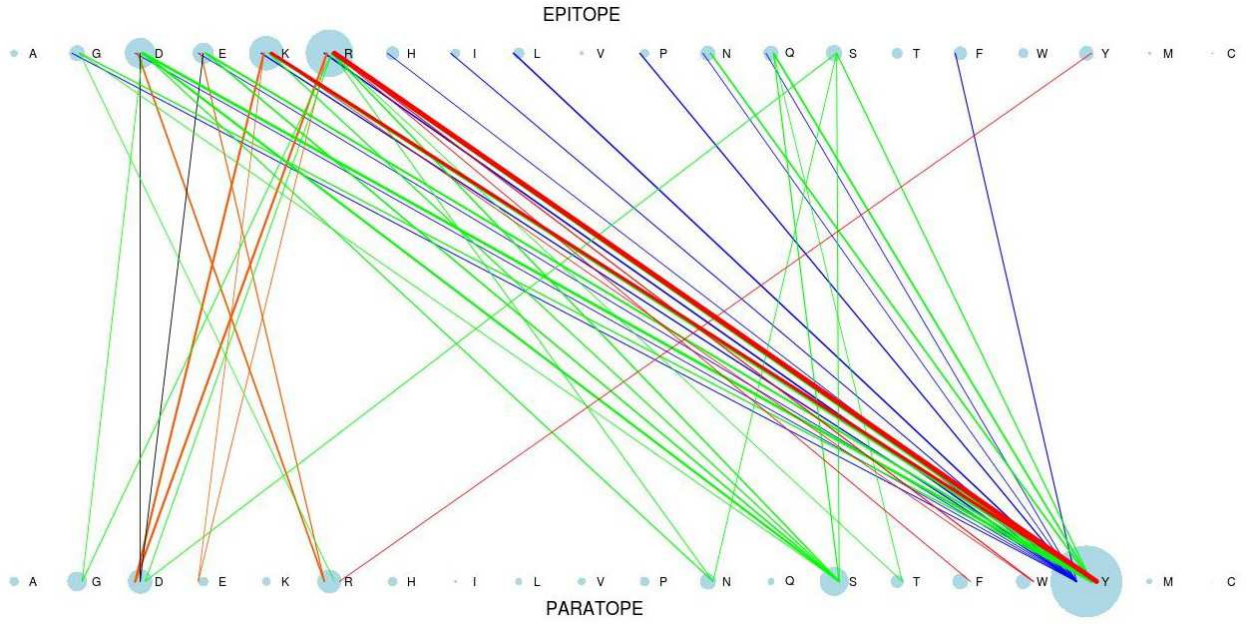

Fig. 4. The bipartite graph representation of the molecular interactions between I2R paratopes and I2R epitopes highlight the strong asymmetric pattern of epitope-paratope interactions. The sphere size of each residue is proportional to the amino acid occurrence in its respective side. The vertex width is proportional to the occurrence of the specific type of interaction; green, hydrogen bonds; blue, hydrophobic interactions; orange, attractive salt bridges; black, repulsive salt bridges; red, aromatic stacking. Only vertices with an occurrence higher than 25 are represented.

with the charged Lys and Arg in the epitope. Indeed, paratopic Tyr interacted with positively charged epitopic residues via cation-π interactions and with negatively charged epitopic residues via hydrogen bonds. The Ser in the paratope seemed important for establishing a network of hydrogen bonds with charged amino acids and also with Gln and Ser in the epitope. Among the charged amino acids in the paratope, a high prevalence of salt bridges done by Arg and Asp was observed. More heterogeneous interactions were observed among the epitope residues. Although Arg was less frequent than in other kinds of protein-protein interactions (Fig.4), it was the most frequent residue in epitopes and was involved in all kinds of interactions. Epitopic Arg interacted mostly with Tyr residues in the paratope via aromatic stacking, hydrogen bonds and hydrophobic interactions. It also formed salt bridges preferentially with Asp, but also with Glu, and repulsive salt bridges with Arg in the paratope. Lys in the epitope formed a similar network with Tyr in the paratope.

The most conserved subgraphs highlight the importance of cation-π interactions in the epitope-paratope interface

The extraction of the most conserved subgraphs from the complete dataset with two of the three nodes showed that paratopic aromatic residues (Tyr) predominantly interacted with positively charged residues in the epitope through an aromatic stacking interaction (cation-π interaction) (Fig.5A). Specifically, 84 of the 101 selected structures contained at least one cation-π interaction in which the positive charge was hold by the epitope. In addition 51 structures contained a double cation-π interaction (Fig.5B) composed of a positively charged residue in the epitope that interacted with two aromatic amino acids from the paratope. The subgraphs also showed that salt bridges often involved three residues: two negatively charged from the paratope with one positively charged from the epitope. Hydrogen bonds had a low score, although they were the second most frequent type of interaction observed in Ab-Ag interfaces. This can be explained by the variety of amino acid group couples that can form such interaction, thus reducing the frequency of same residue group - same interaction couples.

Assessment of the paratope residue prediction

Using these antibody-antigen graph patterns, we then developed a new methodology to design antibody mimetics using the antigen sequence Fig.6. First, we computed the Bayesian probability of all kinds of interactions to predict the residue-interaction couples. Then, to test the predictions, we used the 21 antibody-peptide interfaces from our dataset and a leave-one-out cross-validation method with all the interactions and the seven residue groups (Table 1). Using a cut-off of 5%, meaning that a paratope-residue interaction couple had to have a Bayesian probability of 0.05 to be added, we obtained a sensitivity of 23% and a specificity of 95%, with an accuracy of 92%.

EPI-Peptide Designer tool

From a set of user-defined Ab-Ag complexes (Fig.6A), the EPI-Peptide Designer computed the graph representation of the interfaces (Fig.6B). Then, from the set of graphs, the program computed the amino acid occurrence in the second side (in our study the paratope) and the interaction probability (Fig.6C and Fig.6D). To demonstrate how the EPI-Peptide Designer works, we used the epitope from the PDB structure 1TET that contains the choleric

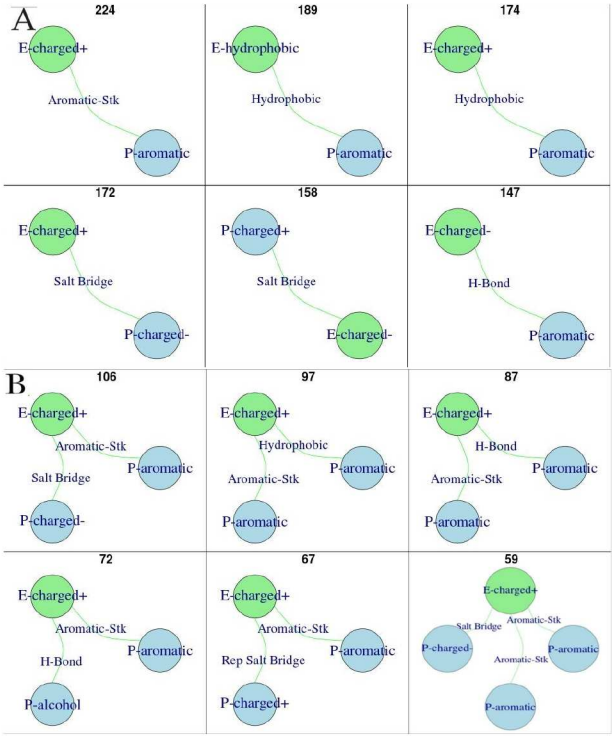

Fig. 5. Each cell contains one of the six most common subgraphs with two (**A**) or three nodes (**B**) from the interface graphs based on the 101 PDB structures dataset. The title indicates in how many interfaces the motif was observed at least once.

toxin complexed with an antibody. We chose this example because it was not in our dataset and is a small linear epitope composed of only one segment: VEVPGSQHIDSQKKA. We used our non-redundant 101 PDB structures as dataset input and as epitope input the 5Å epitope extracted from the 1TET structure. Using a score of 50% (representing the importance of the epitope sequence in the design), the percentage of occurrence of five amino acids from the BRL was modified by at least 2% (see Fig.6E).

Predicted EPI-peptides are able to bind to target LiD1 protein

To test the ability of EPI-peptide designer to successfully predict peptides able to bind to the epitope ($_{37}$FDDNANPEYTYHGIP$_{51}$) of LiD1 protein, 460 peptides were predicted, chemically synthesized and assayed against biotinylated LiD1. As a control for aspecific binding, the membrane was probed with AvidinAlkaline Phosphatase alone and no reactive peptides were observed (supplementary Figure 1). From 460 sequences synthesized, 218 were considered highly positive (47% of sequences, squares on Fig.7**A**) and 83 had a lower reactivity (18%). 159 peptides (35%) presented no reactivity. Highly positive peptides were clusterized in two groups and a graphical representation of the patterns from each multiple sequence alignment was computed (Fig.7**B**,**C**). Cluster 1 (supplementary table 1) contains 107 sequences and shows two conserved tyrosine at position 9 (59%) and 10 (41%) as well as two conserved arginines (positions 14 and 15). The second most frequent amino acid in those positions is also aromatic (Trp). Similarly, the

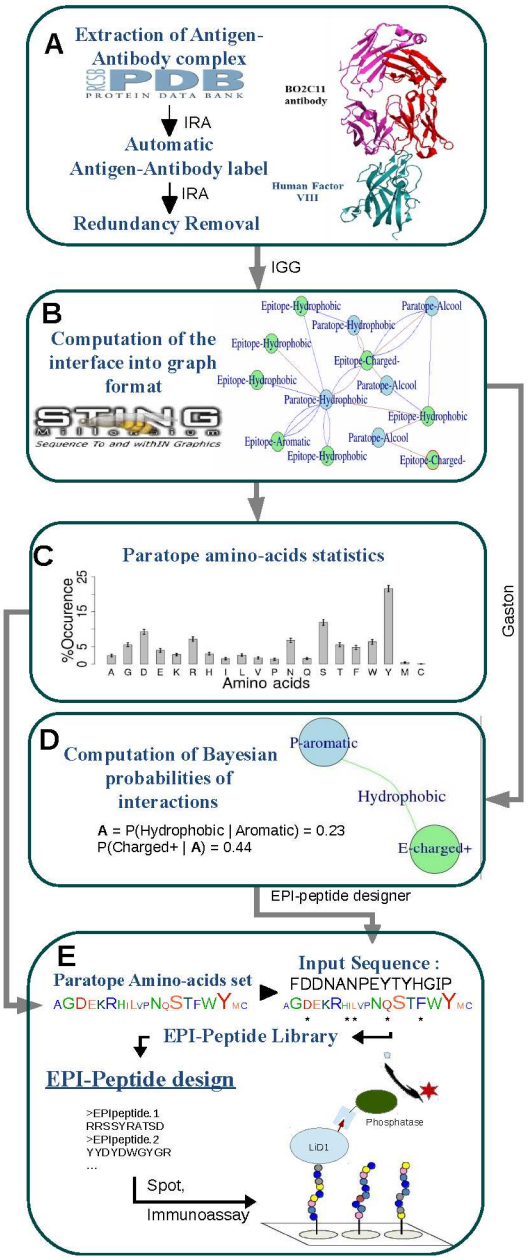

Fig. 6. Schematic of the EPI-peptide Design method to design targeted libraries of peptide ligands. **A**. The Interface Research Algorithm (IRA). **B**. Interface Graph Generator transform the epitope-paratope interface into a graph format using computed molecular interactions extracted the BlueStar STING database. **C**. Computation of the paratope amino acid occurrence using Gaston (Nijssen and Kok, 2004). **D**. Interaction probability. **E**. The epitope sequence modifications are entered in the Based Residue Library (BRL) using the previously computed probabilities. The size of the amino acid font represents the occurrence percentage in the libraries. The star represents amino acid frequencies that have been modified by at least 2% based on the epitope sequence specificity (biased library). EPI-Peptide are synthesized and binding is validated using immunoassay techniques

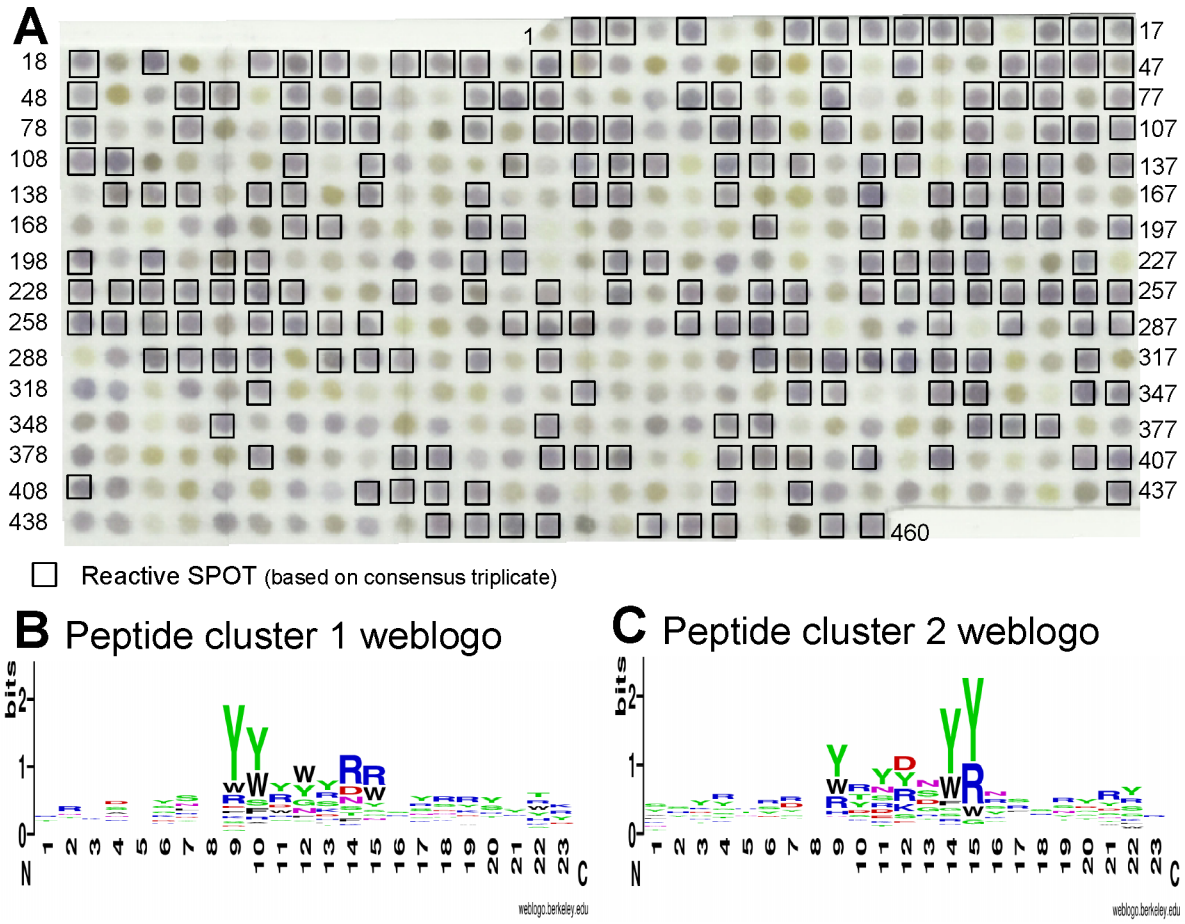

Fig. 7. Experimental validation and analysis of EPI-peptides prepared by SPOT method (**A**). 460 EPI-peptides predicted against LiD1 protein epitope was synthesized. 20 μ/ml of LiD1-biotin followed by alkaline phosphatase-conjugated avidin. 1:10,000 revealed binding peptides. Black boxes represent highly reactive peptides. Weblogo representation of the alignment obtained from reactive peptides grouped in 2 clusters: cluster 1 (**B**) and cluster 2 (**C**).

cluster 2 (supplementary table 2) includes two conserved tyrosines at position 14 and 15 (53 and 54% respectively).

4 DISCUSSION

To overcome the many antibody limitations, such as their inadequate pharmacokinetics, poor tissue accessibility and adverse immunogenicity including high production costs, enormous efforts have been focused on finding alternative strategies (Yin and Hamilton, 2005), such as non-peptidic protein binders (Margulies and Hamilton, 2010), smaller antibody fragments that retain the original binding property (Hudson and Souriau, 2003; Holliger and Hudson, 2005; Nelson and ert, 2009) and even peptidomimetics inferred from the antibody Complementarity Determining Region (CDR) (Wada, 2013). The generation of CDR-derived peptidomimetics is challenging, but it would pave the way to ample biomedical (therapeutic and diagnostic) applications (Timmerman *et al.*, 2010, 2009; Fontenot *et al.*, 1998; Ponde *et al.*, 2011; Park *et al.*, 2000). However, it has been shown that some positions within the CDRs never participate in antigen binding and some off-CDR residues often contribute critically to the interaction with the antigen (Sela-Culang *et al.*, 2012). For this reason, the present work proposes a new *in silico* methodology to design targeted libraries of ligand peptides that is not based on CDRs, but on the amino acids that are important for the interaction with the antigen. The design of these peptides is not arbitrary, but based on the antigen sequence.

The first step to develop this methodology was to better understand the Ab-Ag interactions. Specifically, we identified the amino acids that are most frequently present in the epitope-paratope interactions, the most frequent physico-chemical types of interactions and the most frequent partners in these interactions.

The amino acid frequency in the Ab-Ag interface was analysed in several previous works. However, different cut-offs and methodologies were used to determine the interface boundaries,

such as the distance between atoms of the antigen and the antibody (DBS) and the difference of solvent-accessible surface (ΔSAS). Here, we developed a new method based on the interface molecular contact (I2R) to extract from the Ab-Ag interface only the amino acids that make interactions, using the STING database (Neshich *et al.*, 2006). By comparing the selections obtained using the I2R, DBS and ΔSAS methods, we show that DBS and ΔSAS missed part of the interacting residues that are important for the interface. Indeed, with a distance cut-off of 8Å, 60% of the amino acids that do not interact are selected in addition to the amino acids that do interactions. With a distance cut-off of 4Å, more than 10% of interacting residues are not selected and more than 20% of selected residues are not involved in interactions.

The I2R method also allowed studying the type of interactions and gave an approximation of the residue energetic contribution to the interface in a fast and easy way. Moreover, this selection method could be used to select targets for free-energy perturbation (FEP) (Xia *et al.*, 2012), or to identify binding hot-spots to facilitate the humanization of mouse antibodies (Hanf *et al.*, 2013). As previously noted with other selection techniques (Rubinstein *et al.* (2008); Kringelum *et al.* (2012); Ramaraj *et al.* (2012)), we found that the paratope was significantly enriched in Tyr, Ser and Trp residues. However, by comparing the occurrence of the I2R-selected amino acids and of all protein-protein interactions found in the STING database (Neshich *et al.*, 2006), we found that the occurrence of most of the Ab-Ag interface residues was significantly different (but not for Ala, Glu and Phe), thus characterizing the antigen-antibody interface as a special kind of protein-protein interaction. Concerning the extraction of the most frequent partners, we highlighted the importance of the cation-π interaction. Dalkas and colleagues (Dalkas *et al.*, 2014) previously reported that this type of interaction represents only 5% of the Ab-Ag interfaces, whereas in our study 84 of the 101 structures contained at least one cation-π interaction, where the positive charge is hold by the epitope. Moreover, 51 of them contained a double cation-π interaction composed of a positively charged residue in the epitope that interacted with two aromatic amino acids from the paratope. These results suggest that the cation-π interaction is highly conserved interaction in antigen-antibody interfaces but with low frequency as showed by Dalkas *et al.*

Besides gaining insights into the antigen-antibody interface characteristics, in this work we also describe a methodology to design peptide binders based on the epitope-paratope interface. In addition, this methodology was experimentally validated showing that 65% of the predicted peptides are reactive. Those peptides contain two consecutive conserved Tyr, a key residue in paratopes. Moreover, those Tyr could interact with hydrophobic amino acids from LiD1 epitope sequence (Phe37, Pro 43, Gly 49, Pro 51) or positively charged residue (Hys 48) via cation-π or even negatively charged residues via hydrogen bond (Asp 38 and Asp 39). The computational design protocol is far from perfect because it does not take into account the antibody structural properties. However, strategies, such as cysteine-constrained peptides, could be employed to mimic antibody loops as shown by Burns *et al.* and thus force a constrained conformation of our predicted peptides. In conclusion, our study provides insights into the principles that guide Ab-Ag interactions and describes an original methodology (EPI-Peptide Designer) to design ligand peptide libraries, based on a given antigen sequence. These targeted peptide ligand libraries might be useful for proteomic and high-throughput analyses for antigen characterization because they minimize the work to produce antibodies *in vivo*. Finally, this methodology might guide the development of a new generation of biosensors as well as therapeutic and diagnostic molecules.

Funding

This research was supported by Coordenação de Aperfeioamento de Pessoal de Nível Superior, Brazil (CAPES), Fundação de Amparo a Pesquisa do Estado de Minas Gerais, Brazil (FAPEMIG) and by funds of the Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (CNPq).

REFERENCES

Andreatta, M., Lund, O., and Nielsen, M. (2013). Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics*, **29**(1), 8–14.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**(1), 235–242.

Burns, V. A., Bobay, B. G., Basso, A., Cavanagh, J., and Melander, C. (2008). Targeting RNA with cysteine-constrained peptides. *Bioorg. Med. Chem. Lett.*, **18**(2), 565–567.

Chakrabarti, P. and Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, **47**(3), 334–343.

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, **256**(5520), 705–708.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.*, **14**(6), 1188–1190.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Dalkas, G. A., Teheux, F., Kwasigroch, J. M., and Rooman, M. (2014). Cation-, amino-, -, and H-bond interactions stabilize antigen-antibody interfaces. *Proteins*, **82**(9), 1734–1746.

Felicori, L., Araujo, S. C., de Avila, R. A., Sanchez, E. F., Granier, C., Kalapothakis, E., and Chavez-Olortegui, C. (2006). Functional characterization and epitope analysis of a recombinant dermonecrotic protein from Loxosceles intermedia spider. *Toxicon*, **48**(5), 509–519.

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., Wilson, I. A., and Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**(6031), 816–821.

Fontenot, J. D., Tan, X., and Phillips, D. M. (1998). Structure-based design of peptides that recognize the CD4 binding domain of HIV-1 gp120. *AIDS*, **12**(12), 1413–1418.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**(23), 3150–3152.

Hanf, K. J., Arndt, J. W., Chen, L. L., Jarpe, M., Boriack-Sjodin, P. A., Li, Y., van Vlijmen, H. W., Pepinsky, R. B., Simon, K. J., and Lugovskoy, A. (2013). Antibody humanization by redesign of complementarity-determining region residues proximate to the acceptor framework. *Methods*.

Harris, T. K. and Mildvan, A. S. (1999). High-precision measurement of hydrogen bond lengths in proteins by nuclear magnetic resonance methods. *Proteins*, **35**(3), 275–282.

Holliger, P. and Hudson, P. J. (2005). Engineered antibody fragments and the rise of single domains. *Nat. Biotechnol.*, **23**(9), 1126–1136.

Hudson, P. J. and Souriau, C. (2003). Engineered antibodies. *Nat. Med.*, **9**(1), 129–134.

Kringelum, J. V., Nielsen, M., Padkjæ r, S. B., and Lund, O. (2012). Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol. Immunol.*, **53**(1-2), 24–34.

Kunik, V., Peters, B., and Ofran, Y. (2012). Structural consensus among antibodies defines the antigen binding site. *PLoS Comput. Biol.*, **8**(2), e1002388.

Laune, D., Molina, F., Ferrieres, G., Villard, S., Bes, C., Rieunier, F., Chardes, T., and Granier, C. (2002). Application of the Spot method to the identification of peptides and amino acids from the antibody paratope that contribute to antigen binding. *J. Immunol. Methods*, **267**(1), 53–70.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**(5), 2177–2198.

Mancini, A. L., Higa, R. H., Oliveira, A., Dominiquini, F., Kuser, P. R., Yamagishi, M. E., Togawa, R. C., and Neshich, G. (2004). STING Contacts: a web-based

application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics*, **20**(13), 2145–2147.

Margulies, D. and Hamilton, A. D. (2010). Combinatorial protein recognition as an alternative approach to antibody-mimetics. *Curr Opin Chem Biol*, **14**(6), 705–712.

Morea, V., Lesk, A. M., and Tramontano, A. (2000). Antibody modeling: implications for engineering and design. *Methods*, **20**(3), 267–279.

Nelson, A. L. and ert, J. M. (2009). Development trends for therapeutic antibody fragments. *Nat. Biotechnol.*, **27**(4), 331–337.

Neshich, G., Mazoni, I., Oliveira, S. R., Yamagishi, M. E., Kuser-Falcao, P. R., Borro, L. C., Morita, D. U., Souza, K. R., Almeida, G. V., Rodrigues, D. N., Jardine, J. G., Togawa, R. C., Mancini, A. L., Higa, R. H., Cruz, S. A., Vieira, F. D., Santos, E. H., Melo, R. C., and Santoro, M. M. (2006). The Star STING server: a multiplatform environment for protein structure analysis. *Genet. Mol. Res.*, **5**(4), 717–722.

Nijssen, S. and Kok, J. (2004). A quickstart in frequent structure mining can make a difference. proceedings of the sigkdd.

Park, B. W., Zhang, H. T., Wu, C., Berezov, A., Zhang, X., Dua, R., Wang, Q., Kao, G., O'Rourke, D. M., Greene, M. I., and Murali, R. (2000). Rationally designed anti-HER2/neu peptide mimetic disables P185HER2/neu tyrosine kinases in vitro and in vivo. *Nat. Biotechnol.*, **18**(2), 194–198.

Pei, D. and Wavreille, A. S. (2007). Reverse interactomics: decoding protein-protein interactions with combinatorial peptide libraries. *Mol Biosyst*, **3**(8), 536–541.

Ponde, D. E., Su, Z., Berezov, A., Zhang, H., Alavi, A., Greene, M. I., and Murali, R. (2011). Development of anti-EGF receptor peptidomimetics (AERP) as tumor imaging agent. *Bioorg. Med. Chem. Lett.*, **21**(8), 2550–2553.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramaraj, T., Angel, T., Dratz, E. A., Jesaitis, A. J., and Mumey, B. (2012). Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta*, **1824**(3), 520–532.

Reichert J., Pechon P., T. A. D. M. K. (2010). Report summary: development trends for peptide therapeutics. *Pept. Ther. Found.*, pages 1–11.

Rubinstein, N. D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J. M., and Pupko, T. (2008). Computational characterization of B-cell epitopes. *Mol. Immunol.*, **45**(12), 3477–3489.

Sela-Culang, I., Alon, S., and Ofran, Y. (2012). A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J. Immunol.*, **189**(10), 4890–4899.

Smith, T. F., Waterman, M. S., and Fitch, W. M. (1981). Comparative biosequence metrics. *J. Mol. Evol.*, **18**(1), 38–46.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**(4), 327–332.

Swindells, M. B. (1995). A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci.*, **4**(1), 93–102.

Timmerman, P., Barderas, R., Desmet, J., Altschuh, D., Shochat, S., Hollestelle, M. J., Hoppener, J. W., Monasterio, A., Casal, J. I., and Meloen, R. H. (2009). A combinatorial approach for the design of complementarity-determining region-derived peptidomimetics with in vitro anti-tumoral activity. *J. Biol. Chem.*, **284**(49), 34126–34134.

Timmerman, P., Shochat, S. G., Desmet, J., Barderas, R., Casal, J. I., Meloen, R. H., and Altschuh, D. (2010). Binding of CDR-derived peptides is mechanistically different from that of high-affinity parental antibodies. *J. Mol. Recognit.*, **23**(6), 559–568.

Vanhee, P., van der Sloot, A. M., Verschueren, E., Serrano, L., Rousseau, F., and Schymkowitz, J. (2011). Computational design of peptide ligands. *Trends Biotechnol.*, **29**(5), 231–239.

Vlieghe, P., Lisowski, V., Martinez, J., and Khrestchatisky, M. (2010). Synthetic therapeutic peptides: science and market. *Drug Discov. Today*, **15**(1-2), 40–56.

Wada, A. (2013). Development of Next-Generation Peptide Binders Using In vitro Display Technologies and Their Potential Applications. *Front Immunol*, **4**, 224.

Xia, Z., Huynh, T., Kang, S. G., and Zhou, R. (2012). Free-energy simulations reveal that both hydrophobic and polar interactions are important for influenza hemagglutinin antibody binding. *Biophys. J.*, **102**(6), 1453–1461.

Yin, H. and Hamilton, A. D. (2005). Strategies for targeting protein-protein interactions with synthetic agents. *Angew. Chem. Int. Ed. Engl.*, **44**(27), 4130–4163.

Yin, H., Slusky, J. S., Berger, B. W., Walters, R. S., Vilaire, G., Litvinov, R. I., Lear, J. D., Caputo, G. A., Bennett, J. S., and DeGrado, W. F. (2007). Computational design of peptides that target transmembrane helices. *Science*, **315**(5820), 1817–1822.