Bioinformatics, 32(10), 2016, 1462–1470
doi: 10.1093/bioinformatics/btw014
Advance Access Publication Date: 18 January 2016
Original Paper

Structural bioinformatics

EPI-peptide designer: a tool for designing peptide ligand libraries based on epitope–paratope interactions

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

[1]Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil, [2]Sys2Diag, FRE3690-CNRS/ALCEDIAG, Montpellier, France and [3]Computational Biology Research Group, Embrapa Informática Agropecuária, Campinas, SP, Brazil

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on 26 May 2015; revised on 5 December 2015; accepted on 8 January 2016

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 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 (Ab–Ag) 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 Ab–Ag 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), an enrichment of 22% compared to randomly generated 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 and implementation: All software developed in this work are available at http://www.biocomp.icb.ufmg.br/biocomp/

Contact: liza@icb.ufmg.br

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

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 *et al.*, 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; Vanhee *et al.*, 2011; Yin *et al.*, 2007). *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 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. In addition, 460 random peptide were designed in which 43% of them were positives. These results indicate an enrichement of 22% in ligand peptides designed using EPI-Peptide Desginer. This is, to our knowledge, the first generator of peptide ligand libraries based on epitope–paratope interface.

2 Methods

2.1 Dataset extraction and interface selection

Structures from Ab–Ag used in this work were extracted from SAbDab (Dunbar *et al.*, 2014). Were selected all structure with protein/peptide antigen and from this set, only structures resolved by X-ray and a resolution lower or equal to 3 Å were kept. To analyze 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 Å). 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.

2.2 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 (i) classifying the atoms in groups according to their electrostatic behavior and position in the amino acid (main or side chain) and (ii) 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).

2.3 Redundancy removal

To extract meaningful information from the interface dataset, one firstly removed structure with the same antigen name and keeping the lowest resolution one and secondly removed redundancies by selecting only the DSE and DSP sequences from the complex (with a cut-off of 6 Å). 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 272 PDB structures, 51 antibody–peptide complexes (here, peptides are defined as molecules smaller than 30 amino acids) and 221 antibody–protein complex.

2.4 Interface statistical analysis

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

$$\%\mathrm{Occ}_n = \frac{\mathrm{Occ}_n}{\mathrm{Occ}_{\mathrm{total}}} \times 100,$$

where n is an amino acids, $\%\mathrm{Occ}_n$ is the percentage of occurrence of n, Occ_n is the occurrence of n and $\mathrm{Occ}_{\mathrm{total}}$ 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.

2.5 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:

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

and the false discovery rate (FDR) as:

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

where TP is the true positive, FP the false positive and FN the false negative.

2.6 Computation of the most frequent interface partners using graph analysis

To analyze 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), the most conserved subgraphs were extracted 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).

2.7 Assessment of paratope residue prediction

Based on the Bayesian probabilities extracted from the epitope–paratope graphs, the amino acid sequence and the interaction of a given paratope were predicted using a given epitope sequence. To evaluate the prediction of residues and interactions, using a 10-fold cross validation on the complete dataset. 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.

2.8 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.

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

2.9 EPI-peptide design and random peptides generation

In order to test the effectiveness of the method, 800 EPI-Peptides were generated 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). Random peptide sequences were generated using the expasy tool RandSeq (Gasteiger *et al.*, 2003) using the average amino acid composition from the Uniprot/Swiss-Prot database and filtered following the same methodology than the EPI-pepitde. The statistic used by the program for the different amino acids are: Ala: 8.26, Gln: 3.93, Leu: 9.66, Ser: 6.59, Arg: 5.53, Glu: 6.74, Lys: 5.83, Thr: 5.34, Asn: 4.06, Gly: 7.08, Met: 2.41, Trp: 1.09, Asp: 5.46, His: 2.27, Phe: 3.86, Tyr: 2.92, Cys: 1.37, Ile: 5.94, Pro: 4.71, Val: 6.87.

2.10 Peptide synthesis on cellulose membranes and binding assay

Two membranes with four hundred and sixty peptides each were synthesized on a cellulose membrane as previously described by Laune *et al.* (2002). One of the membranes corresponding to Epi-peptides and other to random peptides. The membranes were blocked by incubation with 137 mM NaCl, 2.68 mM KCl, 3% sacarose, 5% bovine serum albumin, 0.1% tween 20 in 50 mM Tris base at room temperature overnight. 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 analyzed by GibbsCluster (Andreatta *et al.*, 2013) and Weblogo (Crooks *et al.*, 2004) tools. Statistical analysis comparing the success of binding were done using binomial test.

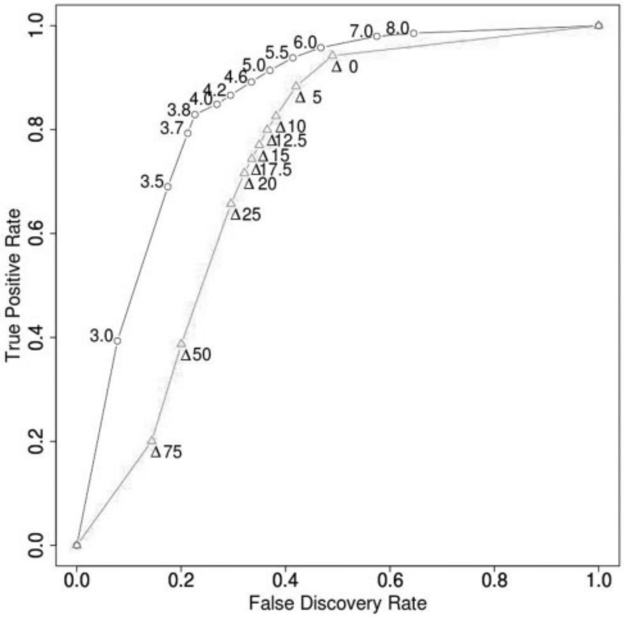

Fig. 1. Comparison of DBS (black circles) and ΔSAS (red triangles) residue selection using different cut-offs relative to the I2R method

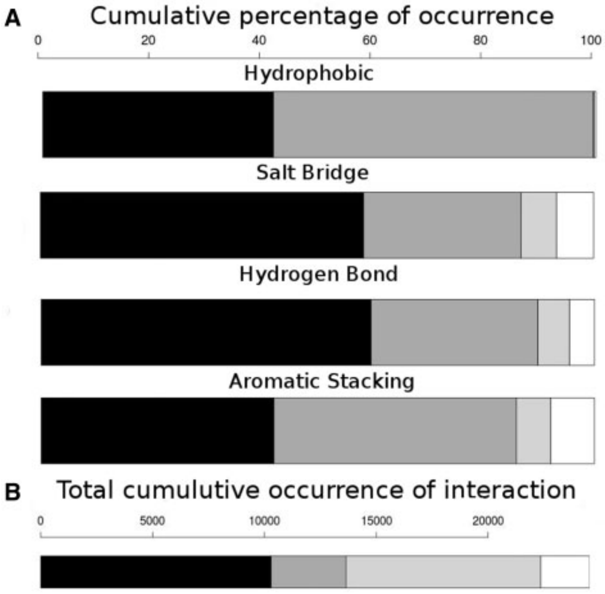

Fig. 2. (**A**) Percentage of molecular interactions by type using DSB from 0 to 3 Å (black), from 3 to 4 Å (dark gray), from 4 to 5 Å light gray) and from 5 to 8 Å (white). (**B**) Cumulative occurrence of hydrophobic interactions (black), salt bridges (dark grey), hydrogen bonds (light grey) and aromatic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

3 Results

3.1 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 272 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 3 Å distance, while the maximum TPR (also called Recall) was reached with 8 Å. The DBS had a higher surface under the curve and the highest value of TPR – FDR was reached for a distance of 3.8 Å. Most DBS-based Ab–Ag structure studies use a cut-off between 4 and 6 Å. For a distance of 5 Å, with this plot, 91.4% (TPR) of interacting residues were selected; however, 36% of the selected residues did not do any kind of interaction. Surprisingly, to reach the maximum TPR, a distance cut-off of 8 Å was needed. As most of the molecular interaction maximum distances are lower than 6 Å, we further investigated the interaction repartition.

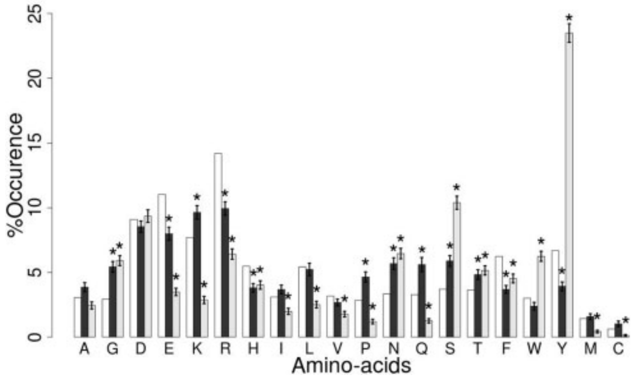

Fig. 3. Comparison of the occurrence (in percentage) of all interacting residues in STING RDB (white), I2R epitopes (black) and I2R paratopes (grey). Error bars are calculated as the standard deviation divided by the root square of the set size. Stars represent statistically significant differences compared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

As all interface interactions are not selected by the 5 Å 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 5 Å, as expected based on the previous results, allowed the selection of most interactions, but still missed 3.9% of them, specifically 6.7% of all salt bridges, 4.5% of all hydrogen bonds and 7.8% of all aromatic stacking, but none of the hydrophobic interactions. The hydrogen bonds with a distance bigger than 5 Å 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.

3.2 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.

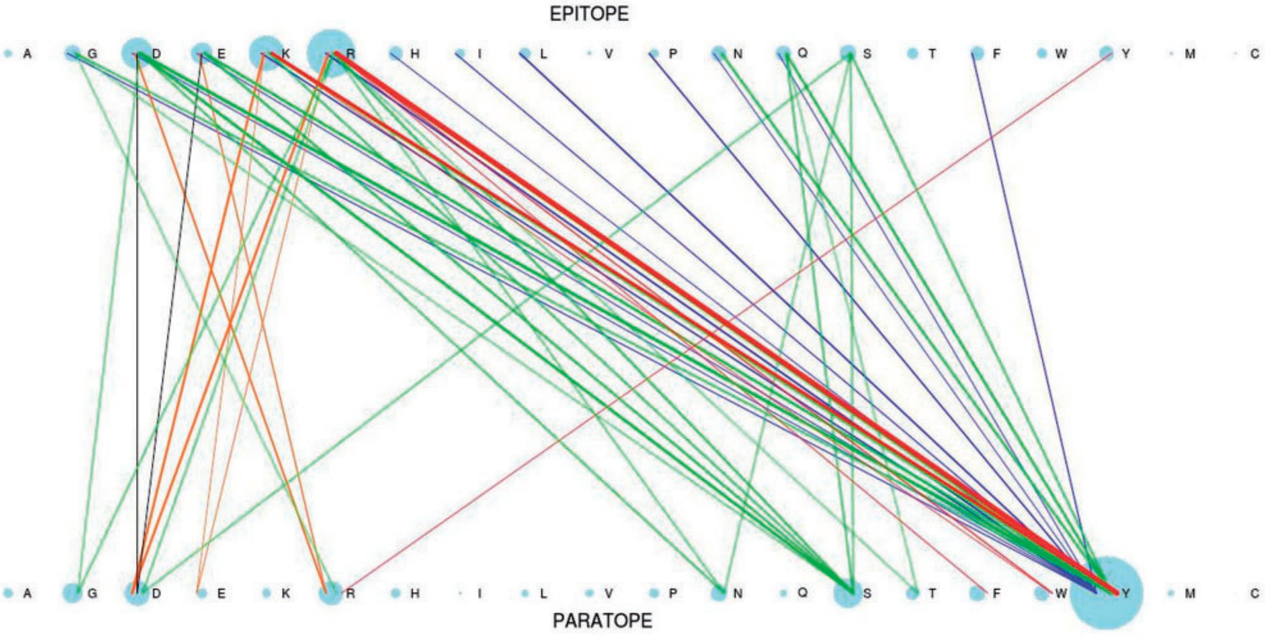
EPITOPE
A G D E K R H I L V P N Q S T F W Y M C
A G D E K R H I L V P N Q S T F W Y M C
PARATOPE

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 50 are represented (Color version of this figure is available at *Bioinformatics* online.)

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 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.

3.3 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, 224 of the 272 selected structures contained at least one cation-π interaction in which the positive charge was hold by the epitope. In addition the six most conserved subgraphs with three or more nodes all contains a cation-π with another molecular interaction. 59 structures contain 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.

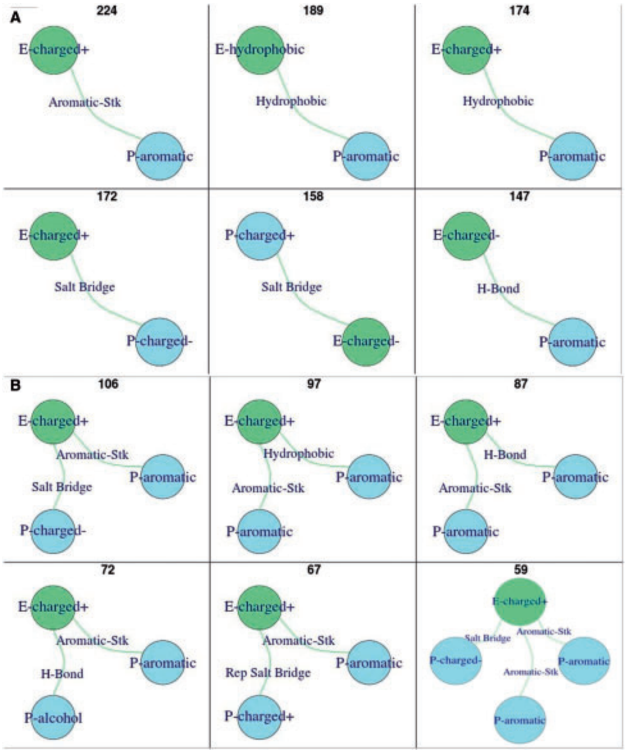

Fig. 5. Each cell contains one of the six most common subgraphs with two (A) or three or more nodes (B) from the interface graphs based on the 272 PDB structures dataset. The title indicates in how many interfaces the motif was observed at least once

3.4 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 Figure 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 full dataset and a 10-fold cross validation method with all the interactions and the seven residue groups (Table 1). The prediction program includes two tuning parameters. First, the probability cut-off which determine if an interaction is added or not and the 'influence' corresponding to the weight of the neighbour residues on the probability value. Those two parameters were tuned in order to obtain maximum sensitivity. The best results was reached with a probability cut-off of 0.1, and an influence of 3.5 obtaining a sensitivity of 66% and a specificity of 70%. These results indicate that the residues selected to compose the predicted ligand peptides are not random.

3.5 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 D). To demonstrate how the EPI-Peptide Designer works, we used the epitope from the LiD1 protein. 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).

3.6 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 (Supplementary Table S1), chemically synthesized and assayed against biotinylated LiD1. From 460 sequences synthesized, 218 were considered highly positive (47% of sequences, squares on Fig. 7A) and 83 had a lower reactivity (18%), a total of 301 reactive peptides. 159 peptides (35%) presented no reactivity. Highly positive peptides were clustered in two groups and a graphical representation of the patterns from each multiple sequence alignment was computed (Fig. 7B and C). Cluster 1 (Supplementary Table S1) 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 cluster 2 (Supplementary Table S2) includes two conserved tyrosines at position 14 and 15 (53 and 54% respectively). In order to address the ability of EPI-peptide Designer to perform a better peptide prediction compared to random generate peptides, 460 random peptides were generated using Swiss-Prot/Uniprot amino acid statistic (Supplementary Table S2) and tested against LiD1-biotin using the same experimental conditions. From these peptides, 128 were considered highly positive (28% of sequences) and 71 presented lower reactivity (15%), a total of 199 reactive peptides, corresponding to 43% of reactivity (102 less than Epi-peptide prediction). 261 peptides (57%) were not reactive (Supplementary Fig. 1S). The peptides were also probed against biotin alone (Supplementary Fig. 2S) without any reactivity observed. With these results, we can conclude that EPI-peptides show an increase of 22% of the success in term of binding corresponding to a P-value of 2.2e^{-16} using a binomial test compared to the randomly generated peptides.

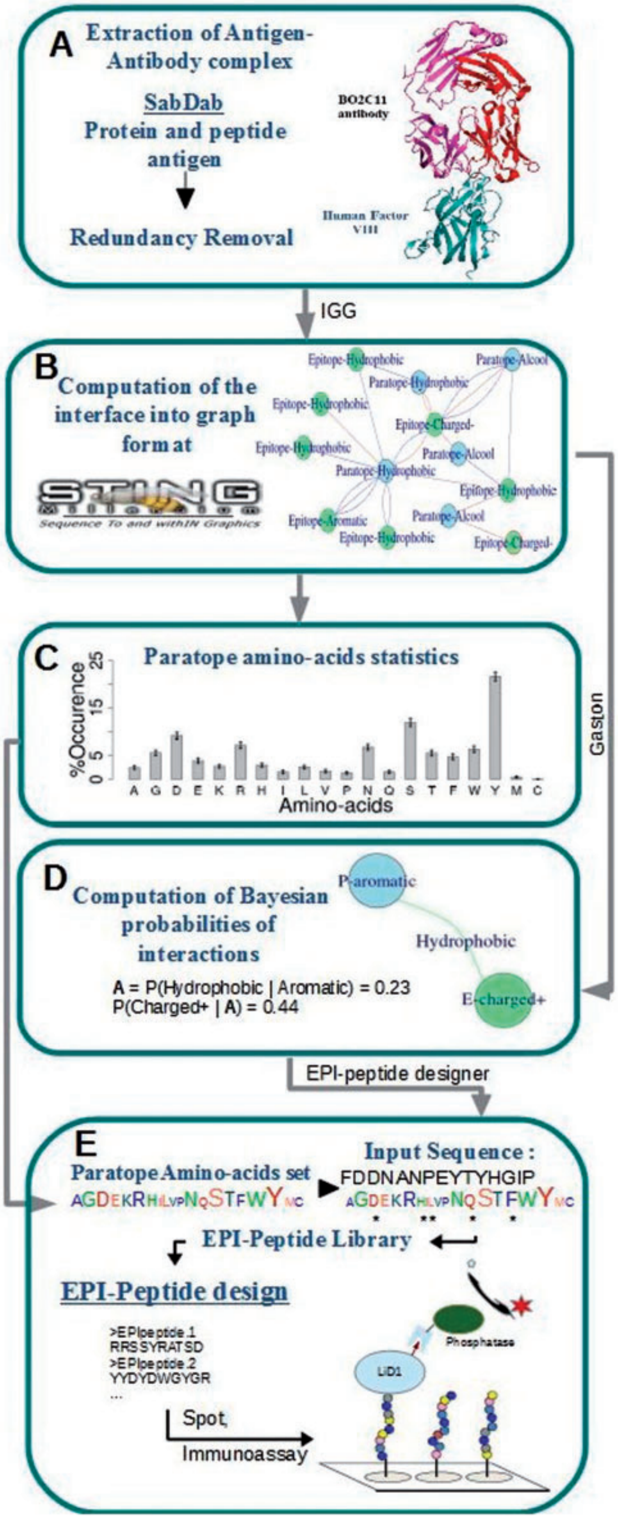

Fig. 6. Schematic of the EPI-peptide Design method to design targeted libraries of peptide ligands. (**A**) Extraction of the Ab–Ag complexed structure dataset. (**B**) Interface Graph Generator transform the epitope–paratope interface into a graph format using computed molecular interactions extracted the BlueStar STING database. (**C**) Computation of the paratope amino acid occurrence using Gaston (Nijssen and Kok, 2004). (**D**) Interaction probability. (**E**) Based Residue Library (BRL) is modified acording to epitope sequence 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

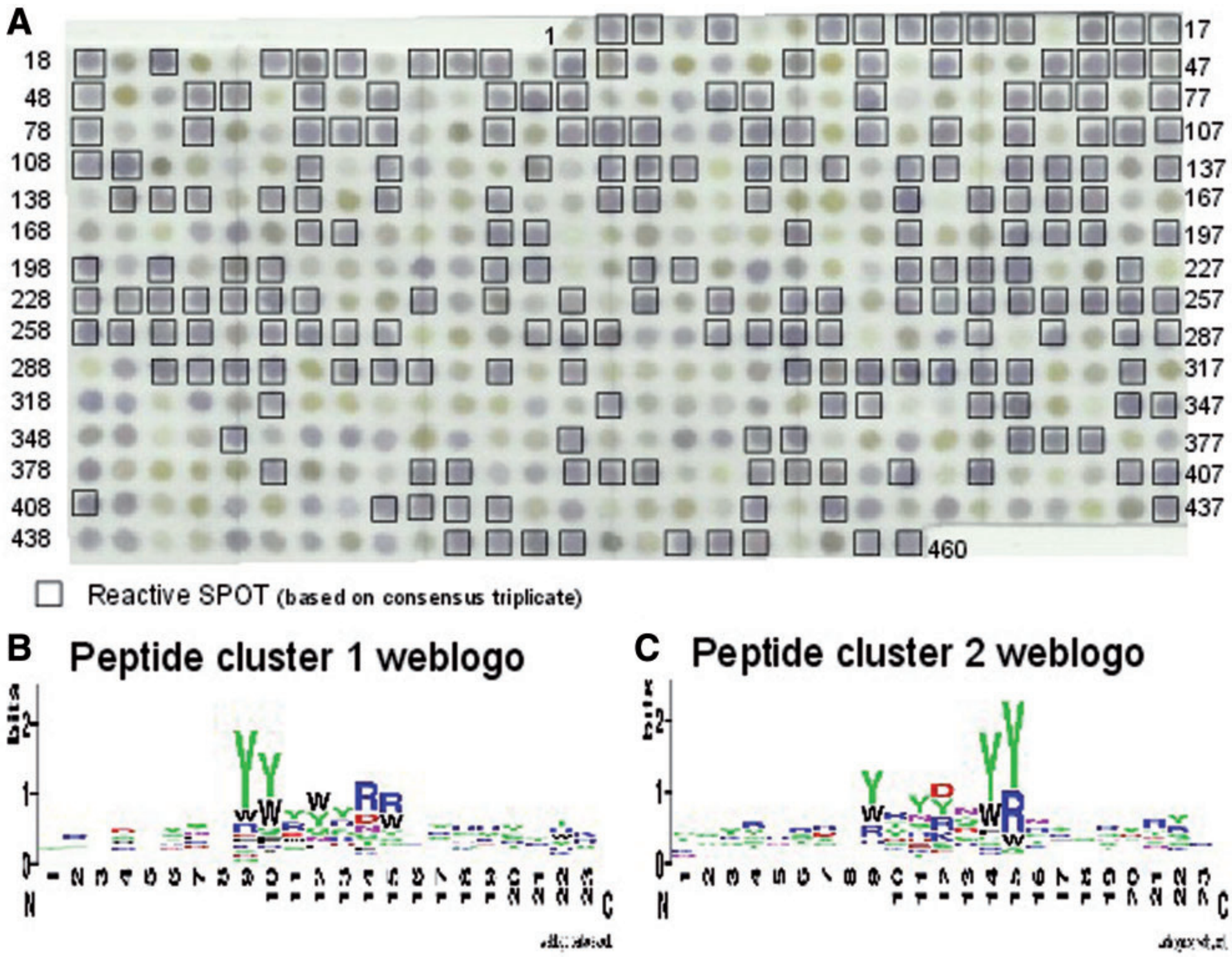

Fig. 7. Experimental validation and analysis of EPI-peptides prepared by SPOT method (**A**). 460 EPI-peptides predicted against LiD1 protein epitope was synthesized. 20 μg/ml of LiD1-biotin followed by alkaline phosphatase-conjugated avidin. 1:10,000 revealed binding peptides. Black boxes represent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

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 (Holliger and Hudson, 2005; Hudson and Souriau, 2003; 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 larger biomedical (therapeutic and diagnostic) applications (Fontenot *et al.*, 1998; Park *et al.*, 2000; Ponde *et al.*, 2011; Timmerman *et al.*, 2009, 2010). 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 physicochemical types of interactions and the most frequent partners in these interactions.

The amino acid frequency in the Ab–Ag interface was analyzed 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 8 Å, more than 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 4 Å, more than 8% 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 (Kringelum *et al.*, 2012; Ramaraj *et al.*, 2012; Rubinstein *et al.*, 2008), 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 *et al.* (2014) previously reported that this type of interaction represents only 5% of the Ab–Ag interfaces, whereas in our study 224 of the 272 structures contained at least one cation-π interaction, where the positive charge is hold by the epitope. Moreover, 59 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 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 ligands based on the epitope–paratope interface. In addition, this methodology was experimentally validated showing an enrichment of 22% of reactive peptides comparing to random generate peptides. The most reactive EPI-peptides designed 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.* (2008) 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.

Acknowledgements

We thank Ivan Mazoni for his help with the extraction of the BLUE Star Sting data and also Pr. Loïc Cerf for his help on the subgraphs and statistic analysis.

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).

Conflict of Interest: none declared.

References

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

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Burns,V.A. *et al.* (2008) Targeting RNA with cysteine-constrained peptides. *Bioorg. Med. Chem. Lett.*, **18**, 565–567.

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

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

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

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

Dalkas,G.A. *et al.* (2014) Cation-, amino-, and H-bond interactions stabilize antigen-antibody interfaces. *Proteins*, **82**, 1734–1746.

Dunbar,J. *et al.* (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146.

Felicori,L.F. *et al.* (2006) Functional characterization and epitope analysis of a recombinant dermonecrotic protein from Loxosceles intermedia spider. *Toxicon*, **48**, 509–519.

Fleishman,S.J. *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–821.

Fontenot,J.D. *et al.* (1998) Structure-based design of peptides that recognize the CD4 binding domain of HIV-1 gp120. *Aids*, **12**, 1413–1418.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Gasteiger,E. *et al.* (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.

Hanf, K.J. *et al.* (2014) Antibody humanization by redesign of complementarity-determining region residues proximate to the acceptor framework. *Methods*, **65**, 68–76.

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

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

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

Kringelum,J.V. *et al.* (2012) Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol. Immunol.*, **53**, 24–34.

Laune,D. *et al.* (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**, 53–70.

Lo Conte,L. *et al.* (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

Mancini,A.L. *et al.* (2004) STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics*, **20**, 2145–2147.

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

Morea,V. *et al.* (2000) Antibody modeling: implications for engineering and design. *Methods*, **20**, 267–279.

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

Neshich,G. *et al.* (2006) The Star STING server: a multiplatform environment for protein structure analysis. *Genet. Mol. Res.*, **5**, 717–722.

Nijssen,S. and Kok,J. (2004). A quickstart in frequent structure mining can make a difference. In: Proceedings of the SIGKDD.

Park,B.W. *et al.* (2000) Rationally designed anti-HER2/neu peptide mimetic disables P185HER2/neu tyrosine kinases in vitro and in vivo. *Nat. Biotechnol.*, **18**, 194–198.

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

Ponde,D.E. *et al.* (2011) Development of anti-EGF receptor peptidomimetics (AERP) as tumor imaging agent. *Bioorg. Med. Chem. Lett.*, **21**, 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. *et al.* (2012) Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta*, **1824**, 520–532.

Reichert,J. *et al.* (2010) Report summary: development trends for peptide therapeutics. *Pept. Ther. Found*, 1–11. http://www.biomodulation.com/attachments/article/104/development%20%20trends%20for%20peptide%20terapeutics.pdf.

Rubinstein,N.D. *et al.* (2008) Computational characterization of B-cell epitopes. *Mol. Immunol.*, **45**, 3477–3489.

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

Sobolev,V. *et al.* (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.

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

Timmerman,P. *et al.* (2009) A combinatorial approach for the design of complementarity-determining region-derived peptidomimetics with in vitro anti-tumoral activity. *J. Biol. Chem.*, **284**, 34126–34134.

Timmerman,P. *et al.* (2010) Binding of CDR-derived peptides is mechanistically different from that of high-affinity parental antibodies. *J. Mol. Recognit.*, **23**, 559–568.

Vanhee,P. *et al.* (2011) Computational design of peptide ligands. *Trends Biotechnol.*, **29**, 231–239.

Vlieghe,P. *et al.* (2010) Synthetic therapeutic peptides: science and market. *Drug Discov. Today*, **15**, 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. *et al.* (2012) Free-energy simulations reveal that both hydrophobic and polar interactions are important for influenza hemagglutinin antibody binding. *Biophys. J.*, **102**, 1453–1461.

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

Yin,H. *et al.* (2007) Computational design of peptides that target transmembrane helices. *Science*, **315**, 1817–1822.