PLOS | ONE

RESEARCH ARTICLE

Proteochemometric Modeling of the Antigen-Antibody Interaction: New Fingerprints for Antigen, Antibody and Epitope-Paratope Interaction

Tianyi Qiu[1], Han Xiao[2], Qingchen Zhang[1], Jingxuan Qiu[1], Yiyan Yang[1], Dingfeng Wu[1], Zhiwei Cao[1,3]*, Ruixin Zhu[1,4]*

1 Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China, **2** Department of Computer Science, University of Helsinki, Helsinki, FI-00014, Finland, **3** Shanghai Center for Bioinformation Technology, Shanghai 201203, China, **4** School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian 116600, Liaoning, China

* rxzhu@tongji.edu.cn (RZ); zwcao@tongji.edu.cn (ZC)

Citation: Qiu T, Xiao H, Zhang Q, Qiu J, Yang Y, Wu D, et al. (2015) Proteochemometric Modeling of the Antigen-Antibody Interaction: New Fingerprints for Antigen, Antibody and Epitope-Paratope Interaction. PLoS ONE 10(4): e0122416. doi:10.1371/journal.pone.0122416

Academic Editor: Serge Muyldermans, Vrije Universiteit Brussel, BELGIUM

Received: November 23, 2014

Accepted: February 20, 2015

Published: April 22, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by grants from National Natural Science Foundation of China (31200986, 31171272) and Ministry of Science and Technology China (2010CB833601). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Despite the high specificity between antigen and antibody binding, similar epitopes can be recognized or cross-neutralized by paratopes of antibody with different binding affinities. How to accurately characterize this slight variation which may or may not change the antigen-antibody binding affinity is a key issue in this area. In this report, by combining cylinder model with shell structure model, a new fingerprint was introduced to describe both the structural and physical-chemical features of the antigen and antibody protein. Furthermore, beside the description of individual protein, the specific epitope-paratope interaction fingerprint (EPIF) was developed to reflect the bond and the environment of the antigen-antibody interface. Finally, Proteochemometric Modeling of the antigen-antibody interaction was established and evaluated on 429 antigen-antibody complexes. By using only protein descriptors, our model achieved the best performance ($R2 = 0.91, Q^2_{test} = 0.68$) among peers. Further, together with EPIF as a new cross-term, our model ($R2 = 0.92, Q^2_{test} = 0.74$) can significantly outperform peers with multiplication of ligand and protein descriptors as a cross-term ($R2 \leq 0.81, Q^2_{test} \leq 0.44$). Results illustrated that: 1) our newly designed protein fingerprints and EPIF can better describe the antigen-antibody interaction; 2) EPIF is a better and specific cross-term in Proteochemometric Modeling for antigen-antibody interaction. The fingerprints designed in this study will provide assistance to the description of antigen-antibody binding, and in future, it may be valuable help for the high-throughput antibody screening. The algorithm is freely available on request.

Introduction

Antigen-antibody interaction is an important and fundamental biochemical function in immune system. By recognizing the epitope area on the surface of protein antigen, antibodies

secreted by B-cell are able to interact with those invasive antigens and then neutralize them to keep our body safe [1,2]. However, for the new emerging antigens caused by mutation, previous antibody may not work effectively due to the antigenicity variance. Since the mechanism of antigen-antibody interaction remains elusive, when a new antigen emerges, experimental methods are most frequently used to test whether the previous antibody or antiserum can still recognize the new antigen or not [3], or to produce functional antibody molecules corresponding to the antigen through mass clonal cell screening [4]. As a special protein-protein interaction, antigen-antibody interaction occurs neither in the whole protein nor in the entire surface, but in the specific "binding site" [5]. For the antigen-antibody interaction, these specific "binding site" can be called as "epitope-paratope interaction site" [1,2]. It has been frequently reported that one or several mutation in "binding site" often lead to large binding affinity changes [3,6,7,8,9]. This may correspond to two interesting phenomenon in antigen-antibody interaction, one is that antigens may change a few amino acids to produce a new epitope through continual antigenic drift [10]; the other one is that antibodies can recognize millions of different antigens through minor amino acid changes in paratope area [11]. Both "antigenic drift" mutations in epitopes and "adaptive" mutations in paratopes are caused by amino acid sequence or structure variations. Despite of the high specificity between antigen and antibody binding, different studies have showed that similar epitopes can still be recognized or cross-neutralized by the same antibody [12] or biological trigger [13]. Therefore, how to accurately characterize the interface of "epitope-paratope interaction" and how to handle multi-target screening problems is the key issue to study the mechanisms of interaction between those biological macromolecules [5].

Till now, many methods have been developed to characterize the interface features of protein, which can be roughly divided into three categories: 1) Geometry-based [14,15,16]; 2) Energy-based [17] and 3) Signature-based [18,19] methods. "Geometry-based" methods contain three aspects: "amino acid-based" [14], "atom-based" [15] and "Geometric & Physical-chemical-based" [16] method. These kinds of method utilize three-dimensional coordinates of atoms, pseudo-atoms and residues to superimpose two structures and quantify their similarity. "Energy-based" methods refer to those decomposition methods after molecular dynamic simulations. Those methods can decompose the binding free energy in the interaction interface into specific residues, and quantitatively characterize the contribution of various residues for the entire protein-protein interaction [20]. Compared to those two above methods, "Signature-based" methods [18,19,21] do not require numerous computing resources and precise three-dimensional coordinate information, which may make it more robust when dealing with slight structural changes occurs in the "binding site" [5].

These methods greatly promoted development for "protein binding site" analysis. However, above methods exist several limitations in the case of "epitope-paratope interaction": "Geometric-based" methods and "Signature-based" methods only derive relevant features either from receptor side or ligand side without considering interaction features. This is not able to completely describe the features of "epitope-paratope interaction". As for the "energy-based" methods, molecular dynamic simulation process took the information of interaction into account, however, due to the time-consuming simulations, it is not able to achieve high-throughput screen analysis. Moreover, the "energy-based" methods may often unable to extract geometric features in the interaction interface, which makes it can only be used to build explanatory models. Therefore, developing a new descriptor for "protein binding site", which can reflectboth spatial geometric features and interaction forces with robustness, accuracy and operational efficiency, is highly desired. A recent idea of "interaction fingerprint" developed in the area of drug design makes it possible to analyze the interaction between two molecular structures [22]. By taking features of antigen-antibody interaction into consideration, a new set

of epitope-paratope interaction fingerprint (EPIF) has been firstly generated to describe the antigen-antibody interaction. Meanwhile, a new set of protein descriptors has been established to describe the residue layout and physical-chemical features for both antigen and antibody proteins.

As an extension of the quantitative structure-activity relationship (QSAR) methods, Proteochemometric (PCM) Modeling has been widely used to study the cross-interactions between a series of ligands and a series of receptors [23,24]. Different from QSAR, PCM contains information from both the ligand and the target descriptors to correlate with activity data. Moreover, an additional term 'cross-term' was introduced to describe interaction features and most of the previous studies defined the cross-terms of PCM model as the Multiplication of Ligand and Protein Descriptors (MLPD) [25,26]. It is worthy of compliment that MLPD contains information from both side of the interaction interface, which can be considered as candidates of cross-term. However, MLPD is generated by the multiplication of ligand and protein descriptors, which has higher time-complexity (n^2) than single side descriptors (n), also the significance of MLPD is not easy to interpret. Thus, our new invented epitope-paratope interaction fingerprints (EPIF) which describes the antigen-antibody interaction can be used as "cross-terms" to address this issue. By combining our new protein fingerprint with EPIF, Proteochemometric Modeling was constructed to simulate the relationship between multiple antigen and antibody proteins in this study.

Results and Discussion

Kernel Selection

Our PCM modeling was performed by employing support vector regression (SVR) methods with different kernels. As a widely used regression model, SVR has a number of advantages over the conventional linear regressions, especially for its robustness to avoid over-fitting [27,28]. By the use of non-linear kernel, SVR projects the data into a high-dimensional space and constructs a set of hyperplanes in it for regression. The construction of learning machine is based on how the inner-product kernel is generated. Therefore, the selection of the kernel function is very important. In our study, four commonly used kernels (Table 1) were implemented in SMOreg of Weka (version 3.7) with default parameters. Previous studies indicated that kernel may perform differently on different datasets, and the adaptation of kernels were based on the type of the dataset [29]. In our PCM modeling, 10-fold cross-validation was evaluated on all four kernels to select effective kernel functions. The cross-validation results (Q^2_{CV}) of each kernel with different combination of fingerprints were listed in Table 2. The results showed that Normalized Poly Kernel function obtains better predictive ability than the other three kernel functions. Therefore, Normalized Poly Kernel function was selected for PCM modeling and performance evaluation.

Table 1. Summary of Kernels.

Type of Kernels	Functions
Normalized Poly Kernel	$k(x,y) = (x^T y + c)^d / \sqrt{(x^{T+1} + y^{T+1})}$
Polynomial Kernel	$k(x,y) = (x^T y + c)^d$
Puk	$k(x,y) = \frac{1}{\left[1+\left(\frac{2\sqrt{\|x-y\|^2\sqrt{2^{(1/\omega)}-1}}}{\sigma}\right)^2\right]^\omega}$
RBF Kernel	$k(x,y) = \exp(-\gamma\|x-y\|^2)$

doi:10.1371/journal.pone.0122416.t001

Table 2. Q^2_{CV} of our three fingerprint combinations with different SVR methods and kernels.

Fingerprint\Kernel	Normalized Poly Kernel	Polynomial Kernel	Puk	RBF Kernel
Fab-Fag-EPIF[a]	**0.52**	0.35	0.40	0.51
Fab-Fag-MLPD[b]	**0.49**	0.36	0.26	0.49
Fab-Fag[c]	**0.47**	0.31	0.38	0.43
Average	**0.49**	0.34	0.35	0.48

[a]Models created using antibody fingerprint and antigen fingerprint with EPIF as cross-term

[b]Models created using antibody fingerprint and antigen fingerprint with the multiplication of antibody fingerprint and antigen fingerprint as cross-term

[c]Models created using only antibody fingerprint and antigen fingerprint.

doi:10.1371/journal.pone.0122416.t002

Development and evaluation of Proteochemometric Modeling

Proteochemometric model with different combination of descriptors were summarized in Table 3. To evaluate the performance of our antigen-antibody interaction fingerprint in Proteochemometric Modeling, three fingerprint combinations (Fab-Fag-EPIF, Fab-Fag-MLPD, Fab-Fag) were tested. Results indicated that Fab-Fag-EPIF obtained better predictive ability than those without cross-terms or those using MLPD as cross-terms. Also, the prediction performance of Fab-Fag-EPIF and Fab-Fag were better than the model with MLPD as cross-terms, which illustrated that the conventional cross-term of MLPD was not only being outperformed by new introduced cross-term of PLIF but also being surpassed by our protein fingerprints without cross-terms.

The original idea of cross-terms is to add information from both sides of ligand-target interaction [30], which intended to describe the features of the interface between ligand and protein. For protein-protein interaction, especially antigen-antibody interaction, the interface features maybe more related to the interaction forces and environments of the binding site. Thus, interaction fingerprint which is generated from the antigen-antibody complexes and could directly describe the interaction between antigen and antibody from different aspects of important

Table 3. Goodness-of-fit (R^2) and predictive ability (Q^2_{test}) of the models which were obtained by different model.

Fingerprint\Kernel	R^2	Q^2_{test}	*MAE*	*RMSE*	*RAE*	*RRSE*
Fab-Fag-EPIF[a]	0.92	**0.74**	**124.10**	**164.28**	**69.12%**	**69.41%**
Fab-Fag-MLPD[b]	0.99	0.61	139.44	187.92	77.66%	79.39%
Fab-Fag[c]	0.91	**0.68**	**131.17**	**175.30**	**73.06%**	**74.06%**
Sab-Sag[d]	0.79	0.50	137.86	188.21	82.13%	86.26%
Gab-Gag[e]	0.39	0.21	150.17	193.58	94.85%	97.70%
Sab-Sag-EPIF[f]	0.81	0.44	150.11	214.66	84.44%	92.70%
Gab-Gag-EPIF[g]	0.57	0.42	137.56	179.80	86.88%	90.75%
Gab-Gag-MLPD[h]	0.41	0.22	149.66	193.45	94.52%	97.64%

[a]Models created using antibody fingerprint and antigen fingerprint with EPIF as cross-term

[b]Models created using antibody fingerprint and antigen fingerprint with the multiplication of antibody fingerprint and antigen fingerprint as cross-term

[c]Models created using only antibody fingerprint and antigen fingerprint.

[d]Models created using only sequence similarity descriptor of antibody and sequence similarity descriptor of antigen

[e]Models created using only geometry descriptor of antibody and geometry descriptor of antigen

[f]Models created using sequence similarity descriptor of antibody and sequence similarity descriptor of antigen with EPIF as cross-term

[g]Models created using geometry descriptor of antibody and geometry descriptor of antigen with EPIF as cross-term

[h]Models created using geometry descriptor of antibody and geometry descriptor of antigen with the multiplication of antibody descriptor and antigen descriptor as cross-term

doi:10.1371/journal.pone.0122416.t003

features may be more suitable for cross-terms [31]. Cross-terms calculated by the multiplication of ligand and target descriptors may not be a reliable reflection of the binding side, sometimes performed even worse than those only use fingerprints of both antibody and antigen side [31]. Therefore, it may indicate that, in the case of antigen-antibody recognition, only when a suitable cross-term such as EPIF is used in Proteochemometric Modeling, the model performance can be significantly improved.

Compared with peers

Existed protein descriptors can be divided into sequence similarity descriptors and geometric structure descriptors [32]. In this study, both sequence similarity descriptor and geometry descriptor were compared with our fingerprints. For sequence similarity descriptor, the amino acid sequences of all the antigen and antibody proteins were retrieved from PDB [33]. BLAST (version2.2.28) was used to calculate sequence identities of all the antigen and antibody structures. Finally, a 429-bit sequence similarity descriptor was obtained. For geometric descriptor, three different aspects were taken into considerations: bond length, bond angel and dihedral angle. 41-bit of protein geometry descriptors were obtained for each antigen-antibody proteins in our dataset. Types of protein geometry descriptors could be seen in S1 Table.

The performance of our antigen-antibody fingerprints compared with peers can be found in **Fig 1 and Table 3**. Here, 8 different combinations of descriptor were used to establish the PCM model (The MLPD of sequence similarity descriptors contains 429*429 bits, which were not adopted in this study). For using protein descriptor only, results indicated that the fingerprint of Fab-Fag ($R^2 = 0.91, Q^2_{test} = 0.68$) outperformed other descriptors($R^2 \leq 0.79, Q^2_{test} \leq 0.50$). After added cross-terms, the introducing of EPIF as cross-terms combined with our protein fingerprints (Fab-Fag-EPIF) can achieve the best predictive ability ($Q^2_{test} = 0.74$) among all others($Q^2_{test} \leq 0.50$). This demonstrated that, Proteochemometric Modeling with our new invented antigen-antibody structure fingerprint and EPIF may be more appropriate than existed protein sequence similarity descriptors or structure geometric descriptors in the case of antigen-antibody interaction. Results also indicated that the prediction ability of using only the antibody and antigen geometric descriptors (Gab-Gag) is the bottom line of the PCM model as well as the prediction ability of added multiplication of antibody and antigen geometric descriptors as cross-term (Gab-Gag-MLPD). However, by adding EPIF as cross-term to geometric descriptors (Gab-Gag-EPIF), predictive ability can be further increased. On the other hand, the result of sequence similarity descriptors seems performed better than those with EPIF as cross-terms. It might be caused by the fact that sequence similarity descriptor describe the sequence features of protein, but the EPIF focuses on those structure characteristics in the binding interface, so the EPIF can increase the predictive ability of structure descriptors but does not apply well with sequence descriptors.

Conclusions

Currently, we can only rely on experimental methods to test the binding affinity of mutated antigens with certain antibody or antiserum. Considering the time-consuming experimental methods, computational methods which can accurately describe the antigen-antibody interaction and further help the measurement of binding affinity is highly desired. In this work, a series of protein fingerprint with epitope-paratope interaction fingerprint (EPIF) were firstly introduced and successfully tested on benchmark dataset through Proteochemometric Modeling. The results indicated that our new established protein fingerprint achieved a better predictive ability than peers. In addition, when cross-terms were introduced into Proteochemometric model, the newly established EPIF not only significantly improved the prediction ability, but

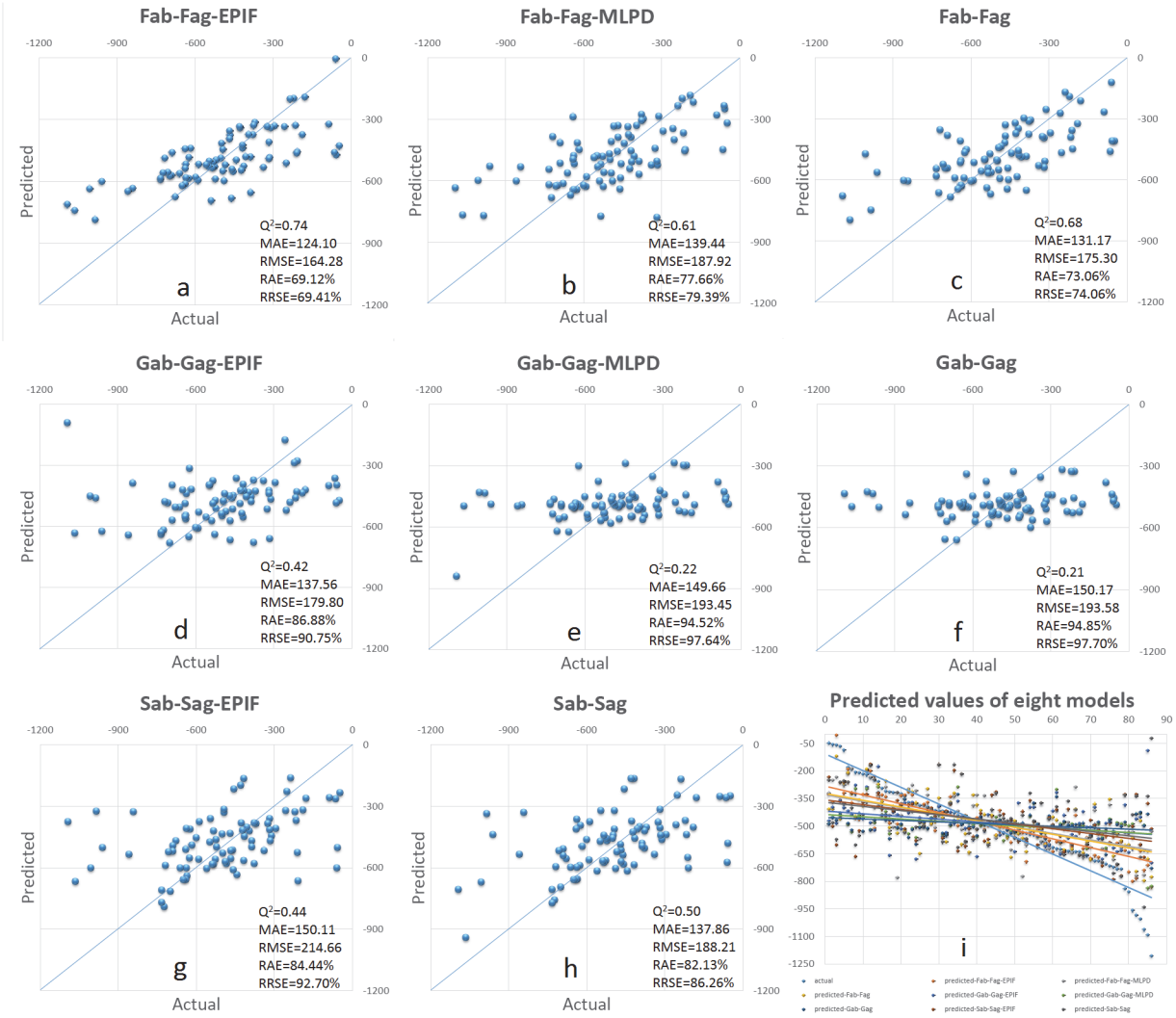

Fig 1. Predicted binding energy of all antigen-antibody in our testing set. Panels a~h represent the predicted value compared with actual value simulated by Hex. Panel i represents the graphical illustrations of the predictive ability of all 8 obtained models with the selected kernel.

doi:10.1371/journal.pone.0122416.g001

also outperformed the pervious cross-terms of MLPD. Results also proposed that EPIF as a structure descriptor can increase the predictive performance of the Proteochemometric model based on conventional structure descriptors, but may not be suitable for sequence descriptor. Moreover, our recommended model based on support vector regression with descriptor combination of Fab-Fag-EPIF showed the ability to simulate bonding affinities for antigen-antibody complexes. With known or simulated conformational structures of antigen-antibody complexes, this new established fingerprint will be able to simulate binding affinity, and further, provide assistance for antibody screening.

Materials and Methods

Data set

Training and validation dataset of antigen-antibody complexes were extracted from Protein Data Bank [33]. We artificially excluded the inappropriate searching results such as: structures

containing only antigen or antibody, T cell epitope-antibody complex structure. Also, structures with low crystalline precision and short sequence length had been excluded to ensure the quality of our dataset. Specific steps and parameters are given as follows:

1. Searching Keywords: antibody, antigen, Fab, Fv, Fc, IgG and immu*
2. Resolution better than 3.0 Å
3. Antigen length with more than 50 residues
4. Two structures share identical sequence and conformational in both epitope and paratope, one of them were removed from our dataset

After these four steps, crystal structures of 429 antigen-antibody complexes including 343 as training data and 86 as testing data were collected. The PDB IDs in our dataset can be found in the Supplementary Data (S2 and S3 Tables).

Epitope and Paratope determination

For each antigen-antibody complex structure in our dataset, epitope and paratope residues were distinguished by Solvent Accessible Surface Area (SASA) based methods. SASA values were calculated (Naccess V2.1.1) for each residue in antigen-antibody complexes and the single molecule structure with probe radius set as 1.4 Å. Surface residues were those more than $1Å^2$ SASA while those loss in binding of more than $1Å^2$ were classified as epitope on the antigen side and as paratope on the antibody side.

Interaction energy simulation

To create Proteochemometric models with different descriptors, binding affinity values of each antigen-antibody complex were simulated by Hex [34]. To guarantee the antigen-antibody complexwa maintain the combination position, Receptor Rotation Range, Ligand Rotation Range, Twist Range were set as 0 and Distance Range was set as 1 (minimum); Correlation type was set as shape & electrostatics. The interaction energies of 429 antigen-antibody complexes were calculated and listed in S2 and S3 Tables.

Interaction interface coordinate system generation

To build the protein fingerprint and EPIF, interaction coordinate system was firstly established (**Fig 2**). Here, residue r_i of the antigen-antibody complex was simplified as a point P_i by averaging its atoms' coordinates. Then, the geometric center of epitope (C_e) and paratope (C_p) were calculated by averaging the coordinate of epitope residue and paratope residue respectively. Later, the geometric center (C) of interaction interface was calculated by averaging the coordinate of all the residues from both epitope side and paratope side. Based on those three points, our coordinate system can be generated.

Protein fingerprint generation

There exist server protein description methods [35,36,32] which contains structure information mainly focusing on coordinate information, distance information and bond type/angel information of protein structures. However, previous studies illustrated that the interface features of epitope-paratope interaction may relate more to the amino acid composition, local structural and physical-chemical environment on the interaction surface [10]. It is widely reported that physical-chemical features such as hydrophobic interaction, hydrogen-bond interaction and electrostatic interaction play essential roles in the antigen-antibody interaction

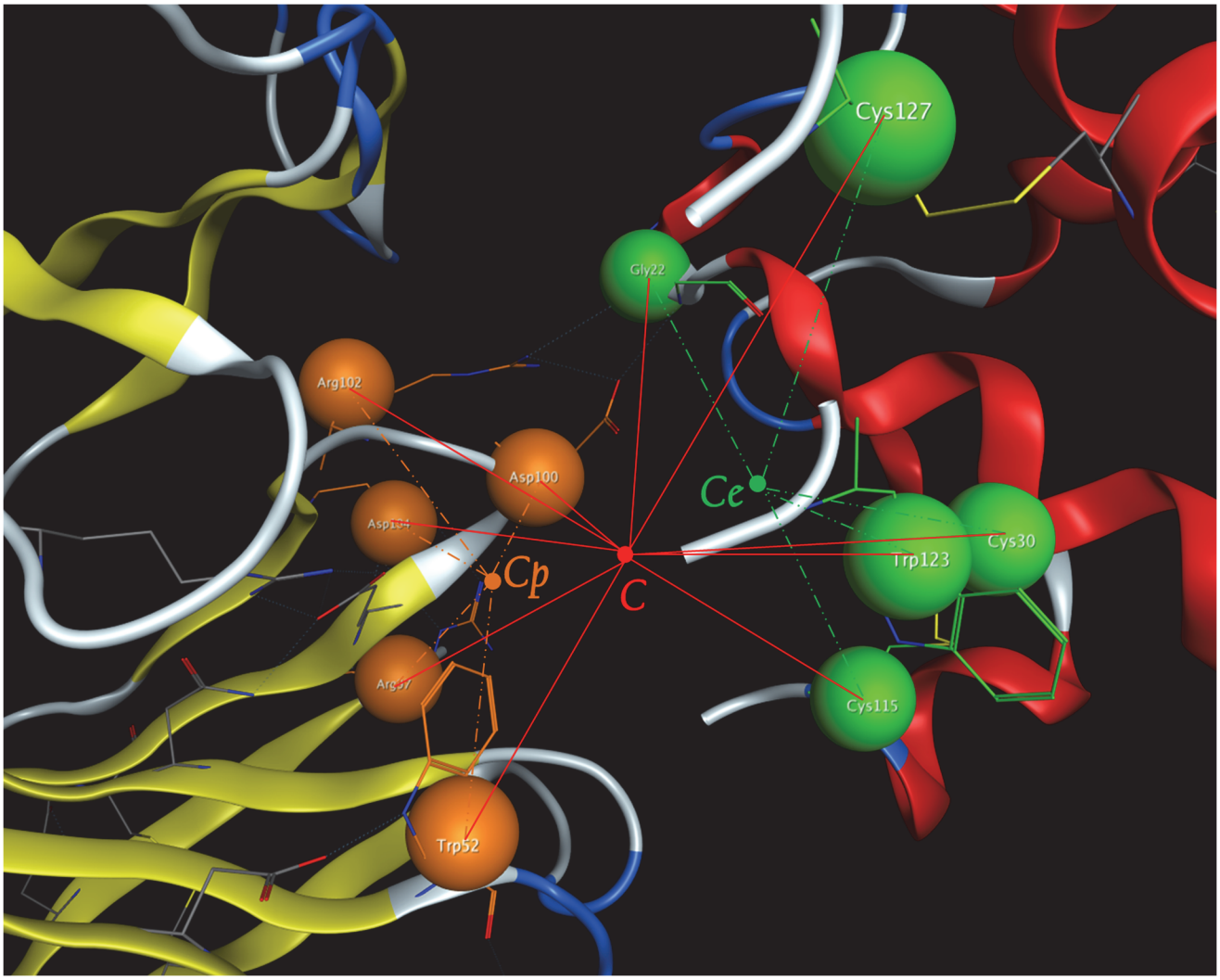

Fig 2. Illustration of antigen-antibody interaction coordinate system. Yellow (paratope side) and green (epitope side) balls represent the simplified point P_i of each residue r_i in the coordinate system; point C_p represents the geometric center of the paratope side while point C_e represents the geometric center of the epitope side; point C represents the geometric center of the interaction interface.

doi:10.1371/journal.pone.0122416.g002

interface [37,38]. Here, fingerprints containing both structural features and physical-chemical environment features were established to describe the structure features of antibody in the interaction interface.

Structure fingerprint generation through cylinder model. By setting a plane through point C and perpendicular to Vector $\overrightarrow{C_eC_p}$, a virtual interaction interface was generated. This "virtual interaction interface" was set as the X-Y axis plane. With point C set as the origin, the Z axis was settled by the normal vector $\overrightarrow{n}$ of X-Y plane towards to the paratope side. Then the rotating plane was established by the X-Y-Z axis to generate the structure fingerprints. Along with a size-defined rotating plane revolving around axis Z, each of the surface residues can be punched into the certain position of the cylinder model (Fig 3). In order to contain enough residues in interaction interface, different plane size and grid resolution were tested. By setting 20 Å as rotating radius and 0 to 40 Å for Z axis, more than 95% of the residues on both epitope and paratope side can be projected into the structure profiles. After setting the radius pixel as 2

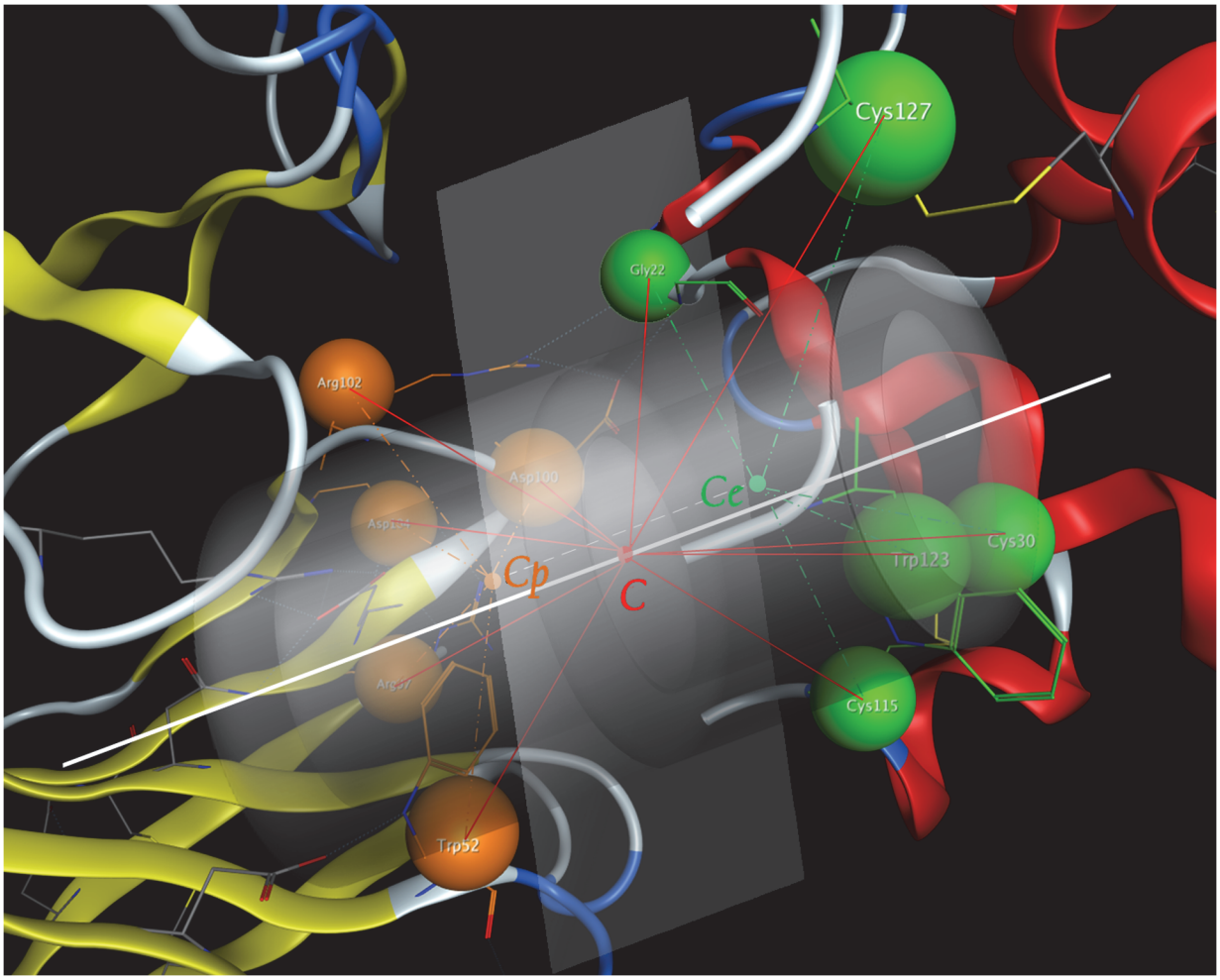

Fig 3. Illustration of structure fingerprint generation. Graphic definition of "virtual interaction interface" and size-defined cylinder was generated perpendicular to this virtual interaction interface.

doi:10.1371/journal.pone.0122416.g003

Å and Z axis pixel as 5 Å, a 2-dimensinal grid which contains 80 (20/2 * 40/5) bit was screened to generate the antibody protein fingerprint.

The antigen fingerprint was generated on the same system with several modifications, an idea of unit patch of residue triangle was introduced in the epitope area [39]. Unit patch of residue triangle was defined among any three surface residues where the distances for each two of them was within 4 Å, only those contain three residues were termed as epitope unit patches. For antigen structure fingerprint, the Z axis was towards to the epitope side. The averaged coordinate of three residues in a unit patches point (UP_i) is to replace the role of residue point P_i in the coordinate system.

Physical-chemical fingerprint generation through shell structure model. To characterize the physical-chemical environment of the protein in interaction interface, a series of shells have been generated with appropriate pixel starting from the geometric center point (C_p & C_e) of each side (**Fig 4**). All neighbor residues within 20 Å from the geometric center (C_p & C_e) have been counted [10] and can be inputted into different layers based on their geometric distances towards geometric center. By setting pixel distance as 2 Å, the encoding array of each physical-chemical property contains 10 independent bits. Three sets of values describing the

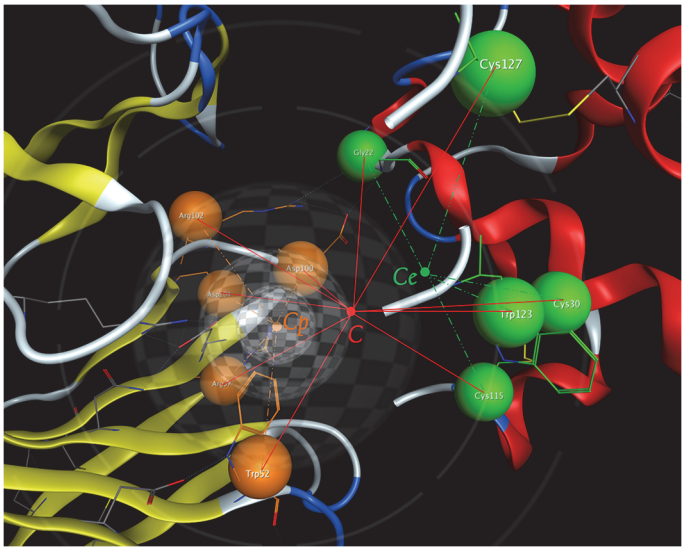

Fig 4. Graphic definition of shell structure model. For the shell structure model of center C_p, 2 residues can be inputted into the second layer while 3 residues can be inputted into the third layer.

doi:10.1371/journal.pone.0122416.g004

physical-chemical properties including hydrophobic interaction, hydrogen-bond interaction and electrostatic interaction (ARGP820101, FAUJ880109 and FAUJ 880108) were derived from AAindex database [40] and led to a 30 length physical-chemical fingerprint. Different from the paratope side, the averaged AAindex of each unit patch of residue triangle was calculated as the physical-chemical index for each shell in epitope. After that, two 110-bits fingerprints for antigen and antibody side were generated respectively to characterize the unit patches layout and physical-chemical environment in the interaction interface.

Epitope-Paratope Interaction fingerprint (EPIF) generation. Antigen-antibody interaction interface is composed of residues from both antigen and antibody sides, appropriate spatial layout and interaction force will lead to a successful binding. To analyze an antigen-antibody complex, an epitope-paratope interaction fingerprint (EPIF) which contains both different interaction forces and environment information in 3-dimensional level is firstly established to demonstrate the interaction features of antigen-antibody complex.

Here, our approach expands the original idea of interaction fingerprint to make it suitable for the large amount of available antigen-antibody complexes data or complexes produced by docking into 3-Dimensional structures. Since EPIF is a bit string representing interactions between antigen and antibody, both the interaction force and interface environment have been fully take into consideration. Here, based on a new shell structure starting from geometric center C, a 15-bit interaction fingerprint of each residue can be inputted into 10 layers (*see "shell structure model"*). Thus, a 150-bit EPIF of each antigen-antibody complexes have been generated. The definition of 15 bits interaction fingerprint is given as follows:

Interaction fingerprint generation. EPIF contains eight different types of interaction: back bone, side chain, polar, hydrophobic, H-bond receptor, H-bond donor, Aromatic and Charged. Our algorithm is designed to determine those interactions by calculating atom distances and residue types. The first bit represents for any contact, if the first bit is 0 means all 14 remains are 0. For 6 strong interactions: back bone, side chain, polar, hydrophobic, aromatic and charged, an additional bit was followed to describe the interaction level of the certain

interaction types, as formula 1 shows.

$$\begin{cases} EPIF^{aa} = S_{1:15}(epix)^{aa} \\ EPIF_i^{aa} = [0 \vee 1] i \in [1, 15] \\ EPIF_i^{aa} = \{any; backbone(exist); backbone(strong); sidechain(e); \\ sidechain(s); polar(e); polar(s); hydrophobic(e); hydrophobic(s); \\ aromatic(e); aromatic(s); charged(e); charged(s); h-bond\ donor; \\ h-bond\ receptor; \} \end{cases} \quad (1)$$

Here, $EPIF^{aa}$ represents an epitope-paratope interaction fingerprint for each epitope amino acid, which contains 15 bits for any amino acid x in the epitope side. Each bit can only be count as 0 or 1. For 8 interaction type sites (side 2,4,6,8,10,12,14,15), 1 means there exist at least one residue from paratope side which can form this type of interaction within distance cutoff, while 0 means the opposite. For 6 force strength identification sites (side 3,5,7,9,11,13), it can be count as 1 only when the same interaction type site defined as 1 and there are enough numbers of this type of interaction appeared around amino acid x, otherwise, it is count as 0. According to our statistical analysis, the number of residues within the distance cutoff of target ranged from 0 to 10 with the median as 4 in our dataset. Considering that the charged force is relatively stronger than other interaction forces, the number cutoff for charged was set as 1 while the others were set as 4. The distance cutoff of each site was set as 4 Å in our study [22].

Proteochemometric Modeling

In our study, 3 Proteochemometric models were created from training set based on different combinations of fingerprints (Fab-Fag-EPIF, Fab-Fag-MLPD, Fab-Fag). All models were implemented in SMOreg of Weka (Version 3.7) by using support vector regression (SVR). The efficacy of all kernels was assessed by Q^2 (predictive ability) with 10-fold cross-validation, and two Kernels (Normalized Poly Kernel and RBF Kernel) were selected (Table 1). Additional 5 Proteochemometric models (Gab-Gag-EPIF, Gab-Gag-MLPD, Gab-Gag, Sab-Sag-EPIF, Sab-Sag) based on peers widely used sequence (S) and geometric descriptors (G) [32] with two selected kernels were established to test the performance of our fingerprints (Table 2). Also, the cross-term was tested for both EPIF and the previous multiplication of the antigen and antibody protein descriptors. Our Proteochemometric Modeling of the antigen-antibody interaction by new protein and epitope-paratope interaction fingerprints is illustrated in **Fig 5**.

Model Evaluation

Statistical parameters for evaluating the PCM models were defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - t_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (2)$$

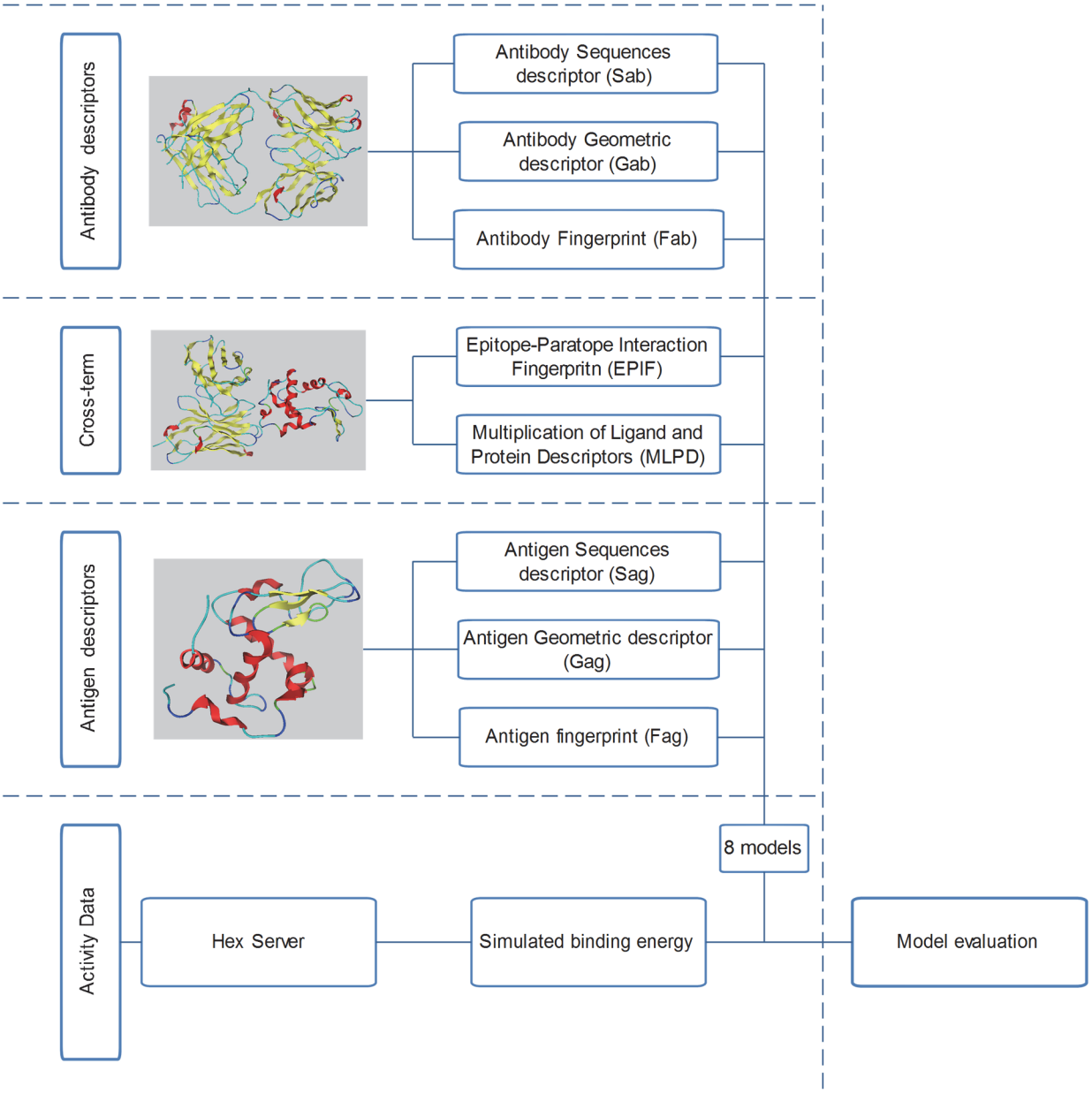

Fig 5. Illustration of our Proteochemometric Modeling of the antigen-antibody interaction by new protein and epitope-paratope interaction fingerprints. More detail information can be seen in *Method* & S1 Fig.

doi:10.1371/journal.pone.0122416.g005

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - t_i)^2}{n}} \quad 3$$

$$RAE = \frac{\sum_{i=1}^{n}|p_i - t_i|}{\sum_{i=1}^{n}|t_i - \bar{t}|} \quad 4$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - t_i)^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2}} \quad 5$$

MAE represents Mean Absolute Error, RMSE represents Root Mean Squared Error, RAE represents Relative Absolute Error and RRSE represents Root Relative Squared Error. p_i represents predicted activity data calculated by different models, t_i represents true activity data simulated by Hex server while $-t$ represents the mean of true activity data.

Supporting Information

S1 Fig. Illustration of structure model for protein fingerprint and EPIF generation.
(DOCX)

S1 Table. Protein geometry descriptors of each protein structure.
(DOCX)

S2 Table. Training dataset.
(DOCX)

S3 Table. Testing dataset.
(DOCX)

Acknowledgments

This work was supported in part by grants from National Natural Science Foundation of China (31200986, 31171272) and Ministry of Science and Technology China (2010CB833601).

Author Contributions

Conceived and designed the experiments: RXZ ZWC TYQ. Performed the experiments: TYQ HX JXQ YYY. Analyzed the data: TYQ HX JXQ YYY. Contributed reagents/materials/analysis tools: TYQ JXQ YYY. Wrote the paper: TYQ QCZ RXZ. Designed the figure: DFW TYQ.

References

1. Goldsby (2003) Antigen. Immunology (Fifth ed). New York: W.H.Freeman and Company. pp. 57–75.
2. Roitt IM BJ, Male DK (1996) Immunology. London: Mosby.
3. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305: 371–376. PMID: 15218094

4. Georgiev IS, Doria-Rose NA, Zhou TQ, Do Kwon Y, Staupe RP, Moquin S, et al. (2013) Delineating Antibody Recognition in Polyclonal Sera from Patterns of HIV-1 Isolate Neutralization. Science 340: 751–756. doi: 10.1126/science.1233989 PMID: 23661761
5. Nisius B, Sha F, Gohlke H (2012) Structure-based computational analysis of protein binding sites for function and druggability prediction. J Biotechnol 159: 123–134. doi: 10.1016/j.jbiotec.2011.12.005 PMID: 22197384
6. Muller YA, Chen Y, Christinger HW, Li B, Cunningham BC, Lowman HB, et al. (1998) VEGF and the Fab fragment of a humanized neutralizing antibody: crystal structure of the complex at 2.4 A resolution and mutational analysis of the interface. Structure 6: 1153–1167. PMID: 9753694
7. Chen Y, Wiesmann C, Fuh G, Li B, Christinger HW, McKay P, et al. (1999) Selection and analysis of an optimized anti-VEGF antibody: crystal structure of an affinity-matured Fab in complex with antigen. J Mol Biol 293: 865–881. PMID: 10543973
8. Ysern X, Fields BA, Bhat TN, Goldbaum FA, Dall'Acqua W, Schwarz FP, et al. (1994) Solvent rearrangement in an antigen-antibody interface introduced by site-directed mutagenesis of the antibody combining site. J Mol Biol 238: 496–500. PMID: 8176740
9. Tharakaraman K, Raman R, Viswanathan K, Stebbins NW, Jayaraman A, Krishnan A, et al. (2013) Structural Determinants for Naturally Evolving H5N1 Hemagglutinin to Switch Its Receptor Specificity. Cell 153: 1475–1485. doi: 10.1016/j.cell.2013.05.035 PMID: 23746829
10. Sun J, Xu T, Wang S, Li G, Wu D, Cao Z (2011) Does difference exist between epitope and non-epitope residues? Analysis of the physicochemical and structural properties on conformational epitope from B-cell protein antigens. Immunome Res 7: 1–11.
11. Wu D, Sun JS, Xu T, Wang S, Li G, Li Y, Cao Z (2010) Stacking and energetic contribution of aromatic islands at the binding interface of antibody proteins. Immunome Res 6(Suppl 1): S1. doi: 10.1186/1745-7580-6-S1-S1 PMID: 20875152
12. Corti D, Lanzavecchia A (2013) Broadly neutralizing antiviral antibodies. Annu Rev Immunol 31: 705–742. doi: 10.1146/annurev-immunol-032712-095916 PMID: 23330954
13. Ivanciuc O, Schein CH, Braun W (2002) Data mining of sequences and 3D structures of allergenic proteins. Bioinformatics 18: 1358–1364. PMID: 12376380
14. Xie L, Bourne PE (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc Natl Acad Sci U S A 105: 5441–5446. doi: 10.1073/pnas.0704422105 PMID: 18385384
15. Hoffmann B, Zaslavskiy M, Vert JP, Stoven V (2010) A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC Bioinformatics 11: 99. doi: 10.1186/1471-2105-11-99 PMID: 20175916
16. Konc J, Janezic D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26: 1160–1168. doi: 10.1093/bioinformatics/btq100 PMID: 20305268
17. Tian C, Zhu L, Yu D, Cao Z, Kang T, Zhu R, et al. (2014) The stereoselectivity of CYP2C19 on R- and S-isomers of proton pump inhibitors. Chem Biol Drug Des 83: 610–621. doi: 10.1111/cbdd.12274 PMID: 24350826
18. Schalon C, Surgand JS, Kellenberger E, Rognan D (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. Proteins 71: 1755–1778. doi: 10.1002/prot.21858 PMID: 18175308
19. Weill N, Rognan D (2010) Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. J Chem Inf Model 50: 123–135. doi: 10.1021/ci900349y PMID: 20058856
20. Xue W, Liu H, Yao X (2012) Molecular mechanism of HIV-1 integrase-vDNA interactions and strand transfer inhibitor action: a molecular modeling perspective. J Comput Chem 33: 527–536. doi: 10.1002/jcc.22887 PMID: 22144113
21. Wu D, Huang Q, Zhang Y, Zhang Q, Liu Q, Gao J, et al. (2012) Screening of selective histone deacetylase inhibitors by proteochemometric modeling. BMC Bioinformatics 13: 212. doi: 10.1186/1471-2105-13-212 PMID: 22913517
22. Mordalski S, Kosciolek T, Kristiansen K, Sylte I, Bojarski AJ (2011) Protein binding site analysis by means of structural interaction fingerprint patterns. Bioorg Med Chem Lett 21: 6816–6819. doi: 10.1016/j.bmcl.2011.09.027 PMID: 21974955
23. Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg JE (2008) Proteochemometric modeling of HIV protease susceptibility. BMC Bioinformatics 9: 181. doi: 10.1186/1471-2105-9-181 PMID: 18402661
24. Lapins M, Wikberg JE (2009) Proteochemometric modeling of drug resistance over the mutational space for multiple HIV protease variants and multiple protease inhibitors. J Chem Inf Model 49: 1202–1210. doi: 10.1021/ci800453k PMID: 19391634

25. Lapinsh M, Prusis P, Lundstedt T, Wikberg JE (2002) Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. Mol Pharmacol 61: 1465–1475. PMID: 12021408

26. Lapinsh M, Prusis P, Uhlen S, Wikberg JE (2005) Improved approach for proteochemometrics modeling: application to organic compound—amine G protein-coupled receptor interactions. Bioinformatics 21: 4289–4296. PMID: 16204343

27. Strombergsson H, Daniluk P, Kryshtafovych A, Fidelis K, Wikberg JE, Kleywegt GJ, et al. (2008) Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. J Chem Inf Model 48: 2278–2288. doi: 10.1021/ci800200e PMID: 18937438

28. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, et al. (2004) Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 44: 1257–1266. PMID: 15272833

29. Muller KR, Ratsch G, Sonnenburg S, Mika S, Grimm M, Heinrich N, et al. (2005) Classifying 'drug-likeness' with kernel-based learning methods. J Chem Inf Model 45: 249–253. PMID: 15807485

30. Van Westen GJPW, JK, Ijzerman AP, Van Vlijmen HWT, Bender A (2011) Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. Med Chem Commun 2: 16–30.

31. Huang Q, Jin H, Liu Q, Wu Q, Kang H, Cao Z, et al. (2012) Proteochemometric modeling of the bioactivity spectra of HIV-1 protease inhibitors by introducing protein-ligand interaction fingerprint. PLoS One 7: e41698. doi: 10.1371/journal.pone.0041698 PMID: 22848570

32. Lapinsh M, Prusis P, Mutule I, Mutulis F, Wikberg JES (2003) QSAR and proteo-chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes. Journal of Medicinal Chemistry 46: 2572–2579. PMID: 12801221

33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. (2000) The Protein Data Bank. Nucleic Acids Research 28: 235–242. PMID: 10592235

34. Macindoe G, Mavridis L, Venkatraman V, Devignes MD, Ritchie DW (2010) HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 38: W445–449. doi: 10.1093/nar/gkq311 PMID: 20444869

35. Bock ME GC, Guerra C (2007) Discovery of Similar Regions on Protein Surface. J comput Biol 14: 285–299. PMID: 17563312

36. Yin SY, Proctor EA, Lugovskoy AA, Dokholyan NV (2009) Fast screening of protein surfaces using geometric invariant fingerprints. Proceedings of the National Academy of Sciences of the United States of America 106: 16622–16626. doi: 10.1073/pnas.0906146106 PMID: 19805347

37. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285: 2177–2198. PMID: 9925793

38. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ (2007) Spatial chemical conservation of hot spot interactions in protein-protein complexes. Bmc Biology 5.

39. Sun J, Wu D, Xu T, Wang X, Xu X, Tao L, et al. (2009) SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Res 37: W612–616. doi: 10.1093/nar/gkp417 PMID: 19465377

40. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M, et al. (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Research 36: D202–D205. PMID: 17998252