Structural Modeling Extends QSAR Analysis of Antibody-Lysozyme Interactions to 3D-QSAR

Eva K. Freyhult,*† Karl Andersson,†* and Mats G. Gustafsson‡

*The Linnaeus Centre for Bioinformatics, Uppsala University and the Swedish University of Agricultural Sciences, Sweden;
†Biacore AB, Uppsala, Sweden; and ‡Signals and Systems, Uppsala University, Sweden

ABSTRACT This work shows that quantitative multivariate modeling is an emerging possibility for unraveling protein-protein interactions using a combination of designed mutations with sequence and structure information. Using this approach, it is possible to stereochemically determine which residue properties contribute most to the interaction. This is illustrated by results from modeling of the interaction of the wild-type and 17 single and double mutants of a camel antibody specific for lysozyme. Linear multivariate models describing association and dissociation rates as well as affinity were developed. Sequence information in the form of amino acid property scales was combined with 3D structure information (obtained using molecular mechanics calculations) in the form of coordinates of the α-carbons and the center of the side chains. The results show that in addition to the amino acid properties of the mutated residues 101 and 105, the dissociation rate is controlled by the side-chain coordinate of residue 105, whereas the association is determined by the coordinates of residues 99, 100, 105 (side chain), 111, and 112. The great difference between the models for association and dissociation rates illustrates that the event of molecular recognition and the property of binding stability rely on different physical processes.

INTRODUCTION

Three-dimensional quantitative structure-activity relationship (3D-QSAR) modeling has become a common technique for elucidating the stereochemical features important for function of small ligands. Several successful experiments have been reported (Ortiz et al., 1995; Stanton, 2000; Xing et al., 1999). Extending the 3D-QSAR approach to protein-protein interactions would be attractive, but is nontrivial. One problem is the relatively expensive production of mutated proteins, a second is the need for a reliable characterization of the interaction, a third is the limited lifetime of thawed proteins, and a fourth is to obtain relevant descriptors of a protein structure in silico, which includes finding the binding configuration and taking into account the spatial constraints. The QSAR analysis performed by De Genst et al. (2002) proved that the first three problems can be resolved. In this work the specific problem with three-dimensional descriptors is addressed and, when combined with experiences from De Genst et al. (2002), is compiled into a methodology for applying three-dimensional QSAR methods to a protein-protein interaction.

Several authors have performed successful QSAR studies on proteins and peptides. Eriksson et al. (1990) used peptide amino acid sequence, to predict function for substance P analogs, enkephalins, and bradykinins. Andersson et al. (2001) and Choulier et al. (2002) characterized the interaction of designed multimutated peptides with an antibody and developed predictive models for both the association and the dissociation rate of the interaction. Wikberg and co-workers are studying several different ligand-receptor interactions (Lapinsh et al., 2001; Lapinsh et al., 2002; Prusis et al., 2001; Prusis et al., 2002). Another of the most recent reports (De Genst et al., 2002) is an investigation of the interaction between the camel antibody cAb-lys3 and its natural antigen lysozyme. The antibody has a protruding loop, consisting of residues 99–108, that inserts into the active site of lysozyme (Fig. 1) and inhibits its enzymatic function (Transue et al., 1998). Two residues (positions 101 and 105) in this loop were mutated (ten single mutations, seven double mutations) essentially according to a multivariate experimental design (Haaland, 1989). The binding of the wild-type and the 17 mutants of cAb-lys3 to lysozyme was characterized using a surface plasmon resonance (SPR) sensor system (Biacore 3000, Biacore AB, Uppsala, Sweden) giving affinity, association rate and dissociation rate constants in duplicate. In De Genst et al., (2002) the measured binding characteristics were related to a description of the sequences only.

Descriptions of three-dimensional protein structures can be obtained both experimentally and in silico. Skerra et al. (Schiweck and Skerra, 1997; Skerra, 2000) is one group that has worked extensively with both experimental determination and in silico prediction of protein structures. The goal has often been to design artificial binders in silico by use of a known protein template e.g., lipocalin. Such research is clearly related to 3D-QSAR for proteins; it tries to predict protein structure and properties (such as binding), but has one important difference: Design of binders is based on hypotheses relating to how individual amino acid residues will affect the desired function, whereas in this QSAR study, collected data is used to interpret function.

In this work, the data presented in De Genst et al. (2002) mentioned above were reanalyzed using a new set of

Submitted June 6, 2002, and accepted for publication December 17, 2002.

Address reprint requests to Eva K. Freyhult, The Linnaeus Center for Bioinformatics, Box 598, Uppsala, Sweden 75124. Tel.: +46-18-4716688; E-mail: eva.freyhult@lcb.uu.se.

0006-3495/03/04/2264/09 $2.00

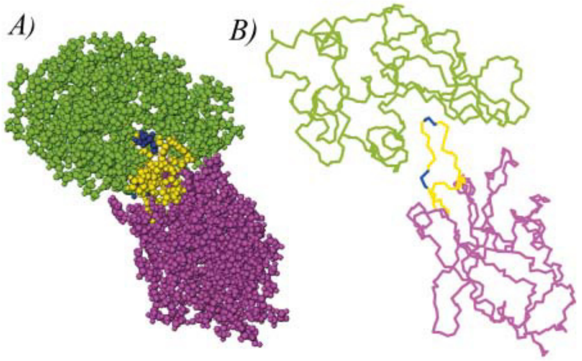

FIGURE 1 The interaction between the antigen cAb-lys3 (shown in *violet*) and its antigen lysozyme (shown in *green*). Residues 99–112 are marked yellow and the mutated residues 101 and 105 are marked blue. The structure shown is the wild-type structure. In *A* the entire molecules are shown, in *B* only the backbone structures are shown.

descriptors. In addition to sequence information, in the form of amino acid property scales as in De Genst et al. (2002) compact 3D structure descriptors based on molecular mechanics calculations were introduced. This represents an example of a general methodology in which designed mutations are combined with sequence and structural information to build quantitative models of protein-protein interactions. Results presented show that the models obtained can give information of the contribution to the interaction of positions of amino acids which have not been mutated. Moreover, the results show that both side-chain and backbone coordinates influence binding characteristics and support the earlier findings (De Genst et al., 2002) that the event of molecular recognition and the property of binding stability rely on different physical processes (Roos et al., 1998).

METHODS

Experimental data

Table 1 shows the experimental values of the association rate (k_a), dissociation rate (k_d), and affinity ($K_d = k_d/k_a$) constants for the 17 mutants and the wild-type (from De Genst et al., 2002). The mutants are referred to by a two letter code, corresponding to the amino acids at the mutation sites (position 101 and 105, respectively) e.g., the wild-type, with threonine at position 101 and serine at position 105, is referred to as TS. For every molecule, duplicate values are presented.

Geometry optimizations

The wild-type antibody, as well as the mutated camel antibodies, in complex with lysozyme were energy minimized. The starting structure for all minimizations was the x-ray crystal structure of the wild-type antibody (Desmyter et al., 1996), PDB-code 1mel. There are two slightly different antibody-lysozyme complexes in the asymmetric unit, one of them was selected at random as the template complex. The optimization was performed with respect to the entire molecular system and the water molecules present in the crystal structure were retained during the optimization.

In the design of the experiment (De Genst et al., 2002) the mutations were chosen so as to avoid large sterical clashes with the antigen (De Genst et al., 2002). Therefore, here it is assumed that the mutations do not cause any large movements of the entire protein and thus that the above described template structure for all the mutants is reasonable. This of course also means that the statistical multivariate quantitative models built should not be applicable to the subset of all the 400 possible mutations in which sterical clashes may occur.

The geometry optimizations were performed in HyperChem (HyperChem, 2000) on a personal computer (Pentium-4 1400 MHz). The standard molecular mechanics method used was based on the amber94 force field (HyperChem, 2000; Cornell et al., 1995) and used a conjugate gradient method with a termination value of the RMS gradient of 0.1 kcal/mol/Å, in general according to the HyperChem manual.

Molecular descriptors

The descriptors used in this work may be regarded as extensions to the ZZ-scales (Sandberg et al., 1998) used in De Genst et al. (2002). The ZZ-scales

TABLE 1 Association (k_a), dissociation (k_d), and affinity (K_d) constants of all mutants and the wild-type camel antibodies

Mutant	k_a (M^{-1} s^{-1})	k_d (s^{-1})	K_d (M)
HD	2.20×10^3, 2.16×10^3	3.77×10^{-3}, 3.47×10^{-3}	1.71×10^{-6}, 1.61×10^{-6}
LS	3.63×10^4, 3.68×10^4	5.28×10^{-4}, 5.68×10^{-4}	1.45×10^{-8}, 1.54×10^{-8}
MV	7.59×10^3, 7.74×10^3	2.77×10^{-2}, 3.05×10^{-2}	3.65×10^{-6}, 3.94×10^{-6}
PG	4.79×10^3, 4.06×10^3	2.17×10^{-2}, 2.10×10^{-2}	4.53×10^{-6}, 5.17×10^{-6}
PS	4.21×10^3, 3.06×10^3	4.80×10^{-3}, 4.76×10^{-3}	1.14×10^{-6}, 1.56×10^{-6}
QP	1.13×10^3, 9.76×10^2	4.13×10^{-4}, 3.06×10^{-4}	3.65×10^{-7}, 3.14×10^{-7}
RT	1.31×10^4, 1.25×10^4	2.59×10^{-2}, 2.42×10^{-2}	1.98×10^{-6}, 1.94×10^{-6}
SQ	1.44×10^4, 1.46×10^4	2.64×10^{-1}, 3.08×10^{-1}	1.83×10^{-5}, 2.11×10^{-5}
SS	5.47×10^4, 4.34×10^4	9.98×10^{-3}, 9.75×10^{-3}	1.82×10^{-7}, 2.25×10^{-7}
TA	4.32×10^4, 4.68×10^4	3.74×10^{-3}, 3.77×10^{-3}	8.66×10^{-8}, 8.06×10^{-8}
TG	1.12×10^5, 1.26×10^5	1.57×10^{-3}, 1.53×10^{-3}	1.40×10^{-8}, 1.21×10^{-8}
TH	6.97×10^4, 6.46×10^4	6.40×10^{-4}, 7.17×10^{-4}	9.18×10^{-9}, 1.11×10^{-8}
TM	2.32×10^4, 2.46×10^4	1.04×10^{-2}, 1.07×10^{-2}	4.48×10^{-7}, 4.35×10^{-7}
TN	3.26×10^4, 3.48×10^4	4.14×10^{-3}, 4.24×10^{-3}	1.27×10^{-7}, 1.22×10^{-7}
TP	3.75×10^3, 3.67×10^3	1.95×10^{-4}, 2.40×10^{-4}	5.20×10^{-8}, 6.54×10^{-8}
TQ	4.09×10^4, 3.71×10^4	3.70×10^{-2}, 3.92×10^{-2}	9.05×10^{-7}, 1.06×10^{-6}
TS	9.10×10^4, 7.00×10^4	8.14×10^{-4}, 8.89×10^{-4}	8.95×10^{-9}, 1.27×10^{-8}
VN	3.47×10^4, 2.99×10^4	2.04×10^{-3}, 2.04×10^{-3}	5.88×10^{-8}, 6.82×10^{-8}

describe, respectively, hydrophobicity (ZZ1), size and polarizability (ZZ2), and polarity and electrophilicity (ZZ3). The previously used descriptors, three ZZ-scales for the two mutated residues (position 101 and 105) each, were extended with structural information of 14 amino acids (residue 99–112) in the antigen binding loop. The amino acid sequence 99–112 is DS□IYA□YYECGHG, where □ indicates one of the two mutation sites 101 and 105.

Coordinates of the backbone and the side chains described the loop structure. The backbone was represented by the coordinates of the 14 α-carbons (C_α). The side chains were represented by the coordinates of the center of each side chain, where the center was defined as the average of the coordinates of all atoms in the side chain. A similar representation (Kleywegt, 1999) has previously been adopted in relation to recognition of spatial motives in protein structures. More detailed descriptors, like CoMFA, were avoided in order to keep the number of parameters to fit relatively small.

A potential problem with the coordinate descriptors used here is the fact that they are dependent on a geometry optimization step. This is a well known potential problem also in classical 3D-QSAR where a slightly different geometry optimization may yield different coordinates which in turn may create a quite different multivariate prediction model than the original. For the particular antibody-lysozyme interaction considered in this work, the robustness of the coordinates was validated qualitatively by recalculating new multivariate models after randomly perturbing all the energy optimized coordinates by ±0.5 Å in a random direction and then comparing the resulting models with the original models. The coordinates which were identified as most important in the original model also turned out to be the most significant in the perturbation based models (data not shown).

Regression

Before the regression, both the descriptors, **x**, and the activities, y, were transformed to have average values of zero.

The partial least squares (PLS) regression algorithm (Geladi and Kowalski, 1986; Höskuldsson, 1988; Gustafsson, 2001) was used to determine linear regression models, describing the relation between the activities, y (ln k_a, ln k_d, and ln K_d), and the molecular descriptors, **x**, of the form

$$y = \mathbf{w}^T\mathbf{x} + e_y, \tag{1}$$

where e_y is the error in y and **w** are the weights computed with the PLS algorithm. The PLS algorithm used was implemented in the PLS toolbox in MATLAB 5.3 (MathWorks Inc., www.mathworks.com).

Validation

The predictivity of each model was measured by the cross-validated regression coefficient (Q^2) defined as $Q^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_{i,CV})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$, where n is the number of predictions, y_i is the experimental activity value, $\hat{y}_{i,CV}$ is the activity value as predicted by cross-validation and $\bar{y}$ is the average of the activities, and the fitted regression coefficient ($R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$, where $\hat{y}_i$ is the fitted activity value). The cross-validation was performed according to the leave-one-out principle (Golbraikh and Tropsha, 2002). Since the activity was measured twice for each molecule, the examples were left out in pairs. (Each training example is defined as a descriptor vector and a corresponding activity value. However, due to duplicate measurements, each molecule is represented by two examples with equal descriptor vectors, but slightly different activity values.) In De Genst et al. (2002), the examples were left out one at a time during cross-validation, as compared to one molecule at a time in this study. This means that the Q^2 values presented in De Genst et al. (2002) are not comparable to the Q^2 values presented here.

During model optimization (variable subset selection or optimization with respect to the number of latent variables used in PLS), the Q^2 value was optimized. This made the Q^2 value likely to be a too optimistic measure of the true predictivity and therefore the blind cross-validated regression coefficient, P^2, (Ortiz et al., 1995) was introduced. Recent results show that Q^2 in general is not a sufficient measure of the model predictivity (Golbraikh and Tropsha, 2002; Kubinyi et al., 1998), but that an independent test data set is required to confirm the predictivity of the model.

In blind cross-validation the data set is divided into a large training data set and a small test data set. The model optimization is performed on the molecules in the training data set only using cross validation and the achieved optimal model (highest Q^2 value) is used to predict the activity of the molecules in the test data set. The blind cross-validated regression coefficient, P^2, is computed similarly to the cross-validated regression coefficient, Q^2; $P^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_{i,BCV})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$. The only difference is the predicted activity value, $\hat{y}_{i,BCV}$, that here is the predicted value of an example in the test set.

In the work presented here the test set consisted of two molecules (four training examples) and, thus, the training set consisted of 16 molecules. To achieve predicted activity values ($\hat{y}_{i,BCV}$) for all 18 molecules, nine such test sets were selected. The test sets were randomly selected as: (TG, SS), (TA, SQ), (QP, LS), (VN, TQ), (RT, PS), (TM, HD), (PG, TN), (TP, TS), and (TH, MV).

To validate the statistical significance of the models obtained, 1000 randomization tests were performed in which the target values were permuted (y-shuffling). Histograms of the P^2 values of the resulting models were then analyzed to obtain a confidence level for each of the models obtained using the true target values.

Model weight analysis

The model weights, **w** (see Eq. 1), were used as a measurement of the relative importance of the descriptor variables. The coordinates were all measured in Å, but the ZZ-scales were measured in different scales and their weights could therefore not be compared without rescaling.

The variance of the activity, $y = \mathbf{w}^T\mathbf{x}$, can be expressed as

$$\sigma_y^2 = \sum_{i=1}^{k} w_i^2\sigma_i^2 + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k} w_i w_j E[x_i x_j], \tag{2}$$

where E denotes the expectation operator, σ_i^2 is the variance of descriptor x_i and k is the total number of descriptors. Both the activity y and the descriptors x_i are assumed to have an average value of zero. If the descriptors are uncorrelated the last term in Eq. 2 is zero and the relative importance of the descriptors x_i can be measured by $\sigma_i w_i$. However, the descriptors are correlated and a study of the relative sizes of the $w_i w_j E[x_i x_j]$ terms was used only to validate the conclusions from the following weight analysis.

Position i (either the position of an α-carbon or a side chain) is described by its x-, y-, and z-coordinates, and the total importance of position i is estimated by $(w_{i,x}^2 + w_{i,y}^2 + w_{i,z}^2)^{1/2}$. The direction of the vector $(w_{i,x}, w_{i,y}, w_{i,z})$ shows how the position should be changed to give a higher activity value. The length of the vector is a measurement of how much the activity value would change if the position was changed by a unit distance (1 Å in the x-, y-, and z-directions).

Variable selection

Variable selection was used only to verify the conclusions from the model weight analysis and not to improve the model, as is the common usage of variable selection methods (Tropsha, 2001; Baroni et al., 1993; Hoffman et al., 1999; Tropsha and Zheng, 2001).

The variable selection was performed once for each of the nine training sets, defined above. The resulting nine variable selections together gave the consensus variable selection, defined by the most frequently selected variables.

For each of the nine training sets, the output from the GA-PLS (genetic algorithm-partial least squares) algorithm (Tropsha, 2001; Hoffman et al., 1999) (see description below) was 100 suggested different variable selections, which all had approximately the same fitness value. A natural assumption is that the variables that really are important to the model would be included in a large fraction of the suggested variable selections. Therefore, instead of selecting the variables included in the very best variable selection, the variables that were included in 80% (or more) of the 100 variable selections were included in the final variable selection.

The GA-PLS method (Tropsha, 2001; Hoffman et al., 1999) applies a genetic algorithm to search the subset of descriptor variables that gives the PLS regression model with the highest fitness (predictivity). In the application of GA-PLS considered here a population of 100 individuals (binary vectors, of the same size as the descriptor vectors, describing inclusion or exclusion of each descriptor variable) was evolved, by over-crossing and mutations, toward higher fitness values.

The fitness function included the cross-validated regression coefficient (Q^2) and was defined as; Fitness $= 1 - (1 - Q^2)(n - 1)/(n - c)$, where n is the number of molecules and c is the optimal number of latent variables in the PLS with respect to Q^2. The GA-PLS algorithm terminated when the difference between the fitness score for the least fit and the most fit individual was smaller than 0.05.

RESULTS

Geometry optimization

The molecular mechanics computations resulted in 18 slightly different structures. A closer look at residues 99–112 in the antibody structures showed how the positions of the residues were affected by the mutations (Fig. 2). The α-carbon of the mutated residue 105 was almost unchanged, as were both the main chain and the side chain in a close proximity of this mutation site. The other mutation site at position 101 seems to affect its neighbors, the positions of residues 99–102 do all vary markedly between the different mutants. Residues 111 and 112 are also affected by the mutations.

The loop structure of one of the mutants differ significantly from the other structures at position 99–102 and 107–112. This mutant, RT, has an arginine at position 101 and a threonine at position 105.

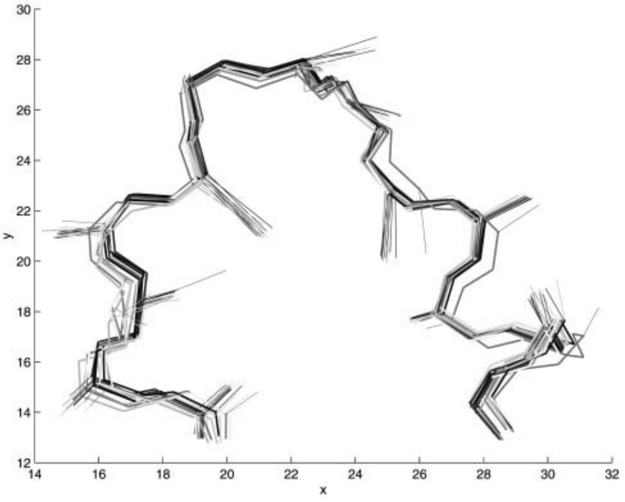

FIGURE 2 A superposition of the loop structure (residues 99–112) for each of the mutants (and the wild-type). One structure is strongly deviating from the others, this structure has an arginine at position 101 and a threonine at position 105. The large side chain of arginine is likely to be the reason for this strong deviation. The coordinates in this figure are the same as in Fig. 7.

Regression

PLS regression models were built using up to the predefined maximum of three latent variables. The final k_a, k_d, and K_d models (see Eq. 1),

$$\ln k_a = \mathbf{w}_{k_a}^{T}\mathbf{x} \tag{3}$$

$$\ln k_d = \mathbf{w}_{k_d}^{T}\mathbf{x} \tag{4}$$

$$\ln K_d = \mathbf{w}_{K_d}^{T}\mathbf{x}, \tag{5}$$

with maximized Q^2 values, used two, three, and three latent variables, respectively. Their corresponding Q^2 values were 0.72, 0.68, and 0.68 (Table 2). The predictivities of the models are shown by the P^2 values in Table 3 to be at the same level as the Q^2 values: 0.62, 0.64, and 0.70, respectively. The mean and standard deviation of the Q^2 values of the nine different k_a, k_d, and K_d models (based on the nine different training sets) are also shown in Table 3. The predictivities of the models are illustrated in Fig. 3.

As already mentioned, the significance of the models was validated using permutations of the target values (y-shuffling). A histogram of the P^2 values for 1000 different k_a-models built with randomly permuted target vectors are shown in Fig. 4, the histograms for k_d and K_d models were similar, also having a slightly asymmetric i.e., non-Gaussian form (not shown). Based on the histograms, the one-sided 99.9% confidence intervals for the distributions of P^2 values were computed. For the k_a, k_d, and K_d models, these intervals were found to be $[-\infty, 0.6]$, $[-\infty, 0.5]$, and $[-\infty, 0.6]$, respectively. Hence all the three models are statistically significant at the 99.9% confidence level.

Model weight analysis

The relative importance of the ZZ-scales to the three models (k_a, k_d, K_d) is shown in Fig. 5. In Fig. 6 the relative importance, $(w_{i,x}^2 + w_{i,y}^2 + w_{i,z}^2)^{1/2}$, of the coordinates of the 14 α-carbons and 14 side chains are shown for the same three models. The most important residues to the k_a model are found to be (in order of importance, the most important first)

TABLE 2 The R^2 and Q^2 values of the PLS regression models and the models based on the GA-PLS selected variables

	PLS			GA-PLS		
Activity	lv*	R^2	Q^2	lv*	R^2	Q^2
ln k_a	2	0.87	0.72	3	0.92	0.85
ln k_d	3	0.86	0.68	3	0.86	0.82
ln K_d	3	0.86	0.68	3	0.80	0.71

*Number of latent variables.

TABLE 3 The mean and standard deviation of Q^2_{max} of nine models based on nine different training sets and P^2 values calculated using the nine training set models

	PLS			GA-PLS		
Activity	Mean (Q^2_{max})	SD (Q^2_{max})	P^2	Mean (Q^2_{max})	SD (Q^2_{max})	P^2
$\ln k_a$	0.69	0.065	0.62	0.76	0.058	0.53
$\ln k_d$	0.62	0.12	0.64	0.79	0.040	0.58
$\ln K_d$	0.64	0.070	0.70	0.71	0.070	0.52

number 100, 101, 112, 111, 105, and 99. Both the position of the α-carbon and the side chain of these residues are found to be important, except for residue 105, where only the side-chain position affects the model. How the positions should be changed to improve the binding is illustrated in Fig. 7, where the vectors $(w_{i,x}, w_{i,y}, w_{i,z})$ are shown for the k_a model and the vectors $(-w_{i,x}, -w_{i,y}, -w_{i,z})$ are shown for the k_d and K_d models.

In the k_d model the importance of the side-chain position of the mutated residue 105 is dominating. The side-chain position of the other mutated residue (101) is also important. For the K_d model the side-chain position of residue 105 is again the most important, but it is not as dominating as in the k_d model. Here also residues 100, 101, 99, 112, and 111 are important.

Variable selection

The variable selection with the GA-PLS method selected in general the ZZ-scales most frequently, but other descriptors were selected as well. The most frequently selected descriptors were (in order of importance, the most important first)

k_a **ZZ2 105**, **ZZ1 105**, **ZZ2 101**, **ZZ3 105**, **S.C. 101y**, ZZ3 101, **Ca 100z**, S.C. 102y, **S.C. 101x**

k_d **ZZ3 105**, **ZZ3 101**, **ZZ1 101**, Ca 106z, S.C. 107x, Ca 111x, S.C. 101z

K_d **ZZ3 105**, **ZZ3 101**, **ZZ2 101**, **S.C. 105x**, **ZZ1 101**

S.C. indicates the position of a side chain and Ca the position of an α-carbon. The selections that confirmed the conclusions from the weight analysis are bolded. The predictivity of the GA-PLS models was not very high ($P^2 \approx 0.5$, see Table 3), but leave-one-out cross validation performed with the consensus selection gave Q^2 values of 0.73, 0.69, and 0.60 (not shown in table), for k_a, k_d, and K_d, respectively.

DISCUSSION

In this paper we have performed a 3D-QSAR analysis of a protein-protein interaction by combining experimental data and sequence description for 18 similar proteins from De Genst et al. (2002) with 3D descriptors for all proteins derived from a single crystal structure (Desmyter et al.,

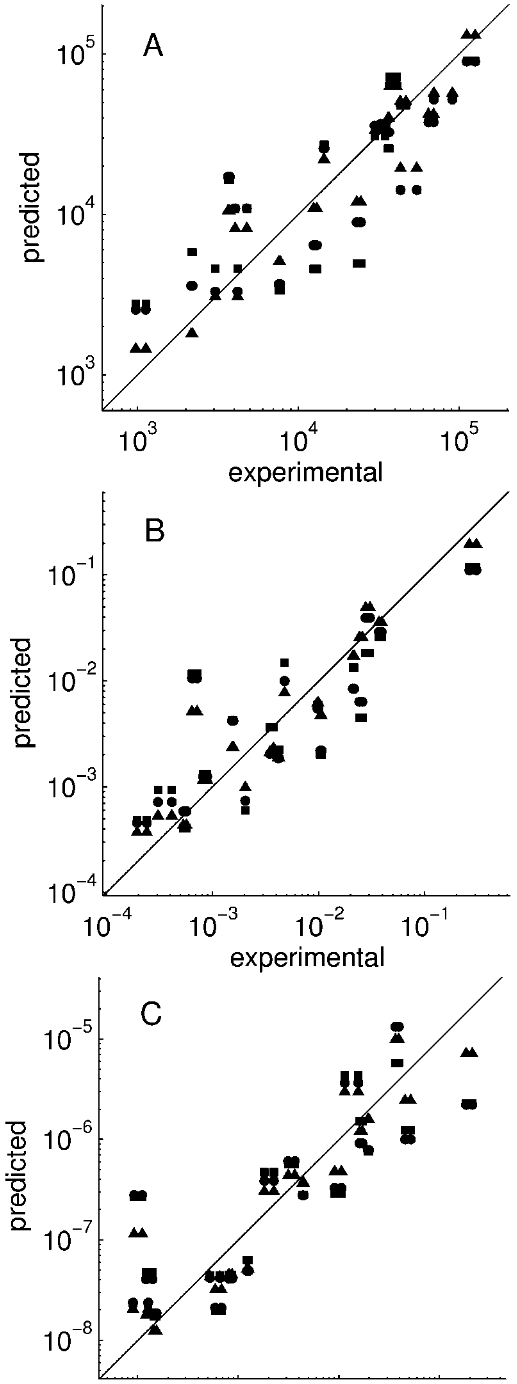

FIGURE 3 Predicted versus experimental activities. The fitted values are shown as triangles, the cross-validated as circles and the blind cross validated as squares. (*A*) Predicted versus experimental k_a. The model was derived using two latent variables. $R^2 = 0.87$, $Q^2 = 0.72$, and $P^2 = 0.62$. (*B*) Predicted versus experimental k_d. The model was derived using three latent variables. $R^2 = 0.86$, $Q^2 = 0.68$, and $P^2 = 0.64$. (*C*) Predicted versus experimental K_d. The model was derived using three latent variables. $R^2 = 0.86$, $Q^2 = 0.68$, and $P^2 = 0.70$.

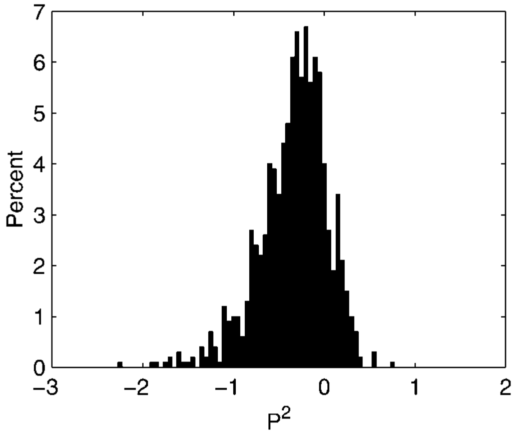

FIGURE 4 Histogram of P^2 values for models of k_a designed using permuted target values. The one-sided 99.9% confidence interval is [$-\infty$, 0.6].

1996). The 3D-QSAR models have statistically significant (at the 99.9% level) predictivities and can resolve how different amino acid residues contribute to the interaction. The models describing association rate and dissociation rate are clearly different.

Geometry optimization

The mutant and wild-type structures used in this study were achieved by energy minimization in silico using the wild-type crystal structure as starting structure. The optimization was performed in vacuo with the water molecules included in the crystal structure present. Since the actual binding takes place in the center of the complex and since the energy minimized structures of all proteins were close to the starting

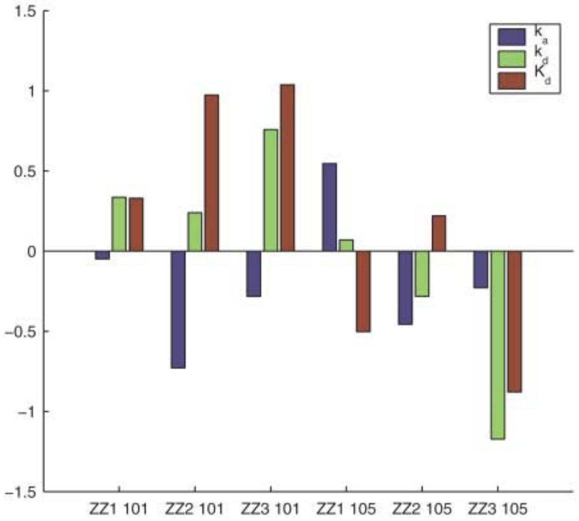

FIGURE 5 The scaled weights, $\sigma_i w_i$, of the six ZZ-scales in the three disjoint models describing $\ln k_a$, $\ln k_d$, and $\ln K_d$, respectively.

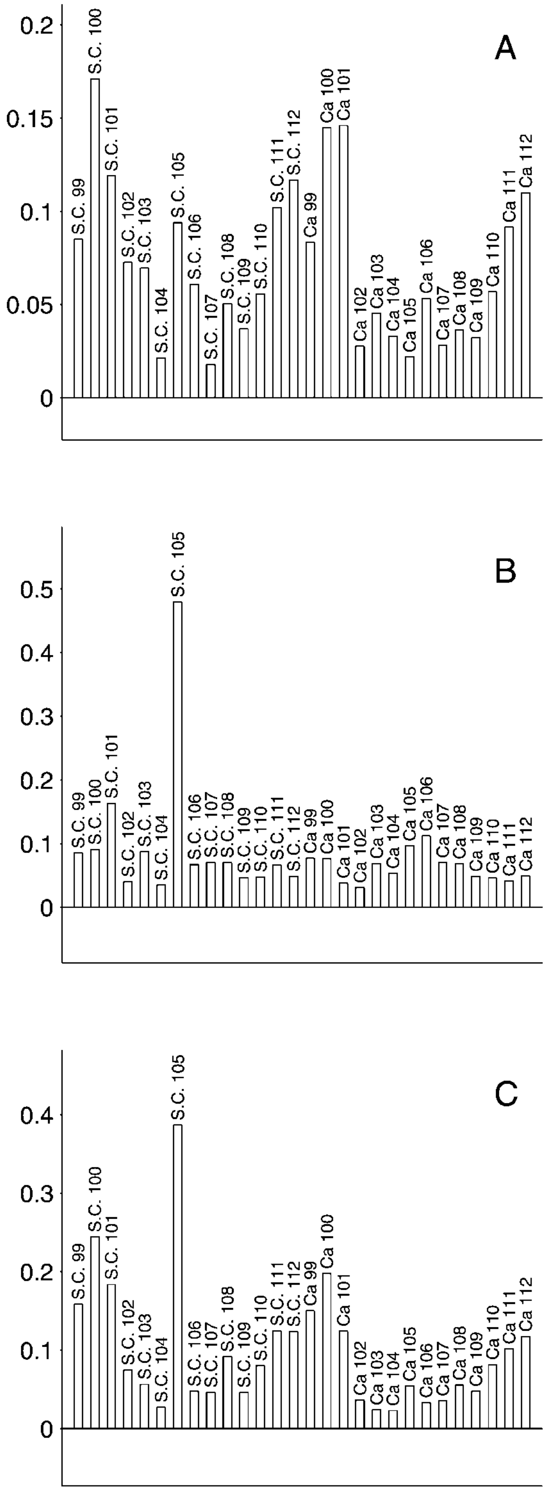

FIGURE 6 The importance of the coordinate descriptors. Ca denotes the position of the α-carbon and S.C. denotes the position of the center of the side chain, the number that follows is the residue number. For residue i, the importance is computed as; $(w_{i,x}^2 + w_{i,y}^2 + w_{i,z}^2)^{1/2}$.

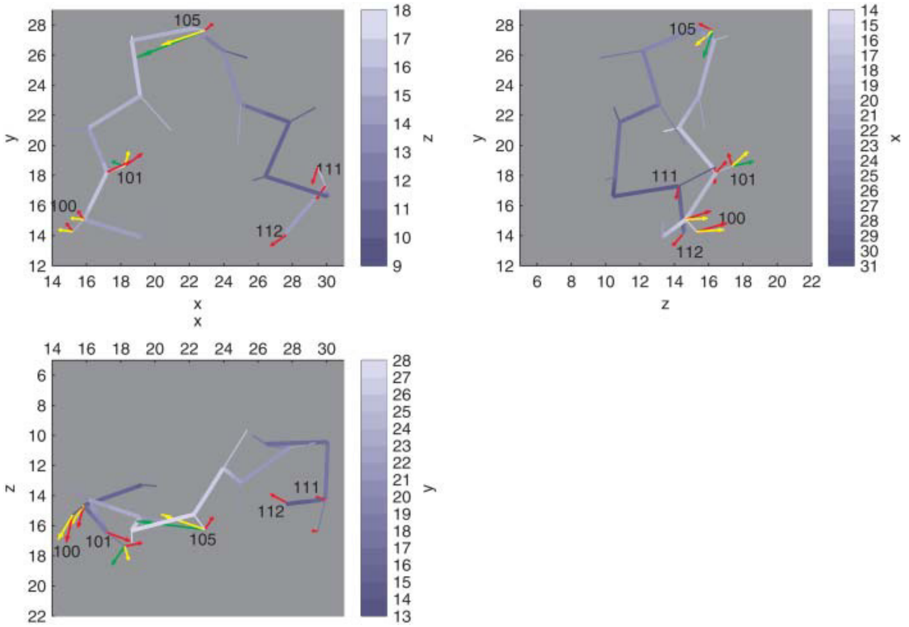

FIGURE 7 2D-projections of the 14 amino acids included in the molecular description (amino acid 99–112 in the antigen binding loop). The loop is shown as it is represented structurally in the molecular descriptors, i.e., it is described by the coordinates of the α-carbons and the center of the side chains, only. The figures show the loop from three perpendicular viewpoints. In each figure the third dimension is shown as a blue-shading of the backbone. Additionally, the directions of movement suggested to give a faster association (higher k_a), a slower dissociation (lower k_d), and a higher affinity (lower K_d) are shown as red, green, and yellow arrows, respectively. The longer the arrow the more a small movement in the indicated position affects the activity. The structure shown is the wild-type structure.

structure, one expects the minimized structures to constitute proper estimates of the true structures.

Structural differences

The structural differences of the mutants for position 99–112 are shown in Fig. 2. The structural changes in the proximity of the mutation site at position 105 are small. In the wild-type antibody, the backbone of residues 103, 104, and 106 are involved in hydrogen bonds to residues in the lysozyme (Transue et al., 1998) and alanine 104 is known to contribute to the interaction by filling a small hydrophobic pocket in the lysozyme (Transue et al., 1998). Therefore, these residues are likely to be fixed in their positions. Thus, regardless of how residue 105 is mutated the backbone is spatially restrained.

The larger structural deviations of residues 99–102 indicate that they were less restrained by the lysozyme than the residues around 105. It has been suggested that residues 99 and 100 are not in contact with the lysozyme, but that residues 101 and 102 are (Desmyter et al., 1996). However, being in contact with the lysozyme does not mean being restrained by the lysozyme.

The deviating structure of the mutant RT was probably due to the large arginine pointing toward the center of the loop. This affected not only position 99–102, but also position 107–111 at the opposite side of the loop (see Fig. 2).

Regression

The descriptors used for QSAR analysis were based on three ZZ-scales (Sandberg et al., 1998; De Genst et al., 2002) describing each of the two mutated residues and structural information of 14 residues (99–112) in the antigen binding loop. Applied to the data set used in this work the ZZ-scales alone gave predictive models ($Q^2 \approx 0.7$ and $P^2 \in [0.6, 0.7]$ (data not shown)). The structural information added here did not improve the predictivity of the models significantly, but more information of the interacting residues was gained. In particular, information of nonmutated residues was obtained.

The GA-PLS models had relatively high Q^2 values, higher than the PLS models (except for K_d). However, the P^2 values achieved with GA-PLS were much lower than the P^2 for the PLS models. This shows that even though the Q^2 values were higher for the GA-PLS models, the predictivity was lower (model overfitting).

Model weight analysis

A comparison of the weights of the ZZ-scales showed that their relative importance in general agreed with the results presented in De Genst et al. (2002). In Fig. 5 the scaled weights, $\sigma_i w_i$, for the ZZ-scales are shown.

To increase the k_a value, the ZZ2 and ZZ3 values for both the mutated residues (101 and 105) should be lowered, and the ZZ1 value of residue 105 should be increased. To slow down the dissociation (lower k_d) the ZZ1, ZZ2, and ZZ3 values should be decreased for residue 101, and ZZ2 and ZZ3 should be increased for residue 105. The same changes are desirable to lower the K_d value, except for the ZZ2 value for residue 105 that should be lowered, and additionally the ZZ1 value for residue 105 should be increased.

Analysis of the coordinate weights indicates that the coordinates of mainly five amino acids influence the interaction (see Fig. 7). The most prominent effect is seen for the side chain of 105 which should move closer to the backbone in a direction toward residue 103 in order to achieve a slow dissociation and a low K_d value. Residue 100 should move away from the loop, i.e., make it wider, for increasing binding strength, but residue 101 should move toward the center of the loop to increase the association rate. The side chain should move away from the backbone slightly to increase the affinity and slow down the dissociation. The coordinates of position 111 and 112 influence mainly k_a and a movement toward the center of the loop would increase the association rate. The model weight analysis shows that the model describing the association rate is different from model describing the dissociation rate.

Residue 102 in the flexible part could not be correlated to changes in interaction characteristics. In the study of the crystal structure of cAb-lys3 in complex with lysozyme (Desmyter et al., 1996) residue 102 is not found to be in contact with the lysozyme. This might explain why residue 102 can move rather freely without affecting the interaction.

Residues 107 to 110 are too restrained to influence the binding characteristics. This could be due to the tight binding of residue 106 to the lysozyme and the S-S binding between cystein 109 and cystein 32. This, however, does not mean that they are not important to the interaction, it merely means that their positions are not changed by the mutations.

Variable selection

As a complement to the weight analysis, a variable selection method was applied to separate the informative descriptors from the uninformative. The blind cross validation showed low predictivity (see Table 3), but the Q^2 values achieved for models based on the consensus selections were higher than achieved with PLS based on all variables or the ZZ-scales only. GA-PLS could be improved by using a larger population size and a more strict termination criterion, but it would also slow the speed of the analysis down considerably.

The variable selections clearly show that the ZZ-scales in general are more important than the coordinate descriptors. For all three models (k_a, k_d, K_d) the three most frequently selected variables (selected eight or nine times out of nine) were ZZ-scales.

Although the majority of the coordinates correlated to binding strength were side-chain coordinates, several backbone coordinates were included in the model. This indicates that it might be erroneous to consider only side-chain properties when analyzing interactions on a mutational level.

SUMMARY

In this paper, molecular mechanics methods, measured kinetics parameters, and 3D-QSAR analysis were combined. This resulted in models that can predict the kinetic properties of mutants and explain what structural properties are important to the model and how these properties should be changed to improve the binding. The investigation shows that the residues that are important to the association rate model are different from those important to the dissociation rate model.

The results from both the ZZ-scale and the coordinate weight analysis are now summarized. Together these results give more information of the interaction than either of them do alone, and illustrate the kind of quantitative information that one may extract using the methodology presented in this paper.

To increase the association rate the mutated residue 101 should move closer to the center of the loop, as should residues 111 and 112. When the residue moves closer to the center of the loop, residue 100 has to move aside. A movement of residue 100 further away from the loop center is suggested to increase the association rate. The association rate is also favored by a small residue at position 101 that preferably is nonelectrophilic and nonpolar.

The dissociation rate and the affinity are both affected mostly by the properties of the mutated residues 101 and 105 and the position of the side chain of residue 105. A small, hydrophobic, nonelectrophilic and nonpolar residue at position 101 and a polar and electrophilic residue at position 105 are preferred, if a fast and stable interaction is desired. A movement of side-chain 105 closer to the backbone in a direction toward residue 103 also gives a slower dissociation.

We thank Erwin De Genst, Magnus Malmqvist, and Markku Hämäläinen for critical reading of the manuscript. We also thank the anonymous reviewers for their helpful comments.

REFERENCES

Andersson, K., L. Choulier, M. D. Hämäläinen, M. H. V. van Regenmortel, D. Altschuh, and M. Malmqvist. 2001. Predicting the kinetics of peptide-antibody interactions using a multivariate experimental design of sequence and chemical space. *J. Mol. Recognit.* 14:62–71.

Baroni, M., G. Costantini, G. Cruciani, D. Riganelli, R. Valigi, and S. Clementi. 1993. Generating optimal linear PLS estimation (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat.* 12:9–20.

Choulier, L., K. Andersson, M. D. Hämäläinen, M. H. V. van Regenmortel, M. Malmqvist, and D. Altschuh. 2002. QSAR studies applied to the prediction of antigen-antibody interaction kinetics as measured by BIACORE. *Protein Eng.* 15:373–382.

Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.

De Genst, E., D. Areskoug, K. Decanniere, S. Muyldermans, and K. Andersson. 2002. Kinetic and affinity predictions of a protein-protein interaction using multivariate experimental design. *J. Biol. Chem.* 277:29897–29907.

Desmyter, A., T. R. Transue, M. A. Ghahroudi, M.-H. D. Thi, F. Poortmans, R. Hamers, S. Muyldermans, and L. Wyns. 1996. Crystal structure of a camel single-domain Vh antibody fragment in complex with lysozyme. *Nat. Struct. Biol.* 3:803–811.

Eriksson, L., J. Jonsson, S. Hellberg, F. Lindgren, B. Skagerberg, M. Sjöström, and S. Wold. 1990. Peptide QSAR on substance P analogues, enkephalins and bradykinins containing L- and D-amino acids. *Acta Chem. Scand.* 44:50–55.

Geladi, P., and B. Kowalski. 1986. Partial least-squares regression: A tutorial. *Anal. Chim. Acta.* 185:1–17.

Golbraikh, A., and A. Tropsha. 2002. Beware of q2! *J. Mol. Graph. Model.* 20:269–276.

Gustafsson, M. G. 2001. A probabilistic derivation of the partial least-squares algorithm. *J. Chem. Inf. Comput. Sci.* 41:288–294.

Haaland, P. D. 1989. Experimental design in biotechnology. Marcel Dekker, New York.

Hoffman, B., S. J. Cho, W. Zheng, S. Wyrick, D. E. Nichols, R. B. Mailman, and A. Tropsha. 1999. Quantitative structure-activity relationship modeling of dopamine D_1 antagonists using comparative molecular field analysis, genetic algorithms-partial least squares, and k nearest neighbor methods. *J. Med. Chem.* 42:3217–3226.

Höskuldsson, A. 1988. PLS regression methods. *J. Chemometrics.* 2: 211–228.

HyperChem. 2000. HyperChem(TM). HyperCube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA.

Kleywegt, G. J. 1999. Recognition of spatial motifs in protein structures. *J. Mol. Biol.* 285:1887–1897.

Kubinyi, H., F. A. Hamprecht, and T. Mietzner. 1998. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* 41:2553–2564.

Lapinsh, M., P. Prusis, T. Lundstedt, and J. E. Wikberg. 2002. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharmacol.* 61:1465–1475.

Lapinsh, M., P. Prusis, A. Gutcaits, T. Lundstedt, and J. E. Wikberg. 2001. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta.* 1525: 180–190.

Prusis, P., T. Lundstedt, and J. E. S. Wikberg. 2002. Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. *Protein Eng.* 15:305–311.

Prusis, P., R. Muceniece, P. Andersson, C. Post, T. Lundstedt, and J. E. S. Wikberg. 2001. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand -receptor interactions. *Biochim. Biophys. Acta.* 1544: 350–357.

Ortiz, A. R., M. T. Pisabarro, F. Gago, and R. Wade. 1995. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* 38:2681–2691.

Roos, H., R. Karlsson, H. Nilshans, and A. Persson. 1998. Thermodynamic analysis of protein interactions with biosensor technology. *J. Mol. Recognit.* 11:204–210.

Sandberg, M., L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold. 1998. New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. *J. Med. Chem.* 41:2481–2491.

Schiweck, W., and A. Skerra. 1997. The rational construction of an antibody against cystatin: lessons from the crystal structure of an artificial Fab fragment. *J. Mol. Biol.* 268:934–951.

Skerra, A. 2000. Lipocalins as a scaffold. *Biochim. Biophys. Acta.* 1482: 337–350.

Stanton, D. T. 2000. Developement of a quantitative structure-property relationship model for estimationg normal boiling points of small multifunctional organic molecules. *J. Chem. Inf. Comput. Sci.* 40: 81–90.

Transue, T. R., E. De Genst, M. A. Ghahroudi, L. Wyns, and S. Muyldermans. 1998. Camel single-domain antibody inhibits enzyme by mimicking carbohydrate substrate. *Proteins.* 32:515–522.

Tropsha, A. 2001. Rational combinatorial library design and database mining using inverse QSAR approach. *In* Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications in Drug Discovery. A.K. Ghose, and V.N. Viswanadhan, editors. Marcel Dekker Inc, New York. 363–378.

Tropsha, A., and W. Zheng. 2001. Identification of the descriptor pharmacophores using variable selection QSAR: applications to database mining. *Curr. Pharm. Des.* 7:599–612.

Xing, L., W. J. Welsh, W. Tong, R. Perkins, and D. M. Sheehan. 1999. Comparison of estrogen receptor alpha and beta subtypes based on comparative molecular field analysis (CoMFA). *SAR QSAR Environ. Res.* 10:215–237.