Protein Engineering, Design & Selection vol. 20 no. 6 pp. 301–307, 2007
Published online June 22, 2007 doi:10.1093/protein/gzm022

Proteochemometric modelling of antibody–antigen interactions using SPOT synthesised peptide arrays

Ilona Mandrika, Peteris Prusis, Sviatlana Yahorava, Medya Shikhagaie and Jarl E.S. Wikberg[1]

Department of Pharmaceutical Biosciences, Pharmacology, Uppsala University, PO Box 591, BMC, SE-751 24 Uppsala, Sweden

[1]To whom correspondence should be addressed. E-mail: jarl.wikberg@farmbio.uu.se

Proteochemometrics is a technology for the study of molecular recognition based on chemometric techniques. Here we applied it to analyse the amino acids and amino acid physico-chemical properties that are involved in antibodies' recognition of peptide antigens. To this end, we used a study system comprised by a diverse single chain antibody library derived from the murine mAb anti-p24 (HIV-1) antibody CB4-1, evaluated on peptide arrays manufactured by SPOT synthesis. The binding pattern obtained was correlated to physico-chemical descriptors (*z*-scales) of antibodies and peptides amino acids using partial least-squares projections to latent structures. Cross terms derived from antibody and antigen descriptors were included, which substantially improved the proteochemometric model. The final model was statistically highly satisfactory with a correlation coefficient $R^2 = 0.73$ and predictive ability $Q^2 = 0.68$. The physico-chemical properties of each interacting amino acid residue of both the peptides and the antibodies being essential for the antigen–antibody recognition could be retrieved from the model. The study shows for the first time the feasibility of using proteochemometrics to analyse the molecular recognition of antigens by antibodies.

Keywords: antibody interactions/peptide array/proteochemometrics/scFv library/SPOT synthesis

Introduction

The unique ability of antibodies to bind with high affinity to diverse antigens makes them attractive for medical and scientific research. The rational engineering of antibodies with proper affinity and specificity requires the identification of residues and interactions that influence antibody–antigen binding. Points of interaction can be predicted using sequence comparisons, mutagenesis and molecular modelling. To improve antibody affinity, various *in vitro* strategies have been devised to mimic the mammalian *in vivo* process of somatic hypermutation and selection (Hudson and Souriau, 2003).

Quantitative structure activity relationship modelling (QSAR) combined with multivariate experimental design has been used in attempts to rationally engineer antibodies. Reliable predictions of binding kinetics and affinity were obtained for antibodies interacting with peptides (Andersson *et al.*, 2001; Choulier *et al.*, 2002) and proteins (De Genst *et al.*, 2002; Freyhult *et al.*, 2003). A multivariate QSAR approach, in which ligands and the chemical environment were varied simultaneously was also applied and resulted in valuable information about kinetic parameters of the interaction (Andersson *et al.*, 2001; De Genst *et al.*, 2002).

Recently, we developed a new bioinformatics approach, proteochemometrics, which is useful for analysing molecular recognition (Lapinsh *et al.*, 2001; Wikberg *et al.*, 2004). Proteochemometrics is derived from chemometrics and is related to QSAR. In proteochemometrics, simultaneous modifications of all interacting molecules are applied, whereas in QSAR only the interacting ligand is modified. In proteochemometrics, descriptions of both proteins and the interacting ligands are correlated to experimentally measured interaction data by applying multivariate data analysis, thereby modelling the so-called protein–ligand interaction space. The method was initially applied to model G-protein coupled receptor (GPCR)-peptide interactions (Prusis *et al.*, 2002), as well as interactions of GPCRs with organic compounds (Lapinsh *et al.*, 2003). Proteochemometrics was shown to yield very robust models giving a multitude of information on the nature of the molecular recognition processes. It was also useful for the analysis of drug receptor interactions and in design of new ligands (Prusis *et al.*, 2002; Lapinsh *et al.*, 2005).

The purpose of the present study was to evaluate proteochemometrics for the modelling of antibody–antigen interactions. We used a small library of antigenic peptides and a small diverse library of mutated single chain Fv (scFv) antibodies derived from a murine anti HIV-1 p-24 antigen antibody CB4-1 (Winkler *et al.*, 2000) as a test case. The epitope-homologous peptide, h-peptide, derived by Kramer *et al.* (1997) was chosen as a representative of the 'wild type' antigenic peptide and used for substitution analysis. To this end, peptide arrays based on systematic alterations of the h-peptide were synthesised by the SPOT method, an approach that was earlier applied to identify peptide epitopes and to map protein–protein interaction sites (Reineke *et al.*, 2001). Interaction activities for peptides and scFv mutants were measured and the data were used to create proteochemometric models, which were of direct use to gain insights into the molecular nature of the antibody–peptide interactions.

Materials and methods

Antibody library

From the previously published set of antibodies, which had been produced by statistical experimental design (Mandrika *et al.*, 2007), we choose the six antibodies that had been found to bind the h-peptide, GATPEDLNQKLAGN-amide (wild-type peptide; Kramer *et al.,* 1997). The varied sequence fragments of these antibodies are listed in Table I.

Table I. scFv antibodies used herein for substitution analysis

scFv	V_L: 91	V_L: 92	V_L: 93	V_L: 94
2	**F**	D	**N**	**Y**
4	**F**	**N**	**N**	**W**
7	Y	**E**	**E**	**P**
11	**F**	**Q**	**Q**	F
12 (wt)	Y	D	D	F
14	Y	**N**	**Q**	**L**

The antibodies amino acids were mutated at positions 91–94 of the light chain fragment variable (V_L) region. (Amino acids varied from the wt antibody (12) are indicated with bold face letters.)

Construction of CB4-1 single chain mutants and their expression

The expression plasmid pHEN 4-1, coding for the murine mAb scFv CB4-1, originated from Myriam Ben Khalifa, Institute de Biologie Structurale, Grenoble, France, by courtesy of Dr Michael Pavlov, Department of Molecular Biology, Uppsala University, Sweden. scFv CB4-1 was recloned from the original plasmid to a new vector, pET 22b(+) (Novagen). This vector carries the sequence of a signal peptide PelB, which permits periplasmic expression of the scFv CB4-1 in *Escherichia coli*, and a His-tag for detection and purification purposes. The mutated scFvs were created by site-directed mutagenesis as described (Mandrika *et al.*, 2007). All mutations were confirmed by sequence analysis.

ScFvs were grown in the BL21 strain of *E.coli* in 2 × YT medium containing 100 μg ml^{-1} ampicillin and 1% glucose. After reaching OD_{600} 0.5–0.6, antibody expression was induced by adding 0.05 mM IPTG, followed by overnight incubation at room temperature. (Prior to induction the cells had been washed to remove glucose.) The soluble fractions of the scFvs were obtained from the periplasms by osmotic shock according to the pET vector manual (Novagen). The fractions were dialyzed against Na_2HPO_4 buffer, pH 8.0, and purified on Ni-NTA agarose gel (Qiagen). The eluted scFvs were dialyzed overnight against 50 mM Tris–HCl, pH 8.0; 5 mM EDTA and further purified on a Mono-Q FPLC column (Amersham Biosciences). The scFv fractions were concentrated to 0.2–0.8 mg ml^{-1} on 10 kDa Vivaspin concentrators (Vivascience AG). Antibody purity was checked by SDS–PAGE. Protein concentrations were determined using the BCA protein assay kit (Pierce).

SPOT synthesis of peptide arrays

Sets of cellulose-bound single substitution analogues of the h-peptide were prepared by SPOT synthesis. SPOT synthesis was carried out on an automated multiple peptide synthesizer (MultiPep, Intavis AG Bioanalytical Instruments, Germany) using amino-PEG_{500} cellulose membranes (Intavis AG Bioanalitical Instruments, Germany). The peptides were synthesised using N^{α}-Fmoc amino acids. The side chains protecting groups were the following: Pbf (Arg), Trt (Asn, Gln, Cys, His), *t*Bu ether (Ser, Thr, Tyr), O*t*Bu ester (Asp and Glu), Boc (Lys, Trp). Fmoc amino acids were pre-activated daily by incubating 0.15 mM of the amino acid with 330 μl 1-hydroxybenzotriazole ester (0.75 M) followed by dilution to 450 μl with 150 μl of *N,N*-diisopropylcarbodiimide (1.1 M). All solutions were in *N*-methylpyrrolidone. The coupling efficiency was monitored after each cycle by bromphenol blue staining (0.005% solution in DMF). After the peptide sequences had been assembled, the side chain protecting groups were removed by treatment with deprotection mixture (trifluoroacetic acid–dichloromethane (DCM)–triisopropylsilane–water, 7.5:7.5:0.45:0.3 ml). This treatment was followed by washes; four with DCM, four with DMF and two with EtOH (2 min). The membrane was then dried and kept in a refrigerator (−20°C) before use.

Binding of scFvs to cellulose-bound peptides

The membrane arrays were rinsed with ethanol for 1 min and washed three times with Tris-buffered saline buffer (TBS, 50 mM Tris-HCl, pH 8.0, 137 mM NaCl, 2.7 mM KCl) for 10 min. The membranes were then blocked overnight with TBS containing 0.05% Tween 20 (T-TBS), 3% dry milk, 1% BSA and 1% sucrose. After washing with T-TBS, 0.5 μg ml^{-1} or 1.5 μg ml^{-1} of scFvs were added in blocking buffer and incubated for 3 h at RT. The membranes were then washed three times with T-TBS, and peroxidase-labelled anti-rabbit-His antibody (Sigma, Sweden) was applied at 1:9000 dilution for 2 h at RT. After washing with T-TBS, the membranes were incubated with ECL™ chemiluminescence substrate (Amersham Biosciences, Sweden). Spots were visualised using a CCD-camera.

Evaluation of spot intensity

Spots were graded into five classes: no spot, traces, weak spot, dark spot and ring effect, as described (Kramer *et. al*, 1999) (Fig. 1). Four antibodies (Ab2, Ab4, Ab7, Ab11) were tested at a concentration of 1.5 μg ml^{-1}. For each spot, a number approximating the interaction activity was then assigned as follows: 0 was assigned for 'no spot', 1 for 'traces', 2 for 'weak spot', 3 for 'dark spot' and 4 for 'ring effect'. The remaining two antibodies (Ab12 and Ab14) gave very strong signals at 1.5 μg ml^{-1} and were therefore re-tested using a lower concentration, 0.5 μg ml^{-1}. As one can assume that equally intense spots for the two test concentrations of antibody correspond to higher interaction activity we assigned, in the later case, larger numbers to the tests of Ab12 and Ab14 as follows: 0 for 'no spot', 2 for 'traces', 3 for 'weak spot', 4 for 'dark spot' and 6 for 'ring effect'. No other attempts in electing scales were performed in order not to induce statistical bias into the modelling.

Numerical description of antibodies for proteochemometrics modelling

The physico-chemical properties of the antibodies were described using the five physico-chemical property scales (z-scales, z_1–z_5) of Sandberg *et al.* (1998). These scales are the principal components derived from principal component analysis of the physico-chemical properties of 87 natural and non-natural amino acids. z_1 is correlated to the hydrophobicity of amino acids, z_2 to size and polarisability, z_3 to

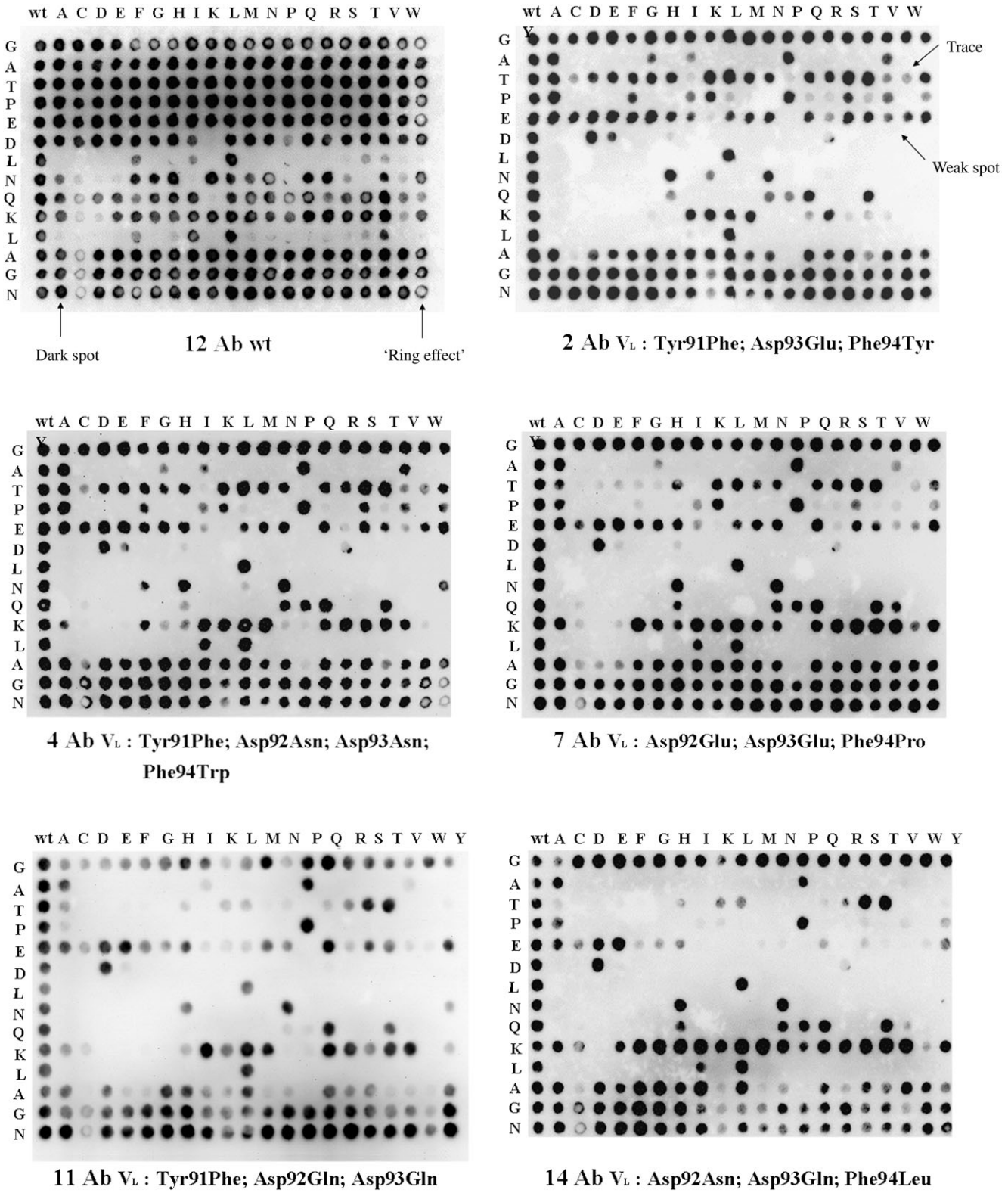

Fig. 1. Substitution analysis of the interaction of the h-peptide with multiple mutated scF versus each position of h-peptide was substituted by all other amino acids (rows). The sequence corresponding to the first left column represents the h-peptide and are identical. Binding of scFv library was visualised using a chemiluminescence detection system. Arrows indicate how spots were graded for modelling purposes.

polarity, z_4 to electronegativity and heat of formation, and z_5 to electrophilicity and hardness (Sandberg *et al.*, 1998). All five z-scales represent more than 95% of the measured and computed properties of the amino acids. Since, four positions were modified in the antibodies, 20 descriptors were obtained for each antibody (i.e. four positions times five z-scales = 20). All descriptors were mean centred and scaled to unit variance prior to further uses.

Numerical description of antigens for proteochemometrics modelling

Antigens were described in the same way as antibodies, using *z*-scales. Since all 14 positions were modified in the peptides, their descriptions required 14*5 = 70 descriptors. All descriptors were mean centred and scaled to unit variance prior to further uses.

Cross terms

In order to account for specific interactions between antibodies and antigens cross terms were computed. Thus, cross terms were computed by multiplying each antibody descriptor with each antigen descriptor. Since we have 20 antibody descriptors and 70 antigen descriptors 20*70 = 1400 cross terms were obtained. Mathematically, cross terms represent a non-linear approximation of interaction effects (see Wikberg *et al.*, 2004).

Block scaling

One could assume that each of the three above descriptor blocks (i.e. antibody, antigen and cross-term descriptor blocks) should have approximately equal importance for activity. On the other hand, there are 1400 cross terms but only 20 and 70 antibody and antigen descriptors, respectively. Therefore, in order to give equal weight for each descriptor block, we applied block scaling for each of the three blocks, as implemented in Simca-P+ 10.0.2 (Umetrics AB, Umeå, Sweden).

Proteochemometric modelling and validation

The numerical descriptions representing the physico-chemical properties of the ligands and antibodies and their interrelation (represented by the cross terms) were correlated with the experimentally determined spot intensities. We considered two models, one without (M1) and one with cross terms (M2). The correlation was done using partial least squares (PLS) as implemented in Simca-P + 10.0.2 (Umetrics AB, Umeå, Sweden; Wold, 1995). The goodness of fit was assessed using R^2 and root mean square error of estimate (RMSEE) values (SIMCA-P+ 10.0.2 User Guide, 2002). The models were further validated using cross-validation (Wold, 1978; Eriksson *et al.*, 1997). For the latter purpose, the data set was randomly divided into seven groups. Each group was excluded from the data once and a new model was computed. Using this new model, the excluded data were predicted and compared with the measured spot intensity. In this way, the performance of the modelling could be assessed by computing a cross-validated regression coefficient, Q^2, as described (Wold, 1978; Eriksson *et al.*, 1997). A model is considered acceptable for biological data if $R^2 > 0.7$ and $Q^2 > 0.4$ (Lundsted *et al.*, 1998). The cross-validation was used to estimate the number of PLS components to be used in the final model according to the rules implemented in SIMCA (SIMCA-P+ 10.0.2 User Guide).

We also performed permutation validation (Efron, 1987). For this purpose, the response data are randomly permutated and new models are induced. The process is repeated several times (twenty was used herein) and for each case R^2 and Q^2 values are computed. The obtained R^2 and Q^2 values are then plotted against the correlation coefficient between original response data and the permutated response data. The intercepts (denoted as iR^2 and iQ^2) of the lines drawn through R^2 and Q^2 points represent R^2 and Q^2 values of an entirely random model (Prusis *et al.*, 2002). A negative iQ^2 value indicates that the original Q^2 value was not obtained by pure chance (SIMCA; Prusis *et al.*, 2002).

Table II. Results of proteochemometric modelling of the peptide array binding to wt and mutated scFv to peptide arrays

Model	NC	R^2	Q^2	RMSEE	iR^2	iQ^2
M1	4	0.66	0.62	1.03	0.03	−0.17
M2	6	0.73	0.68	0.93	0.16	−0.15

Shown are the performances of models based on descriptors of antibodies and antigens (M1), and descriptors of antibodies and antigens, and their cross-terms (M2). NC, number of PLS components; R^2, correlation coefficient; Q^2, cross-validated correlation coefficient using seven cross-validation groups; RMSEE, root mean square error of estimate; iR^2, R^2 intercept from permutation validation; iQ^2, Q^2 intercept from permutation validation.

Results

Substitution analysis of a scFv library

The six mutants shown in Table I were selected from a diverse library of scFvs (see Materials and Methods for details). Each of these antibodies had been mutated in the complementarity determining region of the light chain at positions 91 to 94. The amino acids of the antibodies were mutated at three or four positions simultaneously. In order to maximise the chemical space information of the antibodies, while keeping the number of experiments as low as possible, statistical molecular design had been applied to select the amino acid mutations in the library (Mandrika *et al.*, 2007). The scFvs selected had been earlier tested for their ability to bind to the h-peptide, GATPEDLNQKLAGN. The wt and no. 14 scFvs bound the h-peptide in the nanomolar affinity range while the other four (no. 2, 4, 7 and 11) bound the h-peptide with a micromolar affinity (Mandrika *et al.*, 2007).

In order to compare the binding patterns of the wt and mutated scFvs, we carried out a substitution analysis of the h-peptide using peptide libraries SPOT synthesised on cellulose membranes. Each residue of the h-peptide was exchanged by all other 19 possible amino acids providing data for 266 peptide variants. The substitution analysis made it possible to evaluate the importance of each position in the h-peptide sequence and reveal key amino acids essential for binding. As seen from Fig. 1, key residues (i.e. positions which could not be substituted at all, or only by amino acids with similar physico-chemical properties) were located in the core of the h-peptide. The peptide arrays revealed that the key positions for antibody binding remain the same for both wt and mutated antibodies. Thus, substitution of Leu at position 7 of the h-peptide with any other amino acid led to loss of binding for all the mutated scFvs, while for the wt antibody it could only be replaced by Ile and Phe, Val, Thr and then yielding lower binding signals. Moreover, Leu at position 11 could be substituted by the physico-chemically similar amino acid Ile, which gave binding signals for the wt antibody and antibodies no. 4, 7 and 14. Substitutions in the N-terminus of the h-peptide seemed to afford more specific

effects among the mutated antibodies, however. Thus, Ala at position 2 could be substituted only by Pro for Ab no. 11 and 14, while for antibodies no. 2, 4 and 7 it could be substituted only by Pro, Gly and Val. Changes of amino acids at the C-terminus did not affect the binding much (except in a few are cases) and accordingly would not contribute much to the binding of the h-peptide to the antibodies.

Proteochemometric modelling

In order to bring the analysis further proteochemometric modelling was performed. For the purpose, the intensities of the peptide array spots were converted into numbers and correlated with the physico-chemical descriptions of the peptides and antibodies (see Materials and Methods for details). The results of the modelling are summarised in Table II. As can be seen, quite good models were obtained already by using only antibody and antigen descriptors, the RMSEE being only 1.03 units (M1 model). These results are indeed good, considering that we assigned the spots discrete values on a scale where each level differs by 1 unit. However, most importantly, the addition of cross terms yielded a marked increase in both R^2 and Q^2 (by 0.07 and 0.06 units respectively), as well as a decrease of RMSEE, which reached below unity (0.93) (M2 model). This finding thus indicates that the model takes advantage of information on specific antigen–antibody interactions. Moreover, the permutation validation testing indicated that for both models the R^2 and Q^2 were not obtained by chance. We can therefore conclude that the model, within its scope, can be used for predictions and interpretations of the antibody–antigen recognitions.

Interpretation of the M2 model

Since the M2 model showed better predictability and applied a richer description, we used it for interpretations. For this purpose, the PLS regressions coefficients were used (Prusis *et al.*, 2006), interpreting the coefficients for each descriptor block separately (i.e. interpreting antibody descriptors, antigen descriptors and cross terms separately). In interpreting these coefficients, it should be noted that coefficients derived from the peptide descriptors can be used to explain the average binding activity for a peptide for all the antibodies in the library, while the coefficients derived from the antibody descriptors can be used to explain the average binding activity for an antibody for all the peptides in the library. The cross terms derived from antibody and peptide descriptors, on the other hand, reflect specific interactions between peptides and antibodies that determine the preference for interaction of particular antibody–antigen combinations (i.e. the 'specificity' or 'selectivity' part of the interactions).

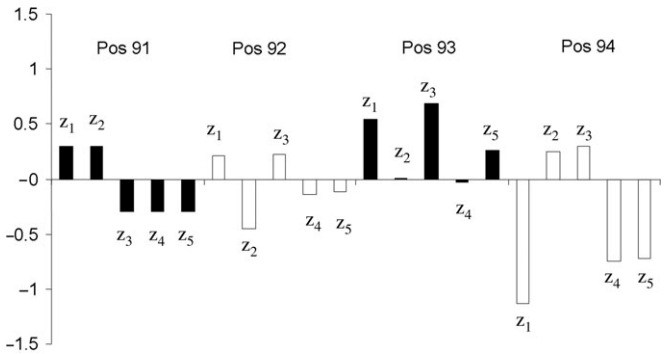

Fig. 2. PLS coefficients of antibody descriptors derived from the proteochemometric model.

Interpretation of descriptors for antibodies The regression coefficients for the antibody descriptors are shown in Fig. 2. As can be seen, the largest coefficients (both positive and negative ones) are found for position 93 (z_1 and z_3) and position 94 (z_1, z_4 and z_5). Considering the z-scales of natural amino acids, this means that a higher affinity antibody is likely to be obtained if there is an Asp (i.e. the amino acid of the wt antibody) or Asn in position 93, while a Gln at this position would have a quite negative impact on affinity.

For position 94, results are even more interesting. Firstly, the large negative value of the regression coefficient for z_1 indicates a preference for Trp, Leu and Phe for affording high affinity. On the other hand, the negative value of the coefficient for z_4 at position 94 indicates that Leu would be beneficial, while it strongly rejected Trp for an antibody to achieve high affinity. Furthermore, the negative value of z_5 indicates preferences for Trp, Tyr and Phe. One can therefore conclude that the likelihood to achieve an antibody with high affinity is the largest for the wt Phe, and for Leu, while a Pro would be the least favourable amino acid at this position.

The absolute values of coefficients for position 91 are of equal magnitude, which is due to the fact that there are only two amino acids present in the dataset (Tyr and Phe). The directions of the coefficients confirm that a native Tyr at this position gives better chance for a high affinity antibody.

For position 92, the largest coefficient is seen for the z_2 descriptor. The negative value for this coefficient indicates a slight preference for the non-native Glu at position 92, as it has the smallest value for the z_2 property compared to the other amino acids used at this position (Asp, Asn and Gln).

Interpretation of descriptors for peptides The PLS coefficients for the antigens are shown in Fig. 3. As can be seen from the figure, several of the PLS coefficients take their largest values for antigen positions 1, 7, 11 and 13. Position 1 shows a very similar pattern as position 13, with preference for large, hydrophobic, aromatic and rigid amino acids for affording a high affinity antigen. Also positions 7 and 11 share similar patterns, giving preference towards hydrophobic, non-aromatic, flexible amino acids. Somewhat smaller PLS coefficients are found for positions 2, 6, 8, 10, 12 and 14, but still giving clear clues to properties needed to increase the chance to obtain an antigen with high affinity. For example, for positions 6 and 8 hydrophilic, neutral or acidic amino acids are preferred, although for position 6 there is a slight preference towards acidic amino acids. Positions 2 and 12 show quite opposing patterns—position 2 preferring small, hydrophobic amino acids, and position 12 preferring large and hydrophilic amino acids. The patterns for positions 10 and 14 are fairly similar to the patterns of positions 7 and 11, although the considerably larger negative value for z_5 of position 14 indicates some preference for sulphur containing amino acids. For positions 4 and 9, only the z_5 property has a notable large value, a situation that is insufficient to clearly assign amino acid preferences for these positions, although proline may have certain prevalence for these positions. Also, for positions 3 and 5, the PLS

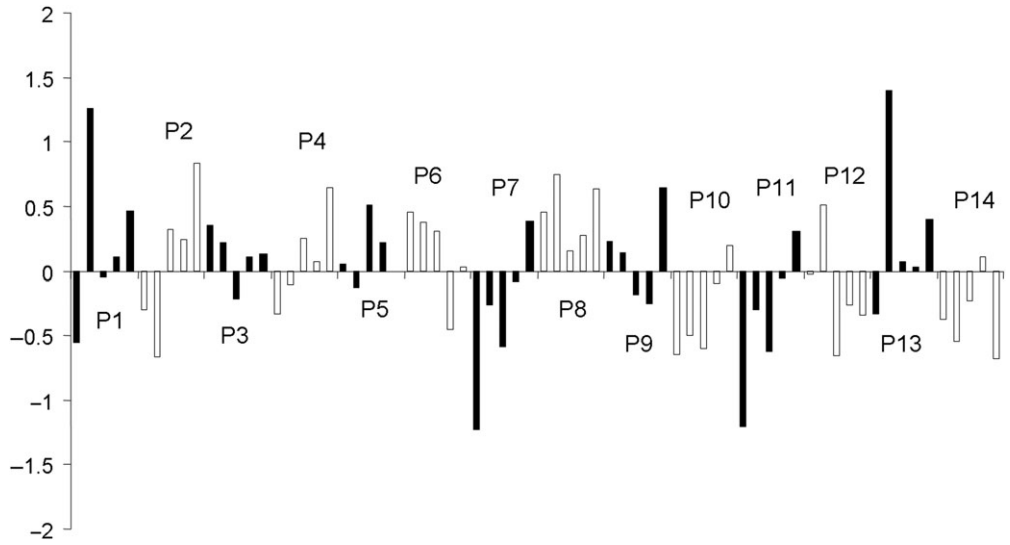

Fig. 3. PLS coefficients of antigen descriptors derived from the proteochemometric model. P1–P14 corresponds to the h-peptide amino acid positions. Bars represent the coefficients for the respective five z-scale descriptors (z_1–z_5).

coefficients show quite small absolute values, indicating that changes in this position are rather unimportant for the antigens' affinities.

Interpretation of cross terms The model comprised 1400 cross terms and most of these had much smaller PLS coefficients compared to those for the original antibody and antigen descriptors. On other hand, a cross term describes the influence of a simultaneous change in an antibody and an antigen and the existence of a correlation of a cross term to the interaction activity can be directly interpreted as the existence of a specific, direct or indirect interaction between amino acids at the positions described by the cross term (Prusis *et al.*, 2002, 2006). Here, each cross-term described the physico-chemical properties of two amino acids, one from the antibody and another from the antigen, which in conjunction are of importance for the interaction.

For the present analysis, we considered all cross terms and identified those which had absolute values of their PLS coefficients larger than 0.15 units. Amino acid pairs with at least one such large cross-term are shown schematically in Fig. 4. As one can read from the figure, the simultaneous changes of peptide amino acids at positions 7 and 10, together with amino acids in the antibodies influence the affinities more often than any other concomitant changes. Thus, positions 7 and 10 of the peptide may be the major contributors to afford specificity of the peptides for the antibodies. Figure 4 shows furthermore that amino acids in positions 92 and 93 of the antibodies accounted for most of the important cross terms, thus indicating that these antibody positions have the largest impact on specific antigen–antibody recognitions.

Discussion

Proteochemometrics was introduced as a general approach for the study of molecular recognitions by bio-molecules. It was successfully developed and validated for studies of the molecular recognition of peptides and organic compounds by physiological receptors (Prusis *et al.*, 2002; Lapinsh *et al.*, 2003; Wikberg *et al.*, 2004). The present study was directed to evaluate the usefulness of proteochemometrics to probe antigen–antibody recognition.

In a previous study, we analysed the binding of an epitope-related peptide with a statistically designed library of scFv antibodies using QSAR (Mandrika *et al.*, 2007). From

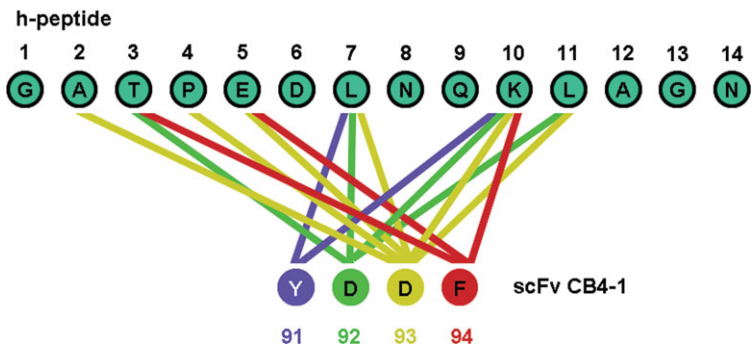

Fig. 4. Schematic representation of the most important specific antibody–antigen interactions found by analysis of PLS coefficients of cross terms. Indicated by the lines are the amino acid pairs for which the absolute value of at least one of their cross-term PLS coefficients were larger than 0.15 units.

the QSAR model, it was possible to deduce which physicochemical properties of the amino acids varied in the antibody library that influence the antibody–peptide interactions. However, since the variation of the peptides was not a part of such a QSAR study, it was not possible to gain any information on which peptide properties that are essential for binding to the antibodies. In the present study, we evaluated six scFvs from our previous antibody library for their interaction with a set of h-peptide substitutions using SPOT arrays. The simultaneous modification of both interacting molecules allowed us to provide a much richer description of the antibody–antigen interaction space and accordingly allowing us to gain more information on the nature of the antibody–peptide interactions.

Several studies were previously performed using X-ray diffraction methods and substitution analysis of the epitope-related peptide in attempts to identify key residues of the antigen participating in the interaction (Kramer *et al.*, 1997; Keitel *et al.*, 1997). However, by combining the substitution and modelling approaches used herein, a deeper insight into the mechanism of antibody–antigen recognition was obtained.

By interpreting the *z*-scale descriptors of the amino acids of the antigens and antibodies, we were able to clarify the contribution of the amino acids' physico-chemical properties, at each of the peptide and antibody position varied in the libraries, to the interaction. Thus, for the antibody, the wt hydrophobic amino acid residues Tyr and Phe in positions 91 and 94, respectively, were determined by the model to be the ones that would afford the highest average interaction activity, while for positions 92 and 93 Glu and Asp or Asn could be used to afford the highest interaction activity.

The proteochemometric model further reveals that key amino acid residues at positions 7 and 11 of the peptide have influence for the average peptide affinity. The model shows that for both positions hydrophobic, non-aromatic and flexible amino acids are favourable. The relative importance is in general agreement with the results obtained by Kramer *et al.* from substitutional analysis (Kramer *et al.*, 1997) and crystal structure analysis of the h-peptide complexed with the monoclonal antibody CB4-1 (Keitel *et al.*, 1997).

The addition of the cross terms between physicochemical descriptors of amino acids of peptides and antibodies resulted in a significant improvement of our model. This means that information about specific interactions between antigens and antibodies is included into the model. Analysis of the cross terms revealed that h-peptide selectivity is mainly determined by the interactions of the Leu at position 7 and Lys at position 10 with the antibody residues. Moreover, for the antibodies amino acid residues at positions 92 and 93 seems to be most influential for selectivity.

In summary, we have shown that it is indeed possible to create good proteochemometrics models using information gained through substitutional analysis of peptide antigens interacting with multiple mutated scFv, and that proteochemometric can be used to identify the important features of amino acid residues in antibodies and their peptide antigens. Such information is difficult or even impossible to dissect by analysing data with the aid of the human brain only, and is so in particular for cases where multiple mutated proteins and antigens are applied.

Acknowledgements

We thank K. Tars for technical assistance. This research was supported by grant from the Swedish VR (04X-05957) and the Swedish Animal Welfare Agency.

References

Andersson,K., Choulier,L., Hamalainen,M.D., van Regenmortel,M.H., Altschuh,D. and Malmqvist,M. (2001) *J. Mol. Recognit.*, **14**, 62–71.
Choulier,L., Andersson,K., Hamalainen,M.D., van Regenmortel,M.H., Malmqvist,M. and Altschuh,D. (2002) *Protein Eng.*, **15**, 373–382.
De Genst,E., Areskoug,D., Decanniere,K., Muyldermans,S. and Andersson,K. (2002) *J. Biol. Chem.*, **277**, 29897–29907.
Efron,B. (1987) *J. Am. Stat. Assoc.*, **78**, 171–200.
Eriksson,L., Johansson,E. and Wold,S. (1997) In Schuurmann,G. and Chen,F. (eds.), *Quantitative Structure-Activity Relationships in Environmental Sciences-VII.* SETAC, Pensacola, pp. 381–397.
Freyhult,E.K., Andersson,K. and Gustafsson,M.G. (2003) *Biophys. J.*, **84**, 2264–2272.
Hudson,P.J. and Souriau,C. (2003) *Nat. Med.*, **9**, 129–134.
Keitel,T., Kramer,A., Wessner,H., Scholz,G., Schneider-Mergener,J. and Hohne,W. (1997) *Cell*, **91**, 811.
Kramer,A., Keitel,T., Winkler,K., Stocklein,W., Hohne,W. and Schneider-Mergener,J. (1997) *Cell*, **91,** 799–809.
Kramer,A., Reineke,U., Dong,L., Hoffmann,B., Hoffmüller,U., Winkler,D., Volkmer engert,R. and Schneider Mergener,J. (1999) *J. Pept. Res.*, **54**, 319–327.
Lapinsh,M., Prusis,P., Gutcaits,A., Lundstedt,T. and Wikberg,J.E.S. (2001) *Biochim. Biophys. Acta*, **1525**, 180–190.
Lapinsh,M., Prusis,P., Mutule,I., Mutulis,F. and Wikberg,J.E.S. (2003) *J. Med. Chem.*, **46**, 2572–2579.
Lapinsh,M., Prusis,P., Uhlén,S. and Wikberg,J.E.S. (2005) *Bioinformatics*, **21**, 4289–4296.
Lundsted,T., Seifert,E., Abramo,L., Thelin,B., Nyström,A., Pettersen,J. and Bergman,R. (1998) *Chemometr. Intellig. Lab. Syst.*, **42**, 3–40.
Mandrika,I., Prusis,P., Yahorava,S., Tars,K. and Wikberg,J.E.S. (2007) *J. Mol. Recognit.* **20**, 97–102.
Prusis,P., Lundstedt,T. and Wikberg,J.E.S. (2002) *Protein Eng.*, **15,** 305–311.
Prusis,P., Uhlen,S., Petrovska,R., Lapinsh,M. and Wikberg,J.E.S. (2006) *BMC Bioinformatics*, **7**, 167.
Reineke,U., Volkmer-Engert,R. and Schneider-Mergener,J. (2001) *Curr. Opin. Biotechnol.*, **12**, 59–64.
Sandberg,M., Eriksson,L., Jonsson,J., Sjöström,M. and Wold,S. (1998) *J. Med. Chem.*, **41,** 2481–2491.
SIMCA-P+ 10.0.2. (2002) *User Guide and Tutorial*. Umetrics AB, Umeå.
Wold,S. (1978) *Technometrics*, **20,** 397–405.
Wold,S. (1995) In van de Waterbeem,D. (ed.), *Chemometric Methods in Molecular Design*. VCH Verlagsgesellschaft, Weinheim, pp. 195–218.
Winkler,K., Kramer,A., Kuttner,G., Seifert,M., Scholz,C., Wessner,H., Schneider-Mergener,J. and Hohne,W. (2000) *J. Immunol.*, **165**, 4505–4514.
Wikberg,J.E.S., Lapinsh,M. and Prusis,P. (2004) In Muller,G. and Kubinyi,H. (eds.), *Methods and Principles in Medicinal Chemistry.* Wiley-VCH, Weinheim, Germany, pp. 289–309.

Received November 4, 2006; revised March 31, 2007; accepted April 30, 2007

Edited by Lars Baltzer