Dissertation
am Fachbereich Statistik der Universität Dortmund

Extensions of the Partial Least Squares approach for the analysis of biomolecular interactions

Nina Kirschbaum

Erstgutachter: Prof. Dr. W. Urfer
Zweitgutachter: Prof. Dr. G. Trenkler

Tag der mündlichen Prüfung: 02.07.2007
Ort der Einreichung: Dortmund
Jahr der Einreichung: 2007

Contents

Chapter 1

Introduction

Many important processes in our organism are based on interactions between biomolecules. The consequences of these interactions depend on the characteristics of the binding behaviour of the corresponding biomolecules. Therefore, the understanding of the influence of diverse factors on particular interactions is of special relevance. The analysis of biomolecular interactions is a potential field of application of Partial Least Squares (PLS) regression since only a limited number of experiments can be performed but a large number of possibly relevant factors are taken into account.

The literature dealing with PLS regression and its application show a lack of completeness in the statistical description of the details of the method, not least the principles underlying its derivation and the statistical properties of the resulting models and their estimates. It is especially in publications using PLS regression to investigate a particular interaction that the underlying computations involved in the data analysis are inadequately explained. This dissertation aims to provide a detailed explanation, using a uniform notation, of the methodology of the PLS procedure as well as its application to the analysis of biomolecular interactions. Consequently, it can be considered a contribution to a comprehensive and advanced presentation of PLS regression within the context of one of its most important areas of application. Particular emphasis is given to the problem of the occurrence of mutants of viruses.

The objective of the analysis of biomolecular interactions is to obtain an understanding of the influences on the interaction under study by modelling the effects of a number of factors on the binding behaviour between certain biomolecules. Often, interactions between an enzyme and its substrate or an antigen and its antibody are under investigation. Potential factors that

the binding process might depend upon are physico-chemical properties or structural features of amino acids at particular positions in the sequence of one or both of the binding partners as well as the composition of the buffer, i.e. the chemical environment in which the interaction takes place.

Usually, in publications on biomolecular interactions, the explanations refer to a particular binding process of interest. Based on these descriptions, a general and comprehensive presentation of the performance of modelling the influences on the binding between interacting biomolecules has been derived.

The binding behaviour can be characterized by kinetic parameters such as the association and dissociation rate constant as well as the affinity constant. These binding parameters can be measured with a high accuracy by biosensor systems relying on the physical process of surface plasmon resonance (SPR). In practice, the biosensor systems most frequently used are the Biacore instruments that are described in detail later for a comprehensive insight into the biochemical and technological background of biomolecular interactions.

The measurements of the kinetic parameters are performed under various conditions in order to obtain knowledge of the influences on the binding process. In detail, the concentrations of diverse chemical additives and the levels of the pH-value of the buffer are varied. Further, amino acid substitutions are realized at specific positions in the sequence of the wild-type protein. Consequently, binding parameters with respect to different modified proteins and diverse buffers are obtained. The choice of possibly relevant buffer components and appropriate mutation sites where amino acid replacements might influence the interaction without preventing the binding requires the expertise of biochemists. The exact experimental settings are based on a statistical design plan.

In order to determine the effect of the factors of interest on the interaction under examination, regression models are established with the help of the available data. The measured binding parameters are used as values of the response variables. Accordingly, variables representing the physico-chemical properties or structural features of the amino acids at the mutation sites as well as the buffer composition are incorporated as descriptor variables in the regression models. By estimating the unknown model parameters, i.e. the intercepts and the regression coefficients, of the respective regression models, the influence of the modifications in the amino acid sequence and the chemical environment on the interaction can be quantified.

In practice, univariate regression models are developed individually for each of the response variables under consideration, i.e. the association or dissociation rate constant or the affinity constant. In general, the regression models are established separately with respect to the different subgroups of descriptor variables being considered. Consequently, the regression models developed in the analysis of biomolecular interactions can be assigned to one of the following types: quantitative buffer-kinetics relationship, quantitative sequence-kinetics relationship, quantitative structure-activity relationship or 3D-quantitative structure-activity relationship. These are referred to respectively as (QBKR), (QSKR), (QSAR) or (3D-QSAR)-models. This dissertation presents a comprehensive formal description of these different kinds of regression models that cannot be found in the articles published with respect to a particular interaction.

After having specified the corresponding regression models, the measured data can be interpreted. Beyond this, values of the response variables, i.e. the kinetic parameters, can be predicted for specified values of the descriptor variables. The information obtained with the help of the regression models leads to an improved understanding of the interaction under investigation allowing a basis for the explanation of the molecular regulation and function of the interacting biomolecules. Gaining knowledge of the influences on the interaction is of special relevance in pharmacology. In particular, the conclusions drawn from the results of the regression analysis can be used in the development of drugs. For example, the binding properties of a potential antibody can be optimized with respect to a specific antigen by determining those amino acid substitutions resulting in a particular desired profile of the corresponding binding parameters.

However, the study presented here deals with a different and important application of the analysis of biomolecular interactions. In fact, the objective of the current investigation is to obtain knowledge of the circumstances, especially the amino acid sequence of an antigen of a particular virus, under which the corresponding available antibody can still be expected to be effective. Consequently, changes in the experimental settings were performed by modifying the amino acid sequence of the antigen instead of that of the antibody. Further, by considering different buffer compositions in the experiments, the statements concerning the influence on the binding behaviour of the mutations at the mutation sites can be determined with respect to several chemical environments occurring in the cells of an organism.

The motivation for this kind of research is the fact that mutations in the RNA or DNA of viruses might take place which lead to the expression of

modified proteins of the capsid. Accordingly, as a result of the mutation, the features of the coat protein might be altered. In unfavourable cases, the existing antibodies are not able to bind to the virus any longer or at least not as effectively as normally because of these changes in the capsid.

The binding characteristics of the interaction between an occurred mutant of a virus and the corresponding available antibody can be predicted with the help of the results of a regression analysis involving the virus of interest or a similar one. These predictions can be easily computed since the amino acid sequence of a particular mutant can be determined in laboratories, and the values respecting the relevant physico-chemical properties of the amino acids at the mutation sites can be used as values of the corresponding descriptor variables in the established regression models. Comparing these predictions of the association and dissociation rate constant with the binding parameters of the wild-type virus yields conclusions on the efficacy of the existing antibody with respect to the occurred mutant.

In this dissertation, the binding process under study is the interaction between a peptide of the antigen of the tobacco mosaic virus protein (TMVP) and a Fab fragment of the monoclonal antibody 57P. Improved understanding of this interaction is of special importance because the TMV is a virus infecting a number of species such as tobacco, tomato, pepper and cucumber and hence it might lead to enormous crop losses. Beyond this, information on the TMV can be useful as well to elucidate the binding behaviour of other similar viruses. For example, it is hoped that conclusions concerning the Orthomyxovirus causing influenza may be drawn from the research on the TMV. In the context of the Orthomyxovirus, the possibility of predicting the binding characteristics between the occurrence of a mutant and the available antibody in order to evaluate the efficacy of the antibody is especially relevant because the application of ineffective vaccinations might be prevented.

The importance of being prepared for possible mutants of viruses is reflected by the fact that several worldwide flu epidemics have happened in the last 100 years. The application of regression methods to measurements of particular biomolecular interactions contributes considerably to this ambitious task.

In practical applications of regression analysis, a common problem is collinearity amongst the descriptor variables, e.g. if the number of descriptor variables exceeds the number of objects in the sample. In biomolecular interaction studies, it is often the case that a limited number of observations is available

but a relatively large number of descriptor variables is taken into account. The reason for this situation is the fact that the production of modified proteins is very expensive.

If the problem of collinearity among the descriptor variables arises, the classical Ordinary Least Squares (OLS) estimation of the unknown model parameters cannot be performed. In order to specify the regression model nevertheless, alternative methods, which reduce the dimensionality of the data, have to be applied. One of these alternative procedures is Partial Least Squares (PLS) regression.

By applying PLS regression, it is possible to estimate the intercepts and regression coefficients of models in cases when the OLS method fails. PLS regression can be considered to have the advantage over the other available alternatives since the information inherent in the descriptor variables is compressed effectively with incorporation of the information of the response variables. Consequently, regression models established by the PLS method provide reliable predictions, the most important characteristic of a good regression model in practice. Therefore, PLS regression is the appropriate procedure to apply to data obtained in biomolecular interaction studies.

The basic idea of PLS regression is to extract a few but relevant latent variable vectors comprising the information of the descriptor variables. The latent variable vectors are obtained by considering the information in the response variables as linear combinations of the standardized descriptor variables. The extraction of the information by the successive construction of latent variable vectors can be represented by decomposition models referring to the descriptor and response variables, respectively.

The latent variable vectors as well as further terms are computed iteratively by applying an algorithm to the standardized available data. Several variants of the PLS algorithm exist. These differ principally from each other in the normalizations that are performed. In this dissertation, the NIPALS-(Non-Iterative PArtial Least Squares-) algorithm is described. The term PLS regression comes from the fact that a number of "partial" regressions are performed in the course of the algorithm. Since these partial regression models show a reduced dimensionality compared with the original regression problem, the OLS procedure is applied to them.

In contrast to the OLS method, the results obtained from a multivariate regression analysis differ from those of several corresponding univariate re-

gressions in the PLS procedure. Often, in PLS regression publications, either the multivariate or the univariate PLS algorithm is explained. However, in this dissertation, both the multivariate and the univariate algorithm are presented in detail, where the univariate algorithm is shown to be derived as a special case of the multivariate case.

The different computations performed in the iterations of the PLS algorithm are described and interpreted in detail. These comprehensive explanations of the relationships between the obtained terms cannot be found in the cited literature. In particular, the motivation for the calculation of the weight vectors used to extract the latent variable vectors is generally presented incorrectly. Therefore, in this thesis, the maximization problem in respect of a certain covariance related to this context is explained by regarding the elements of particular column vectors as realizations of corresponding random variables.

The PLS algorithm has been reported in the literature in a different form using a different formulation of the decomposition models. The equivalence of this unnecessarily complicated presentation to the one described below is shown.

With the help of the terms computed in the PLS algorithm, the unknown model parameters of the regression model involving the standardized variables can be estimated. The derivation and computation of the formulae of this estimation of the regression coefficients corresponding to the standardized variables is explained in detail, in fact in a comprehensiveness that cannot be found in the articles listed in the references.

The estimates of the regression coefficients of the standardized variables can be used to calculate the estimates of the regression coefficients of the original variables. It is shown how the derivation of the formulae for the estimation of the original regression coefficients is based upon the formulae for the estimation of the standardized regression coefficients. The relationship between these estimation formulae for the regression coefficients belonging to the standardized and original variables and thus the expression of the estimated original regression coefficients and its dependence on the standardized regression coefficients, has not been given before in any of the literature cited. This presentation of the estimation formulae can be considered as important contributions to the completeness of the statistical description of the PLS regression.

Generally, the explanations in the literature cited in the references of the prediction of values of the response variables are vague. It is not described accurately whether the descriptor variables are used in the standardized or original form and accordingly nor whether the predictions refer to the standardized or original response variables. In this dissertation, two ways of presenting the computation of predictions of the original response variables are described on the basis of the previously derived estimation formulae. Furthermore, a proof is given that the predictions obtained by these prediction procedures coincide with each other.

The prediction accuracy of a regression model obtained after the performance of the PLS algorithm depends upon the number of realized iterations since the model parameters specifying the regression equation are estimated with the help of the terms computed in the algorithm. The prediction accuracies of models derived on the basis of different numbers of iterations are compared so as to determine the optimal number of iterations resulting in the regression model providing the most exact predictions. Usually, the calculation of the prediction accuracy is performed by applying leave-one-out cross validation.

In the application of biomolecular interactions, the evaluation of the prediction accuracy is of special relevance. The reliability of a specific prediction of the binding parameters referring to specific values of the descriptor variables concerning an occurred mutant can be judged additionally by computing a prediction interval for the particular prediction. However, in contrast to the OLS method, prediction intervals for PLS regression can only be computed approximately.

In order to analyze the available data on the interaction under study, the initial regression analysis is first summarized and reproduced. In detail, the univariate subgroup models without interaction terms, i.e. the QBKR models per peptide and the QSKR models are established. After comparing the results of the reproduced models with the earlier ones, explanations of the differences in the specifications of the models are given.

A significance test of the regression coefficients is not implemented in the software used for the application of the PLS regression. Since its performance is not reported in the literature referred to, it could not be programmed. Instead, another criterion usually applied in the context of PLS regression was used for the evaluation of the relevance of the descriptor variables. In fact, the VIP-values (Variable Importance for Projection) that are not automatically computed by the software employed were programmed for the

descriptor variables included in the different regression models. Further, a formula of the computation of the VIP-values was derived on the basis of the descriptions in software manuals.

Usually, in a biomolecular interaction study, univariate regression models without incorporation of interaction terms are established separately with respect to the different subgroups of descriptor variables, though an alternative modelling procedure might be advantageous. Therefore, several novel aspects of the analysis of binding processes are proposed, performed and evaluated. The aim of this research is to improve the prediction accuracy of the resulting regression models. By realizing this advanced and comprehensive regression analysis to the available data, the knowledge of the interaction between the TMVP and the existing antibody 57P is extended.

Broadly, the novel aspects refer to the multivariate modelling, a unified modelling, the incorporation of interaction terms as well as variables representing a more detailed quantification of physico-chemical properties of amino acids. In detail, it is proposed to perform multivariate regression analysis instead of the usual univariate analysis. Additionally, a unified regression model including all potential descriptor variables is presented. This is proposed as a replacement for the diverse subgroup models presenting special cases of this unified model. With the help of a unified model, the effects of both the amino acid and buffer variables can be modelled and presented simultaneously in one single compact model. Consequently, the novel unified modelling approach contributes to a facilitation of the biochemical interpretation of the results of the regression analysis.

The incorporation of interaction terms in the regression models is proposed in order to obtain information on the importance of interactions between the included descriptor variables. The additional information contributes notably to an improved understanding of the influences on the binding process under investigation. The investigation of interaction terms in a unified regression model is especially useful since this permits the modelling of the interactions between the amino acid and buffer variables beyond those within the amino acid variables or within buffer variables separately. Previous analyses of the available data suggest that such interactions exist.

The physico-chemical properties of amino acids are usually quantified by regression models involving only a few variables. A more detailed representation of the possibly relevant physico-chemical properties of amino acids is proposed for use in the regression analysis since it permits a sophisticated

modelling of the influences of the features of amino acids on the interaction. Therefore, the consideration of the more detailed descriptor variables can be expected to lead to an improved accuracy of prediction with the resulting models.

In order to determine the benefit of the different novel modelling approaches, the PLS regression is applied in several ways to the available data. The evaluation of the novel modelling aspects is predominantly based on the comparison of the prediction accuracies computed for the corresponding regression models because of the importance of obtaining accurate predictions in practice. The measure reported in general with respect to the prediction accuracy is not automatically calculated by the applied software and so was programmed.

The regression models determined to be optimal for the interaction under investigation are presented as well as further regression models being specified in order to obtain some additional information. Finally, the results of the established regression models are interpreted and used to draw novel and sophisticated biochemical conclusions concerning the interaction between the tobacco mosaic virus protein and the corresponding antibody 57P. These statements giving some first hints about functional domains in which the TMVP might be helpful in the case of occurrence of a mutant of the tobacco mosaic virus.

The dissertation comprises five chapters, an appendix and the bibliography. Following the introduction, multivariate multiple linear regression is explained. The next chapters deal with the methodology of PLS regression and its general application to the analysis of biomolecular interactions. Finally, the analysis of the available data relating to the interaction of the TMVP and the antibody 57P is presented. In the following, the contents of the different chapters are listed.

In the next chapter, multivariate multiple linear regression is briefly described, i.e. the general form of a regression model is given with various initial considerations of regression analysis. Estimation of the unknown model parameters by applying the method of Ordinary Least Squares is explained and it is shown how univariate linear regression can be derived as a special case.

The third chapter deals with the methodology of Partial Least Squares regression. In particular, the corresponding sections refer to the basic idea of this method, both the multivariate and the univariate PLS algorithm and predic-

tion intervals. With respect to multivariate PLS regression, diverse aspects regarding the performance of the algorithm are considered. In detail, the standardization of the descriptor and response variables, the representation of the extraction of the information inherent in the data by latent variable vectors in the form of decomposition models and the general realization of the multivariate PLS algorithm are presented. Subsequently, the computations performed in the course of the multivariate PLS algorithm are listed and explained in detail. Furthermore, the equivalence of the described multivariate algorithm to a reported alternative algorithm is shown. The formulae for the estimation of the standardized regression coefficients are given and used to derive the estimation formula for the original regression coefficients. A description of the computations required to obtain predictions of values of the original response variables is given and, finally, the determination of the optimal number of iterations that should be performed during the PLS algorithm is explained. In the section on univariate PLS regression, descriptions of the performance of the univariate PLS algorithm are provided by analogy with the multivariate case, where the differences between the multivariate and univariate PLS algorithms are explained. With respect to the prediction intervals, different approaches for approximate computations of approximate PLS regression prediction intervals are presented and evaluated.

The application of the PLS method to the investigation of biomolecular interactions is described in a general form in the fourth chapter. First, considerations of biochemical interactions are provided and a detailed description of surface plasmon resonance biosensors is given, consisting of aspects of the biochemical background, the components of a Biacore instrument, the course of the measurements, the mode of operation of the surface plasmon resonance and the output obtained by interaction experiments. Next, the novel unified multivariate regression model is introduced in detail, and the different subgroup models usually established in biomolecular interaction studies are described theoretically by giving the forms of the corresponding regression models.

The last chapter presenting the data analysis is introduced by a biochemical motivation of the subsequent investigation. After a detailed description of the data to be analyzed, the previous regression analysis relating to the interaction under study is summarized and reproduced, and the corresponding results are compared. Diverse novel approaches are motivated and explained with the objective of optimizing the modelling procedure. Further, the established regression models to be compared in terms of a particular novel modelling aspect are described. Subsequently, the prediction accuracies as

well as further measures of interest referring to the different regression models are given. Using these model descriptions, the novel procedures are evaluated to determine the optimal modelling approach. The optimal and other useful regression models are specified. Finally, sophisticated biochemical conclusions are presented on the basis of the developed regression models.

In the appendix, a number of additional tables is given. The literature referred to in the dissertation is listed in the bibliography.

Chapter 2

The multivariate multiple linear regression

2.1 General considerations

Multivariate multiple linear regression is a method that is applied to model the functional relationship between k response variables $y_1^o, \ldots, y_k^o$ and m descriptor variables $x_1^o, \ldots, x_m^o$. In subsequent presentations, the descriptor and response variables are used in a standardized form. For this reason, the original variables and terms referring to the original variables are marked accordingly by the superscript o in order to distinguish them from the standardized variables and their corresponding terms.

The response variables are presumed or known to be influenced by the descriptor variables. This dependence can be expressed mathematically by the following equation, the model for the regression of the l-th response variable y_l^o on m descriptor variables $x_1^o, \ldots, x_m^o$:

$$y_l^o = f_l \begin{pmatrix} x_1^o \\ \vdots \\ x_m^o \end{pmatrix} \quad \text{for} \quad l = 1, \ldots, k.$$

In this equation, the term f_l denotes a function of the descriptor variables that reflects their functional relationship to the l-th response variable. The objective of regression analysis is to specify the functions $f_1, \ldots, f_k$ referring to the k response variables. In linear regression the connection between the response and descriptor variables is considered to be linear and therefore the functions $f_1, \ldots, f_k$ are modelled accordingly.

In order to determine these functions, values for both the response variables and the descriptor variables are measured on a sample of n objects. Thus, the available data consist of $(m + k)n$ observations since the realizations of the response variables constitute kn values and mn measurements refer to the descriptor variables.

With the help of these data, the functional relationship between the response and descriptor variables is determined by establishing a regression model as explained in the next section. Then, not only can the measured data be interpreted but also values of the response variables can be predicted for combinations of values of the descriptor variables.

The special case of a single response variable, namely univariate multiple linear regression, is dealt with in subsection **??**.

2.2 The regression model

In the following, the regression model relating the response variables to the descriptor variables is presented in detail. This is realized by describing the general form of the function f_l in the multiple linear regression.

The dependence of the i-th observation y_{il} of the l-th response variable on the values of the m descriptor variables can be modelled as

$$y_{il}^o = b_{0l}^o + \sum_{j=1}^{m} b_{jl}^o x_{ij}^o + e_{il}^o, \tag{2.1}$$

$$\text{for} \quad i = 1, \ldots, n; \quad j = 1, \ldots, m \quad \text{and} \quad l = 1, \ldots, k.$$

In this equation, the term x_{ij}^o represents the i-th observation of the j-th descriptor variable, and the expression e_{il} denotes the error corresponding to the i-th observation of the l-th response variable. Further, the terms $b_{0l}^o, b_{1l}^o, \ldots, b_{ml}^o$ represent the unknown model parameters of the regression model of the original response and descriptor variables.

The regression coefficient b_{jl}^o relates to the j-th descriptor variable and the l-th response variable. It can be interpreted as the increase in value of the l-th response variable if all of the values of the descriptor variables were to be kept constant with the exception of the j-th descriptor variable which is increased by one unit. Beyond this, the constant term b_{0l}^o for the l-th response variable is that value that the l-th response variable would acquire, if all of the values

of the descriptor variables were to be set to zero. Hence, the parameter b^o_{0l} represents the intercept and the regression coefficients $b^o_{1l}, \ldots, b^o_{ml}$ are the parameters of the gradient of the regression equation for the l-th response variable.

The multivariate regression model, incorporating the observations of the k response and m descriptor variables from a sample of n objects, can be expressed in matrix notation as

$$Y^o = 1_n b^{o'}_{0.} + X^o B^o + E^o,$$

where the components of this model are of the following dimensions:

$$Y^o \sim n \times k, \quad X^o \sim n \times m, \quad 1_n \sim n \times 1, \quad b^o_{0.} \sim k \times 1, \quad B^o \sim m \times k,$$

$$\text{and} \quad E^o \sim n \times k.$$

Further comments on the notation are that the term 1_n is a column vector of length n with all elements equal to 1 and that the expression $b^{o'}_{0.}$ denotes the transposition of the vector $b^o_{0.}$.

The observed values of the k response variables on the n sample objects are summarized in the matrix Y^o. Each column of this matrix refers to one of the response variables. Consequently, the i-th row of the matrix Y^o consists of the measurements of the k response variables obtained on the i-th object of the sample.

The observations of the m descriptor variables are given in the matrix X^o, where the j-th column of this matrix contains the values of the j-th descriptor variable. Hence, the i-th row of the matrix X^o gives the measurements of the m descriptor variables of the i-th object of the sample.

All of the regression coefficients of the corresponding response and descriptor variables are included in the matrix B^o, whereas the column vector $b^o_{0.}$ comprises the intercepts relating to the k response variables. Further, the matrix E^o contains the errors.

The classical procedure for estimating the unknown intercepts and regression coefficients is the application of the method of Ordinary Least Squares (OLS) that is briefly explained in the following section.

2.3 Ordinary Least Squares estimation

The aim of regression analysis is to estimate the unknown model parameters, i.e. the k intercepts and the mk regression coefficients, with the help of the available data. These data are the matrix Y^o comprising the measurements of the response variables and the matrix X^o summarizing the observations of the descriptor variables.

If the rank of the matrix X^o, defined as the number of linear independent columns, equals the number m of descriptor variables, the matrix X^o has full rank and is called regular. In this case, the matrix $X^{o'}X^o$ is invertible, i.e. the inverse $(X^{o'}X^o)^{-1}$ of the matrix $X^{o'}X^o$ exists. Then, the unambiguously defined Ordinary Least Squares estimate $\hat{B}^o_{OLS}$ for the matrix B^o of regression coefficients can be obtained by calculating

$$\hat{B}^o_{OLS} = (X^{o'}X^o)^{-1}X^{o'}Y^o. \tag{2.2}$$

Using this result, the intercepts summarized in the column vector $b^o_{0.}$ can be estimated as

$$\hat{b}^o_{0.,OLS} = \bar{y}^o_{..} - \hat{B}^{o'}_{OLS}\bar{x}^o_{..},$$

where the terms $\bar{y}^o_{..}$ and $\bar{x}^o_{..}$ denote the column vectors of mean values of the response or descriptor variables, respectively, i.e.:

$$\bar{y}^o_{..} = \begin{pmatrix} \bar{y}^o_{.1} \\ \vdots \\ \bar{y}^o_{.k} \end{pmatrix} \sim k \times 1 \quad \text{and} \quad \bar{x}^o_{..} = \begin{pmatrix} \bar{x}^o_{.1} \\ \vdots \\ \bar{x}^o_{.m} \end{pmatrix} \sim m \times 1.$$

In detail, the means are calculated with the help of the values of the respective columns of the matrices X^o and Y^o, where the mean $\bar{x}^o_{.j}$ of the j-th descriptor variable and the mean $\bar{y}^o_{.l}$ of the l-th response variable, respectively, are obtained by computing:

$$\bar{x}^o_{.j} = \frac{1}{n}\sum_{i=1}^{n} x^o_{ij} \quad \forall j = 1, \ldots, m \quad \text{and}$$

$$\bar{y}^o_{.l} = \frac{1}{n}\sum_{i=1}^{n} y^o_{il} \quad \forall l = 1, \ldots, k.$$

2.4 Univariate linear regression as a special case

If only one response variable is taken into account, the multivariate regression model reduces to the univariate model that presents the functional re-

lationship between this single response variable and m descriptor variables. Therefore, univariate regression can be considered as a special case of multivariate regression.

The regression equation of the univariate case,

$$y^o = 1_n b_0^o + X^o b^o + e^o,$$

can be derived from the multivariate model by modifying some terms. In detail, the column vector $b_{0.}^o$ is substituted by the scalar b_0^o, and the matrix Y^o of response variables, the matrix B^o of regression coefficients and the matrix E^o of errors are replaced by the corresponding vectors y^o, b^o and e^o with the following dimensions:

$$y^o \sim n \times 1, \quad b^o \sim m \times 1 \quad \text{and} \quad e^o \sim n \times 1.$$

Just as in the multivariate case, Ordinary Least Squares estimates for the regression coefficients summarized in the vector b^o can be obtained by computing

$$\hat{b}_{OLS}^o = (X^{o\prime} X^o)^{-1} X^{o\prime} y^o,$$

presuming that the matrix X^o has full rank. Further, the estimate of the intercept b_0^o can be obtained by the following formula:

$$\hat{b}_{0,OLS}^o = \bar{y}^o - \hat{b}_{OLS}^{o\prime} \bar{x}_{..}^o.$$

In this equation, the scalar $\bar{y}^o$ represents the mean of the values of the single response variable, i.e.:

$$\bar{y}^o = \tfrac{1}{n} \sum_{i=1}^{n} y_i^o.$$

Instead of modelling the relationship between k response variables and m descriptor variables simultaneously by one multivariate regression, an alternative procedure is to perform k separate univariate regressions, one for each of the k individual response variables. However, both procedures lead to the same results.

Chapter 3

Methodology of Partial Least Squares regression

3.1 Basic idea of PLS regression

If the matrix X^o of descriptor variables is not of full rank, it is called singular and the matrix $X^{o'}X^o$ is not invertible. In this case of collinearity among the descriptor variables, problems concerning the application of the Ordinary Least Squares procedure arise. The reason is that the classical estimation of the regression coefficients summarized in the matrix B^o cannot be performed according to formula **??** with respect to a singular matrix X^o.

This problem of collinearity occurs frequently in practical applications, especially when the number m of descriptor variables exceeds the number n of objects in the sample. In this situation, the matrix X^o of descriptor variables consists of more columns than rows. Therefore, the columns of the matrix X^o cannot be linear independent and consequently, the matrix X^o is not of full rank.

In order to solve this problem and nevertheless obtain reliable estimates of the model parameters, it is necessary to reduce the dimensionality of the data used for the regression. Methods that could be applied to a dataset showing collinearity among the descriptor variables include the so-called ridge regression technique, principal component regression as well as stepwise selection of variables. Neither will be dealt with in the following.

Another available procedure is Partial Least Squares (PLS) regression, whose performance guarantees the effective compression of the information inherent

in the available data. An advantage of PLS regression is that the reduction of the amount of data is accompanied by a minimal loss of information. In the following, the implementation of this method will be explained in detail.

During the operation of the PLS regression algorithm, the observed data is used in a standardized form. For this reason, the original response and descriptor variables are mean-centered and variance-scaled.

The basic idea of the PLS methodology is to construct so-called latent variable vectors with the help of the standardized data summarized in the matrix X of standardized descriptor variables and the matrix Y of standardized response variables. The objective of the procedure is to derive only a few but relevant latent variable vectors that contain the information inherent in the matrix X of standardized descriptor variables in a compact form.

The latent variable vectors are computed as linear combinations of the standardized observations of the descriptor variables. They are extracted from the matrix X by decomposing it using the information of the standardized response variables. Since the construction of the latent variable vectors is influenced by the standardized measurements of the response variables, the effect of variations occurring in the descriptor variables is reduced if they show no relevance to variations in the response variables.

The idea behind the calculations of PLS regression can be explained with the help of decomposition models describing the extraction of the information of the standardized measured data by the successive extraction of the latent variable vectors. Both the matrix X of standardized descriptor variables and the matrix Y of standardized response variables can be expressed in terms of the latent variable vectors by decomposition models.

The latent variable vectors and further required components are calculated iteratively by performing an algorithm incorporating the standardized data. With the help of the terms computed in the course of the iterations of the algorithm, the unknown parameters of the regression model involving the standardized variables can be estimated. Subsequently, these estimates can be used to compute those of the regression coefficients and intercepts referring to the original variables.

In this way, the application of the PLS algorithm leads to the establishment of the originally required regression model in cases when the Ordinary Least Squares regression cannot be applied. Regression models developed by the

PLS method provide reliable predictions of values of the response variables because the information of the descriptor as well as the response variables is used in the computations. This is a property of special importance in practical applications.

The name of the PLS procedure is based on the fact that several "partial" regressions involving only one descriptor variable are obtained during the PLS algorithm. These descriptor variables are particular terms obtained in the course of the iterations of the algorithm. As the partial regression models show a reduced dimensionality compared with the original regression problem, the Ordinary Least Squares procedure is applied to them.

The following section deals with several important aspects of multivariate PLS regression. In the two subsequent sections, the univariate special case and the computation of prediction intervals are presented.

After subsection **??** describing the standardization of the data, the decomposition models of the matrices X and Y are given. The implementation of the multivariate PLS algorithm is explained in subsection **??** and the exact calculations, interpretations and relevant properties of all of the terms obtained during the performance of the multivariate PLS algorithm are presented in the following two subsections. Geladi et. al. (1986) gives an alternative but more complicated version of the multivariate PLS algorithm. Subsection **??** contains the derivation of this alternative presentation of the PLS algorithm and how it corresponds to the one below.

Subsection **??** describes in detail the estimation of the model parameters of the original regression model, followed by explanations of two alternative ways of obtaining predictions of the original response variables in subsection **??**.

The usefulness of a regression model results mainly from its ability to provide sufficiently correct predictions. Therefore, the determination of the optimal number of iterations is based on a criterion judging the accuracy of the predictions obtained by regression models resulting from different numbers of iterations. Subsection **??** deals with the choice of the number of iterations which is also referred to as the number of dimensions of the PLS regression.

3.2 The multivariate PLS algorithm

3.2.1 Scaling and centering

Before PLS regression can be applied to the dataset, both the original response variables and the original descriptor variables are variance-scaled and mean-centered. Then, the PLS algorithm is performed using these transformed data.

Though mean-centering is not necessary for the implementation of the PLS procedure, it is nevertheless conventionally performed. Mean-centering results in variables all having zero means. The original variables are mean-centered by calculating the mean of the values of each original variable, i.e. of each column of the matrices X^o and Y^o, and subtracting these from the corresponding original variables, i.e.:

$$x^o_{ij} - \bar{x}^o_{.j} \quad \forall i = 1, \ldots, n; \forall j = 1, \ldots, m \quad \text{and}$$

$$y^o_{il} - \bar{y}^o_{.l} \quad \forall i = 1, \ldots, n; \forall l = 1, \ldots, k.$$

The variance-scaling of the original variables is required since the PLS method is not invariant to scaling, i.e. the effect of a variable during the calculations of the PLS regression would be influenced by the value of its variance. Variance-scaling of the original variables results in variables all having variances of one.

To obtain variance-scaled variables, the standard deviation of the values of each original variable is calculated. These standard deviations are denoted $s^o_{x.j}$, $s^o_{y.l}$, respectively, and are computed as follows:

$$s^o_{x.j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x^o_{ij} - \bar{x}^o_{.j})^2} \quad \forall j = 1, \ldots, m \quad \text{and}$$

$$s^o_{y.l} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y^o_{il} - \bar{y}^o_{.l})^2} \quad \forall l = 1, \ldots, k.$$

Then the mean-centered values of each of the original variables are divided by the corresponding standard deviation, giving:

$$x_{ij} := \frac{x^o_{ij} - \bar{x}^o_{.j}}{s^o_{x.j}} \quad \forall i = 1, \ldots, n; \forall j = 1, \ldots, m \quad \text{and}$$

$$y_{il} := \frac{y^o_{il} - \bar{y}^o_{.l}}{s^o_{y.l}} \quad \forall i = 1, \ldots, n; \forall l = 1, \ldots, k.$$

In what follows, the observed responses and descriptors are used in their standardized form, i.e. mean-centered and variance-scaled.

3.2.2 The decomposition models

In the following, the decomposition models of the matrix X of standardized descriptor variables and the matrix Y of standardized response variables are given. These models describe the representation of the information of the standardized descriptor variables and the standardized response variables in terms of the latent variable vectors. Hence, the decomposition models contribute to the motivation as well as the understanding of the computations performed during the PLS algorithm.

Decomposition of the matrix X of standardized descriptor variables

The matrix X of standardized descriptor variables can be decomposed into a sum of weighted latent variable vectors and an additional residual matrix X_A as follows:

$$X = \sum_{a=1}^{A} t_{.a} p'_{.a} + X_A.$$

The number A represents the number of iterations of the PLS algorithm. This number has to be determined in each case for the respective application of the PLS regression (see subsection **??**).

In the decomposition model, the term $t_{.a}$ denotes the a-th latent variable vector that contains the so-called factor scores for the matrix X, i.e.:

$$t_{.a} = \begin{pmatrix} t_{1a} \\ \vdots \\ t_{na} \end{pmatrix} \sim n \times 1, \quad \text{where} \quad a = 1, \ldots, A.$$

The i-th scalar t_{ia} of the a-th latent variable vector is that score that corresponds to the i-th object of the sample.

The factor scores for the matrix X of standardized descriptor variables are summarized in the matrix T_A whose a-th column is determined by the a-th latent variable vector, i.e.:

$$T_A = (t_{.1}, \ldots, t_{.A}) \sim n \times A.$$

Further, the expression $p_{.a}$,

$$p_{.a} = \begin{pmatrix} p_{1a} \\ \vdots \\ p_{ma} \end{pmatrix} \sim m \times 1,$$

denotes the column vector of loadings for the matrix X of standardized descriptor variables. Each of the elements of the vector $p_{.a}$ refers to one of the m standardized descriptor variables. In particular, the j-th scalar p_{ja} is that loading of the a-th latent variable vector that belongs to the j-th standardized descriptor variable.

The loadings can be given in the matrix P_A, whose a-th column equals the a-th vector of loadings as follows:

$$P_A = (p_{.1}, \ldots, p_{.A}) \sim m \times A.$$

That part of the matrix X of standardized descriptor variables that remains unexplained by the linear combination of the latent variable vectors weighted by their corresponding loadings is summarized in the residual matrix X_A which is of the following form:

$$X_A = \begin{pmatrix} x_{11,A} & \ldots & x_{1m,A} \\ \vdots & \vdots & \vdots \\ x_{n1,A} & \ldots & x_{nm,A} \end{pmatrix} \sim n \times m.$$

Using matrix notation, the decomposition of the matrix X can be expressed as well as

$$X = T_A P_A' + X_A.$$

The decomposition model of the matrix X shows that each observation of each standardized descriptor variable can be represented in terms of scores, loadings and a residual scalar. In particular, the i-th observation x_{ij} of the j-th standardized descriptor variable can be written as the following expression:

$$x_{ij} = \sum_{a=1}^{A} t_{ia} p_{ja} + x_{ij,A},$$

i.e. as the sum of the corresponding residual term $x_{ij,A}$ and the linear combination of the i-th scores of the A latent variable vectors that are weighted by the j-th values of the A loading vectors.

Decomposition of the matrix Y of standardized response variables

By analogy with the decomposition of the matrix X of standardized descriptor variables, the matrix Y of standardized response variables can be presented as a sum of the latent variable vectors multiplied by corresponding weights and an additional residual matrix Y_A, i.e.:

$$Y = \sum_{a=1}^{A} t_{.a} q'_{.a} + Y_A. \tag{3.1}$$

In this decomposition of the matrix Y of standardized response variables, the column vector $q_{.a}$ of loadings,

$$q_{.a} = \begin{pmatrix} q_{1a} \\ \vdots \\ q_{ka} \end{pmatrix} \sim k \times 1,$$

contains the weights of the a-th latent variable vector. In detail, the l-th scalar q_{la} denotes the loading respecting the a-th latent variable vector and the l-th standardized response variable.

The loading vectors are summarized as columns in the matrix Q_A that can thus be written as

$$Q_A = (q_{.1}, \ldots, q_{.A}) \sim k \times A.$$

The residual matrix Y_A,

$$Y_A = \begin{pmatrix} y_{11,A} & \cdots & y_{1k,A} \\ \vdots & \vdots & \vdots \\ y_{n1,A} & \cdots & y_{nk,A} \end{pmatrix} \sim n \times k,$$

represents that part of the matrix Y of standardized response variables that is not explained by the latent variable vectors and the loading vectors.

With the help of these introduced expressions, the decomposition of the matrix Y can be expressed in matrix notation as

$$Y = T_A Q'_A + Y_A.$$

Incorporating the scores of the latent variable vectors, their corresponding loadings and the residual scalar $y_{il,A}$, the i-th observation y_{il} of the l-th standardized response variable can be given as

$$y_{il} = \sum_{a=1}^{A} t_{ia} q_{la} + y_{il,A}.$$

3.2.3 Performance of the multivariate PLS algorithm

Several variants of the PLS algorithm exist that differ mainly from each other in the normalizations of the computed terms. In the following, the multivariate PLS algorithm is presented essentially according to Martens et. al. (1989) who describe the NIPALS-algorithm (Non-Iterative PArtial Least Squares-algorithm) developed by Svante Wold in 1983. Complementary explanations are derived from Höskuldsson (1988), Manne (1987), Helland (1988), Geladi et. al. (1986) and Gustafsson (2001).

The performance of the PLS algorithm begins with initializing settings and ends after each iteration with an increase of the index number a of the iterations.

A number of steps are performed successively in every iteration of the PLS algorithm. Thus, each completed iteration leads to the computation of several terms. In the a-th iteration of the algorithm, the following components are obtained: the weight vector $\hat{w}_{.a}$, the latent variable vector $\hat{t}_{.a}$, the loading vector $\hat{p}_{.a}$ for the matrix X of standardized descriptor variables, the loading vector $\hat{q}_{.a}$ relating to the matrix Y of standardized response variables and the sequential latent variable vector $\hat{u}_{.a}$. These terms are denoted as estimates since they can be shown to be equal or proportional to Ordinary Least Squares estimates of corresponding partial regression models, as explained in detail in subsection **??**. The interpretation as OLS estimates is especially important in relation to the loading vectors $\hat{p}_{.a}$ and $\hat{q}_{.a}$ referring to the standardized data matrices X and Y.

Further terms that are calculated in the a-th iteration of the algorithm are the residual matrix X_a corresponding to the standardized descriptor variables and the residual matrix Y_a related to the standardized response variables. These residual matrices represent those parts of the standardized data that have not yet been expressed with the help of the components obtained in the earlier iterations of the algorithm. Their calculation and interpretation results directly from the corresponding decomposition models.

Within the a-th iteration of the multivariate PLS algorithm, several subiterations are performed in which the sequential latent variable vectors $\hat{u}^1_{.a}, \hat{u}^2_{.a}, \ldots$, the latent variable vectors $\hat{t}^1_{.a}, \hat{t}^2_{.a}, \ldots$, the weight vectors $\hat{w}^1_{.a}, \hat{w}^2_{.a}, \ldots$ and the loading vectors $\hat{q}^1_{.a}, \hat{q}^2_{.a}, \ldots$ referring to the matrix Y of standardized response variables are generated iteratively until convergence. Therefore, in the a-th iteration, the elements of the latent variable vector $\hat{t}^h_{.a}$ calculated in the h-th

subiteration are compared with those of the latent variable vector $\hat{t}_{.a}^{h-1}$ obtained in the previous subiteration. If the elements of these vectors $\hat{t}_{.a}^{h}$ and $\hat{t}_{.a}^{h-1}$ are approximately the same, convergence is achieved. Then, the vectors $\hat{w}_{.a}^{h}, \hat{t}_{.a}^{h}, \hat{q}_{.a}^{h}$ and $\hat{u}_{.a}^{h}$ computed in this final subiteration are denoted $\hat{w}_{.a}, \hat{t}_{.a}, \hat{q}_{.a}$ and $\hat{u}_{.a}$, respectively, and are used in the following calculations of the a-th iteration. This means that they are incorporated in the computations leading to the loading vector $\hat{p}_{.a}$ respecting the matrix X of standardized descriptor variables as well as the residual matrices X_a and Y_a.

3.2.4 The computations of the multivariate algorithm

Set $X_0 : X, \quad Y_0 : Y, \quad a : 1; \quad h : 1, \quad \hat{u}^1_{.a}$: any column of Y_{a-1}

$$
\begin{aligned}
\tilde{w}^h_{.a} &= X'_{a-1}\hat{u}^h_{.a} \\
\hat{w}^h_{.a} &= \frac{X'_{a-1}\hat{u}^h_{.a}}{\sqrt{\hat{u}^{h\prime}_{.a} X_{a-1} X'_{a-1}\hat{u}^h_{.a}}} = \frac{\tilde{w}^h_{.a}}{||\tilde{w}^h_{.a}||} \\
\hat{t}^h_{.a} &= X_{a-1}\hat{w}^h_{.a} \\
\hat{q}^h_{.a} &= \frac{Y'_{a-1}\hat{t}^h_{.a}}{\hat{t}^{h\prime}_{.a}\hat{t}^h_{.a}} = \frac{Y'_{a-1}\hat{t}^h_{.a}}{||\hat{t}^h_{.a}||^2}
\end{aligned}
$$

check for convergence

$\rightarrow$ no convergence

$$
\hat{u}^{h+1}_{.a} = \frac{Y_{a-1}\hat{q}^h_{.a}}{\hat{q}^{h\prime}_{.a}\hat{q}^h_{.a}} = \frac{Y_{a-1}\hat{q}^h_{.a}}{||\hat{q}^h_{.a}||^2}
$$

Set $h : h+1$

$\rightarrow$ convergence

$$
\begin{aligned}
\hat{w}^h_{.a} &:= \hat{w}_{.a} \\
\hat{t}^h_{.a} &:= \hat{t}_{.a} \\
\hat{q}^h_{.a} &:= \hat{q}_{.a} \\
\hat{u}^h_{.a} &:= \hat{u}_{.a}
\end{aligned}
$$

$$
\begin{aligned}
\hat{p}_{.a} &= \frac{X'_{a-1}\hat{t}_{.a}}{\hat{t}'_{.a}\hat{t}_{.a}} = \frac{X'_{a-1}\hat{t}_{.a}}{||\hat{t}_{.a}||^2} \\
X_a &= X_{a-1} - \hat{t}_{.a}\hat{p}'_{.a} \\
Y_a &= Y_{a-1} - \hat{t}_{.a}\hat{q}'_{.a}
\end{aligned}
$$

Set $a : a+1$

3.2.5 Considerations concerning the components of the algorithm

In this subsection, important explanations concerning the individual terms obtained by the multivariate algorithm are given to provide a comprehensive motivation for the computations performed in the algorithm. In particular, the operation of the reduction of the dimension of the data by computing latent variable vectors is presented in detail.

The calculations related to the weight vectors are described in the context of a maximization problem respecting a particular covariance. The need to perform subiterations in each iteration is explained in connection with the computation of the sequential latent variable vectors. The calculation of the latent variable vectors is described with reference to the influence of the standardized response variables. The extraction of information of the standardized data matrices by the latent variable vectors is explained by presenting the decompositions of the matrices X and Y. These decompositions are performed by calculating corresponding residual matrices.

Further, the different partial regression models on which the computations of the PLS algorithm are based are given. Their interpretation is of special relevance in what concerns the loading vectors for the matrices X and Y. In these regression models referring to the calculation of the respective loading vectors, the latent variables take on the role of the descriptor variables. Therefore, these regression models are required to quantify the extent of information of the standardized response and descriptor variables, respectively, that can be explained by the corresponding latent variable vectors. This determination is obtained by calculating the loading vectors representing the unknown model parameters.

The weight vector

In the h-th subiteration of the a-th iteration, the weight vector $\tilde{w}_{.a}^h$ being of the form

$$\tilde{w}_{.a}^h = \begin{pmatrix} \tilde{w}_{1a}^h \\ \vdots \\ \tilde{w}_{ma}^h \end{pmatrix} \sim m \times 1,$$

is obtained. Its elements are calculated as linear combinations of the values of the residual matrix X_{a-1} computed in the previous iteration. These values are weighted by the elements of the corresponding sequential latent variable vector. In particular, the j-th element $\tilde{w}_{ja}^h$ of the weight vector $\tilde{w}_{.a}^h$

is computed as follows:

$$\tilde{w}_{ja}^{h} = \sum_{i=1}^{n} x_{ij,a-1} \hat{u}_{ia}^{h} \quad \text{with} \quad j = 1, \ldots, m.$$

In this formula, the expression $x_{ij,a-1}$ presents that value of the residual matrix X_{a-1} obtained in the $(a-1)$-th iteration that is related to the i-th observation of the j-th standardized descriptor variable. Further, the term $\hat{u}_{ia}^{h}$ denotes that value of the sequential latent variable vector computed in the h-th subiteration of the a-th iteration that refers to the i-th object. In the first iteration, the linear combination described above involves the values of the standardized descriptor variables instead of those of the corresponding residual matrix.

In the h-th subiteration of the a-th iteration, the weight vector $\tilde{w}_{.a}^{h}$ is divided by its norm $||\tilde{w}_{.a}^{h}||$, resulting in the scaled weight vector $\hat{w}_{.a}^{h}$ having unit length.

With respect to the following explanations, it is important to note that in the context of PLS regression, the elements of a column vector can be considered as realizations of a particular random variable.

In detail, the j-th column vector $x_{.j}^{o}$ of the matrix X^{o} contains the n realizations of the j-th descriptor variable x_j. Accordingly, the j-th column vector $x_{.j}$ comprises the n elements of the j-th standardized descriptor variable. Further, the j-th column vector $x_{.j,a}$ of the residual matrix X_a summarizes the n elements of the a-th residual variable $x_{j,a}$ of the j-th standardized descriptor variable.

The interpretations of the response variables can be described analogously to those of the descriptor variables. In particular, the l-th column vector $y_{.l}^{o}$ consists of the n realizations of the l-th response variable y_l. The corresponding standardized realizations are summarized in the l-th column vector $y_{.l}$, and the l-th column vector $y_{.l,a}$ of the residual matrix Y_a contains the realizations of the a-th residual variable $y_{l,a}$ of the l-th standardized response variable.

The n elements of the latent variable vector $\hat{t}_{.a}^{h}$ and the sequential latent variable vector $\hat{u}_{.a}^{h}$ can be considered as realizations of random variables, in fact of the a-th latent variable t_a^h and the a-th sequential latent variable u_a^h, respectively. The sequential latent variable vector $\hat{u}_{.a}^{h}$ represents the information of the standardized response variables as explained in more detail in

the corresponding subsubsection.

With the help of these considerations, the statement that the j-th element $\tilde{w}_{ja}^h$ of the weight vector $\tilde{w}_{.a}^h$ is proportional to the empirical covariance between the j-th residual variable $x_{j,a-1}$ and the sequential latent variable u_a^h can be derived. This empirical covariance is defined as

$$\hat{cov}_{x_{j,a-1},u_a^h} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij,a-1} - \bar{x}_{.j,a-1})(\hat{u}_{ia}^h - \bar{\hat{u}}_{.a}^h).$$

According to Geladi et. al. (1986), the sequential latent variables are centered around zero, i.e.:

$$\sum_{i=1}^{n}\hat{u}_{ia}^h = 0 \quad \forall a = 1,\ldots,A$$

and hence, the corresponding mean $\bar{\hat{u}}_{.a}^h$ is zero as well. Moreover, the mean $\bar{x}_{.j,a-1}$ of the $(a-1)$-th residual variable $x_{j,a-1}$ of the j-th standardized descriptor variable is zero. This fact can be shown by expressing this mean as follows:

$$\begin{aligned}\bar{x}_{.j,a-1} &= \frac{1}{n}\sum_{i=1}^{n}x_{ij,a-1} = \frac{1}{n}\sum_{i=1}^{n}x_{ij,a-2} - \hat{t}_{i(a-1)}^h\hat{p}_{j(a-1)} \\ &= \frac{1}{n}\sum_{i=1}^{n}x_{ij,a-2} - \frac{1}{n}\hat{p}_{j(a-1)}\sum_{i=1}^{n}\hat{t}_{i(a-1)}^h = \bar{x}_{.j,a-2}.\end{aligned}$$

In this reformulation, the centering of the latent variables around zero is used (see Geladi et. al. (1986)), i.e.:

$$\sum_{i=1}^{n}\hat{t}_{ia}^h = 0 \quad \forall a = 1,\ldots,A.$$

In the first iteration, the mean $\bar{x}_{.j,0} = \bar{x}_{.j}$ equals zero because of the mean-centering of the original descriptor variables. Further, in the second iteration, the mean $\bar{x}_{.j,1}$ equals this mean $\bar{x}_{.j,0}$ which is zero. In subsequent iterations, the mean $\bar{x}_{.j,a}$ is zero since it coincides with the mean $\bar{x}_{.j,a-1}$ of the mean-centered residual variable $x_{j,a-1}$ obtained in the previous iteration. Consequently, the residual variables referring to each of the descriptor variables are mean-centered in each of the A iterations and thus, the following equation is valid:

$$\bar{x}_{.j,a} = 0 \quad \forall j = 1,\ldots,m \quad \text{and} \quad a = 1,\ldots,A.$$

Using these results, the proportionality of the j-th element $\tilde{w}^h_{ja}$ of the weight vector $\tilde{w}^h_{.a}$ and accordingly, of the j-th element $\hat{w}^h_{ja}$ of the normalized weight vector $\hat{w}^h_{.a}$, to the empirical covariance between the j-th residual variable $x_{j,a-1}$ and the sequential latent variable u^h_a can easily be shown as follows:

$$\begin{aligned} \hat{cov}_{x_{j,a-1},u^h_a} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij,a-1} - \bar{x}_{.j,a-1})(\hat{u}^h_{ia} - \bar{\hat{u}}^h_{.a}) = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij,a-1}\hat{u}^h_{ia} \\ &= \frac{1}{n-1}x'_{.j,a-1}\hat{u}^h_{.a} = \frac{1}{n-1}\tilde{w}^h_{ja}. \end{aligned}$$

As explained below, the j-th element $\hat{w}^h_{ja}$ of the weight vector $\hat{w}^h_{.a}$ is used to weight the residual values of the j-th standardized descriptor variable in the computation of the latent variable vector $\hat{t}^h_{.a}$. Therefore, the residual values of those descriptor variables exhibiting a high covariance with the sequential latent variable u^h_a receive a corresponding high weight in the calculation of each element of the latent variable vector $\hat{t}^h_{.a}$.

Since the residual values of the standardized descriptor variables contribute to the computation of the elements of the latent variable vector $\hat{t}^h_{.a}$ to an extent proportional to their empirical covariance with the sequential latent variable u^h_a, the latent variable t^h_a shows a high covariance with the sequential latent variable u^h_a. Consequently, the computation of the elements of the weight vector $\hat{w}^h_{.a}$ leads to a maximization of the covariance between the latent variable t^h_a and the sequential latent variable u^h_a.

The weight vector $\tilde{w}^h_{.a}$ is proportional to the Ordinary Least Squares solution of the regression of the residual matrix X_{a-1} on the sequential variable vector $\hat{u}^h_{.a}$, i.e. of the partial regression model

$$X_{a-1} = \hat{u}^h_{.a}w'_{.a} + E_{w,a}.$$

The proportionality is proved easily by converting the OLS estimator $\hat{w}^h_{.a,OLS}$ of this regression equation as follows into a term involving the weight vector $\tilde{w}^h_{.a}$:

$$\hat{w}^h_{.a,OLS} = \frac{X'_{a-1}\hat{u}^h_{.a}}{\hat{u}^{h\prime}_{.a}\hat{u}^h_{.a}} = \frac{\tilde{w}^h_{.a}}{\hat{u}^{h\prime}_{.a}\hat{u}^h_{.a}}.$$

The A weight vectors obtained in those subiterations of the A iterations for which convergence could be ascertained are summarized in the matrix $\hat{W}_A$, whose a-th column gives the a-th weight vector $\hat{w}_{.a}$, i.e.:

$$\hat{W}_A = (\hat{w}_{.1}, \ldots, \hat{w}_{.A}) \sim m \times A.$$

A notable property of the weight vectors computed in the course of the PLS algorithm is their mutual orthonormality. This characteristic can be expressed mathematically as:

$$\hat{w}'_{.a}\hat{w}_{.a^*} = \begin{cases} 1 & a = a^* \\ 0 & a \neq a^*, \quad \text{where} \quad a, a^* \in \{1, \ldots, A\}. \end{cases}$$

The orthonormality of the weight vectors results in the fact that $\hat{W}'_A\hat{W}_A$ is the identity matrix. Hence, the following equation is valid:

$$\hat{W}'_A\hat{W}_A = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \sim A \times A.$$

The latent variable vector

The latent variable vector $\hat{t}^h_{.a}$ obtained in the h-th subiteration of the a-th iteration can be presented in form of the following column vector:

$$\hat{t}^h_{.a} = \begin{pmatrix} \hat{t}^h_{1a} \\ \vdots \\ \hat{t}^h_{na} \end{pmatrix} \sim n \times 1.$$

The i-th element $\hat{t}^h_{ia}$ of this vector is calculated as

$$\hat{t}^h_{ia} = \sum_{j=1}^{m} \hat{w}^h_{ja} x_{ij,a-1} = x_{i.,a-1}\hat{w}^h_{.a} \quad \text{with} \quad i = 1, \ldots, n.$$

This equation illustrates that the i-th element $\hat{t}^h_{ia}$ of the latent variable vector $\hat{t}^h_{.a}$ is given as the linear combination of the i-th observations of the residuals of the standardized descriptor variables. These residual scalars are multiplied by the corresponding elements of the weight vector $\hat{w}^h_{.a}$. In the first iteration the values of the standardized descriptor variables are used instead of the residual terms in the calculation of the latent variable vector $\hat{t}^h_{.1}$.

Obviously, the weight vectors that are derived with respect to the sequential latent variable vectors representing the measurements of the standardized response variables determine the computation of the latent variable vectors. This fact reflects the important aspect of the PLS methodology that the construction of the latent variable vectors and consequently, the decomposition of the matrix X, is influenced by the information inherent in the standardized response variables.

The calculation of the latent variable vectors can be considered as the dimensionality reduction of the matrix X of standardized descriptor variables. The reason is that the latent variable vector $\hat{t}^h_{.a}$ consists of the scalars into which the values of the residuals of the standardized descriptor variables are projected. In detail, the i-th row of the residual matrix X_{a-1} is projected into the i-th element of the latent variable vector $\hat{t}^h_{.a}$ in the corresponding subiteration of the a-th iteration.

As already explained, the elements of the weight vectors determine the contribution of the different standardized descriptor variables to the construction of the corresponding latent variable vectors. Therefore, comparing the elements of the weight vectors of each iteration of the algorithm permits statements concerning the importance of each of the standardized descriptor variables during the dimensionality reduction. The larger the absolute value of an element of a weight vector is, the more does the corresponding standardized descriptor variable contribute to the projection of the residual matrix into the latent variable vector computed in the respective iteration.

To illustrate this kind of examination, the elements of the weight vectors can be plotted against the descriptor variables for each iteration. Further, the elements of two weight vectors calculated in different iterations can be plotted against each other to investigate whether the significance of a particular descriptor variable is consistent among the iterations.

An interesting though not very important fact is that the latent variable vector $\hat{t}^h_{.a}$ calculated in the h-th subiteration of the a-th iteration equals the Ordinary Least Squares estimation of the regression of the residual matrix X_{a-1} on the transposed weight vector $\hat{w}^h_{.a}$, i.e. of the partial regression model

$$X_{a-1} = t^h_{.a}\hat{w}^{h'}_{.a} + E_{t,a}.$$

The proof of this correspondence is performed easily by taking into account that the weight vector $\hat{w}^h_{.a}$ has unit length since

$$\hat{t}^h_{.a,OLS} = \frac{X_{a-1}\hat{w}^h_{.a}}{||\hat{w}^h_{.a}||^2}.$$

The A latent variable vectors that are obtained in the iterations of the multivariate PLS algorithm can be summarized as columns in the matrix $\hat{T}_A$ that is hence defined as

$$\hat{T}_A = (\hat{t}_{.1}, \ldots, \hat{t}_{.A}) \sim n \times A.$$

Since the latent variable vectors are orthogonal to each other, i.e.:

$$\hat{t}'_{.a}\hat{t}_{.a^*} = 0 \quad \text{for} \quad a \neq a^*, \quad \text{where} \quad a, a^* \in \{1, \ldots, A\},$$

the matrix product $\hat{T}'_A\hat{T}_A$ can be presented as the following diagonal matrix:

$$\hat{T}'_A\hat{T}_A = \begin{pmatrix} \hat{t}'_{.1}\hat{t}_{.1} & & 0 \\ & \ddots & \\ 0 & & \hat{t}'_{.A}\hat{t}_{.A} \end{pmatrix} \sim A \times A.$$

The sequential latent variable vector

The initializing sequential latent variable vector $\hat{u}^1_{.a}$ used in the first subiteration of the a-th iteration is randomly chosen as one of the columns of the residual matrix Y_{a-1} of the previous iteration. Thus, this sequential latent variable vector $\hat{u}^1_{.a}$ represents the observations concerning one standardized response variable in the first iteration or its corresponding residual in subsequent iterations, respectively.

In the h-th subiteration of the a-th iteration, the sequential latent variable vector being of the form $\hat{u}^{h+1}_{.a}$,

$$\hat{u}^{h+1}_{.a} = \begin{pmatrix} \hat{u}^{h+1}_{1a} \\ \vdots \\ \hat{u}^{h+1}_{na} \end{pmatrix} \sim n \times 1$$

is calculated. The i-th element of this column vector is a weighted sum of the measurements of the i-th object of the k standardized response variables in the first iteration or their residuals in subsequent iterations, respectively. The weights are the corresponding elements of the loading vector $\hat{q}^h_{.a}$ obtained in the h-th subiteration of the a-th iteration that are standardized by division by the squared norm of the respective loading vector. Hence, the i-th element $\hat{u}^{h+1}_{ia}$ of the sequential latent variable vector $\hat{u}^{h+1}_{.a}$ can be computed as

$$\hat{u}^{h+1}_{ia} = \frac{1}{||\hat{q}^h_{.a}||^2} \sum_{l=1}^{k} \hat{q}^h_{la} y_{il,a-1} = \frac{1}{||\hat{q}^h_{.a}||^2} y_{i.,a-1}\hat{q}^h_{.a} \quad \text{with} \quad i = 1, \ldots, n.$$

This formula reflects the fact that in the a-th iteration, each row of the residual matrix Y_{a-1} is projected into a single scalar. These n projections of the residuals of the standardized response variables concerning a particular object are summarized in the sequential latent variable vector $\hat{u}^{h+1}_{.a}$.

Consequently, the computation of the sequential latent variable vectors leads to the dimensionality reduction in the standardized response variables. This dimensionality reduction is required for the calculation of the weight vectors that are relevant in the construction of the latent variable vectors.

The projection of the standardized response variables is performed by incorporating corresponding loading vectors. In contrast to the weight vectors used in the projection of the standardized descriptor variables, these loading vectors are not obtained on the basis of an optimization problem. Therefore, different projections of the standardized response variables are obtained successively until stable results for the projections of the standardized descriptor variables are obtained. That is the reason why the subiterations are necessary in each iteration until convergence of the latent variable vectors is reached.

In the derivation of the latent variable vectors, the sequential latent variable vectors can be expressed as Ordinary Least Squares estimates. In particular, the sequential latent variable vector $\hat{u}_{.a}^{h+1}$ calculated in the h-th subiteration of the a-th iteration corresponds to the Ordinary Least Squares solution of the partial regression model

$$Y_{a-1} = u_{.a}^{h+1}\hat{q}_{.a}^{h'} + E_{u,a},$$

i.e. of the regression of the residual matrix Y_{a-1} on the transposed loading vector $\hat{q}_{.a}^{h}$ of the h-th subiteration of the a-th iteration.

The matrix $\hat{U}_A$ summarizes the A sequential latent variable vectors obtained in the course of the iterations of the algorithm as follows:

$$\hat{U}_A = (\hat{u}_{.1}, \ldots, \hat{u}_{.A}) \sim n \times A.$$

The vector of loadings for the matrix of standardized response variables

In the h-th subiteration of the a-th iteration, the vector $\hat{q}_{.a}^{h}$ of loadings concerning the matrix Y of standardized response variables is obtained. This vector has the form

$$\hat{q}_{.a}^{h} = \begin{pmatrix} \hat{q}_{1a}^{h} \\ \vdots \\ \hat{q}_{ka}^{h} \end{pmatrix} \sim k \times 1,$$

and consists of elements giving linear combinations of the values of the columns of the residual matrix Y_{a-1}. The weights involved in this linear combination are standardized elements of the latent variable vector $\hat{t}_{.a}^{h}$, where the

standardization is performed by division by the squared norm of the respective latent variable vector. Thus, the l-th element $\hat{q}_{la}^h$ of the loading vector $\hat{q}_{.a}^h$ is computed as follows:

$$\hat{q}_{la}^h = \frac{1}{||\hat{t}_{.a}^h||^2} \sum_{i=1}^{n} \hat{t}_{ia}^h y_{il,a-1} = \frac{1}{||\hat{t}_{.a}^h||^2} y'_{.l,a-1} \hat{t}_{.a}^h \quad \text{with} \quad l = 1, \ldots, k.$$

This formula shows that the l-th column $y_{.l,a-1}$ of the residual matrix Y_{a-1} is projected into the l-th element of the loading vector $\hat{q}_{.a}^h$.

The interpretation of the loading vector $\hat{q}_{.a}^h$ is based on the fact that it is the Ordinary Least Squares solution of the regression of the residual matrix Y_{a-1} on the latent variable vector $\hat{t}_{.a}^h$ of the h-th subiteration of the a-th iteration, i.e. of the partial regression model:

$$Y_{a-1} = \hat{t}_{.a}^h q_{.a}^{h'} + E_{q,a}. \tag{3.2}$$

Therefore, the l-th element of the loading vector $\hat{q}_{.a}^h$ gives the OLS estimates of the regression of the l-th column of the residual matrix Y_{a-1} on the latent variable vector $\hat{t}_{.a}^h$, i.e. of the regression equation

$$y_{.l,a-1} = \hat{t}_{.a}^h q_{la}^{h'} + e_{.l_{q,a}}.$$

Consequently, the l-th element $\hat{q}_{la}^h$ of the loading vector $\hat{q}_{.a}^h$ represents the relationship between the l-th column of the residual matrix Y_{a-1} and the latent variable vector $\hat{t}_{.a}^h$. In other words, it describes to what extent the residual of the l-th standardized response variable is explained by the latent variable vector computed in the h-th subiteration of the a-th iteration.

The A loading vectors of the matrix Y of standardized response variables are comprised in the matrix $\hat{Q}_A$ that can hence be given as

$$\hat{Q}_A = (\hat{q}_{.1}, \ldots, \hat{q}_{.A}) \sim k \times A.$$

The vector of loadings for the matrix of standardized descriptor variables

The explanations of the vector $\hat{p}_{.a}$ of loadings for the matrix X can be given following those for the vector $\hat{q}_{.a}^h$ of loadings. In fact, these loading vectors have the same interpretation but refer to different data matrices. However, the loading vectors respecting the standardized descriptor variables are not computed within the subiterations as they are calculated without regard to the sequential latent variable vectors.

The elements of the vector $\hat{p}_{.a}$ of loadings,

$$\hat{p}_{.a} = \begin{pmatrix} \hat{p}_{1a} \\ \vdots \\ \hat{p}_{ma} \end{pmatrix} \sim m \times 1,$$

obtained in the a-th iteration, are calculated as standardized linear combinations of the values of the columns of the residual matrix X_{a-1}. These residuals of the standardized descriptor variables are weighted by the respective elements of the a-th latent variable vector $\hat{t}_{.a}$. In particular, the j-th element $\hat{p}_{ja}$ of the a-th loading vector is computed as a standardized weighted sum of the values of the j-th column of the residual matrix X_{a-1}. This weighted sum is standardized by dividing it by the squared norm of the a-th latent variable vector, i.e.:

$$\hat{p}_{ja} = \frac{1}{||\hat{t}_{.a}||^2} \sum_{i=1}^{n} x_{ij,a-1} \hat{t}_{ia} = \frac{1}{||\hat{t}_{.a}||^2} x'_{.j,a-1} \hat{t}_{.a} \quad \text{with} \quad j = 1, \ldots, m.$$

Therefore, the a-th loading vector $\hat{p}_{.a}$ of the standardized descriptor variables can be interpreted as a projection of the residual matrix X_{a-1} into a column vector, where each column of this matrix is projected into a scalar.

By analogy with the loading vector $\hat{q}^h_{.a}$, the vector $\hat{p}_{.a}$ of loadings corresponds to the Ordinary Least Squares solution of the regression of the residual matrix X_{a-1} on the a-th latent variable vector $\hat{t}_{.a}$, i.e. of the partial regression model

$$X_{a-1} = \hat{t}_{.a} p'_{.a} + E_{p,a}. \tag{3.3}$$

Accordingly, the j-th element $\hat{p}_{ja}$ of the a-th loading vector represents the estimated regression coefficient in the regression of the j-th column $x_{.j,a-1}$ of the residual matrix X_{a-1} on the a-th latent variable vector, i.e. of the model

$$x_{.j,a-1} = \hat{t}_{.a} p_{ja} + e_{.j_{p,a}}.$$

This partial regression model reflects the fact that the j-th element $\hat{p}_{ja}$ of the a-th loading vector $\hat{p}_{.a}$ quantifies the effect of the a-th latent variable vector $\hat{t}_{.a}$ with respect to the explanation of the corresponding residual of the j-th standardized descriptor variable.

The matrix $\hat{P}_A$ summarizes the A loading vectors referring to the standardized descriptor variables and is hence of the following form:

$$\hat{P}_A = (\hat{p}_{.1}, \ldots, \hat{p}_{.A}) \sim m \times A.$$

The residual matrix of the standardized descriptor variables

In the a-th iteration, the $(n \times m)$-dimensional residual matrix X_a of the matrix X of standardized descriptor variables is obtained by subtracting the product of the a-th latent variable vector $\hat{t}_{.a}$ and the transposed a-th loading vector $\hat{p}_{.a}$ from the residual matrix X_{a-1} computed in the $(a-1)$-th iteration. This product can be interpreted as the estimated effect of the a-th latent variable vector $\hat{t}_{.a}$ on the residual matrix X_{a-1} calculated in the former iteration. The interpretation results from the partial regression model **??**, i.e. the regression of the residual matrix X_{a-1} on the a-th latent variable vector $\hat{t}_{.a}$, since the corresponding Ordinary Least Squares solution is the a-th loading vector $\hat{p}_{.a}$.

Thus, by subtraction of this effect from the residual matrix X_{a-1}, that part of the matrix X that is not yet explained after the performance of the a-th iteration of the algorithm remains in form of the residual matrix X_a. Consequently, the residual matrix X_a of the a-th iteration can also be written as

$$X_a = X - \sum_{a^*=1}^{a} \hat{t}_{.a^*} \hat{p}'_{.a^*}.$$

In other words, the residual matrix X_a represents that share of the matrix X that still has to be expressed in terms of the latent variable vectors in subsequent iterations of the algorithm in order to explain the remaining information.

In this way, the matrix X is decomposed successively into the latent variable vectors in the course of the algorithm. Therefore, the computations of the residual matrices result in a gradual exhaustion of the information inherent in the matrix X of standardized descriptor variables.

After A iterations of the PLS algorithm, the matrix X of standardized descriptor variables can be expressed as

$$X = \sum_{a=1}^{A} \hat{t}_{.a} \hat{p}'_{.a} + X_A,$$

where the sum

$$\sum_{a=1}^{A} \hat{t}_{.a} \hat{p}'_{.a}$$

denotes that part of the matrix X that could be explained with the help of the latent variable vectors calculated during the course of the algorithm.

Further, the residual matrix X_A comprises that share of the matrix X that remains unexplained after A iterations.

The total variance inherent in the standardized descriptor variables can be computed as $tr[X'X]$, since the columns of the matrix X are centered around zero. The expression $tr[X]$ denotes the trace, i.e. the sum of the diagonal elements, of the matrix X. The percentage $pctvar_{descr_A}$ of the total variance of the standardized descriptor variables that can be explained after completing the PLS algorithm is obtained as

$$pctvar_{descr_A} = 100 \left(1 - \frac{tr[X'_A X_A]}{tr[X'X]}\right)$$

because the columns of the residual matrix X_A are also centered around zero. Consequently, the percentage of the total variance of the standardized descriptor variables that is not accounted for after the performance of A iterations of the PLS algorithm can be calculated as

$$100 - pctvar_{descr_A} = \frac{100 tr[X'_A X_A]}{tr[X'X]}.$$

The computation of the explained percentage of the total variance of the standardized descriptor variables can be used to describe not only a characteristic of the final regression model, but also to follow the process of extraction of the information inherent in the data. Accordingly, after each iteration, the percentage of the total variance that is already accounted for can be determined. In particular, the percentage $pctvar_{descr_a}$ of the total variance of the standardized descriptor variables that is explained after a iterations can be calculated with the help of the residual matrix X_a as follows:

$$pctvar_{descr_a} = 100 \left(1 - \frac{tr[X'_a X_a]}{tr[X'X]}\right).$$

In this computation, the fact that the columns of the residual matrices obtained in each iteration of the PLS algorithm are centered around zero is used.

The residual matrix of the standardized response variables

The calculation and interpretation of the $(n \times k)$-dimensional residual matrix Y_a of the standardized response variables computed in the a-th iteration of the algorithm can be presented by analogy with the considerations concerning the residual matrix X_a of the standardized descriptor variables.

The residual matrix Y_a of the standardized response variables is calculated by subtracting that part of the residual matrix Y_{a-1} obtained in the previous iteration that is explained by the a-th latent variable vector. This share is given as the product of the a-th latent variable vector and the corresponding transposed a-th loading vector since this product represents the extent to which the a-th latent variable vector contributes to the explanation of the residual matrix Y_{a-1}. The interpretation can be derived from the partial regression model **??**, i.e. the regression of the residual matrix Y_{a-1} on the a-th latent variable vector, that provides the a-th loading vector $\hat{q}_{.a}$ as Ordinary Least Squares estimation.

Thus, the residual matrix Y_a contains that part of the matrix Y that results after eliminating that share of the matrix Y that is already explained in terms of latent variable vectors after the a-th iteration. Consequently, it can be presented as

$$Y_a = Y - \sum_{a^*=1}^{a} \hat{t}_{.a^*} \hat{q}'_{.a^*}.$$

This means that after the performance of a iterations, the information in the residual matrix Y_a remains to be expressed by the latent variable vectors computed in further iterations. Hence, these calculations of the residual matrices lead to a successive decomposition of the matrix Y of standardized response variables into latent variable vectors.

In the derivatation of the standardized descriptor variables, the decomposition of the matrix Y of standardized response variables can be expressed as

$$Y = \sum_{a=1}^{A} \hat{t}_{.a} \hat{q}'_{.a} + Y_A$$

after the performance of A iterations. In this equation, the unexplained share of the matrix Y is given in form of the residual matrix Y_A. Further, that part of the matrix Y that is expressed in terms of the latent variable vectors obtained in the A iterations is represented by the sum

$$\sum_{a=1}^{A} \hat{t}_{.a} \hat{q}'_{.a}.$$

By analogy with the computations related to the standardized descriptor variables, the percentage $pctvar_{resp_A}$ of the total variance of the standardized response variables that is accounted for after the performance of A iterations

can be obtained by computing

$$pctvar_{resp_A} = 100 \left(1 - \frac{tr[Y_A' Y_A]}{tr[Y'Y]}\right).$$

Further, the percentage of the total variance that cannot be explained after A iterations of the PLS algorithm can be determined as:

$$100 - pctvar_{resp_A} = \frac{100 tr[Y_A' Y_A]}{tr[Y'Y]}.$$

The percentage $pctvar_{resp_a}$ of the total variance explained after the performance of a iterations can be obtained by calculating:

$$pctvar_{resp_a} = 100 \left(1 - \frac{tr[Y_a' Y_a]}{tr[Y'Y]}\right).$$

According to the computations referring to the descriptor variables, these calculations use the fact that the columns of the matrix Y as well as the columns of the residual matrices obtained in each of the iterations are centered around zero.

3.2.6 An alternative presentation of the multivariate PLS algorithm

In Geladi et. al. (1986), an alternative presentation of the multivariate PLS algorithm is given that differs from that one described in subsection **??** in what concerns the decomposition model regarding the matrix Y of standardized response variables. Further, a normalization of the vector $\hat{q}_{.a}^h$ of loadings referring to the standardized response variables is performed during the subiterations. Accordingly, the vector $\hat{q}_{.a}$ of loadings used after the subiterations is also normalized.

The alternative decomposition model involves the sequential latent variable vectors instead of the latent variables. Assuming this decomposition model results in the need to establish a so-called inner relation between the latent variables and the sequential latent variable vectors in order to derive a decomposition model of the matrix Y in terms of the latent variable vectors. This decomposition model is called the mixed relation.

However, in the following, it is shown that this mixed relation equals the previously explained decomposition model of the matrix Y of standardized response variables presented in equation **??**. This proof is performed at first

by incorporating the unnormalized vector $q_{.a}$ of loadings. The following statements may clarify this unnecessarily complicated version of the multivariate PLS algorithm.

The alternative decomposition of the matrix Y of standardized response variables is presented in terms of the sequential latent variable vectors as

$$Y = \sum_{a=1}^{A} u_{.a} q'_{.a} + Y_A,$$

which can also be expressed in matrix notation:

$$Y = UQ' + Y_A.$$

The inner relation between the decomposition models of the matrix X of standardized descriptor variables and the matrix Y of standardized response variables can be represented by the regression of the a-th sequential latent variable vector on the a-th latent variable vector, i.e. by the equation

$$u_{.a} = g_a t_{.a} + e_{.a}.$$

Because of the establishment of this inner relation, the Ordinary Least Squares estimation of the regression coefficient g_a of the inner relation, i.e. the calculation

$$\hat{g}_a = \frac{\hat{u}'_{.a}\hat{t}_{.a}}{||\hat{t}_{.a}||^2},$$

is performed additionally during the computations of the algorithm presented by Geladi et. al. (1986).

Using the inner relation in the context of the alternative decomposition model of the matrix Y of standardized response variables leads to the so-called mixed relation. This model is obtained by substituting the sequential latent variable vectors of the alternative decomposition model by the vector $g_a t_{.a}$ of the inner relation. Accordingly, the mixed relation represents the dependence of the matrix Y on the latent variable vectors and can be described by the following formula:

$$Y = \sum_{a=1}^{A} g_a t_{.a} q'_{.a} + Y_A.$$

In matrix notation, the mixed relation can be expressed as well as

$$Y = T_A G_A Q'_A + Y_A,$$

where the matrix G_A is a diagonal matrix containing the A regression coefficients of the inner relation on its diagonal, i.e.:

$$G_A = \begin{pmatrix} g_1 & & 0 \\ & \ddots & \\ 0 & & g_A \end{pmatrix} \sim A \times A.$$

Consequently, in the a-th iteration of the algorithm from Geldai et. al. (1986), the residual matrix Y_a of the standardized response variables is computed using the mixed relation as follows:

$$Y_a = Y_{a-1} - \hat{g}_a \hat{t}_{.a} \hat{q}'_{.a}.$$

However, the estimated regression coefficient of the inner relation attains the value 1 in every iteration as can easily be proved as follows:

$$\hat{g}_a = \frac{\hat{u}'_{.a}\hat{t}_{.a}}{||\hat{t}_{.a}||^2} = \frac{\hat{q}'_{.a}Y'_{a-1}\hat{t}_{.a}}{||\hat{q}_{.a}||^2||\hat{t}_{.a}||^2} = \frac{\hat{q}'_{.a}\hat{q}_{.a}}{||\hat{q}_{.a}||^2} = 1.$$

Therefore, it is shown that the mixed relation can be converted into the previously presented decomposition model **??** of the matrix Y of standardized response variables. Further, the residual matrix defined by Geladi et. al. (1986) equals the residual matrix Y_A obtained in the algorithm described in subsection **??** since the expression $\hat{g}_a \hat{t}_{.a} \hat{q}'_{.a}$ coincides with the matrix $\hat{t}_{.a}\hat{q}'_{.a}$.

In order to normalize the vector $\hat{q}^h_{.a}$ of loadings, the vector $\hat{\tilde{q}}^h_{.a}$ having unit length is calculated as

$$\hat{\tilde{q}}^h_{.a} = \frac{\hat{q}^h_{.a}}{||\hat{q}^h_{.a}||}.$$

Correspondingly, the vector $\hat{\tilde{q}}_{.a}$ is also normalized. If the sequential latent variable vectors are calculated with the help of the normalized vector of loadings, the estimation of the regression coefficient of the inner relation is obtained as

$$\hat{g}_a = \frac{\hat{u}'_{.a}\hat{t}_{.a}}{||\hat{t}_{.a}||^2} = \frac{\hat{\tilde{q}}'_{.a}Y'_{a-1}\hat{t}_{.a}}{||\hat{t}_{.a}||^2} = \frac{\hat{q}'_{.a}\hat{q}_{.a}}{||\hat{q}_{.a}||} = \frac{||\hat{q}_{.a}||^2}{||\hat{q}_{.a}||} = ||\hat{q}_{.a}||.$$

In this case, the mixed relation incorporating the normalized vector of loadings also equals the decomposition model introduced previously. This fact is easily proved since the term $\hat{g}_a \hat{t}_{.a} \hat{\tilde{q}}'_{.a}$ can be shown as follows to correspond to the matrix $t_{.a}\hat{q}'_{.a}$:

$$\hat{g}_a \hat{t}_{.a} \hat{\tilde{q}}'_{.a} = ||\hat{q}_{.a}||\hat{t}_{.a}\frac{\hat{q}'_{.a}}{||\hat{q}_{.a}||} = \hat{t}_{.a}\hat{q}'_{.a}.$$

Accordingly, the resulting residual matrix coincides with that used in the algorithm presented in subsection **??**.

3.2.7 Derivation of the estimation of the model parameters for the original regression model

In the following, the formula for estimating the regression coefficients referring to the standardized variables is given. Further, the derivation and computation of this formula is explained in detail, in fact in a comprehensiveness that cannot be found so far in the PLS literature.

On the basis of this formula, the estimation of the regression coefficients with the original variables is derived. This relationship between the estimation formulae of the regression coefficients of the standardized and original variables and thus the expression of the estimated original regression coefficients in terms of the standardized regression coefficients, is not given in any of the publications listed in the references. Therefore, the following presentations can be considered as important contributions to the completeness of the statistical description of the PLS regression.

The regression model involving the standardized response and descriptor variables is of the following form:

$$Y = 1_n b'_{0.} + XB + E.$$

The estimates of the intercepts that are summarized in the vector $\hat{b}_{0.}$,

$$\hat{b}_{0.} = \bar{y}_{..} - \hat{B}'\bar{x}_{..},$$

are zero since the means of the standardized response and descriptor variables being the elements of the vectors $\bar{y}_{..}$ and $\bar{x}_{..}$ are zero. As a result of this, the intercepts can be omitted from the presentation leading to the simplified model

$$Y = XB + E.$$

After the application of the PLS algorithm to the standardized measured data, the unknown regression coefficients of the regression model involving the standardized response and descriptor variables can be obtained by calculating the following formula:

$$\hat{B}_{PLS} = \hat{W}_A(\hat{P}'_A\hat{W}_A)^{-1}\hat{Q}'_A. \tag{3.4}$$

In this equation, the term $\hat{W}_A$ denotes the matrix containing the weight vectors, the expression $\hat{P}'_A$ represents the transposed matrix summarizing the loading vectors referring to the standardized descriptor variables, and the matrix $\hat{Q}'_A$ is the transposed matrix comprising the loading vectors for the standardized response variables.

For the derivation of this formula, the following relationship is used in order to find an expression for the matrix $\hat{T}_A$ of latent variable vectors respecting the matrix X of standardized descriptor variables:

$$\begin{aligned} X\hat{W}_A &= \hat{T}_A\hat{P}'_A\hat{W}_A \\ \Leftrightarrow \hat{T}_A &= X\hat{W}_A(\hat{P}'_A\hat{W}_A)^{-1}. \end{aligned} \tag{3.5}$$

The assumption for relationship **??** is that the matrix X is entirely expressed in terms of latent variable vectors, i.e. it can be written as:

$$X = \hat{T}_A\hat{P}'_A.$$

However, this condition cannot be considered as a severe restriction of the validity of the derivation because the sufficiently exact representation of the matrix X by the matrix product $\hat{T}_A\hat{P}'_A$ can be achieved by incorporating sufficient latent variable vectors. Then, the residual matrix X_A resulting after the performance of A iterations can be neglected.

The element $[(\hat{P}'_A\hat{W}_A)^{-1}]_{aa^*}$ in the a-th row and a^*-th column of the inverse of the $(A \times A)$-dimensional matrix $\hat{P}'_A\hat{W}_A$ is calculated as:

$$[(\hat{P}'_A\hat{W}_A)^{-1}]_{aa^*} = \frac{\delta_{aa^*} - \sum\limits_{z<a^*}[(\hat{P}'_A\hat{W}_A)^{-1}]_{az}[\hat{P}'_A\hat{W}_A]_{za^*}}{[\hat{P}'_A\hat{W}_A]_{a^*a^*}} \quad \text{with} \quad a, a^* \in \{1, \ldots A\},$$

where the expression δ_{aa^*} denotes Kronecker's delta.

The elements of the matrix $(\hat{P}'_A\hat{W}_A)^{-1}$ can be computed in this way since the matrix $\hat{P}'_A\hat{W}_A$ is upper bidiagonal, i.e.: the elements of this matrix equal zero except for the elements whose index a for the row equals the index a^* or $a^* - 1$, respectively, for the column, and the elements on the diagonal have the value 1 (see Manne (1987) and Helland (1988)).

As it is shown below, formula **??** can be proved easily by taking into account expression **??** for the matrix $\hat{T}_A$ of latent variable vectors and the decomposition model of the standardized response variables. The term $\hat{Y}$ used in the derivation denotes the matrix containing the estimated values of the standardized response variables.

$$\hat{Y} = \hat{T}_A\hat{Q}'_A = X\hat{W}_A(\hat{P}'_A\hat{W}_A)^{-1}\hat{Q}'_A = X\hat{B}_{PLS}$$

$$\Rightarrow \hat{B}_{PLS} = \hat{W}_A(\hat{P}'_A\hat{W}_A)^{-1}\hat{Q}'_A \quad \square$$

The estimated regression coefficients of the matrix $\hat{B}_{PLS}$ refer to the standardized, i.e. the variance-scaled and mean-centered data. Consequently, it is necessary to reverse these transformations accordingly in order to obtain regression coefficients with respect to the original response and descriptor variables. Subsequently, these estimates of the regression coefficients corresponding to the original variables can be used to compute the intercepts of the original regression model.

The matrix $\hat{B}^o_{PLS}$ gives the estimates of the regression coefficients belonging to the original response and descriptor variables. These estimates can be obtained from the estimates of the regression coefficients for the standardized variables by calculating

$$\hat{B}^o_{PLS} = \begin{pmatrix} \frac{1}{s_{x^o_{\cdot 1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{x^o_{\cdot m}}} \end{pmatrix} \hat{B}_{PLS} \begin{pmatrix} s_{y^o_{\cdot 1}} & & 0 \\ & \ddots & \\ 0 & & s_{y^o_{\cdot k}} \end{pmatrix}. \tag{3.6}$$

In particular, the regression coefficient b^o_{jl} referring to the j-th original descriptor variable and the l-th original response variable can be estimated as

$$\hat{b}^o_{jl} = \frac{\hat{b}_{jl} s_{y^o_{\cdot l}}}{s_{x^o_{\cdot j}}},$$

using the estimate $\hat{b}_{jl}$ of the regression coefficient for the corresponding standardized variables. This relationship between the estimated regression coefficients concerning the original and standardized variables can be derived as follows by transforming the estimated regression model incorporating the original observations into the estimated regression model involving the standardized data.

$$
\begin{aligned}
& \hat{Y}^o = 1_n \hat{b}^{o'}_{0.,PLS} + X^o \hat{B}^o_{PLS} \\
\Leftrightarrow \quad & \hat{Y}^o = 1_n \bar{y}^{o'}_{..} + (X^o - 1_n \bar{x}^{o'}_{..}) \hat{B}^o_{PLS} \\
\Leftrightarrow \quad & \hat{Y} = (X^o - 1_n \bar{x}^{o'}_{..}) \hat{B}^o_{PLS} \begin{pmatrix} \frac{1}{s_{y^o_{.1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{y^o_{.k}}} \end{pmatrix} \\
\Leftrightarrow \quad & \hat{Y} = (X^o - 1_n \bar{x}^{o'}_{..}) \begin{pmatrix} \frac{s_{x^o_{.1}}}{s_{x^o_{.1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{s_{x^o_{.m}}}{s_{x^o_{.m}}} \end{pmatrix} \hat{B}^o_{PLS} \begin{pmatrix} \frac{1}{s_{y^o_{.1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{y^o_{.k}}} \end{pmatrix} \\
\Leftrightarrow \quad & \hat{Y} = X \begin{pmatrix} s_{x^o_{.1}} & & 0 \\ & \ddots & \\ 0 & & s_{x^o_{.m}} \end{pmatrix} \hat{B}^o_{PLS} \begin{pmatrix} \frac{1}{s_{y^o_{.1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{y^o_{.k}}} \end{pmatrix} \\
\Leftrightarrow \quad & \hat{Y} = X \hat{B}_{PLS} \\
\Rightarrow \quad \hat{B}_{PLS} = & \begin{pmatrix} s_{x^o_{.1}} & & 0 \\ & \ddots & \\ 0 & & s_{x^o_{.m}} \end{pmatrix} \hat{B}^o_{PLS} \begin{pmatrix} \frac{1}{s_{y^o_{.1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{y^o_{.k}}} \end{pmatrix} \\
\Leftrightarrow \quad \hat{B}^o_{PLS} = & \begin{pmatrix} \frac{1}{s_{x^o_{.1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{x^o_{.m}}} \end{pmatrix} \hat{B}_{PLS} \begin{pmatrix} s_{y^o_{.1}} & & 0 \\ & \ddots & \\ 0 & & s_{y^o_{.k}} \end{pmatrix} \quad \square
\end{aligned}
$$

With the help of the estimated regression coefficients referring to the original variables, the intercepts of the original regression model can be obtained by computing

$$\hat{b}^o_{0.,PLS} = \bar{y}^o_{..} - \hat{B}^{o'}_{PLS} \bar{x}^o_{..}. \tag{3.7}$$

Therefore, the regression model involving the original response and descriptor variables can be established as follows:

$$\begin{aligned}
\hat{Y}^o &= 1_n \hat{b}^{o\prime}_{0.,PLS} + X^o \hat{B}^o_{PLS} \\
&= 1_n \bar{y}^{o\prime}_{\cdot\cdot} + (X^o - 1_n \bar{x}^{o\prime}_{\cdot\cdot}) \hat{B}^o_{PLS} \\
&= 1_n \bar{y}^{o\prime}_{\cdot\cdot} + (X^o - 1_n \bar{x}^{o\prime}_{\cdot\cdot}) \begin{pmatrix} \frac{1}{s_{x^o_{\cdot 1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{x^o_{\cdot m}}} \end{pmatrix} \\
&\quad \hat{B}_{PLS} \begin{pmatrix} s_{y^o_{\cdot 1}} & & 0 \\ & \ddots & \\ 0 & & s_{y^o_{\cdot k}} \end{pmatrix}.
\end{aligned}$$

In this way, the application of the PLS methodology permits the calculation of the model parameters related to the original descriptor and response variables. Hence, PLS regression provides a procedure to specify regression models in cases where the requirements of the Ordinary Least Squares method cannot be met.

3.2.8 Predictions of the original response variables

The main objective of the application of PLS regression is to obtain accurate predictions. There are two possibilities for computing predictions of the original response variables corresponding to combinations of values of the original descriptor variables that have not been used during the algorithm performed in order to establish a regression model. These two alternative prediction procedures result in the same predictions as is shown below.

The prediction formulae described in the following differ from those given in the publications in the references. The reason is that the presentations refer to predictions of the original response variables that are consequently based on the novel estimation formula introduced in the previous subsection.

The values of the original descriptor variables relating to the r objects for which the values of the original response variables are to be predicted are summarized in the matrix

$$\tilde{X}^o = \begin{pmatrix} \tilde{x}^o_{11} & \dots & \tilde{x}^o_{1m} \\ \vdots & \vdots & \vdots \\ \tilde{x}^o_{r1} & \dots & \tilde{x}^o_{rm} \end{pmatrix} \sim r \times m.$$

For the application of one of the alternative prediction procedures, these values of the descriptor variables need to be standardized. Therefore, the observations of the original descriptor variables are mean-centered and variance-scaled with those means and standard deviations from the original descriptor variables used in the PLS algorithm. This normalization leads to the matrix $\tilde{X}$ containing the values of the standardized descriptor variables, i.e.:

$$\tilde{X} = \begin{pmatrix} \tilde{x}_{11} & \dots & \tilde{x}_{1m} \\ \vdots & \vdots & \vdots \\ \tilde{x}_{r1} & \dots & \tilde{x}_{rm} \end{pmatrix} \sim r \times m.$$

The variance-scaling can be obtained in this way though it does not result in variables with variance one because in the corresponding prediction procedure, the variances of the variables do not affect the computations as becomes clear in the following.

Further, the matrix $\tilde{Y}^o$ being of the form

$$\tilde{Y}^o = \begin{pmatrix} \tilde{y}^o_{11} & \dots & \tilde{y}^o_{1k} \\ \vdots & \vdots & \vdots \\ \tilde{y}^o_{r1} & \dots & \tilde{y}^o_{rk} \end{pmatrix} \sim r \times k$$

denotes that matrix that comprises the unknown values of the original response variables that have to be predicted. Thus, these values refer to the r objects whose values of the original descriptor variables are summarized in the matrix $\tilde{X}^o$.

The predictions of these unknown values of the original response variables are summarized in the matrix $\hat{\tilde{Y}}^o$:

$$\hat{\tilde{Y}}^o = \begin{pmatrix} \hat{\tilde{y}}^o_{11} & \dots & \hat{\tilde{y}}^o_{1k} \\ \vdots & \vdots & \vdots \\ \hat{\tilde{y}}^o_{r1} & \dots & \hat{\tilde{y}}^o_{rk} \end{pmatrix} \sim r \times k.$$

The unknown values of the original response variables can be predicted with the help of the established original regression equation involving the estimates of the model parameters referring to the original variables.

Alternatively, the predictions can be calculated by applying two steps of the PLS algorithm to the standardized values of the descriptor variables for which the predictions are required. This procedure incorporates some

terms computed in the iterations of the algorithm performed to establish a regression model. The objective of the reduced PLS algorithm is to obtain a decomposition of the matrix $\tilde{X}$ similar to that of the matrix X involved in the modelling process. Then, the predictions are calculated using the latent variable vectors resulting from the application of the reduced PLS algorithm.

Predictions using the model parameters estimated by the PLS procedure

After applying the PLS algorithm, the regression model incorporating the original variables can be developed using the estimates of the intercepts and regression coefficients obtained from the formulae **??** and **??**. Consequently, by applying this regression equation, the values of the k original response variables corresponding to the r objects whose values of the original descriptor variables are summarized in the matrix $\tilde{X}^o$ can be predicted as follows:

$$\begin{aligned} \hat{\tilde{Y}}^o &= 1_r \hat{b}^{o'}_{0.,PLS} + \tilde{X}^o \hat{B}^o_{PLS} \\ &= 1_r \bar{y}^{o'}_{..} + (\tilde{X}^o - 1_r \bar{x}^{o'}_{..}) \hat{B}^o_{PLS}. \end{aligned} \tag{3.8}$$

Predictions by applying computations of the PLS algorithm

The prediction of values of the k original response variables can also be obtained by incorporating latent variable vectors that are derived with respect to the matrix $\tilde{X}$. These latent variable vectors are obtained by decomposing the matrix $\tilde{X}$ in the same way as the matrix X of standardized descriptor variables is decomposed in the PLS algorithm performed during the process of modelling the observed data.

In detail, the a-th latent variable vector $\hat{\tilde{t}}_{.a}$ of the matrix $\tilde{X}$ is computed using the a-th weight vector $\hat{w}_{.a}$ obtained previously during the algorithm which derived the regression model. This is the reason why the descriptor variables of the matrix $\tilde{X}$ do not have to be scaled to variance one since no new weight vectors that would have been sensitive to the variances of the descriptor variables are constructed. Further, the a-th loading vector $\hat{p}_{.a}$, already calculated in the course of the modelling, is incorporated in the computation of the a-th residual matrix $\tilde{X}_a$ of the matrix $\tilde{X}$.

Specifically, the decomposition of the matrix $\tilde{X}$ is performed by repeating, after initializing settings, the following two steps of the PLS algorithm A times:

$$\begin{array}{rcl} \text{Set} \quad \tilde{X}_0 & : & \tilde{X}, \quad a:1 \end{array}$$

$$\begin{array}{rcl} \hat{\tilde{t}}_{.a} & = & \tilde{X}_{a-1}\hat{w}_{.a} \\ \tilde{X}_a & = & \tilde{X}_{a-1} - \hat{\tilde{t}}_{.a}\hat{p}'_{.a} \end{array}$$

$$\begin{array}{rcl} \text{Set} \quad a & : & a+1 \end{array}$$

The number A equals the number of iterations obtained previously in the course of the PLS algorithm applied to the standardized measurements of the matrix X.

The resulting A latent variable vectors that are used to decompose the matrix $\tilde{X}$ can be summarized in the matrix $\hat{\tilde{T}}_A$ that can hence be presented as follows:

$$\hat{\tilde{T}}_A = (\hat{\tilde{t}}_{.1}, \ldots, \hat{\tilde{t}}_A) \sim r \times A.$$

After the decomposition of the matrix $\tilde{X}$ in terms of latent variable vectors, the values of the k original response variables corresponding to the r objects are predicted as follows using these latent variable vectors and the A loading vectors for the matrix Y of standardized response variables calculated earlier:

$$\begin{array}{rcl} \hat{\tilde{Y}}^o = & 1_r \bar{y}^{o\prime}_{..} + \begin{pmatrix} s_{y^o_{.1}} & & 0 \\ & \ddots & \\ 0 & & s_{y^o_{.k}} \end{pmatrix} \sum_{a=1}^{A} \hat{\tilde{t}}_{.a}\hat{q}'_{.a} \\ = & 1_r \bar{y}^{o\prime}_{..} + \begin{pmatrix} s_{y^o_{.1}} & & 0 \\ & \ddots & \\ 0 & & s_{y^o_{.k}} \end{pmatrix} \hat{\tilde{T}}_A \hat{Q}'_A. \end{array}$$

This prediction formula can be derived from the alternative prediction formula **??** using the expression **??** for the matrix $\hat{B}^o_{PLS}$ of estimated original regression coefficients in terms of the matrix $\hat{B}_{PLS}$ of estimated regression coefficients of the standardized variables. It is proved below that the predictions obtained by the two prediction procedures are identical.

$$
\begin{aligned}
\hat{\tilde{Y}}^o &= 1_r \bar{y}_{\cdot\cdot}^{o\prime} + (\tilde{X}^o - 1_r \bar{x}_{\cdot\cdot}^{o\prime}) \hat{B}_{PLS}^o \\
\Leftrightarrow \hat{\tilde{Y}}^o &= 1_r \bar{y}_{\cdot\cdot}^{o\prime} + (\tilde{X}^o - 1_r \bar{x}_{\cdot\cdot}^{o\prime}) \begin{pmatrix} \frac{s_{x_{\cdot 1}^o}}{s_{x_{\cdot 1}^o}} & & 0 \\ & \ddots & \\ 0 & & \frac{s_{x_{\cdot m}^o}}{s_{x_{\cdot m}^o}} \end{pmatrix} \hat{B}_{PLS}^o \\
\Leftrightarrow \hat{\tilde{Y}}^o &= 1_r \bar{y}_{\cdot\cdot}^{o\prime} + \tilde{X} \begin{pmatrix} s_{x_{\cdot 1}^o} & & 0 \\ & \ddots & \\ 0 & & s_{x_{\cdot m}^o} \end{pmatrix} \hat{B}_{PLS}^o \begin{pmatrix} \frac{s_{y_{\cdot 1}^o}}{s_{y_{\cdot 1}^o}} & & 0 \\ & \ddots & \\ 0 & & \frac{s_{y_{\cdot k}^o}}{s_{y_{\cdot k}^o}} \end{pmatrix} \\
\Leftrightarrow \hat{\tilde{Y}}^o &= 1_r \bar{y}_{\cdot\cdot}^{o\prime} + \tilde{X} \hat{B}_{PLS} \begin{pmatrix} s_{y_{\cdot 1}^o} & & 0 \\ & \ddots & \\ 0 & & s_{y_{\cdot k}^o} \end{pmatrix} \\
\Leftrightarrow \hat{\tilde{Y}}^o &= 1_r \bar{y}_{\cdot\cdot}^{o\prime} + \tilde{X} \hat{W}_A (\hat{P}_A' \hat{W}_A)^{-1} \hat{Q}_A' \begin{pmatrix} s_{y_{\cdot 1}^o} & & 0 \\ & \ddots & \\ 0 & & s_{y_{\cdot k}^o} \end{pmatrix} \\
\Leftrightarrow \hat{\tilde{Y}}^o &= 1_r \bar{y}_{\cdot\cdot}^{o\prime} + \hat{\tilde{T}}_A \hat{Q}_A' \begin{pmatrix} s_{y_{\cdot 1}^o} & & 0 \\ & \ddots & \\ 0 & & s_{y_{\cdot k}^o} \end{pmatrix}.
\end{aligned}
$$

3.2.9 Determination of the optimal model complexity

In practical applications, the most important characteristic of an established regression model is its ability to provide reliable predictions of values of the original response variables for combinations of values of the original descriptor variables. Because of the way that the computations of the PLS method incorporate the information of the response variables, a regression model obtained by applying the PLS algorithm can be expected to lead to sufficiently accurate predictions.

The prediction accuracy of a regression model achieved after the performance of the PLS algorithm depends on the number A of iterations since the estimates of the model parameters specifying the regression equation are based on terms obtained in these iterations. Consequently, the accuracy of the predictions is influenced by the number of latent variable vectors that are

used to decompose the data. The number of latent variable vectors incorporated in the computations leading to the establishment of the model, or equivalently the number of iterations performed, represents the complexity of the regression model. To determine the optimal model complexity requires finding the regression model which provides the most exact predictions. So the prediction accuracy of models derived on the basis of different numbers of iterations has to be investigated.

One way to achieve this is to assess the prediction accuracy of the respective resulting models after each iteration. If the prediction accuracy of the model established after the $(a+1)$-th iteration is considerably better than that after the a-th iteration, the iteration is continued. Otherwise, the latent variable vector computed in the $(a + 1)$-th iteration is assumed not to contribute relevantly to the representation of the observed data. Then, the regression model from a iterations can be considered to be the one providing the most correct predictions, indicating that this number a of iterations is the optimal value of A.

Alternatively, the PLS algorithm could be run for a predefined number A_* of iterations. Afterwards, the accuracy of the predictions of the models obtained at each step is investigated simultaneously to find the optimal model which has the best prediction accuracy.

If there is a model that is obtained after fewer than A_* iterations that has a prediction accuracy only slightly different from the optimal, the significance of this difference in prediction accuracy should be investigated. Finally, that number a of iterations is determined to be the required number that results in that regression model being the least complex one showing an insignificant difference in accuracy from the optimal regression model.

The final regression model should be derived from as few latent variable vectors as possible. The reason is that those latent variable vectors calculated in the first few iterations can be expected to contain the most relevant information in the data, whereas those obtained in subsequent iterations are assumed to reflect mainly the noise in the measurements. Additionally, a relatively simple model is preferable because of the resulting interpretability.

In practice, the assessment of prediction accuracy is generally performed using leave-one-out cross-validation. To obtain a more reliable measure of the prediction accuracy, the blind cross-validation, a combination of leave-one-out and random cross validation, can be applied. Depending on the available

data, other validation methods could be chosen instead, e.g. the blocked or the split-sample validation. These will both not be dealt with in the following since they are rarely used in the context of the PLS regression.

During the leave-one-out cross-validation calculations n different regression equations are obtained at each stage, i.e. for each level of complexity corresponding to the a of PLS iterations. The i-th of these n models uses the $n-1$ observations in the sample after dropping the i-th, yielding the i-th regression equation which is used to predict the k responses for the i-th object and these are compared with the observed responses for that case by calculating the residuals.

The residual for the i-th object and the l-th original response variable is defined as the difference between the i-th observed value y_{il}^o of the l-th original response variable and the corresponding predicted value $\hat{y}_{il,CV_a}^o$. This prediction is based on the results of a iterations applied to the reduced dataset.

Omitting successively one object in turn from the sample used for specifying the regression model and predicting the values of the original response variables results in nk residual terms. The sum of squared residual terms gives the prediction residual sum of squares ($PRESS_a$) for a given complexity a, i.e.:

$$PRESS_a = \sum_{i=1}^{n}\sum_{l=1}^{k}(y_{il}^o - \hat{y}_{il,CV_a}^o)^2.$$

Large values of the $PRESS_a$-measure indicate a high prediction inaccuracy. Therefore, the number a of iterations for which the $PRESS_a$-value is smallest is chosen to be the optimal one that can be considered to result in a regression model providing reliable predictions.

In practice, the determination of the optimal model complexity and the description of the prediction accuracy of the final regression model is conventionally not based on the $PRESS_a$-measure but on the Q_a^2-statistic. The reason is that the Q_a^2-statistic provides a standardized measure in contrast to the $PRESS_a$-statistic. The Q_a^2-value for a iterations is essentially a scaled version of the $PRESS_a$-value and is computed as follows:

$$Q_a^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{l=1}^{k}(y_{il}^o - \hat{y}_{il,CV_a}^o)^2}{\sum_{i=1}^{n}\sum_{l=1}^{k}(y_{il}^o - \bar{y}_{.l}^o)^2}.$$

The Q_a^2-value can be interpreted as an estimate of the fraction of the variance of the response variables that could be explained by the model established on the basis of a iterations of the PLS algorithm. To illustrate the results of the validation analysis, the values of the Q_a^2-statistic can be plotted against the corresponding number of iterations, i.e. the complexity of the model.

If the predicted values of the original response variables coincides with the observed ones for all of the n objects and all of the k original response variables, the Q_a^2-statistic would attain the value 1. Therefore, the optimal number a of iterations is that corresponding to the largest value of the Q_a^2-statistic, i.e. to the value closest to 1.

However, assessing a regression model by whether it provides sufficiently accurate predictions, i.e. by using the Q_a^2-value leads to a false, in fact over optimistic idea of its prediction accuracy. This is because the prediction accuracy is computed without using an external test set which is not involved in the modelling process. Consequently, over-fitting is a disadvantage of determining the optimal model by the Q_a^2-statistic. Therefore, models selected by maximizing the Q_a^2-value cannot be considered to be necessarily reliable.

An alternative is to evaluate the prediction accuracy by blind cross-validation. This is performed by incorporating the performance of the leave-one-out cross-validation in a random test set validation procedure. This results in the computation of the P^2-statistic which can be considered to be a more accurate measure of the prediction accuracy compared to the Q_a^2- statistic because it relies on some kind of external validation. Therefore, the P^2-statistic should be calculated as well as the Q_a^2-statistic so as to give a more valid representation of the model's prediction accuracy. It can be expected that the resulting P^2-value is lower than the optimal Q_A^2-value since the P^2-statistic provides a less optimistic and hence more realistic measure of the prediction accuracy.

In blind cross-validation, the available data are split randomly into a large training dataset and a relatively small test dataset. Based on the training dataset, a regression model is derived by applying the ordinary procedure of the leave-one-out cross-validation, i.e. choosing the final regression equation by optimizing the model's complexity with respect to the Q_a^2-value. Subsequently, this model is used to predict the original response variables from the original descriptor variables of the test set. Thus, in contrast to the ordinary leave-one-out cross-validation, the predictions correspond to observations that have not been incorporated at all in the derivation of a regression model used to evaluate the prediction accuracy.

Afterwards, another test set is randomly selected from the whole dataset. From this training set, the procedure of deriving a regression model using the Q_a^2-statistic and calculating the predictions of observations of the test set is repeated. These computations are continued until every object has been included once and only once in the test set.

Subsequently, the blind cross-validated P^2-value can be calculated from these predictions of the k original response variables as follows:

$$P^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{l=1}^{k}(y_{il}^o - \hat{y}_{il,BCV_a}^o)^2}{\sum_{i=1}^{n}\sum_{l=1}^{k}(y_{il}^o - \bar{y}_{.l}^o)^2}.$$

In this equation, the term $\hat{y}_{il,BCV}^o$ denotes the prediction of the l-th original response variable for the i-th object in the testset. As with the Q_a^2-statistic, a P^2-value of 1 indicates that all of the predictions equal the measured values of the respective original response variables.

A disadvantage of the P^2-statistic is that it is based on several models with possibly different numbers of iterations from maximizing the Q_a^2-statistic for the respective training sets. In detail, the computation of the P^2-measure is based on K_1 regression models, where K_1 is the number of random test sets used to obtain the P^2-value. Consequently, the blind cross-validation cannot necessarily be used to determine the optimal number of iterations and hence a single final regression model unambiguously. However, conclusions on the optimal model complexity can be drawn by considering the number of iterations stated to be optimal with respect to the K_1 regression models. If all K_1 optimal regression models are based on the same number a of iterations, this number can be considered to be required to obtain the best prediction accuracy. In this case, the final regression model providing the most accurate predictions can be determined.

Nevertheless, the computation of the P^2-statistic is important anyway for evaluating the prediction accuracy that can be expected on average from the K_1 regression models with the highest Q_a^2-values amongst all models derived during the various maximizations of the Q_a^2-statistic. One single set of blind cross-validation results in one P^2-value and K_1 optimal Q_A^2-values. From these Q_A^2-values, the mean can be calculated and compared with the P^2-value. The difference between the mean of the Q_A^2-values and the P^2-value

indicates the extent of over-fitting inherent in the models selected by maximizing the Q_a^2-statistic. Consequently, the computation of the P^2-statistic can be used to detect models providing unreliable predictions though they may show a high a Q_A^2-value.

Freyhult et. al. (2005) give comprehensive explanations and extensions of the computations performed to obtain the P^2-statistic as a reliable measure of the prediction accuracy. In this paper, the procedure of blind cross-validation is described as a double cross-validation loop. In this context, the model selection using the Q_a^2-statistic is the inner loop within the outer loop calculating the P^2-statistic.

However, maximization of the Q_a^2-statistic is introduced on the basis of the random validation rather than the leave-one-out cross-validation as explained above. Further, within the inner loop, a joint selection of the number of latent variable vectors and subsets of the descriptor variables used in the regression analysis is performed during the determination of the optimal model with respect to the Q_a^2-value. This procedure of choosing additional descriptor variable subsets by a variable ranking algorithm could be applied in cases where the dataset comprises an extremely large number of descriptor variables.

The specific data partitioning used in the blind cross-validation procedure influences the value of the P^2-statistic. Therefore, in Freyhult et. al. (2005), a repetition of the blind cross-validation process is proposed to determine the variability of the resulting P^2-values. After performing K_2 blind cross-validation computations, K_2 corresponding P^2-values are obtained whose mean and standard deviation can be calculated. Further, the mean and standard deviation of the K_2 means of the Q_A^2-values obtained within the inner loops can be computed and used for a comparison with the mean and standard deviation of the P^2-values. If the repeated blind cross-validation procedure results in a relatively high value of the P^2-statistic accompanied by a small standard deviation, it can be assumed that this mean P^2-value gives a reliable measure of the average prediction accuracy of the models maximizing the Q_a^2-statistic that can be considered to provide useful predictions.

3.3 The univariate special case

3.3.1 Comparison with the multivariate case

The univariate PLS algorithm can be considered as a special case of the multivariate PLS algorithm. In the case of only one response variable the matrix Y^o summarizing the measurements of the k original response variables in the multivariate situation is reduced to the vector y^o. This column vector contains the observations of the single original response variable in the univariate case. Accordingly, the matrix Y of standardized response variables is replaced by the vector y of the single standardized response variable. Thus, in the univariate algorithm, the calculation of those terms that are computed using the information of the response variable differs from that one in the multivariate algorithm.

Those computations that refer to the standardized descriptor variables are made identically to the multivariate PLS algorithm. Therefore, the formulae for the a-th latent variable vector $\hat{t}_{.a}$, the a-th loading vector $\hat{p}_{.a}$ and the a-th residual matrix X_a of the standardized descriptor variables correspond to those in the multivariate PLS algorithm.

In the univariate PLS algorithm, the residual column vector y_a for the standardized response variable is calculated in the a-th iteration instead of the residual matrix Y_A in the multivariate case. Consequently, the computation of the a-th loading vector $\hat{q}_{.a}$ for the standardized response variables differs from the loading scalar $\hat{q}_a$ calculated in the a-th iteration in the univariate PLS algorithm. In this case, a scalar is obtained since its computation incorporates the residual vector y_{a-1} which has a reduced dimension compared with the residual matrix Y_{a-1} used to obtain the loading vector $\hat{q}_{.a}$ in the multivariate case. The A loading scalars calculated in the course of the performance of the univariate algorithm are summarized in the $(A \times 1)$-dimensional column vector $\hat{q}_{.A}$.

Further, in the univariate PLS algorithm, no sequential latent variable vectors are computed because the projection of the rows of the residual vector y_{a-1}, that is used instead of the residual matrix Y_{a-1}, into a scalar is unnecessary. Therefore, the subiterations performed within the iterations of the multivariate PLS algorithm are omitted from the calculations in the univariate PLS algorithm. This means that the a-th weight vector $\hat{w}_{.a}$, the a-th latent variable vector $\hat{t}_{.a}$ and the a-th loading scalar $\hat{q}_a$ are calculated directly instead of generating the corresponding vectors $\hat{w}^h_{.a}, \hat{t}^h_{.a}$ and $\hat{q}^h_{.a}$ iteratively within the subiterations of the a-th iteration.

Since the vector y of the standardized response variable or its residual vector does not have to be projected into the sequential latent variable vector, the calculation of the weight vector $\tilde{w}_{.a}$ incorporates the residual vector y_{a-1} instead of the sequential latent variable vector $\hat{u}_{.a}$. Thus the covariance between the latent variable t_a and the $(a-1)$-th residual variable of the standardized response variable y is maximized by the computation of the weight vector $\hat{w}_{.a}$ in the univariate PLS algorithm. This is different from the multivariate case. The derivation of this fact can be presented noting the multivariate case by taking into account that the residual variables of the standardized response variable are centered around zero in every iteration. This property can be shown by analogy with the explanations referring to the residual variables of the standardized descriptor variables.

Theoretically, if k response variables are taken into account, the univariate PLS algorithm might be applied to each of the standardized response variables individually instead of performing the multivariate PLS algorithm, i.e. incorporating the k standardized response variables simultaneously in the calculations. However, the application of k univariate PLS algorithms to the standardized data would result in k different decompositions of the matrix X since the information of the respective standardized response variable influences the decomposition. Therefore, the multivariate PLS algorithm results in different results from those of k univariate PLS algorithms. To decide which procedure should be preferred, the prediction accuracy of the corresponding regression models can be examined.

3.3.2 Scaling and centering

The variance-scaling and mean-centering of the data in the univariate situation is performed as in the multivariate case. The computations referring to the descriptor variables correspond to those in the multivariate situation. However, there is a difference in the standardization since the calculations in the univariate case are only performed for the single original response variable y^o instead of the k different ones in the multivariate case.

Thus, the mean-centering of the response variable is performed by calculating the mean of the values of the original response variable and subtracting it from each of the observations of the original response variable, i.e. by computing

$$y_i^o - \bar{y}^o \quad \forall i = 1, \ldots, n.$$

For the variance-scaling, the standard deviation

$$s_y^o = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i^o - \bar{y}^o)^2}$$

of the original response variable is calculated. Subsequently, the mean-centered values of the original response variable are divided by this standard deviation, resulting in the corresponding standardized values, i.e.:

$$y_i := \frac{y_i^o - \bar{y}^o}{s_y^o} \quad \forall i = 1, \ldots, n.$$

3.3.3 The decomposition models

In the univariate case, the decomposition model of the matrix X of standardized descriptor variables equals that presented for the multivariate situation. The information in the standardized response variable y can be decomposed according to the decomposition of the matrix Y. However, particular terms involved in the decomposition have a different dimension since the decomposition is only related to a single standardized response variable.

Instead of the loading vectors respecting the matrix Y of standardized response variables, loading scalars are used in the univariate situation. Further, that part of the standardized response variable y that remains unexplained by the A latent variable vectors is summarized in the column vector y_A of residuals,

$$y_A = \begin{pmatrix} y_{1,A} \\ \vdots \\ y_{n,A} \end{pmatrix} \sim n \times 1,$$

replacing the matrix Y_A from the multivariate case.

Consequently, the decomposition model of the standardized response variable y can be presented as

$$y = \sum_{a=1}^{A} t_{.a} q_a + y_A,$$

which can be expressed in matrix notation as

$$y = T_A q_A + y_A.$$

In this equation, the term q_A denotes the column vector containing the loading scalars, i.e.:

$$q_A = \begin{pmatrix} q_1 \\ \vdots \\ q_A \end{pmatrix} \sim A \times 1.$$

Thus, the i-th observation y_i of the standardized response variable can be given as a sum of the i-th elements of the A latent variable vectors which are weighted by the corresponding A loading scalars and the additional i-th residual scalar as follows:

$$y_i = \sum_{a=1}^{A} t_{ia} q_a + y_{i,A}.$$

3.3.4 The computations of the univariate algorithm

$$\text{Set} \quad X_0 \;:\; X, \quad y_0 : y, \quad a : 1$$

$$\begin{aligned}
\tilde{w}_{.a} &= X'_{a-1} y_{a-1} \\
\hat{w}_{.a} &= \frac{X'_{a-1} y_{a-1}}{\sqrt{y'_{a-1} X_{a-1} X'_{a-1} y_{a-1}}} = \frac{\tilde{w}_{.a}}{||\tilde{w}_{.a}||} \\
\hat{t}_{.a} &= X_{a-1} \hat{w}_{.a} \\
\hat{p}_{.a} &= \frac{X'_{a-1} \hat{t}_{.a}}{\hat{t}'_{.a} \hat{t}_{.a}} = \frac{X'_{a-1} \hat{t}_{.a}}{||\hat{t}_{.a}||^2} \\
\hat{q}_a &= \frac{y'_{a-1} \hat{t}_{.a}}{\hat{t}'_{.a} \hat{t}_{.a}} = \frac{y'_{a-1} \hat{t}_{.a}}{||\hat{t}_{.a}||^2} \\
X_a &= X_{a-1} - \hat{t}_{.a} \hat{p}'_{.a} \\
y_a &= y_{a-1} - \hat{t}_{.a} \hat{q}_a
\end{aligned}$$

$$\text{Set} \quad a \;:\; a + 1$$

3.3.5 Derivation of the estimates of the model parameters for the original regression model

With the help of the terms obtained in the univariate PLS algorithm, the regression coefficients summarized in the column vector $\hat{b}$ can be estimated by computing

$$\hat{b}_{PLS} = \hat{W}_A (\hat{P}'_A \hat{W}_A)^{-1} \hat{q}_A.$$

This formula for the estimates of the regression coefficients respecting the standardized variables can be proved analogously to the multivariate case. Using these estimates, the regression coefficients for the original variables can be calculated as

$$\hat{b}^o_{PLS} = \begin{pmatrix} \frac{1}{s_{x^o_{\cdot 1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{x^o_{\cdot m}}} \end{pmatrix} \hat{b}_{PLS} \cdot s^o_y.$$

The intercept b^o_0 of the original regression equation is then given as

$$\hat{b}^o_{0,PLS} = \bar{y}^o - \bar{x}^{o\prime}_{\cdot\cdot} \hat{b}^o_{PLS}.$$

Consequently, according to the multivariate case, the univariate original regression equation can be established as follows:

$$\begin{aligned} \hat{y}^o = & \; 1_n \hat{b}^o_{0,PLS} + X^o \hat{b}^o_{PLS} \\ = & \; 1_n \bar{y}^o + (X^o - 1_n \bar{x}^{o\prime}_{\cdot\cdot}) \hat{b}^o_{PLS} \\ = & \; 1_n \bar{y}^o + (X^o - 1_n \bar{x}^{o\prime}_{\cdot\cdot}) \begin{pmatrix} \frac{1}{s_{x^o_{\cdot 1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_{x^o_{\cdot m}}} \end{pmatrix} \hat{b}_{PLS} \cdot s^o_y. \end{aligned}$$

3.3.6 Predictions of the original response variable

The predictions of the original response variable from r objects whose values of the original descriptor variables are in the matrix $\tilde{X}^o$ can be obtained by analogy with the multivariate procedure.

The predictions of the original response variable can be obtained using the estimates $\hat{b}^o_{PLS}$ of the regression coefficients, i.e.:

$$\begin{aligned} \hat{\tilde{y}}^o = & \; 1_r \hat{b}^o_{0,PLS} + \tilde{X}^o \hat{b}^o_{PLS} \\ = & \; 1_r \bar{y}^o + (\tilde{X}^o - 1_r \bar{x}^{o\prime}_{\cdot\cdot}) \hat{b}^o_{PLS}. \end{aligned}$$

Alternatively, the predictions can be computed with the help of the following formula:

$$\begin{aligned} \hat{\tilde{y}}^o = & \; 1_r \bar{y}^o + s^o_y \sum_{a=1}^{A} \hat{\tilde{t}}_{\cdot a} \hat{q}_a \\ = & \; 1_r \bar{y}^o + s^o_y \hat{\tilde{T}}_A \hat{q}_A. \end{aligned}$$

According to the multivariate situation, the A latent variable vectors result from the application of two steps of the univariate PLS algorithm to the matrix $\tilde{X}$ comprising the standardized values of the descriptor variables. The A loading scalars and further terms incorporated in the reduced algorithm are taken from the calculations performed during the A iterations of the univariate PLS algorithm for obtaining the regression model.

3.3.7 Determination of the optimal model complexity

In the univariate case, the determination of the optimal number A of iterations and thus the optimal model complexity, can be performed by analogy with the multivariate situation. However, since the computations are based on the single original response variable y^o instead of the k original response variables, the formulae of the validation statistics, i.e. the $PRESS_a$-measure, the Q_a^2-statistic as well as the P^2-statistic, are calculated as follows:

$$\begin{aligned} PRESS_a &= \sum_{i=1}^{n}(y_i^o - \hat{y}_{i,CV_a}^o)^2, \\ Q_a^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i^o - \hat{y}_{i,CV_a}^o)^2}{\sum_{i=1}^{n}(y_i^o - \bar{y}^o)^2} \quad \text{and} \\ P^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i^o - \hat{y}_{i,BCV_a}^o)^2}{\sum_{i=1}^{n}(y_i^o - \bar{y}^o)^2}. \end{aligned}$$

3.4 Prediction intervals in PLS regression

In practice, the main objective of applying regression methods is to obtain reliable predictions of the values of the response variable corresponding to given values of the descriptor variables. After having predicted a value of the response variable, it is of great interest to gain knowledge about the accuracy of the prediction. On the one hand, measures of the overall prediction accuracy of an established regression model can be computed (see subsection **??**). On the other hand, a prediction interval can be calculated with respect to a certain prediction in order to obtain information about the reliability of the predicted value. Therefore, a comprehensive evaluation of the accuracy of a particular prediction can be presented by determining additionally a prediction interval.

In the following, considerations of prediction intervals for PLS regression are presented with respect to the prediction of a single response variable from the values of the descriptor variables for one single object. In the multivariate case, prediction intervals can be established separately for the different response variables by a straightforward extension of the univariate case, the only difference being that the terms used in the calculation of the prediction intervals are obviously those obtained in the multivariate PLS algorithm. If the prediction intervals are meant to be constructed simultaneously for the predictions referring to r objects and not merely to one single object, a Bonferroni correction has to be applied. In detail, the prediction intervals of the values of the response variable for different objects are computed under incorporation of the value $\frac{\alpha}{r}$ instead of the value α in the determination of the respective quantiles.

The true but unknown value $\tilde{y}^o_{i_r}$ of the original response variable for the i_r-th of r potential prediction objects can be expressed as

$$\tilde{y}^o_{i_r} = b^o_0 + \tilde{x}^{o\prime}_{i_r\cdot} b^o + \tilde{e}^o_{i_r} \quad \text{with} \quad i_r = 1, \ldots, r.$$

In this equation, the term $\tilde{x}^o_{i_r\cdot}$ denotes the column vector containing the values of the i_r-th row of the matrix $\tilde{X}^o$, i.e. the values of the m descriptor variables for the i_r-th object for which the value $\tilde{y}^o_{i_r}$ of the response variable is meant to be predicted.

A $(1-\alpha)$-prediction interval for this unknown value $\tilde{y}^o_{i_r}$ is that interval $[lb, ub]$ defined by the lower and upper bound that contains the value $\tilde{y}^o_{i_r}$ with probability $1-\alpha$ for a given value α, i.e.:

$$P(lb \leq \tilde{y}^o_{i_r} \leq ub) = 1 - \alpha.$$

For the Ordinary Least Squares regression, a $(1-\alpha)$-prediction interval is easily obtained under the assumption of normally distributed prediction errors $\tilde{e}^o_{i_r} = \tilde{y}^o_{i_r} - \hat{\tilde{y}}^o_{i_r}$ as

$$PI_{\alpha,OLS}(\tilde{y}^o_{i_r}) = \left[\hat{\tilde{y}}^o_{i_r} \mp t_{1-\frac{\alpha}{2},n-m-1} s_{OLS} \sqrt{1 + \frac{1}{n} + (\tilde{x}^o_{i_r\cdot} - \bar{x}^o_{\cdot\cdot})'(X^{o\prime}X^o)^{-1}(\tilde{x}^o_{i_r\cdot} - \bar{x}^o_{\cdot\cdot})}\right].$$

In this formula, the term

$$s_{OLS} = \sqrt{\frac{1}{n-m-1}\sum_{i=1}^{n}(y^o_i - \hat{y}^o_i)^2}$$

represents the root of the residual sum of squares divided by the term $n - m - 1$. Furthermore, the scalar $t_{1-\frac{\alpha}{2},n-m-1}$ denotes the $(1-\frac{\alpha}{2})$-quantile of a t-distribution with $n - m - 1$ degrees of freedom.

In order to derive a formula for a prediction interval for a particular regression method, it is necessary to know the statistical distribution of the unknown prediction errors $\tilde{y}^o_{i_r} - \hat{\tilde{y}}^o_{i_r}$ with $i_r = 1, \ldots, r$. However, this distribution cannot be obtained exactly for PLS regression because the estimation $\hat{b}^o_{PLS}$ of the original regression coefficients is a non-linear function of the vector y^o of the original response variable. Hence, it is quite a difficult task, compared with the case of the Ordinary Least Squares regression, to provide prediction intervals for PLS regression.

Denham (1997) proposes four different approaches for the computation of prediction intervals in this case: firstly based on naive considerations, secondly a cross-validation procedure, thirdly a local linear approximation and finally a bootstrapping method. The performance of the corresponding calculations are presented in the following with the exception of those of the bootstrapping procedure. The reason is that this approach can be considered to be inappropriate for the application to PLS regression as could be shown by Denham (1997). The inaccuracy of the bootstrapping method is caused by the fact that it is based on the assumption that the distribution of the unknown prediction errors is approximated sufficiently well by the distribution of the observed residuals. This situation can only be expected to be valid if a multitude of objects is taken into account. Consequently, in case of the PLS regression, this approach should not be used for the construction of prediction intervals since usually, merely a relative small sample size is available.

3.4.1 Approaches for the establishment of PLS prediction intervals

The naive approach

The $(1-\alpha)$-prediction interval based on the naive approach is constructed by neglecting the non-linearity of the PLS predictor. Beyond this, an alternative presentation of the estimated regression coefficients referring to the original variables is used. In particular, the estimations of the original regression coefficients can be expressed by the following formula:

$$\hat{b}^o_{PLS} = H_A X^{o'} y^o,$$

where the matrix H_A is given as

$$H_A = W_A(W_A' X^{o'} X^o W_A)^{-1} W_A'.$$

According to Helland (1988), this expression can also be written as

$$H_A = V_A(V_A' X^{o'} X^o V_A)^{-1} V_A'$$

with

$$V_A = (X^{o'} y^o, X^{o'} X^o X^{o'} y^o, \ldots, (X^{o'} X^o)^{A-1} X^{o'} y^o).$$

Therefore, the prediction interval based on the naive approach is obtained by analogy with the prediction interval for Ordinary Least Squares regression. Substituting the number m of descriptor variables by the number A of latent variable vectors in the determination of the quantile of the t-distribution and replacing the matrix $(X^{o'} X^o)^{-1}$ by the matrix H_A results in the following prediction interval formula:

$$PI_{\alpha,naive}(\tilde{y}_{i_r}^o) = \left[\hat{\tilde{y}}_{i_r}^o \mp t_{1-\frac{\alpha}{2},n-A-1} s_{naive} \sqrt{1 + \frac{1}{n} + (\tilde{x}_{i_r\cdot}^o - \bar{x}_{\cdot\cdot}^o)' H_A (\tilde{x}_{i_r\cdot}^o - \bar{x}_{\cdot\cdot}^o)} \right].$$

In this expression, the term

$$s_{naive} = \sqrt{\frac{1}{n-A-1} \sum_{i=1}^{n} (y_i^o - \hat{y}_i^o)^2}$$

denotes the corresponding standardized root of the residual sum of squares.

The cross-validation approach

As explained previously, in case of the PLS regression, using the quantile of the t-distribution in the prediction interval formula is not appropriate in contrast to the situation with Ordinary Least Squares regression. Therefore, this value should be substituted by a more realistic one that is denoted c_α. The scalar c_α can be found with the help of cross-validation. Incorporating the term c_α, the $(1-\alpha)$-prediction interval can be presented as

$$PI_{\alpha,CV}(\tilde{y}_{i_r}^o) = \left[\hat{\tilde{y}}_{i_r}^o \pm c_\alpha \sqrt{1 + \frac{1}{n} + (\tilde{x}_{i_r\cdot}^o - \bar{x}_{\cdot\cdot}^o)' H_A (\tilde{x}_{i_r\cdot}^o - \bar{x}_{\cdot\cdot}^o)} \right].$$

The scalar c_α is determined to be that value that leads to the narrowest prediction interval showing a coverage probability of $1-\alpha$. Accordingly, the

minimum positive value satisfying the following inequality is chosen as the scalar c_α:

$$\frac{1}{n}\sum_{i=1}^{n} I_{\{|r_{i,CV_i}|>c_\alpha\}} \leq \alpha, \quad \text{where}$$

$$r_{i,CV_i} := \frac{y_i^o - \bar{y}_{CV_i}^o - (x_{i.}^o - \bar{x}_{.,CV_i}^o)'\hat{b}_{PLS_{CV_i}}^o}{s_{CV_i}\sqrt{\frac{n}{n-1} + (x_{i.}^o - \bar{x}_{.,CV_i}^o)'H_{A,CV_i}(x_{i.}^o - \bar{x}_{.,CV_i}^o)}}.$$

In this expression, the subscript CV_i indicates terms being computed without the values of the respective i-th object. In the column vector $x_{i.}^o$, the elements of the i-th row of the matrix X^o of original descriptor variables are summarized. The vector $\hat{b}_{PLS_{CV_i}}^o$ contains the estimated original regression coefficients calculated on the basis of A iterations performed using the dataset from which the values referring to the i-th object are excluded. Furthermore, the term s_{CV_i} denotes the corresponding standardized root of the residual sum of squares, and the indicator function is defined as

$$I_{\{|r_{i,CV_i}|>c_\alpha\}} = \begin{cases} 1 & \text{if the argument is true} \\ 0 & \text{else.} \end{cases}$$

The term r_{i,CV_i} is interpretable as a "studentized" residual, because the nominator of this expression equals the prediction error $e_{i,CV_i}^o = y_i^o - \hat{y}_{i,CV_i}^o$. The prediction $\hat{y}_{i,CV_i}^o$ of the response variable for the i-th object is calculated with the help of the data omitting the values from the i-th object. The prediction error e_{i,CV_i}^o is divided by a term representing its estimated standard deviation. Consequently, the required value c_α is given as the $(n(1-\alpha))$-th order statistic of these studentized residuals. If the value $n(1-\alpha)$ is not an integer, the next smallest integer value is chosen for the value c_α.

The local linearization approach

The linear approximation method suggested by Denham (1997) is presented according to the description given by Serneels et. al. (2004). The idea of the local linearization approach is to approximate the vector of original regression coefficients by a linear expression in order to derive an approximate estimate of the residual standard deviation used in the formula for the prediction interval. The local linear approximation of the vector of original regression coefficients is obtained by expanding it as a Taylor series of first order about some given vector y_0, i.e.:

$$\hat{b}_{PLS}^o(y^o) \approx \hat{b}_{PLS}^o(y_0) + \left(\frac{\partial \hat{b}_{PLS}^o}{\partial y^o}\right)_{y_0} (y^o - y_0).$$

In this approximation, the matrix

$$\left(\frac{\partial \hat{b}^o_{PLS}}{\partial y^o}\right)_{y_0} \sim m \times n$$

denotes the partial derivative of the vector of original regression coefficients of the original response variable y^o, evaluated at y_o. An algorithm for the computation of this partial derivative matrix is given both by Denham (1997) and by Serneels et. al. (2004). With the help of local linearization, an approximate $(1-\alpha)$-prediction interval can be obtained as follows:

$$PI_{\alpha,linear}(\tilde{y}^o_{i_r}) = \left[\hat{\tilde{y}}^o_{i_r} \pm t_{\frac{\alpha}{2},df} s_{linear} \sqrt{1 + \frac{1}{n} + (\tilde{x}^o_{i_r\cdot} - \bar{x}^o_{\cdot\cdot})^T \frac{\partial \hat{b}^o_{PLS}}{\partial y^o} \left(\frac{\partial \hat{b}^o_{PLS}}{\partial y^o}\right)^T (\tilde{x}^o_{i_r\cdot} - \bar{x}^o_{\cdot\cdot})}\right].$$

To avoid confusion with partial derivatives in this subsubsection, the transpose of a term is denoted by an index T rather than a superscript $'$. This differs from the usual notation elsewhere.

The estimate s_{linear} of the residual standard error used in the prediction interval is computed as:

$$s_{linear} = \frac{r'r - ||r - r'r||^2}{df}.$$

In this formula, the term r denotes the residual vector

$$r = y^o - \hat{y}^o$$

and the term r' represents the first partial derivative of this residual vector with respect to the original response variable, evaluated at y_o. Furthermore, the degrees of freedom are given as

$$df = Tr\left[\left(I_n - X^o \frac{\partial \hat{b}^o_{PLS}}{\partial y^o}\right)^T \left(I_n - X^o \frac{\partial \hat{b}^o_{PLS}}{\partial y^o}\right)\right],$$

where the expression Tr denotes the trace of a term.

3.4.2 Evaluation of the approaches

The three prediction interval formulae for PLS regression above can be considered merely as approximate intervals because they are not derived on the

basis of exact distributional properties of the prediction error. Thus, to judge the accuracy of these different proposals, Denham (1997) estimated the coverage probabilities for them with desired coverage probabilities of 80%, 90% and 95% on both a real example and simulated data. The results of this comparison of observed and expected coverage probabilities for the different approaches are presented briefly in the following.

Generally, it can be stated that for every approach, the accuracy of the established prediction intervals depends on the number A of iterations of the PLS algorithm used to obtain the relevant terms.

The prediction intervals based on the naive approach tend to over-estimate the coverage probability, especially if they are built using the results of more than two iterations of the PLS algorithm.

The observed coverage probabilities of the prediction intervals of the cross-validation approach depend on the number A of iterations performed to obtain the terms that are included in the computation of the prediction intervals. Cross-validation prediction intervals calculated on the basis of a few iterations over-estimate the coverage probabilities, whereas those based on more iterations under-estimate the coverage probabilities. A drawback of the approach to provide prediction intervals by applying the cross-validation procedure is the fact that the prediction intervals rely on quantiles of the ordered absolute values of studentized residuals. This leads to a certain inexactness in case of a limited number of observations. However, according to Denham (1997), the resulting inaccuracy is not too serious and this approach can be considered to be useful nevertheless.

For prediction intervals calculated by the linear approximation of the vector of the estimated original regression coefficients, the coverage probabilities are over-estimated. But in contrast to the naive approach, the results are better for prediction intervals established on the basis of the terms obtained in a number of iterations.

Chapter 4

Application of PLS regression to the analysis of biomolecular interactions

4.1 General considerations

The following statements concerning the application of PLS regression to the investigation of biomolecular interactions are derived from Andersson (2004), Andersson et. al. (2001), Andersson et. al. (1999), De Genst et. al. (2002) and Choulier et. al. (2002). All of these publications show a lack of description of the theoretical statistical methodology the data analysis is based upon.

The performance of the examination of biomolecular interactions is usually explained for a particular interaction of interest. Based on the examples described in the publications, a general and comprehensive presentation of the performance of modelling characteristics of the binding between two interacting biomolecules has been derived. Often, enzyme-substrate or antibody-antigen systems are of special relevance in practice. However, any type of ligand-receptor interaction could be studied using the procedure presented below.

The binding behaviour of interacting biomolecules can be characterized by three kinetic parameters, the association and the dissociation rate constant as well as the affinity constant. The association rate constant reflects the extent of recognition of the interacting biomolecules, whereas the dissociation rate constant represents the stability of the resulting complexes of bound

biomolecules and the affinity constant indicates the binding strength between the biomolecules involved in the interaction.

These binding parameters can be measured with a high accuracy by a biosensor system in order to record the binding properties of the interacting biomolecules. This biosensor system is based on the physical process of surface plasmon resonance. Details of the technology of surface plasmon resonance biosensors are given in the following section.

In biomolecular interaction studies, the objective of the investigation is the quantification of the effect of a number of factors on the binding behaviour. Potential factors that are usually incorporated in these kind of studies either represent the physico-chemical properties or structural features of amino acids at certain positions in the sequence of one or both of the binding partners or describe the composition of the chemical environment in which the interaction takes place.

The quantification of the effect of these factors is realized by establishing regression models relating the measured binding parameters to the diverse factors by applying PLS methodology. In these regression models, the association rate constant, the dissociation rate constant and the affinity constant are the response variables, whereas the factors supposed to influence these response variables are incorporated as the descriptor variables. Consequently, the data on which the derivation of the regression models is based comprise on the one hand the measurements of the three binding parameters and on the other hand, values of the adjustments of the factors the interaction might depend upon.

In order to obtain these data that can be used to establish the regression models, the binding parameters respecting a particular biomolecular interaction are measured under different conditions. The experiments are performed by varying simultaneously the adjustments of the factors that are suspected to have an effect on the interaction of interest. Therefore, the amino acids situated at relevant positions in the sequence of the wild-type of one or both of the binding parameters are substituted by several amino acids and the composition of the buffers in which the interaction takes place is altered.

Some data might be missing owing to the fact that the binding either does not take place or that the events of association and dissociation occur too rapidly to be measured. This problem might either be caused by the modified protein that does not interact with its binding partner or by the composition of a specific buffer.

By estimating the regression coefficients of the descriptor variables involved, the effects the diverse factors have on the response variables, i.e. on the binding parameters, can be quantified. In detail, if the descriptor variable is increased by one unit the response variable changes by an amount given by its regression coefficient up or down depending upon its sign. Therefore, the extent of the influence of the respective descriptor variable can be determined.

The nature of the effect of a particular descriptor variable on the interaction can be determined by inspecting the sign of the respective regression coefficient. A positive sign indicates that an increase in the value of this variable leads to an increase of the corresponding binding parameter to an extent given by the value of the regression coefficient, whilst a negative sign leads to a decrease.

Therefore, examination of the signs of the regression coefficients reveals whether changes of a particular descriptor variable disturb or favour the association, dissociation or affinity, respectively. This means that statements can be made about the settings of the descriptor variables that cause an improved or disturbed recognition, a less or more stable complex of the interacting biomolecules or a lower or higher binding strength, respectively.

In practice, univariate regression models referring individually to one of the relevant response variables, i.e. the association rate constant, the dissociation rate constant and the affinity constant, are developed instead of one single multivariate model. Thus, the influence of the diverse factors is modelled separately for every response variable. Since the methodology of PLS regression is applied to the data of biomolecular interaction studies, the results received by the performance of one single multivariate regression analysis differ from those obtained by establishing several univariate regression models.

Furthermore, univariate regression models are usually developed separately for the different subgroups of potential descriptor variables. Hence, they can be assigned to one of the following types: quantitative buffer-kinetics relationship (QBKR)-, quantitative sequence-kinetics relationship (QSKR)-, quantitative structure-activity relationship (QSAR)-, three-dimensional quantitative structure-activity relationship (3D-QSAR)- or quantitative sequence-perturbation relationship (QSPR)-model. In this publication, a comprehensive theoretical description of these different kinds of regression models is given in subsections **??**, **??**, **??** and **??** that cannot be found in the articles published with respect to a particular interaction.

In QBKR investigations, the effect of the chemical environment on the interaction is examined. Accordingly, the descriptor variables used in QBKR models are concentrations of chemical additives and the pH value of the buffer. The influence of the physico-chemical properties of some amino acids at particular positions in the sequence of the biomolecules on the binding behaviour is subject to QSKR modelling. Therefore, in QSKR models, variables quantifying the physico-chemical properties of amino acids at certain positions of the sequence of a protein are incorporated as the descriptor variables.

QSAR examinations can either be considered to be equivalent to the QSKR analysis or they are meant to comprise those investigations dealing with the structure of the proteins of interest in addition to the physico-chemical properties. In particular, 3D-QSAR models denote those QSAR models that take three-dimensional features of the biomolecules involved in the interaction into account. Therefore, the descriptor variables used in QSKR models are supplemented by variables quantifying structural features that are additionally considered in QSAR- or 3D-QSAR-models, respectively.

Subsequent to the establishment of a QBKR and a QSKR model for a particular interaction, so-called QSPR models can be developed. QSPR models relate the sensitivity of the binding behaviour concerning a particular chemical additive to the physico-chemical properties of amino acids at certain positions in the sequence.

In this publication, an alternative to the usual procedure of establishing separate univariate regression models for the different subgroups of descriptor variables is proposed. This novel approach for modelling the data obtained in biomolecular interaction studies is introduced by presenting a unified multivariate regression model. The different univariate regression models that are usually referred to in literature can be derived as special cases from this single regression model that is described in subsection **??**.

The application of the PLS method in order to establish regression models is preferred to the computations of the OLS procedure in situations when the descriptor variables are linear dependent or when less experiments are performed than descriptor variables incorporated in the model. In biomolecular interaction studies, it is often the case that only a few observations, i.e. kinetic measurements respecting the interaction under different conditions, are available since the performance of numerous experiments is time consuming and expensive. The limited number of experiments is accompanied

by relative large numbers of descriptor variables that could be taken into consideration. Especially in 3D-QSAR studies, an extremely large number of descriptors is required for a correct representation of the structure of the biomolecules. Therefore, PLS regression is the appropriate method to apply to the data obtained during the investigation of biomolecular interactions.

As the PLS procedure is able to deal with this challenging data situation, it permits the incorporation of a sufficiently large number of descriptor variables in the regression model. Thus, it permits a comprehensive description of the circumstances supposed to influence the binding properties of an interaction, an essential prerequisite for an appropriate determination of the predominant effects on the interaction. Consequently, the regression models established by the PLS method provide an accurate characterization of the binding behaviour of the interacting biomolecules by quantifying the effects of the diverse factors on the interaction.

Obtaining knowledge about the extent of influence the various factors have on the binding behaviour presents the basis for the explanation of the molecular regulation and the function of the interacting biomolecules in vivo. Developing regression models in the context of the analysis of the binding between biomolecules leads to an improved understanding of the interaction under investigation. Furthermore, predictions of values of the association and dissociation rate constant as well as of the affinity constant of the interaction can be computed with respect to specified adjustments of the influencing factors. Gaining this knowledge results in the possibility of determining the conditions that have to be realized in order to obtain a particular predefined profile of the binding parameters. Consequently, with the help of the conclusions drawn from the established regression models, the binding properties between two biomolecules can be optimized according to the purpose of the investigation.

Therefore, the analysis of biomolecular interactions by means of the PLS regression provides results that are of great importance with respect to diverse applications. For example, in pharmacology, this kind of knowledge is required in the development of drugs. Furthermore, in biotechnology, the determination of appropriate adjustments of the affinity chromatography can be based on the obtained conclusions. The affinity chromatography is a method that is based on the specific binding between two biomolecules and is used in order to purify proteins.

4.2 Surface plasmon resonance biosensors

The binding parameters characterizing a biomolecular interaction, the association and the dissociation rate constant, as well as the affinity constant, can be measured sufficiently accuratly with the help of surface plasmon resonance (SPR) biosensors. In practice, the SPR biosensors most frequently used are the Biacore systems produced by Biacore AB in Sweden. The following explanation of this instrument is based on the presentations given in Andersson (2004) and on the homepage of Biacore (www.biacore.com). Furthermore, the illustrations are taken from Wimmer (2005).

Biosensors based on SPR are appropriate for the investigation of biomolecular interactions as they are able to trace the course of the binding process in real time. The measuring method makes use of the fact that the binding partners under examination interact in a specific way.

SPR biosensors show two predominant advantages compared to alternative technologies applied to kinetic measurements. On the one hand, the measurements are obtained with a high accuracy, in fact with respect to interactions with a low as well as with a high binding strength. In detail, the affinity constants may range between millimolar (mM) and picomolar (pM) units, where molar (M) is defined as mol per liter (mol/liter). On the other hand, the experiments can be performed very quickly.

The major prerequisite for the measurement of the kinetic parameters using a SPR biosensor is the ability to immobilize one of the binding partners on a sensor chip as explained below. The binding parameters of a wide range of diverse biomolecules can be measured with SPR biosensors. In particular, the method is applicable to interactions involving either proteins, peptides, small potential drug compounds, DNA, RNA, viruses or whole cells.

4.2.1 The binding parameters

The binding behaviour of the interacting biomolecules of interest can be characterized by the association rate constant k_a, the dissociation rate constant k_d and the affinity constant. These binding parameters represent the speed of the formation and the decay of the complexes of the binding partners, respectively, as well as the binding strength.

The association rate constant k_a, measured in amount per molar and per second ($[1/Ms]$), indicates the number of complexes resulting per second from

the reaction of the binding partners, given a one molar solution of the interacting biomolecules. The dissociation rate constant k_d, given in amount per second ($[1/s]$), determines the fraction of the complexes consisting of the binding partners that decompose each second. Hence, the dissociation rate constant reflects the stability of the formed complexes.

Usually, the values of the association rate constant range between 1000 $[1/Ms]$ and 10000000 $[1/Ms]$, whereas the values of the dissociation rate constant vary between 0.0000001 and 0.10 $[1/s]$. In practice, the values of the association rate constant are often multiplied in the context of equilibrium equations by the concentration of one of the binding partners involved in the interaction. Typically, this concentration attains values between 0.0000000001 and 0.00001 $[M]$. Consequently, the product of this concentration and the association rate constant is measured in $[s^{-1}]$ and has a similar range of values compared with that one of the measurements of the dissociation rate constant.

Summarizing, the interpretation of these two kinetic parameters in the context of the course of the reversible reaction of the binding of the biomolecules can be expressed as follows:

$$A + B \underset{k_d}{\overset{k_a}{\rightleftharpoons}} AB.$$

In this presentation, the term A denotes the amount of free biomolecules of one of the binding partners, the term B represents the amount of immobilized biomolecules of the other binding partner and the term AB reflects the amount of complexes consisting of the interacting biomolecules. High values of the association and dissociation rate constants indicate a fast association or dissociation process, respectively.

The affinity constant K_D measured in molar (M) represents the binding strength, i.e. the tightness of the binding occurring between the interacting biomolecules. It is defined as the ratio of the dissociation rate constant k_d and the association rate constant k_a, i.e.

$$K_D = \frac{k_d}{k_a}.$$

Consequently, a high binding strength of an interaction, i.e. a low rate of complex decay in relation to the forming of the complexes, is indicated by small values of the affinity constant. Since interactions showing different association and dissociation rate constants can have an identical binding

strength, the values of the association and dissociation rate constants should also be determined rather than using just the value of the affinity constant as the only measure to characterize a biomolecular interaction.

The logarithmic values of the association and dissociation rate constants can be interpreted as energy terms. Therefore, the logarithm of the observed binding parameters might be used in the data analysis as the values of the response variables instead of the raw data. The decision of whether to apply the regression analysis to the raw or logarithmic data depends upon the biological process that is to be modelled, i.e. either the kinetic or the energetic one.

4.2.2 Components of a Biacore instrument

Biacore systems are composed of three constituents, an optical detection system, a disposable sensor chip and an integrated microfluidic cartridge (IFC) as it can be seen in illustration **??**. The sensor chip consists of a glass

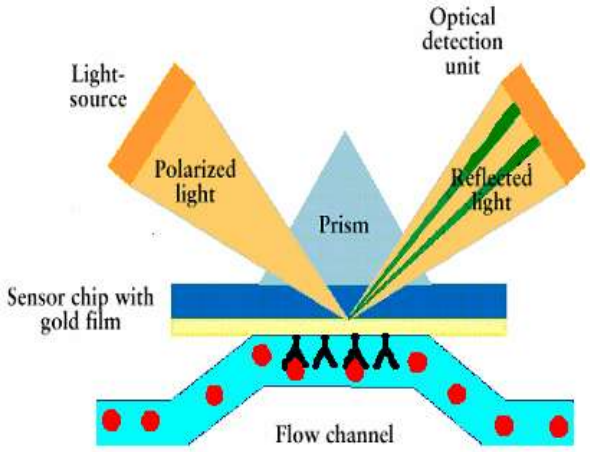

Figure 4.1: The components of a Biacore instrument

slide that is covered with a thin layer of gold. A layer of dextran is attached to this gold film, enabling the immobilization, i.e. the covalent binding, of one of the binding partners to the surface of the sensor. Therefore, this layer of biomolecules present at the surface of the sensor chip is specific in what concerns the interaction of interest.

The sensor chip is linked to the flow cells of the IFC that also comprises channels with valves that allow a solution injected into the system to reach the flow cells. Thus, the sensor chip can be supplied with liquid by the flow cells.

The optical detection system consisting of a light source, a prism and a detector is situated opposite to that side of the sensor chip where the IFC can be found. The light source illuminates the sensor chip with a beam of polarized light that covers a particular range of incident angles. The detector records the response as explained in detail below.

4.2.3 The course of the experiments

In order to prepare the Biacore instrument for the investigation of a particular interaction, biomolecules of one of the binding partners involved in the interaction under study are immobilized on the surface of the sensor chip. This is done by binding the biomolecules covalently to the layer of dextran attached to the gold film of the sensor chip.

Subsequently, biomolecules of the other binding partner are injected into the system and reach the surface of the sensor chip owing to the distribution of the solution within the flow cells. Consequently, the immobilized biomolecules interact with biomolecules in the solution leading to complexes formed of the binding partners bound to each other during the process of association. Afterwards, the solution containing the biomolecules of one of the binding partners is replaced by a standard buffer. Hence, dissociation begins to take place spontaneously, where the complexes between the interacting biomolecules decompose.

In case of a very low dissociation, a regeneration solution such as glycine buffer at pH 2.0 - 3.0 is used. As a result of this change in the chemical environment, an increase of the dissociation is achieved without destroying the biomolecules bound to the dextran. After the process of dissociation is completed, i.e. when all complexes of the binding partners are decayed, the next injection of the solution with biomolecules can be performed.

Altogether, the time such an analysis cycle takes ranges from 5 to 30 minutes, where the solution of biomolecules is injected for a period of 1 to 5 minutes.

4.2.4 Mode of operation of a Biacore instrument

Biacore instruments use the fact that changes in the amount of bound biomolecules lead to a change of the refractive index within the system. By detecting the absorbed light, conclusions concerning the state of the interaction can be drawn. For a better understanding of the following complicated explanations of the Surface Plasmon Resonance, an illustration of this physical process is given in figure **??**.

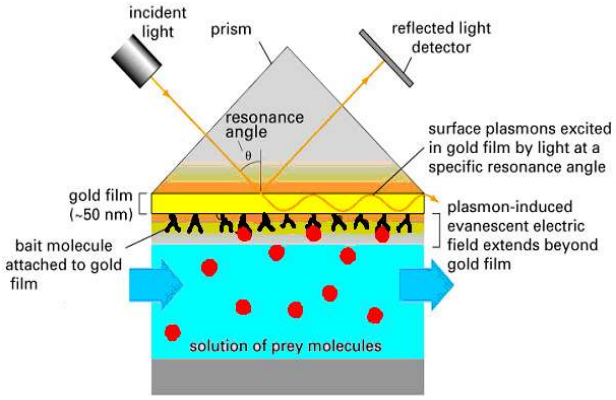

Figure 4.2: Illustration of the Surface Plasmon Resonance

The striking of the polarized light from the light source on the layer of glass causes the generating of an evanescent wave, an electric field intensity. The absorption of this wave by free electron clouds existing in the gold film results in the emergence of plasmons, electron charge density waves. The presence of these plasmons reduces the intensity of the beam that is reflected from the glass slide.

The surface plasmon resonance takes place when the polarized light, in case of total internal reflection, is directed toward the gold film that is electrically conducting and situated at the interface between two media showing different refractive indices. This difference of the refractive indices is present in a Biacore system because the solution flowing in the IFC has a low refractive index compared to the glass slide of the sensor chip showing a high refractive index.

The reflected light that is recorded by the detector shows an intensity minimum with a particular angular position. This resonance angle at which the intensity minimum occurs depends on the refractive index in the environment of the sensor chip.

The interaction of the binding partners at the sensor chip results in an accumulation of biomolecules that leads to an increase of the refractive index at the surface of the sensor chip in most cases. Thus, when the biomolecules in the solution bind at or dissociate from the immobilized biomolecules at the surface of the sensor chip, the refractive index at the interface between the surface of the sensor chip and the solution in the flow cells is changed. This change in the refractive index alters the angle at which the polarized light of reduced intensity is reflected from the glass slide.

As the extent of the change of this angle is proportional to the change of the amount of bound biomolecules, the course of the interaction can be recorded. This is done by the detector that registers the alterations that occur in the refractive index in the environment of the gold layer.

4.2.5 Output of Biacore experiments

The detector of the Biacore system records the angular position of the intensity minimum of the polarized light reflected from the glass slide. This output being proportional to the amount of bound biomolecules is measured in resonance units (RU). It is known that 1000 resonance units correspond approximately to 1 nanogram of biomolecules per quadrat millimeter (ng/mm^2). This relationship permits conclusions concerning the amount of bound biomolecules dependent on time. Consequently, the course of the interaction can be determined with the help of the observed data.

During the process of association that is accompanied by an accumulation of biomolecules at the surface of the sensor chip, the response is increased since the angle of the reflection intensity minimum is increased. When the equilibrium of the interaction is reached, the response presents a constant signal. The reason is that the measured angle remains unchanged when the biomolecules in the solution neither bind to the immobilized biomolecules nor dissociate from them. The state of the equilibrium is characterized by the process of association and dissociation compensating each other. As soon as the dissociation commences, the observed output begins to decrease. Then, the amount of bound biomolecules at the sensor chip is reduced, resulting in a decrease of the angle of the reflection intensity minimum.

The obtained measurements can be illustrated by a so-called sensorgram that represents the recorded response in relation to the time. Therefore, the sensorgram visualizes the course of the interaction between the two kinds of biomolecules of interest, providing a characteristic profile comprising both the process of association and dissociation as it can be seen in figure **??**.

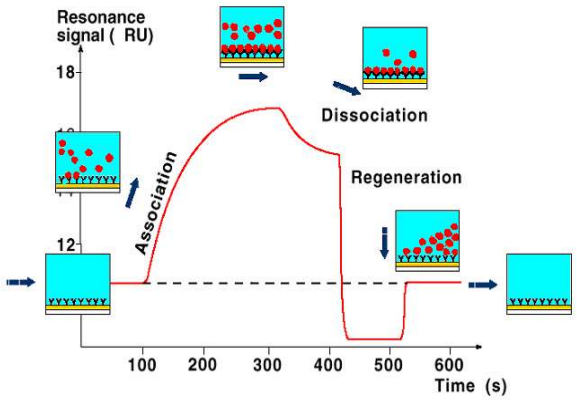

Figure 4.3: A sensorgram showing the dependence of the measured resonance units on the time

The association and dissociation rate constants as well as the affinity constant can be determined with the help of the data measured by the Biacore system. In contrast to the calculation of the dissociation rate constant, a disadvantage of the computation of the association rate constant is that it can only be determined when the concentrations of both of the interacting types of biomolecules are known, which is difficult to measure for the biomolecules in the solution. Thus, problems respecting the exact calculation of the association rate constant might arise. The affinity constant can easily be computed as the ratio of the dissociation and the association rate constants.

4.3 Application of PLS regression

4.3.1 The unified multivariate regression model

In practice, a number of univariate regression models are developed in order to draw conclusions on the influence of the descriptor variables of interest on the binding parameters. In contrast to this usual procedure, the multivariate PLS algorithm could be applied once with the objective to establish one single unified multivariate regression model. Hence, the statements derived from the multitude of univariate regression models might be obtained as well by one single multivariate regression model, where the compact results would facilitate the interpretation.

In the unified multivariate regression model, all of the possible response and descriptor variables relevant to the analysis of biomolecular interactions are incorporated simultaneously. The two response variables taken into account are the association and the dissociation rate constant. The affinity constant is not included as a response variable in the multivariate regression model. The reason is that it represents the ratio of the dissociation and the association rate constant. Therefore, statements concerning the affinity constant can be determined with the help of the results obtained with respect to the association and dissociation rate constant.

All of the descriptor variables that can be assigned to different subgroups are incorporated simultaneously in the unified multivariate regression model. Accordingly, on the one hand, the multivariate regression model comprises buffer descriptor variables, i.e. variables representing the composition of the chemical environment in which the interaction takes place. On the other hand, it involves the amino acid descriptor variables. This subgroup can be divided in those variables that can be determined for every position of interest in the sequence of the protein, e.g. variables describing the two- or three-dimensional characteristics of an amino acid, and in those variables that should be given only for the amino acids at positions of the protein where mutations are carried out. Those variables represent the physico-chemical properties of the amino acids. Examples are the so-called ZZ-scales and the helix-forming tendency (HFT)-scale.

In the following corresponding subsections, the descriptor variables usually used in biomolecular interaction studies are explained in more detail. However, it is possible to include various additional descriptor variables in the regression model to obtain more information about the influences on the

interaction of interest. In this case, the number of descriptor variables obviously increases and thus, the PLS regression is particularly useful.

The data on which the establishment of the unified multivariate regression model is based on are measured for n^p proteins in n^b buffers. Consequently, the total number of experiments, i.e. the sample size n, equals the value $n^p n^b$, presuming that the values of the binding parameters are measured once for the interaction of each of the n^p proteins in each of the n^b buffers. In case of repeated measurements, the sample size increases accordingly.

For clarity, the form of the unified multivariate regression model is introduced by giving two alternative presentations of the regression equation for the term $y_{i^b i^p l}$, the value of the l-th response variable respecting the i^p-th protein in the i^b-th buffer. The value $y_{i^b i^p l}$ can be expressed as the following summation:

$$y_{i^b i^p l} = b_{0l} + \sum_{j_b=1_b}^{m_b} b_{j_b l} x_{i^b j_b} + \sum_{j_p=1_p}^{m_p} \sum_{s=1}^{q} b_{(j_p)_s l} x_{i^p (j_p)_s} + \sum_{j_p=1_p}^{r_p} \sum_{s=q+1}^{z} b_{(j_p)_s l} x_{i^p (j_p)_s} + e_{i^b i^p l}, \tag{4.1}$$

where

$$l = 1, \ldots, k; \quad i^b = 1^b, \ldots, n^b; \quad i^p = 1^p, \ldots, n^p; \quad q < z \quad \text{and} \quad r_p < m_p.$$

As descriptor variables, m_p amino acid descriptor variables and m_b buffer descriptor variables are involved in the model. In what concerns the amino acid descriptor variables, r_p of them can be determined for all of the z positions of interest, whereas $m_p - r_p$ variables are merely given for q mutation sites of the sequence.

The term $x_{i^b j_b}$ denotes the value of the j_b-th buffer descriptor variable respecting the i^b-th buffer. The regression coefficient $b_{j_b l}$ corresponding to the j_b-th buffer descriptor variable refers to the l-th response variable. Thus, this term indicates the quantity the l-th response variable would be increased by, if the j_b-th buffer descriptor variable would be increased by one unit, under the assumption that the values of the other descriptor variables would be kept constant.

Furthermore, the value of the j_p-th amino acid descriptor variable concerning the amino acid at the s-th position of the i^p-th protein is denoted by the term $x_{i^p (j_p)_s}$. The expression $b_{(j_p)_s l}$ represents the regression coefficient respecting

the l-th response variable and the j_p-th amino acid descriptor variable at the s-th position of the sequence. Consequently, the l-th response variable would be increased by the value $b_{(j_p)_s l}$, if the j_p-th amino acid descriptor variable referring to the amino acid at the s-th position would be increased by one unit and the values of the other descriptor variables would be fixed.

The term b_{0l} represents the intercept of the regression model for the l-th response variable. Hence, the l-th response variable would attain this value, if all of the descriptor variables would be set to the value zero. Furthermore, the expression $e_{i^b i^p l}$ denotes the error corresponding to the value $y_{i^b i^p l}$ of the l-th response variable.

With respect to a more compact presentation of equation **??**, the values of the descriptor variables belonging to different subgroups as well as the corresponding regression coefficients can be summarized in column vectors. The vectors containing the values of the descriptor variables are obviously of the same dimension as the respective corresponding vectors including the regression coefficients.

The vector $x_{i^b m_b}$ comprises the values of the m_b buffer descriptor variables respecting the i^b-th buffer. The regression coefficients referring to these descriptor variables are given in the vector $b_{m_b l}$, i.e.:

$$x_{i^b m_b} = \begin{pmatrix} x_{i^b 1_b} \\ \vdots \\ x_{i^b m_b} \end{pmatrix} \quad \text{and} \quad b_{m_b l} = \begin{pmatrix} b_{1_b l} \\ \vdots \\ b_{m_b l} \end{pmatrix} \sim m_b \times 1.$$

The elements of the vector $x_{i^p (m_p)_q}$ are the values of the m_p amino acid descriptor variables regarding both the physico-chemical properties and structural features at the q mutation sites. Accordingly, the corresponding regression coefficients are summarized in the vector $b'_{(m_p)_q l}$ as follows:

$$x_{i^p (m_p)_q} = \begin{pmatrix} x_{i^p (1_p)_1} \\ \vdots \\ x_{i^p (1_p)_q} \\ x_{i^p (2_p)_1} \\ \vdots \\ x_{i^p (m_p)_q} \end{pmatrix} \quad \text{and} \quad b_{(m_p)_q l} = \begin{pmatrix} b_{(1_p)_1 l} \\ \vdots \\ b_{(1_p)_q l} \\ b_{(2_p)_1 l} \\ \vdots \\ b_{(m_p)_q l} \end{pmatrix} \sim (m_p q) \times 1.$$

Furthermore, the vector $x_{i^p (r_p)_{\substack{q+1 \\ \rightarrow z}}}$ consists of the values of the r_p descriptor variables of structural features of amino acids at additional $z - q$ positions

of interest. The vector $b_{(r_p)_{\underset{\rightarrow z}{q+1}}l}$ contains the regression coefficients related to these descriptor variables, i.e.:

$$x_{i^p(r_p)_{\underset{\rightarrow z}{q+1}}} = \begin{pmatrix} x_{i^p(1_p)_{q+1}} \\ \vdots \\ x_{i^p(1_p)_z} \\ x_{i^p(2_p)_{q+1}} \\ \vdots \\ x_{i^p(r_p)_z} \end{pmatrix} \quad \text{and} \quad b_{(r_p)_{\underset{\rightarrow z}{q+1}}l} = \begin{pmatrix} b_{(1_p)_{q+1}l} \\ \vdots \\ b_{(1_p)_z l} \\ b_{(2_p)_{q+1}l} \\ \vdots \\ b_{(r_p)_z l} \end{pmatrix} \sim (r_p(z-q)) \times 1.$$

With the help of these column vectors summarizing the values of particular descriptor variables or the corresponding regression coefficients, respectively, the value $y_{i^b i^p l}$ can be expressed alternatively as follows:

$$y_{i^b i^p l} = b_{0l} + b'_{m_b l} x_{i^b m_b} + b'_{(m_p)_q l} x_{i^p(m_p)_q} + b'_{(r_p)_{\underset{\rightarrow z}{q+1}}l} x_{i^p(r_p)_{\underset{\rightarrow z}{q+1}}} + e_{i^b i^p l}.$$

Compared with the general form of the multivariate regression model presented in equation **??**, the unified multivariate regression model relating to the application of the investigation of biomolecular interactions shows additional indices. In detail, the indices i and j are supplemented by the indices p and b to indicate the reference to a protein or a buffer, respectively. For a better distinction, these additional indices are written above the index i denoting the object and underneath the index j representing the descriptor variable. For the descriptor variables describing features of the proteins, the additional index s is required to denote the position in the sequence the respective variable refers to.

In this presentation of the unified multivariate regression model as well as in the following explanations, it is assumed that only one of the interacting biomolecules is modified and thus, represented by the descriptor variables. Otherwise, if the amino acid sequence of both binding partners is altered and described, two indices, i^{p_1} and i^{p_2}, are necessary to use instead of the index i^p respecting the single modified protein in order to refer to the two different biomolecules.

Furthermore, the model is presented without interaction terms. However, in order to permit a more sophisticated modelling, interaction terms between the buffer descriptor variables, the amino acid descriptor variables and between the buffer and amino acid descriptor variables could be included. The incorporation of interaction terms would result in an enormous number of descriptor variables that would have to be dealt with. But this situation

cannot be considered as a problem owing to the application of the PLS regression that is able to take a multitude of descriptor variables into account.

In practice, several various univariate regression models are established separately with respect to the different subgroups of descriptor variables for the association rate, dissociation rate and affinity constant as response variables. Thus, $3n^p$ QBKR models, 3 QSKR or 3D-QSAR models, respectively, and $3m_b$ QSPR models are obtained that can be derived from this unified multivariate regression model. The connection between these univariate regression models and the single unified multivariate one is described in detail in the following subsections.

4.3.2 Analysis of the quantitative sequence-kinetics relationship

Amino acid replacements

In the analysis of quantitative sequence-kinetics relationships (QSKR), various amino acid replacements are performed at particular positions in the sequence of one or both of the binding partners in order to investigate the effects of these modifications on the binding behaviour. Thus, a number of mutants of the wild-type protein are produced. Sometimes, modified oligopeptides are used instead of whole proteins. In the following, it is assumed that the amino acid substitutions are merely performed in the sequence of one of the binding partners.

In general, two or three mutation sites are selected and usually, up to three amino acid substitutions are realized simultaneously in one mutant. The aim of these modifications is to receive proteins showing relative different physico-chemical properties in order to cover a range as wide as possible of the values of the amino acid descriptor variables.

Potential positions for modifications are those ones where introduced mutations will probably influence the binding parameters up to a limited extent without preventing the interaction completely. This means that only non-essential positions with respect to the binding process are taken into account. Accordingly, the chosen positions for amino acid substitutions are situated at the periphery of the binding interface instead of being located at the center of the interaction interface. The interface of the interaction denotes that area of a protein that gets in direct contact with the binding partner.

Furthermore, those positions are excluded from the choice of mutation sites where modifications will presumably lead to large sterical clashes with the binding partner, folding problems or dimerization. In summary, those positions whose modifications might have an effect on the protein‘s conformation that disturbs the interaction significantly should not be chosen as locations for the amino acid substitutions. There might be also some particular amino acids that are suspected to cause these problems when they are introduced at specific positions. Hence, these amino acids are not included in the choice of conceivable amino acids for the replacements.

The suggestions of relevant positions for modifications and potential amino acids to substitute are based on the results of previously performed investigations of the biomolecules of interest. For example, knowledge about the properties of the amino acids and their expected effect on the interaction might be derived from the crystal structure of the binding partners. Evaluating the crystal structure contributes to a preliminary characterization of the biomolecules involved in the interaction and thus it provides hints for appropriate mutation sites in the amino acid sequence. Detailed explanations about the crystal structure of proteins are given in subsection **??**.

In general, individual modifications of the amino acids at those positions that show a favourable flexibility in what concerns mutations result merely in a moderate modulation of the binding characteristics. The amino acids located at positions in the center of the interface of the biomolecular interaction determine the binding behaviour predominantly. However, a significant change in the binding parameters can be obtained by modifying the amino acids simultaneously at several positions at the edge of the interface. Therefore, these positions can be used for investigations with the objective of optimizing the binding parameters of an interaction, though the effect of the most relevant positions cannot be determined by the QSKR analysis.

The amino acids of the protein under investigation are numbered according to their appearance in the sequence. The modified proteins are denoted by giving the amino acids at the mutation sites in their linear order occurring in the sequence, where the amino acids are represented by their single letter code (see table **??**). For example, if a protein is modified at the two positions x and y, the number x being smaller than the number y, by introducing the amino acid threonine at position x and the amino acid serine at position y, the resulting mutant is designated as TS.

Usually, the modified proteins are produced by substitutions with one of the 20 natural amino acids. However, it is also conceivable to introduce

non-natural amino acids. The use of non-natural amino acids in QSKR examinations would result in the possibility to achieve larger variations in the physico-chemical properties by the modifications.

The predominant restricting factor in QSKR studies is the limited amount of modified proteins that are available owing to the very cost-extensive production. Considering merely the 20 natural amino acids as substituents, 19 modifications are possible for a particular position of the sequence. Thus, if two positions are selected for amino acid replacements, 400 different proteins could be investigated. However, only a small subset of all of these possible mutants can be included in the analysis. Therefore, because of the limited number of amino acid substitutions that can be performed, the appropriate choice of the mutation sites and the amino acids used for the substitutions is of special importance.

Furthermore, the established regression models might be only valid for a subset of all conceivable mutants. Perhaps, the statements derived from the models might only be reliable for proteins with a particular feature. For example, if the amino acid substituents are chosen in order to avoid particular structural features, such as sterical clashes, conclusions based on the obtained models are not valid for those proteins that show this characteristic.

Quantification of the physico-chemical properties of amino acids

The properties of each of the modified proteins vary in accordance with the performed amino acid replacements and distinguish the mutants from each other. Analyzing the effect of the amino acid replacements by means of regression analysis requires a quantification of the physico-chemical properties of the amino acids chosen for the substitutions. In order to parameterize these features of the amino acids, quantitative variables, the so-called ZZ-scales, have been derived that can be used as descriptor variables in a regression model.

These scales have been established by Sandberg et. al. (1998) based on the results of evaluating the data collected for a sample of 87 amino acids, both natural and non-natural ones. These amino acids have been characterized by 26 variables indicating the diverse physico-chemical features of the amino acids. The values of the 26 variables used by Sandberg et. al. (1998) are given for the 20 natural amino acids in table **??**. The three ZZ-scales reflecting different types of properties have been derived from these 26 variables that are listed in the following:

- experimentally determined retention values in seven thin-layer chromatography systems (TL1-TL7)
- three nuclear magnetic resonance shift variables (NM1, NM7, NM12)
- six semi-empirical molecular orbital indices (EHOMO, ELUMO, HOF, POLAR, EN, HA)
- four indicator variables representing hydrogen bond donor and acceptor properties and side chain charge (HDONR, HACCR, Chpos, Chneg)
- total, polar and nonpolar surface area (Stot, Spol, Snp)
- molecular weight (MW)
- logP-value
- van der Waals volume of the side chain (vdW).

The ZZ1-scale represents the hydrophobicity, the ZZ2-scale describes the size and polarizability and the ZZ3-scale corresponds to diverse electronic properties like charge, polarity, electrophilicity and electronegativity of the amino acids. Thus, each amino acid, either a natural or a synthetic one, can be characterized by its values of these ZZ-scales.

Another variable that can be used additionally to describe an important property of the amino acids that might be relevant with respect to a biomolecular interaction is the helix-forming tendency (HFT)-scale. This scale indicates the tendency of an amino acid of being involved in a helical structure. The values of the three ZZ-scales and the HFT-scale taken from Andersson et. al. (2001) for the 20 natural amino acids are presented in table **??**. By giving the values of the three ZZ-scales as well as the HFT-scale for the amino acids at the mutation sites, the specific features of the mutants can be quantified.

QSKR regression models

In QSKR investigations, the data to be analyzed is obtained by measuring the binding parameters between the proteins modified in their amino acid sequence and their binding partner in a defined standard buffer. With the help of these data, a regression model describing the relationship between the physico-chemical properties of the amino acids at the mutation sites and the measured binding parameters is established by applying the PLS methodology.

Amino acid	Code	ZZ1	ZZ2	ZZ3	HFT
Alanine	A, Ala	0.24	-2.32	0.6	1.49
Arginine	R, Arg	3.52	2.5	-3.5	1.22
Asparagine	N, Asn	3.05	1.62	1.04	0.77
Aspartic Acid	D, Asp	3.98	0.93	1.93	0.92
Cysteine	C, Cys	0.84	-1.67	3.71	0.97
Glutamine	Q, Gln	1.75	0.5	-1.44	1.16
Glutamic Acid	E, Glu	3.11	0.26	-0.11	1.50
Glycine	G, Gly	2.05	-4.06	0.36	0.51
Histidine	H, His	2.47	1.95	0.26	1.00
Isoleucine	I, Ile	-3.89	-1.73	-1.71	1.00
Leucine	L, Leu	-4.28	-1.3	-1.49	1.24
Lysine	K, Lys	2.29	0.89	-2.49	1.17
Methionine	M, Met	-2.85	-0.22	0.47	1.36
Phenylalanine	F, Phe	-4.22	1.94	1.06	1.20
Proline	P, Pro	-1.66	0.27	1.84	0.49
Serine	S, Ser	2.39	-1.07	1.15	0.74
Threonine	T, Thr	0.75	-2.18	-1.12	0.79
Tryptophan	W, Trp	-4.36	3.94	0.59	1.09
Tyrosine	Y, Tyr	-2.54	2.44	0.43	0.79
Valine	V, Val	-2.59	-2.64	-1.54	0.99

Table 4.1: Values of the three ZZ-scales and the HFT-scale and code of the 20 natural amino acids

Since the physico-chemical properties of the amino acids incorporated in the substitutions are represented by the ZZ-scales as well as the HFT-scale, these variables are used as the descriptor variables in the regression model. The QSKR models are developed separately for the association rate, the dissociation rate and the affinity constant. Therefore, these binding parameters present the response variables of the univariate models. Consequently, the regression equation for the l-th response variable is of the following form:

$$y_{i^b i^p l} = b_{0l} + \sum_{j_p=(r+1)_p}^{m_p} \sum_{s=1}^{q} b_{(j_p)_s l} x_{i^p (j_p)_s} + e_{i^b i^p l},$$

where

$$i^p = 1^p, \ldots, n^p$$

with $l \in \{1, \ldots, k\}$ and i^b, representing the standard buffer, being fixed for each of the $k = 3$ univariate regression models. Alternatively, the QSKR

model can be presented as

$$y_{i^b i^p l} = b_{0l} + b'_{\underset{\rightarrow(m_p)}{((r+1)_p)_q l}} x_{\underset{\rightarrow(m_p)}{i^p((r+1)_p)_q}} + e_{i^b i^p l}.$$

In this equation, the vector $x_{\underset{\rightarrow(m_p)}{i^p((r+1)_p)_q}}$ denotes the vector summarizing the values of the descriptor variables referring to the physico-chemical properties of the amino acids at the q mutation sites, and the vector $b_{\underset{\rightarrow(m_p)}{((r+1)_p)_q l}}$ represents the vector containing the corresponding regression coefficients, i.e.:

$$x_{\underset{\rightarrow(m_p)}{i^p((r+1)_p)_q}} \begin{pmatrix} x_{i^p((r+1)_p)_1} \\ \vdots \\ x_{i^p((r+1)_p)_q} \\ x_{i^p((r+2)_p)_1} \\ \vdots \\ x_{i^p(m_p)_q} \end{pmatrix} \quad \text{and} \quad b_{\underset{\rightarrow(m_p)}{((r+1)_p)_q l}} = \begin{pmatrix} b_{((r+1)_p)_1 l} \\ \vdots \\ b_{((r+1)_p)_q l} \\ b_{((r+2)_p)_1 l} \\ \vdots \\ b_{(m_p)_q l} \end{pmatrix}.$$

In this QSKR model, the term $y_{i^b i^p l}$ denotes that value of the l-th response variable that is measured for the interaction involving the i^p-th protein and taking place in the standard buffer. Thus, the sample size equals the number n^p of different proteins incorporated in the experiments.

The value $x_{i^p(j_p)_s}$ represents that value of the j_p-th amino acid descriptor variable that refers to the amino acid situated at the s-th mutation site of the i^p-th protein. The numbering of the mutation sites is performed according to their order in the amino acid sequence. For example, substitutions of amino acids are realized at positions 20 and 30. Then, the index $s = 1$ indicates position 20 and the index $s = 2$ denotes position 30. Accordingly, the number q corresponds to the number of mutation sites.

Furthermore, the number $m_p - r_p$ of different descriptor variables is 4 because the three ZZ-scales and the HFT-scale are the descriptors incorporated generally in the regression model. The number r_p refers to the number of descriptor variables referring to structural features that might be included in regression models in biomolecular interaction studies (see the following subsection). In case of the QSKR analysis, the term r_p attains the value zero since only physico-chemical properties are included in the regression model. However, in the presentation of the QSKR model, the index j_p is determined to begin with the value $(r + 1)_p$ in order to show that the QSKR model can be derived from the unified multivariate model that considers r_p structural descriptor variables.

Altogether, the regression models derived in QSKR examinations comprise $4q$ descriptor variables. Thus, the number of descriptor variables increases according to the number of mutation sites. This fact illustrates the usefulness of the application of PLS regression to data obtained in experiments involving a number of mutation sites.

The constant term of the regression equation for the l-th response variable is denoted by b_{0l}. According to the number of descriptor variables, $4q$ corresponding regression coefficients are incorporated in each of the three univariate regression models. The regression coefficient $b_{(j_p)_s l}$ refers to the j_p-th amino acid descriptor variable respecting the amino acid at the s-th mutation site relating to the l-th response variable. Furthermore, the term $e_{i^b i^p l}$ denotes the error regarding the l-th response variable and the i^p-th interaction in the standard buffer i^b.

The univariate QSKR models that are established separately for each binding parameter result from the unified multivariate regression model by fixing the index i^b that refers to the standard buffer used in the course of the experiments. Furthermore, the amino acid descriptor variables respecting the physico-chemical properties of the amino acids at the mutation sites are merely considered. Therefore, the index z of positions of interest is set to the number q of mutation sites. In what concerns the descriptor variables, the number m_b of buffer descriptor variables included in the model equals zero, as the number r_p of amino acid descriptor variables referring to structural characteristics does.

Interpretation of the QSKR models

Subsequent to the establishment of the three univariate QSKR models, the effect of the modifications in the amino acid sequence on the interaction can be analyzed with the help of the estimated regression coefficients. Thus, it is possible to determine which physico-chemical properties at which positions of the amino acid sequence influence the binding between the two interacting biomolecules relevantly.

The contributions of the physico-chemical properties to the interaction can be assessed by relating the descriptor variables, the ZZ-scales as well as the HFT-scale, to the properties they describe, i.e. the hydrophobicity, the size, the electronic properties and the helix-forming tendency of the amino acids. Then, conclusions can be drawn on the amino acids that should be used for the replacements at the positions taken into account as they show the pre-

ferred physico-chemical properties that are required to improve the binding behaviour. For example, it can be suggested whether a small or a large, a hydrophilic or a hydrophobic, a polar or an apolar, an electrophilic or a non-electrophilic amino acid should be present at the particular positions. Consequently, the quantification of the influence of the physico-chemical properties at the mutation sites on the binding process results in an improved understanding of the interaction under investigation.

Furthermore, with the help of the regression coefficients, values for the binding parameters can be predicted with respect to given modifications in the amino acid sequence, i.e. for specified physico-chemical properties of the amino acids at the mutation sites taken into consideration. Therefore, statements concerning the properties that are desirable at particular sites of the sequence in order to obtain specific values of the association and the dissociation rate constant and hence, the affinity constant, can be derived.

If the processes of association, dissociation and affinity are influenced by different physico-chemical properties at the mutation sites, amino acids with features providing a compromise concerning the optimization of the three binding parameters are chosen for the substitutions at the positions of interest. In order to determine the best compromise, the extents and directions of the effects of the various properties are considered. Probably, a number of amino acids can be proposed that are supposed to cause an improved binding behaviour.

The determination of the influence of the amino acid replacements can only be performed for those sites of the sequence that have been submitted to modifications. Since those positions are chosen for the amino acid substitutions that are located at the edge of the interaction interface, it is possible that only limited effects on the binding parameters might be caused by the modifications. Therefore, a QSKR analysis is merely of use when the performed amino acid replacements result in a sufficiently notable change of the binding parameters. However, the major factors contributing to the binding behaviour cannot be detected by the QSKR approach because the effects of amino acid substitutions at the predominantly influencing positions are not investigated.

The conclusions that can be drawn from the QSKR models on the modifications in the amino acid sequence that might lead to an improved binding behaviour can be useful in the context of designing new biomolecules as drugs. In general, new potential drug molecules, for example antibodies, are

obtained using libraries of native or synthetic compounds. Often, the binding properties between these biomolecules and the corresponding binding partner, e.g. the specific antigen, are not sufficiently satisfying. With the help of the knowledge derived from the QSKR models, suggestions can be made on how to alter the amino acid sequence of the potential drug in order to optimize the binding characteristics of the interaction with the binding partner of interest.

4.3.3 Analysis of the 3D-quantitative structure-activity relationship

The aim of 3D-quantitative structure-activity relationship (3D-QSAR) examinations is to determine which stereochemical features and physico-chemical properties of amino acids at which positions have an effect on the interaction under investigation. The 3D-QSAR investigation is performed instead of the more simple QSKR analysis when the influence of structural features of the interacting biomolecules on the binding behaviour is meant to be determined in addition to the effect of the physico-chemical properties considered in QSKR modelling.

Therefore, in the 3D-QSAR analysis, the information about the physico-chemical properties of the amino acids used for the substitutions, i.e. the ZZ-scales and the HFT-scale incorporated in QSKR models, is supplemented by structural information about a number of positions in the amino acid sequence. The structural properties of amino acids located in or near the binding interface are of special interest.

The combination of structural information and knowledge about physico-chemical properties applied in 3D-QSAR modelling permits a more comprehensive characterization of the biomolecules under examination than it would have been obtained had only one type of information been incorporated. Consequently, the inclusion of descriptors of the structure of the proteins might result in a model providing more accurate predictions than a model for only physico-chemical properties would be able provide.

However, the information necessary to deal with in 3D-QSAR investigations is more complex than that used in the QSKR analysis. In contrast to the physico-chemical properties, the structure of a protein is not easy either to quantify by relevant descriptor variables or to alter in a controlled way.

Especially, the performance of 3D-QSAR modelling with respect to protein-protein interactions is more challenging compared with the analysis of interactions involving only small molecules. The reason is that the description of the structure of a complex protein can only be performed partially, whereas the structure of a small compound can be represented more completely. Since the usefulness of regression models relies on taking into account relevant descriptor variables, the adequate representation of structural features as well as the appropriate choice of positions of the sequence whose structure is described is crucial to the success of the research of a protein-protein interaction.

Quantification of structural features

The description of the structure of proteins refers to statements concerning the experimentally determined configuration of the biomolecules. Furthermore, spatial constraints can be taken into consideration in 3D-QSAR investigations.

In general, the variables used to represent the conformation of proteins are the $x-, y-$ and $z-$coordinates, measured in angstrom (Å), referring to the backbone and the side chain, respectively, of the amino acids at the chosen positions. Thus, each amino acid is characterized by 6 values of the corresponding coordinates.

In particular, features of the backbone are quantified by giving the coordinates of the α-carbons (c_α) of the amino acids. Furthermore, the side chain of an amino acid is described by the coordinates of its center that is calculated as the mean of the coordinates of all atoms in the side chain. These descriptors included in the 3D-QSAR analysis are obtained with the help of the crystal structure of the proteins of interest.

The x-ray crystallography provides information concerning the structure of both single proteins and complexes of proteins. It is performed by preparing crystals of the relevant proteins or protein complexes. The arrangement of the atoms in the cells of these crystals is determined by the structure of the proteins. Subsequently, x-rays are directed toward these crystals diffracting this radiation according to the arrangement of the atoms in the crystal. Therefore, a characteristic profile of the diffraction is obtained that presents the basis for the development of a map of the density of the electrons in the crystal. The structure of the examined protein or protein complex can be derived with the help of this map.

A structural characterization is determined experimentally for the wild-type protein and for each mutant. A comparison of these different structures reveals the influence of the performed amino acid substitutions on the structural features, i.e. how the coordinates of the amino acids taken into account are changed as a result of the modifications. The differences in the structural effects observed between the different mutants are caused by particular features of the amino acids used for the replacements.

It is possible that also the coordinates of those amino acids that are not located at the mutation sites are altered owing to the introduced modifications since the amino acid replacements at a particular position might lead to changes in the structure of the proteins at diverse positions. This is the reason why in 3D-QSAR modelling, structural features of amino acids at mutation sites as well as additional positions are represented by descriptor variables. Consequently, the effect of changes of structural features of amino acids at any position of interest can also be determined by 3D-QSAR regression models.

In case of observing only small deviations of the structure at a particular position for the different mutants, the amino acid located there is obviously fixed in its location. If an amino acid is too restrained in its position, i.e. its coordinates are not changed by the introduced modifications, its relevance to the interaction cannot be determined by the 3D-QSAR analysis. Nevertheless, this amino acid might contribute in a significant way to the interaction of interest.

Beyond the quantification of structural features at several positions of interest, diverse additional conclusions concerning structural properties of the proteins can be drawn from the crystal structure of the complex of the interacting biomolecules, For example, it can be determined which amino acids of the binding partners are directly in contact and whether the amino acids are involved in hydrogen bonds or disulfide (S-S) binding. Furthermore, it can be stated whether amino acids can move quite freely without influencing the interaction or if they are spatially restrained and thus, are not able to vary their position. Some amino acids might be found to be located in a pocket of the binding partner, for example. Beyond this, helical structures can be detected. All these considerations provide important insights in the binding process leading to an improved knowledge about the interaction under study. Therefore, the evaluation of the crystal structure contributes to the choice of conceivable mutation sites and amino acids that might be appropriate for the replacements.

An alternative but more complicated and hence not very common approach to obtain descriptors of the 3D-structure of proteins is Comparative Molecular Field Analysis (CoMFA). This method provides a more detailed characterization of the structural features than that one based on the coordinates of the backbone and side chains of the amino acids. Therefore, when the descriptors calculated by CoMFA are used, more descriptor variables have to be dealt with compared with the number of descriptors derived from the crystal structure.

3D-QSAR regression models

The three univariate 3D-QSAR models are similar to the QSKR regression models. However, some terms have to be incorporated additionally in 3D-QSAR models. These supplementary terms refer to the information about structural features of the amino acids at some positions in the sequence. In the following presentation, it is assumed that structural properties of only one of the binding partners are taken into account.

The 3D-QSAR models are built separately for the association and the dissociation rate constant as well as for the affinity constant that present the response variables of the univariate regression models. The 3D-QSAR regression equation respecting the l-th response variable is of the following form:

$$y_{i^b i^p l} = b_{0l} + \sum_{j_p=1_p}^{m_p} \sum_{s=1}^{q} b_{(j_p)_s l} x_{i^p (j_p)_s} + \sum_{j_p=1_p}^{r_p} \sum_{s=q+1}^{z} b_{(j_p)_s l} x_{i^p (j_p)_s} + e_{i^b i^p l},$$

where

$$i^p = 1^p, \ldots, n^p; \quad r_p < m_p; \quad q < z$$

with $l \in \{1, \ldots, k\}$ and i^b, representing the standard buffer, being fixed for each of the k univariate regression models.

Alternatively, the 3D-QSAR regression model can be presented more compactly with the help of the column vectors introduced in subsection **??** as follows:

$$y_{i^b i^p l} = b_{0l} + b'_{(m_p)_q l} x_{i^p (m_p)_q} + b'_{(r_p)_{\underset{\to z}{q+1}} l} x_{i^p (r_p)_{\underset{\to z}{q+1}}} + e_{i^b i^p l}.$$

The value $y_{i^b i^p l}$ of the l-th response variable, the constant term b_{0l} and the error term $e_{i^b i^p l}$ can be interpreted by analogy with the QSKR model.

The descriptors of the physico-chemical and structural properties of the

amino acids at the mutation sites as well as the descriptors of the structural features of the amino acids at a number of additional positions of interest present the descriptor variables of the 3D-QSAR regression models. For an amino acid at a mutation site, 10 descriptor variables are included in the regression model since this amino acid is represented both by its physico-chemical properties and its structural coordinates. An amino acid at a position of the sequence where no modifications are performed is described by its 6 coordinate values. Thus, the total number of descriptor variables and accordingly, of regression coefficients, is $4q + 6q + 6(z - q) = 4q + 6z$, where z is the number of all considered positions in the sequence and q the number of mutation sites. This term shows that in 3D-QSAR modelling, there is a huge number of descriptor variables that have to be dealt with during the establishment of the regression models. Consequently, the application of PLS regression is of particular use in this kind of research.

The numbering of the positions taken into account is based on their order in the protein's sequence according to the procedure performed in the QSKR modelling. Hence, if, for example, amino acid substitutions are introduced at positions 20 and 30, these positions are indicated by $s = 1$ and $s = 2$, respectively, and the number q of mutation sites equals 2. If the structural information of the amino acids at position 10, 25 and 60 is considered additionally, these positions are referred to by the indices $s = 3$, $s = 4$ and $s = 5$, respectively. In this case, the total number z of positions where amino acids are represented by descriptor variables is 5.

The total number m_p of different descriptor variables is given as 10, as the number $r_p = 6$ of structural descriptors, i.e. the coordinates, are added to the value $m_p - r_p$, i.e. the number of descriptors of the physico-chemical properties. For the backbone of the amino acid, the $x-$coordinate is indicated by $j_p = 1_p$, the $y-$coordinate by $j_p = 2_p$ and the $z-$coordinate by $j_p = 3_p$. In what concerns the side chain of the amino acid, the index $j_p = 4_p$, 5_p and 6_p refers to the $x-, y-$ and $z-$coordinate, respectively. Furthermore, the index j_p equals $(r + 1)_p = 7_p$ for the ZZ1-scale, $(r + 2)_p = 8_p$ for the ZZ2-scale, $(r + 3)_p = 9_p$ for the ZZ3-scale, and $(r + 4)_p = m_p = 10_p$ for the HFT-scale.

Therefore, the term $x_{i^p(j_p)_s}$, with $j_p = 1_p, \ldots, r_p$, denotes the value of the coordinate that is indicated by the value j_p and describes the amino acid at the s-th of the z positions of interest. Accordingly, the value $x_{i^p(j_p)_s}$, with $j_p = (r + 1)_p, \ldots, m_p$, refers to the corresponding value of one of the 4 descriptor variables of the physico-chemical properties, the 3 ZZ-scales and the HFT-scale, of the amino acid at the s-th of the q mutation sites.

The univariate 3D-QSAR models can be derived from the unified multivariate regression model by omitting the buffer descriptor variables and the respective regression coefficients. Furthermore, according to the QSKR model, the index i^b is fixed since the measurements are performed in one single standard buffer. Beyond this, the index l determining the response variable is fixed for each univariate regression model.

Interpretation of 3D-QSAR models

Evaluating the regression coefficients of the univariate 3D-QSAR models reveals the influence of the ZZ-scales, the HFT-scale as well as the 3D-descriptors on the association and dissociation rate constant and on the affinity constant. With the help of the regression coefficients, it can be determined whether the structural features taken into consideration have an effect on the interaction under examination beyond the influences of the physicochemical properties. Therefore, additionally to the information about the effect of the ZZ-scales and the HFT-scale as obtained merely by establishing QSKR models, the influence of the localization of the α-carbon and the side chain of amino acids at several positions can be quantified with respect to the interaction of interest.

For the l-th response variable, the regression coefficients relating to the $x-, y-$ and $z-$coordinates of the backbone or the side chain, respectively, of the amino acid at a particular position s can be summarized in the vectors $b_{bb_s,l}$ and $b_{sc_s,l}$, i.e.

$$b_{bb_s,l} := \begin{pmatrix} b_{(1_p)_s l} \\ b_{(2_p)_s l} \\ b_{(3_p)_s l} \end{pmatrix} \quad \text{and} \quad b_{sc_s,l} := \begin{pmatrix} b_{(4_p)_s l} \\ b_{(5_p)_s l} \\ b_{(6_p)_s l} \end{pmatrix}.$$

In order to quantify the relative importance of the backbone, $rel.imp._{bb_{s,l}}$, or the side chain, $rel.imp._{sc_{s,l}}$, of an amino acid at a specific position s with respect to the l-th response variable, it is common to calculate the roots of the sums of the corresponding squared regression coefficients, i.e. the following terms:

$$rel.imp._{bb_{s,l}} = \sqrt{b^2_{(1_p)_s l} + b^2_{(2_p)_s l} + b^2_{(3_p)_s l}}$$

and

$$rel.imp._{sc_{s,l}} = \sqrt{b^2_{(4_p)_s l} + b^2_{(5_p)_s l} + b^2_{(6_p)_s l}}.$$

Comparing the results obtained for the univariate models for the different response variables permits conclusions on whether the recognition of the biomolecules, the stability of the formed complexes or the binding strength are influenced by different factors and are hence dependent on different physical processes.

Consequently, with the help of the regression coefficients, it can be determined which backbone or side chain of the amino acid at which position of the sequence should be moved in which direction in order to obtain either a lower or higher respective binding parameter.

4.3.4 Analysis of the quantitative buffer-kinetics relationship

Descriptor variables in QBKR models

The interaction between two biomolecules is influenced by the chemical environment, the buffer, in which the binding takes place. Thus, the binding properties can be altered by varying the concentrations of the ingredients the buffer contains or the pH value. A number of chemical additives possibly have an effect on the association and dissociation rate constant as well as the affinity constant.

Depending on the biomolecules under investigation, chemical components that might be suspected to influence the binding behaviour without disturbing the interaction completely are for example sodium chloride (NaCl), dimethyl sulfoxide (DMSO), ethylenediaminetetraacetic acid (EDTA), urea, potassium thiocyanate (KSCN), methyl sulfat (Me_2SO) and the pH value. The choice of the factors whose effect on the binding is examined should be based on existing knowledge about the interacting biomolecules and their kinetic parameters.

The number and kind of different concentrations of each of the chemical additives and levels of the pH value that should be incorporated in the experiments are selected according to the supposition concerning the effect that might be caused by and depend on the purpose of the study. In order to obtain knowledge about the binding between two biomolecules in vivo, for example, the buffer composition used during the experiments should be chosen to reflect the chemical environments existing in the corresponding cells of an organism. If it is the case that the variations in the buffer composition result merely in limited changes of the binding behaviour, the experiments are of little use for the derivation of regression models.

In order to quantify the effect of the presumed relevant buffer ingredients, their adjustments are varied simultaneously. This results in a number of so-called perturbation buffers with determined compositions, i.e. concentrations of the chemical additives and levels of the pH value. One buffer that is known to provide a chemical environment in which the interaction takes place is defined to be the standard buffer. Kinetic measurements are performed for the interactions between the wild-type as well as several mutants of this protein and their binding partner in the perturbation buffers and the standard buffer.

In order to determine the reproducibility of the realized measurements, the binding parameters can be collected and compared for buffers with identical concentrations of the ingredients and pH values.

QBKR regression models

The data obtained in the experiments performed during the analysis of quantitative buffer-kinetic relationships (QBKR) are used to establish regression models by the application of the PLS methodology. The QBKR regression equations present the relationship between the measured binding parameters and the concentrations of the chemical additives and the level of the pH value in the different buffers.

In practice, the univariate QBKR regression models are derived separately for each of the different interactions involving one of the various proteins. The association and dissociation rate constants as well as the affinity constant are used as the response variables of the univariate models, whereas the variables quantifying the buffer composition present the descriptor variables.

The value $y_{i^b i^p l}$ of the l-th response variable measured for the interaction involving the i^p-th protein in the i^b-th buffer can be given as:

$$y_{i^b i^p l} = b_{0l,i^p} + \sum_{j_b=1_b}^{m_b} b_{j_b l,i^p} x_{i^b j_b} + e_{i^b i^p l},$$

where

$$i^b = 1^b, \ldots, n^b$$

with $l \in \{1, \ldots, k\}$ and $i^p \in \{1^p, \ldots, n^p\}$ being fixed for each of the $i^p l$ univariate regression models. The QBKR regression model can be expressed as well as the following equation using the column vectors presented in subsection **??**:

$$y_{i^b i^p l} = b_{0l,i^p} + b'_{m_b l,i^p} x_{i^b m_b} + e_{i^b i^p l}.$$

In the context of QBKR regression models, the number of the sample size equals the number n^b of buffers used in the experiments. Furthermore, the number of descriptor variables is given as the number m_b of buffer descriptor variables, i.e. the number of chemical components, including the pH value, whose adjustments are varied. The number of different regression models developed for each of the three binding parameters corresponds to the number of different interactions that are investigated, i.e. the number n^p of different proteins incorporated in the experiments. Consequently, the total number of regression models that are established during the QBKR analysis is $3n^p$.

In the QBKR regression equation, the value $x_{i^b j_b}$ represents the value of the j_b-th buffer descriptor variable, i.e. the concentration of the corresponding chemical additive or the level of the pH value, respectively, in the i^b-th buffer. Furthermore, the expression $e_{i^b i^p l}$ denotes the error concerning the i^b-th buffer, the i^p-th interaction and the l-th response variable.

The constant term b_{0l,i^p} represents the intercept of the univariate regression model of the l-th response variable for the i^p-th interaction. The regression coefficient referring to the j_b-th buffer descriptor variable in the model of the i^p-th interaction referring to the l-th response variable is denoted by $b_{j_b l,i^p}$. Accordingly, the term $b_{j_b l,i^p}$ indicates that quantity the l-th response variable is increased by with respect to the i^p-th interaction by increasing the j_b-th descriptor variable by one unit.

By establishing these regression models relating the buffer composition for a given interaction either to the association rate constant, the dissociation rate constant or the affinity constant, the effect of the different buffer components on the binding behaviour can be quantified with the help of the estimated regression coefficients. Consequently, a characterization of the interaction is obtained since the dependence of the binding behaviour on the chemical environment is determined.

Furthermore, the regression models under study can be used to predict values of the binding parameters for a given interaction respecting particular combinations of values of the buffer descriptor variables that have not been incorporated in the experiments. However, reliable predictions by applying the developed regression equations can only be guaranteed for values of the descriptor variables within the range of adjustments of the buffer components that is covered by the realized experiments.

The QBKR models can be derived from the unified multivariate regression model by fixing the index i^p for each of the n^p univariate models built with respect to the three binding parameters. Furthermore, the different kinds of descriptor variables incorporated in the unified model are reduced to those ones representing the buffer composition.

The establishment of one single multivariate regression model including interaction terms between the buffer and amino acid descriptor variables would provide more sophisticated statements compared to those ones obtained by the $3n^p$ QBKR models. The reason is that these models can only be used to determine the effects of the different buffer descriptor variables on the binding behaviour for each of the interactions incorporating one of the mutants. In contrast to the unified multivariate regression model, the QBKR models are not able to provide explanations concerning the properties of the proteins that cause the observed differences in the influences of the buffer descriptor variables on the kinetic parameters with respect to the different interactions. Therefore, in order to obtain this additional information, quantitative sequence-perturbation relationship (QSPR) models, described more detailed in subsection **??**, are required to establish subsequent to the development of the QBKR models in practice.

Interpretation of the QBKR models

Since the QBKR models relate the binding parameters to the buffer composition, variations of the binding behaviour can be explained by variations of the concentrations of the chemical additives and the levels of the pH value. Analyzing the regression coefficients of the developed regression models leads to conclusions concerning the influence of the buffer components on the interactions involving the different mutants. Hence, the sensitivity of the binding process under examination respecting varying adjustments of the chemical environment can be determined.

Comparing the results obtained in the different regression models referring to the various mutants permits the detection of differences and similarities between the effects of the concentrations of chemical additives and the levels of the pH value on the interactions. Based on the QBKR models derived separately for the association and the dissociation rate constant as well as for the affinity constant, the dependence of these binding parameters on the buffer components can be analyzed. If the association rate and the dissociation rate constant are influenced by different factors, it can be concluded that different binding forces contribute to the process of recognition between

the interacting biomolecules and the stability of the complexes formed by the binding partners.

Four different intermolecular forces can be distinguished that might be relevant to an interaction, namely electrostatic forces, interactions between Lewis acids and Lewis bases, van der Waal's forces and direct hydrogen bonds. The knowledge about the chemical factors having an effect on the binding is the basis of conclusions relating to the forces that contribute to the interaction under study. In the following, some examples describing what kind of information can be derived from QBKR models are given.

If variations of the pH value ranging from values of 7.0 to 7.8 do not have an effect on the binding parameters, it can be concluded that it is unlikely that histidines are involved in the binding process. Furthermore, if EDTA influences the interaction, there are probably metal ions contributing to the binding. Otherwise, it can be suggested that hydrogen bonds or ionic interactions are relevant. The effect caused by increasing concentrations of NaCl permits statements concerning the contribution of electrostatic forces to the interaction. The determination of the influence of KSCN can result in suggestions respecting the water structure present in the complexes of bound biomolecules. The importance of hydrophobicity during the binding process can be judged by evaluating the effect of DMSO on the binding behaviour.

Desirable binding properties between the interacting biomolecules can be achieved by adjusting particular chemical additives to specified levels in the experimental conditions according to the results of the QBKR models. For example, the conclusions from the QBKR analysis are relevant to the affinity chromatography since the knowledge gained with the help of the regression models can be used to improve the settings for this method. In the application of the affinity chromatography, a rapid dissociation is required, i.e. a low stability of the formed complexes of the biomolecules. Thus, respecting this procedure, it is helpful to obtain knowledge in QBKR studies about those chemical additives by which the interaction between the two biomolecules of interest can easily be disturbed.

Furthermore, using the statements derived from the QBKR models concerning the influence of the chemical additives on the interaction, conclusions can be drawn respecting the compounds of the chemical environment whose concentrations should be controlled during an interaction experiment in order to avoid undesired effects on the binding. Beyond this, the characterization of a biomolecular interaction under varying experimental conditions, i.e. for

several levels of the chemical additives instead of only one adjustment of the buffer components, facilitates comparisons of the binding properties between different interactions. This is a desirable performance with respect to a standardized exchange of knowledge about binding characteristics of diverse interactions in data bases.

The chemical sensitivity fingerprint

In order to present the results of the QBKR analysis in a compact manner, the regression coefficients of the numerous univariate regression models are usually summarized in so-called chemical sensitivity fingerprints. These fingerprints can be determined for each regression model, i.e. respecting each mutant under investigation and each of the binding parameters.

The fingerprints are computed by dividing the regression coefficients referring to the different buffer descriptor variables by the constant term of the respective equation and summarizing the resulting values in a vector. Thus, the chemical sensitivity fingerprint fpr_{l,i^p} for the i^p-th interaction referring to the l-th response variable is given as

$$fpr_{l,i^p} := \begin{pmatrix} \frac{b_{1_b l,i^p}}{b_{0l,i^p}} \\ \frac{b_{2_b l,i^p}}{b_{0l,i^p}} \\ \vdots \\ \frac{b_{m_b l,i^p}}{b_{0l,i^p}} \end{pmatrix}.$$

Each element of the vector reflects the extent of the effect of one of the chemical additives or the pH value, respectively, on the corresponding binding parameter in relation to the corresponding intercept. A negative value of the intercept cannot be obtained since the binding parameters do not attain negative values. Thus, a positive sign of an entry of the chemical sensitivity fingerprint indicates a positive effect of the corresponding buffer descriptor variable and vice versa.

The fingerprints derived from the univariate QBKR regression models present a unique characterization of the interactions under study as they contain the relevant information concerning the quantified effects of the buffer descriptor variables in a compact form. Consequently, the fingerprints can be used to compare the results obtained for the different modified proteins. Since the

results of QBKR investigations become comparable by this presentation, differences and similarities respecting the influence of the buffer composition on the interactions, i.e. the sensitivity of the binding process in what concerns changes in the chemical environment, can easily be determined. Therefore, collecting the fingerprints of various kinds of interactions facilitates the exchange of information about biomolecular interactions.

4.3.5 Analysis of the quantitative sequence-perturbation relationship

In order to relate the descriptors of the physico-chemical properties of the amino acids at the mutation sites to the sensitivity of the binding behaviour to changes in the concentration of a particular chemical additive or the pH value of the buffer, quantitative sequence-perturbation relationship (QSPR) models are established in practice. With the help of these models, it is possible to explain on which physico-chemical properties of the amino acids at which of the mutation sites the sensitivity of the interactions to the varying adjustments in the chemical environment depends. The univariate QSPR regression models are derived separately for each of the m_b buffer components by applying the PLS regression. The response variable refers either to the association or the dissociation rate constant or the affinity constant.

The QSPR regression model respecting the j_b-th chemical component and the l-th response variable is of the following form:

$$\frac{b_{j_b l,i^p}}{b_{0l,i^p}} = b_{0l,j_b} + \sum_{j_p=(r+1)_p}^{m_p} \sum_{s=1}^{q} b_{(j_p)_s l} x_{i^p(j_p)_s} + e_{i^p l,j_b},$$

$$\text{where} \quad i^p = 1^p, \ldots, n^p$$

with $l \in \{1, \ldots, k\}$ and $j_b \in \{1_b, \ldots, m_b\}$ being fixed for each of the $m_b k$ univariate regression models. Another presentation of the QSPR model can be given under incorporation of the column vectors defined in subsection **??**, i.e.:

$$\frac{b_{j_b l,i^p}}{b_{0l,i^p}} = b_{0l,j_b} + b'_{\substack{((r+1)_p)_q l \\ \to (m_p)}} x_{\substack{i^p((r+1)_p)_q \\ \to (m_p)}} + e_{i^p l,j_b}.$$

The ZZ-scales as well as the HFT-scale of the amino acids at the q mutation sites present the descriptor variables of the QSPR models. Furthermore, the elements respecting the j_b-th buffer descriptor of the chemical sensitivity fingerprints referring to the l-th binding parameter and the n^p mutants provide the values of the response variable of the corresponding regression model.

The constant term b_{0l,j_b} denotes the intercept of the univariate regression model respecting the l-th binding parameter and the j_b-th buffer component. The values of the descriptor variables and the regression coefficients of the QSPR models regarding to the m_b different buffer components can be interpreted according to those ones of the QSKR model. The error term $e_{i^p l,j_b}$ refers to the i^p-th interaction, the j_b-th buffer component and the l-th binding parameter. Furthermore, equivalently to the QSKR models, the number q represents the number of mutation sites.

Analyzing the regression coefficients of the QSPR models reveals the effects of the different physico-chemical properties of amino acids at the specific positions of the sequence on the sensitivity of the interaction to a particular chemical compound. Based on this information, it is possible to suggest which physico-chemical properties the amino acids at the chosen sites of the sequence should show in order to obtain a biomolecule with a predefined sensitivity to the different buffer components.

The form of the QSPR model cannot be derived directly from the unified multivariate regression model though the descriptor variables are the same as those ones included in the QSKR model. But the individual response variables refer to estimations of regression coefficients of the QBKR models instead of measurements of the binding parameters as it is the case in the multivariate model. However, as already explained in the context of the QBKR analysis, knowledge obtained with the help of the unified multivariate regression model incorporating interaction terms of the buffer and amino acid descriptor variables can be considered to be equivalent to the information the QSPR models provide.

Chapter 5

Data analysis

The data analyzed in the following have been made available to us by Karl Andersson who described and modelled these data in the publication "Predicting the kinetics of peptide-antibody interactions using a multivariate experimental design of sequence and chemical space" from 2001.

The interaction under investigation is the binding between an antigen and its antibody, namely a peptide of the antigen of the tobacco mosaic virus protein (TMVP) and a Fab fragment of the monoclonal antibody 57P. The relevance of studying this interaction is explained in the following section. This is followed by a description of the data obtained by Biacore measurements of the kinetic constants of this interaction under various circumstances.

Section **??** gives the results of the analysis of the interaction between the tobacco mosaic virus protein and a Fab fragment of the antibody 57P performed by Andersson et. al. (2001). The reproduction of the models developed in this publication using the PLS regression is described in section **??** and section **??** investigates the usefulness of the novel aspects of the modelling of biomolecular interactions by applying various modifications of PLS regression analysis.

5.1 Biochemical motivation

The following explanations concerning the components of the tobacco mosaic virus (TMV) as well as its illustration in figure **??** are taken from the internet page of wikipedia (www.wikipedia.org). Viruses consist of a so-called capsid surrounding either RNA or DNA. The TMV belongs to the RNA viruses. Its single RNA-strand comprises 6401 bases.

The capsid is a coat protein whose function is the protection of the RNA from the action of cellular enzymes. Furthermore, it determines the antigenicity of the virus, i.e. which antibodies are able to bind to the virus. The capsid of the TMV contains 2134 protein monomers. Each of these monomers has a length of 158 amino acids. The RNA-strand and its capsid assemble by themselves to functional viruses showing a helical rod-structure as it can be seen in figure **??**.

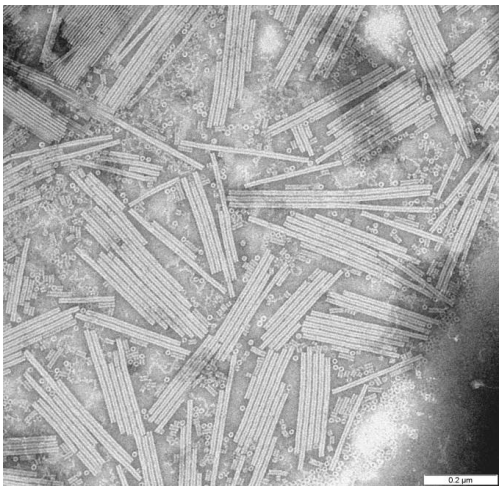

Figure 5.1: Electron microphotograph of TMV particles

As explained in more detail in the following, several mutations were introduced in a peptide of a protein monomer of the TMV capsid by substituting the amino acids at three positions in the sequence. Subsequently, the kinetic parameters of the interaction between these modified peptides and the antibody 57P were measured in different chemical environments. First, this procedure of modifying the amino acid sequence of the antigen instead of that one of the antibody might be considered surprising. Often, a biomolecular interaction analysis is realized in order to identify that modified antibody showing optimal binding features for a particular antigen. Therefore, in these cases, the kinetic experiments incorporate modified biomolecules of the antibody instead of modified peptides of the antigen.

However, for the interaction under investigation, another kind of question is to be answered. In fact, the aim of the current analysis is to obtain knowledge about the circumstances, especially the amino acid sequence of the TMVP,

under which the available antibody 57P can still be expected to be effective. In particular, by modelling the relationship between the physico-chemical properties of the amino acids at the mutation sites and the resulting kinetic constants, it is possible to predict the binding characteristics of this interaction for diverse potential mutations. Further, by considering different buffer compositions as well in the experiments, the statements concerning the influence of the mutations at the chosen mutation sites on the binding behaviour can be determined for several chemical environments that might occur in the cells of an organism.

The motivation for this kind of research is the fact that mutations in the RNA or DNA, respectively, of viruses might take place that result in the expression of modified proteins of the capsid. Consequently, the features of the coat protein might be altered. In unfavourable cases, the existing antibodies are not able to bind to the virus any longer or at least not as effectively as usual because of these changes in the capsid.

Mutations in viruses can be caused either by the so-called antigenic drift or the antigenic shift. If a mutation at a single position in the sequence occurs randomly, this situation is called antigenic drift. Often, these mutations do not influence the expression of the protein monomers and hence the features of the capsid. Furthermore, usually, several mutations are necessary to lead to relevant changes in the coat protein. This is the reason why different mutation sites were taken into account in the kinetic experiments.

The antigenic shift refers to those mutations resulting from the combinations of the RNA or DNA, respectively, of two different types of viruses being present simultaneously in the same organism. The resulting virus particles consist of the RNA or DNA, respectively, containing the new combinations of bases. In general, mutations caused by the antigenic shift lead to notably changes in the expressed protein monomers of the capsid. Consequently, the new viruses show quite different features of the coat protein compared to those of capsids of the original virus types.

With the help of the results of a regression analysis respecting the virus of interest or a similar one, the binding characteristics of the interaction between an occurring mutant of the virus and the existing antibody can be predicted since the amino acid sequence of the mutant can be determined in laboratories. Thus, the relevant physico-chemical properties of the replaced amino acids can be used as values of the descriptor variables in the established regression models. Consequently, if a mutant of the virus occurs, it is

easily and quickly possible to decide whether the available antibody is still effective enough or whether it has to be modified in some way. Therefore, no time must be wasted in testing the effectiveness of the antibody in vivo. Further, the biochemical conclusions drawn from the regression analysis are also useful in suggesting modifications of the antibody required to obtain an effective new antibody, if necessary. Hence, in this case, the results of the regression analysis in biomolecular interaction studies permit the performance of specific relevant experiments for the development of a new antibody.

The TMV is a particularly suitable research object since it can readily be produced in large quantities. Accordingly, it has often been used to study general aspects of virus assembly and disassembly processes. However, the examination of the interaction between modified antigens of the TMV and the existing antibody 57P is of special interest for further reasons. The TMV itself is a severe virus infecting members of nine plant families and approximately 125 individual species, especially tobacco, tomato, pepper and cucumber. Therefore, an infection by this virus can cause enormous crop losses. This is the reason why it is important to obtain information about the efficiency of the available antibody 57P with respect to possible modified viruses.

Furthermore, knowledge about the TMV can be used as well to elucidate the binding behaviour of other similar viruses. For example, conclusions concerning the Orthomyxovirus causing influenza are desired to be drawn from the investigation of the interaction between the modified peptides of the TMV and its antibody. For the Orthomyxoviruses, the possibility of predicting the binding characteristics between an occurred mutant and the available antibody is especially relevant in evaluating the efficacy of the existing antibody and hence to prevent the application of ineffective vaccinations.

In Orthomyxoviruses, antigenic shifts occur quite often. There have been five worldwide flu epidemics since 1890. For example, the so-called Spanish influenza in 1918 cost 20 million of people their life. The frequency of the antigenic shifts in the Orthomyxovirus illustrates the importance of being prepared for possible mutants of viruses. The application of PLS regression to measured binding parameters can contribute considerably to this ambitious task. This is the reason why the optimization of the modelling procedure in biomolecular interaction studies presents an important practical objective.

5.2 Description of the data

As mentioned briefly above, changes in the experimental settings were performed by modifying the amino acid sequence of the antigen of the TMVP as well as by varying the buffer composition. The experiments were performed following a statistical design plan to obtain appropriate data for the modelling with a limited number of measurements.

Altogether, 20 buffers were used, one defined standard buffer and 19 perturbation buffers. The perturbation buffers contain the chemical additives expected to be relevant, namely NaCl, urea, EDTA, KSCN and DMSO, in three different concentrations and show three different levels of the pH value. The exact buffer compositions chosen on the basis of a fractional factorial design are given in table **??** that is taken from Andersson et. al. (2001). Three of the 20 buffers, the buffers 17, 18 and 19, correspond to each other in terms of the concentrations of the ingredients and the pH value. These identical buffers might be used for the determination of the reproducibility of the measurements.

Substitutions of amino acids were realized at three mutation sites, in fact at the positions 142, 145 and 146 of the sequence of the wild-type TMV-peptide. These positions were chosen since modifications at these localizations are suspected to influence the binding behaviour to a limited extent without preventing the interaction completely. The 17 modified peptides were obtained by replacing amino acids at one, two or all of the three mutation sites simultaneously. The amino acid substitutions were determined according to calculations of the condition number of the design matrix respecting the ZZ-scales.

The wild-type peptide is characterized by Serine at position 142, Glutamic Acid at position 145 and Serine at position 146 and is hence denoted by the abbreviation SES. Some of the modified peptides have a length of 16 amino acids, whereas some are 19 amino acids in length. Those mutants containing 16 amino acids comprise the positions 137-151 of the antigen and an additional N-terminal cysteine, and the 19 amino acids long peptides correspond to the positions 134-151 and an additional N-terminal cysteine. With the help of this additional N-terminal cysteine, the peptides are bound to the sensor chip of the Biacore instrument. The different modifications of the wild-type are summarized in table **??** that is taken partially from Andersson et. al. (2001). Those amino acids deviating from the sequence of the wild-type are emphasized by boldface letters.

buffer	concentration [mM] of					pH-value
	NaCl	KSCN	DMSO	EDTA	urea	
standard buffer	150	0	0	3	0	7.4
buffer 1	150	4	30	3	40	7
buffer 2	150	4	300	23	400	7
buffer 3	150	22	30	23	400	7
buffer 4	150	22	300	3	40	7
buffer 5	550	4	30	23	40	7
buffer 6	550	4	300	3	400	7
buffer 7	550	22	30	3	400	7
buffer 8	550	22	300	23	40	7
buffer 9	150	4	30	3	400	7.8
buffer 10	150	4	300	23	40	7.8
buffer 11	150	22	30	23	40	7.8
buffer 12	150	22	300	3	400	7.8
buffer 13	550	4	30	23	400	7.8
buffer 14	550	4	300	3	40	7.8
buffer 15	550	22	30	3	40	7.8
buffer 16	550	22	300	23	400	7.8
buffer 17	350	13	165	13	220	7.4
buffer 18	350	13	165	13	220	7.4
buffer 19	350	13	165	13	220	7.4

Table 5.1: Composition of the buffers used in the experiments

For the interaction of interest, descriptor variables respecting the physico-chemical properties of the amino acids at the mutation sites are considered in the regression analysis. In particular, the influence of the three ZZ-scales and the HFT-scale on the binding behaviour is meant to be quantified. Beyond this, the 26 physico-chemical variables used by Sandberg et. al. (1998) for the derivation of the ZZ-scales are conceivable to incorporate in the modelling. Furthermore, the buffer descriptor variables representing the composition of the chemical environment are taken into account. However, information about the 3D-structure of the peptides under study is not available.

In order to obtain observations of the response variables of the regression analysis, the association and dissociation rate constants of the interactions involving the 18 different peptides and taking place in each of the 19 perturbation buffers and the standard buffer were measured with the help of a Biacore instrument. Consequently, for each of the various mutants, 20 pairs of binding parameters are available.

peptide	amino acid at mutation site			number of modifications
	142	145	146	
wild-type	S	E	S	0
mutant 1	**V**	**Q**	**E**	3
mutant 2	**M**	**Y**	**T**	3
mutant 3	**D**	**Y**	**D**	3
mutant 4	**G**	**R**	**A**	3
mutant 5	**G**	**S**	**Q**	3
mutant 6	**F**	**G**	**R**	3
mutant 7	**D**	**R**	**K**	3
mutant 8	**R**	**V**	**A**	3
mutant 9	**D**	**S**	**A**	3
mutant 10	**R**	**D**	**G**	3
mutant 11	**Q**	**D**	**F**	3
mutant 12	**M**	**G**	S	2
mutant 13	**N**	E	S	1
mutant 14	S	E	**A**	1
mutant 15	S	**A**	S	1
mutant 16	**A**	E	S	1
mutant 17	**E**	E	S	1

Table 5.2: Summary of the amino acid substitutions at the three mutation sites

For the determination of the reproducibility of the data, replicate measurements were performed, i.e. the measurements of the binding parameters were repeated a few times, in fact up to six times, for the interaction involving a particular modified peptide in a specific buffer. Therefore, several values of the association rate constant as well as the dissociation rate constant are given for each mutant. The measurements of each of these binding parameters should differ only slightly for the same adjustments of the experiment.

In table **??**, the number of measurements in the standard buffer and the corresponding average measured values and standard deviations referring to the different peptides are presented. A missing value is indicated by a dot. Apart from the fact that no observations of the association rate constant are available for the mutants FGR and QDF, there are no further missing observations in the standard buffer measurements. For the perturbation buffers, the measurements were repeated three times for the wild-type peptide, twice

for the mutants DYD, SAS, GRA, RDG, RVA, NES and SEA and only one single measurement was obtained for the remaining peptides.

mutant	mean $[1/Ms]$	standard deviation	number of measure-ments	mean $[1/s]$	standard deviation	number of measure-ments
	of k_a			of k_d		
SES	762500	55861	2	0.000424	0.0001575	6
VQE	322000		1	0.004090		1
MYT	479000		1	0.024000		1
DYD	601500	102530	2	0.014250	0.0014849	2
GRA	816000		1	0.004913	0.0003365	3
GSQ	924000		1	0.034725	0.0040672	4
FGR	.		0	0.067550	0.0111016	2
DRK	368000	74953	2	0.003840	0.0004808	2
RVA	378000		1	0.003373	0.0005519	3
DSA	681000		1	0.013350	0.0031820	2
RDG	708000	38183	2	0.012150	0.0000707	2
QDF	.		0	0.004910		1
NES	569000		1	0.001717	0.0000569	3
SEA	824000	66468	2	0.000771	0.0000372	3
SAS	1130000	141421	2	0.005150	0.0002263	2
AES	606000		1	0.000825	0.0001351	2
EES	48000		1	0.003960	0.0005798	2

Table 5.3: Number of measurements, average measured values and standard deviations for each peptide in the standard buffer

The numbers of missing values of both the association and dissociation rate constant are listed in table **??** for each peptide and the perturbation buffers and in table **??** regarding each of the 19 perturbation buffers. The values are missing for several reasons. Either, the interaction did not take place since the mutant of TMVP could not bind to the Fab fragment of the antibody 57P or the association or dissociation, respectively, proceeded too rapidly to be recognized by the Biacore instrument. Furthermore, in a number of cases, the measured value was judged to be false because of biochemical considerations and was hence denoted as a missing value as well. If a value of the response variables is missing, the respective observation is excluded from the analysis. The peptide MGS is omitted completely from the analysis since it did not bind in any of the buffers to the Fab fragment of the antibody 57P.

mutant	number of measurements per perturbation buffer	number of missing values	
		k_a	k_d
SES	3	1	2
VQE	1	2	3
MYT	1	3	3
DYD	2	5	5
GRA	2	2	4
GSQ	1	1	6
FGR	1	4	3
DRK	1	2	2
RVA	2	8	4
DSA	1	3	3
RDG	2	5	10
QDF	1	5	2
NES	2	8	11
SEA	2	2	2
SAS	2	6	4
AES	1	1	2
EES	1	5	5

Table 5.4: Number of measurements and missing values for each peptide regarding the 19 perturbation buffers

Regarding the association rate constant, the minimum observed value is 315000 $[1/Ms]$ in the standard buffer and 42300 $[1/Ms]$ in the perturbation buffers. The maximum measured value of the association rate constant is 1230000 $[1/Ms]$ in the standard buffer and 1170000 $[1/Ms]$ in the perturbation buffers. Furthermore, the mean of the available association rate measurements is 668714 $[1/Ms]$ in the standard buffer and 243497 $[1/Ms]$ in the perturbation buffers. In terms of the measurements of the dissociation rate constant, the values range between 0.00017 and 0.07540 $[1/s]$ with a mean of 0.01095 $[1/s]$ in the standard buffer and between 0.00022 and 0.11000 $[1/s]$ with 0.01165 $[1/s]$ as the average obtained value in the perturbation buffers.

buffer	number of missing values	
	k_a	k_d
buffer 1	4	4
buffer 2	1	1
buffer 3	1	1
buffer 4	2	4
buffer 5	2	4
buffer 6	3	4
buffer 7	2	4
buffer 8	1	4
buffer 9	1	3
buffer 10	3	2
buffer 11	14	10
buffer 12	3	8
buffer 13	1	0
buffer 14	3	2
buffer 15	4	7
buffer 16	3	2
buffer 17	5	4
buffer 18	0	1
buffer 19	10	6

Table 5.5: Number of missing values for each perturbation buffer

These statistical measures and further ones of interest are summarized in table **??** for the association and dissociation rate constant, respectively. The statistical measures are given separately for the dataset of standard buffer measurements and for the dataset of perturbation buffers measurements. The reason for this separate determination is the fact that usually in biomolecular interaction studies, the establishment of the regression models is based individually on the respective dataset. In particular, in QSKR modelling, the values measured in the standard buffer are used, whereas the QBKR modelling relies on the measurements obtained in the perturbation buffers.

In order to get a graphical impression of the available data, boxplots were created. In detail, the measured association and dissociation rate constants, respectively, are presented for the 19 different perturbation buffers in figure **??** and figure **??**.

statistical measure		in the standard buffer	in the perturbation buffers
sample size n		21	494
number of missing values		0	63
minimum	of	315000	42300
maximum	the	1230000	1170000
mean	k_a-	668714	243497
standard deviation	values	235236	152468
range	$[1/Ms]$	915000	1127700
interquartile range		322000	167000
sample size n		41	494
number of missing values		0	71
minimum	of	0.00017	0.00022
maximum	the	0.07540	0.11000
mean	k_d-	0.01095	0.01165
standard deviation	values	0.01668	0.01804
range	$[1/s]$	0.07523	0.10978
interquartile range		0.01128	0.01244

Table 5.6: Statistical measures for both the association and dissociation rate constant for the dataset comprising the measurements obtained in the standard buffer and the dataset containing the values measured in the perturbation buffers, respectively

Examining the boxplots referring to the association rate constant, relatively low measurements were obtained in buffer 16, whereas high values were measured in buffer 1. Since buffer 16 is characterized by high concentrations of all of the chemical additives and high levels of the pH-value, and buffer 1 by low concentrations and low levels of the pH-value, it can be expected that the buffer variables influence the association rate constant negatively. Furthermore, relative high association rate constants were observed in the buffers 2, 3, 4, 5, 9, 10 and 11. However, conclusions concerning the effect of the buffer composition cannot be drawn from this fact as these buffers vary from each other in terms of the concentrations of the chemical additives and the levels of the pH-value. The measured values can be considered to be sufficiently reproducible because the observations approximately coincide with each other in the identical buffers 17, 18 and 19.

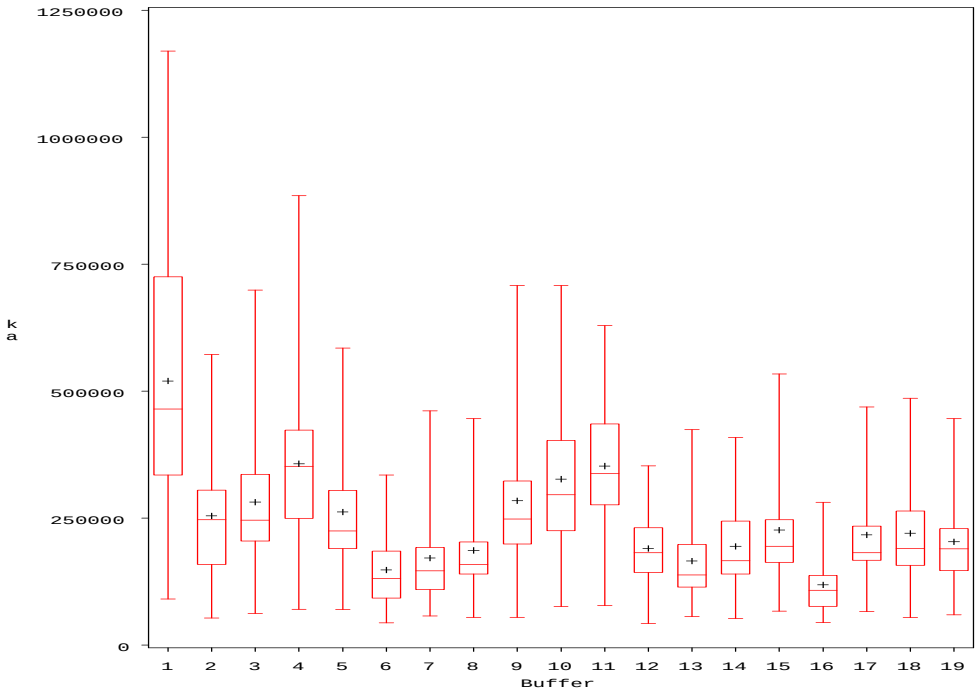

Figure 5.2: Boxplots of the measurements of the association rate constant depending on the 19 different perturbation buffers

The boxplots corresponding to the dissociation rate constant do not show notable deviations between the measured values regarding the different buffers. Especially, the medians and the minimum observations do not vary considerably from each other. Only the maximum measured values of the dissociation rate constant differ with respect to the various buffers.

In figure **??** and figure **??**, the boxplots of the measurements of the association and dissociation rate constant, respectively, obtained in the standard buffer are presented depending on the 17 different peptides.

For the different peptides, the association rate constants measured in the standard buffer vary from each other. The mutants lead both to an increased and a decreased association process compared to that one referring to the wild-type peptide. Consequently, the extent of the binding between the mutants and the Fab fragment of the antibody 57P is influenced relevantly by the performed mutations. Obviously, even one single mutation

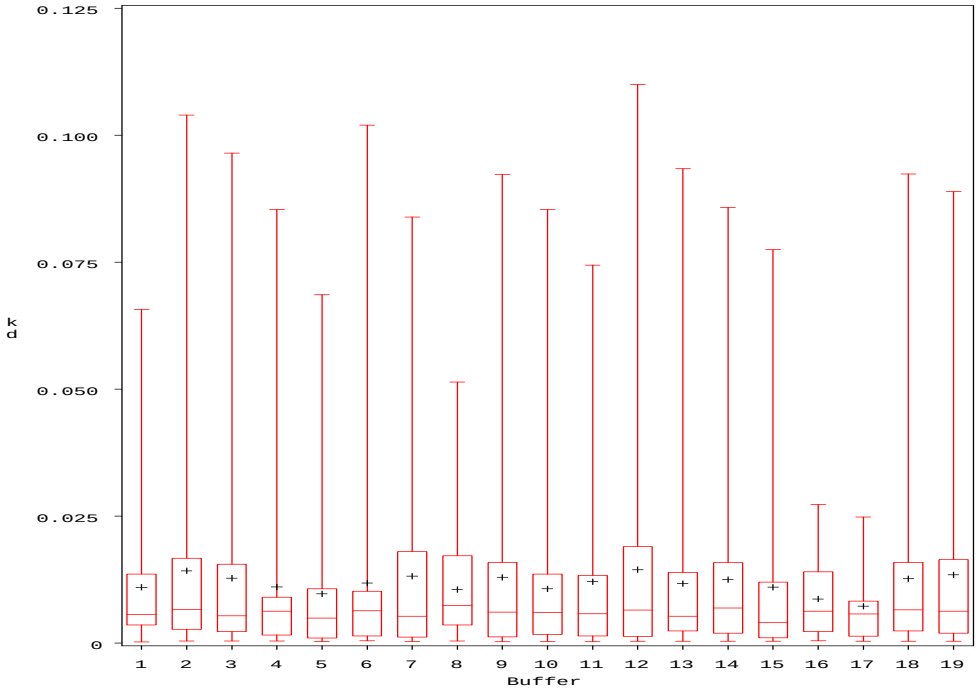

Figure 5.3: Boxplots of the measurements of the dissociation rate constant depending on the 19 different perturbation buffers

might cause notable changes in the association process since the association rate constants of the mutants showing merely one amino acid replacement differ remarkably from the association rate constant of the wild-type. Extremely small association rate constants are observed for the mutants VQE, DRK and RVA, whereas the mutant SAS results in a relatively large value of the association rate constants. However, it is not possible to suggest which physico-chemical properties contribute to this fact.

In terms of the boxplots referring to the dissociation rate constant, notably differences between the observed measurements in the standard buffer can be determined for the different peptides as well. The wild-type peptide corresponds to the lowest observations of the dissociation rate constant. Consequently, all of the produced mutants result in a faster dissociation process. For the mutants NES, SEA and AES, only a limited increase of the dissociation rate constant is observed. These peptides like the mutants RVA and SAS showing also a relatively low dissociation rate constant are those mu-

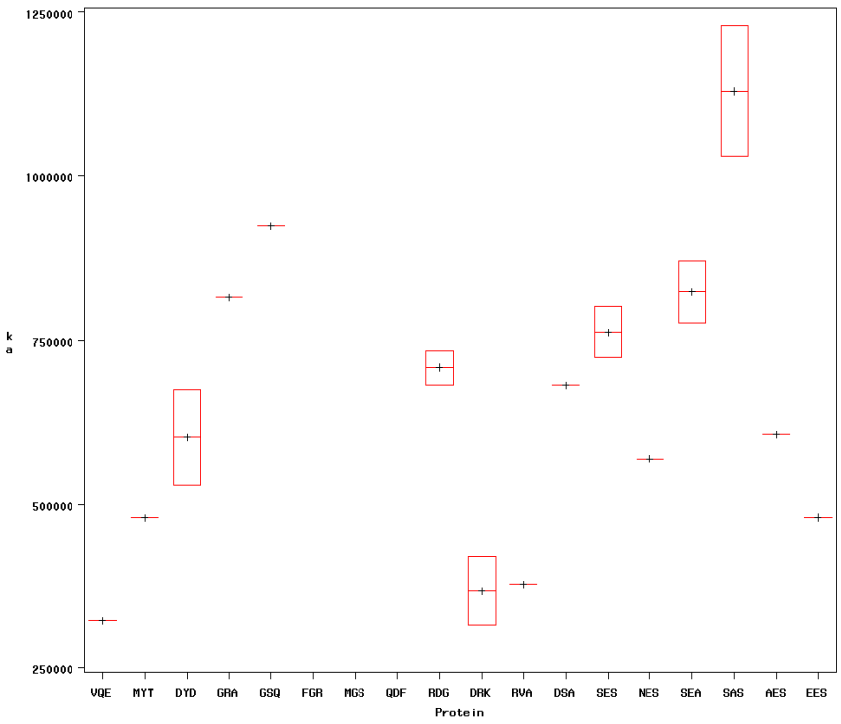

Figure 5.4: Boxplots relating the association rate constants measured in the standard buffer to the 17 different peptides

tants with only one single amino acid substitution. This result illustrates the fact that more than one mutation is necessary to occur in the amino acid sequence in order to lead to considerable changes in the dissociation process. A relatively high increase in the dissociation rate constant compared to that corresponding to the wild-type peptide is obtained for the interactions incorporating the mutants FGR and GSQ. Further mutants resulting in a moderately increased dissociation process are the peptides MYT, DYD, RDG and DSA. Because of the diverse attained values of the amino acid descriptor variables that might contribute to this effect, detailed conclusions cannot be drawn directly from this observation.

The boxplots given in figure **??**, figure **??**, figure **??** and figure **??** give first hints about the variations in the measurements for the different buffers or peptides, respectively. However, graphics showing the relationship between the descriptor variables taken into account in the regression analysis and the respective response variable provide more detailed information. Accordingly, scatterplots illustrating the relationship between either the association or the

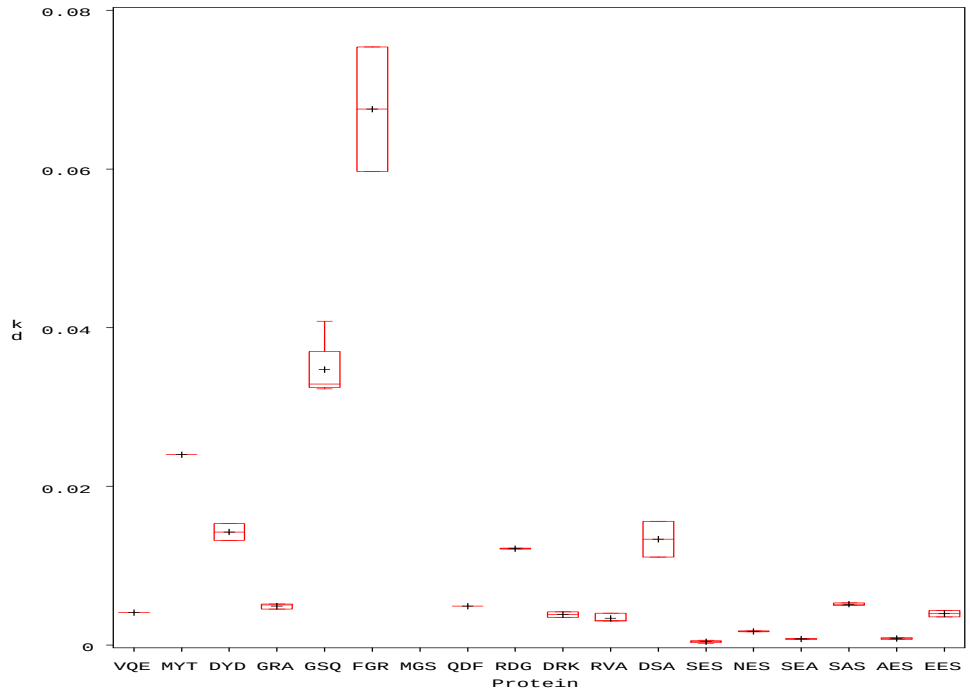

Figure 5.5: Boxplots relating the dissociation rate constants measured in the standard buffer to the 17 different peptides

dissociation rate constant, respectively, and the quantitative descriptor variables could be created, i.e. for the three ZZ-scales, the HFT-scale and the quantitative variables of the 26 physico-chemical properties of amino acids (Sandberg et. al. (1998)) at the three mutation sites.

Furthermore, boxplots showing the measurements of the association and dissociation rate constant, respectively, for the qualitative 26 physico-chemical property variables from Sandberg et. al. (1998) at the three mutation sites as well as on the buffer descriptor variables are conceivable to present. Examining the illustrations regarding the six buffer descriptor variables, boxplots referring separately to the various peptides should be created instead of scatterplots because only three different adjustments of each buffer descriptor variable were investigated for the perturbation buffers.

However, if these boxplots referring to each of the descriptor variables are meant to be presented, would require approximately 800 illustrations and so are not shown here.

5.3 Already realized analysis of the interaction between the TMVP and a Fab fragment of the antibody 57P

In Andersson et. al. (2001), several univariate QBKR models regarding the different peptides and two univariate QSKR models are presented in order to analyze the interaction between the tobacco mosaic virus protein and a Fab fragment of the antibody 57P. These regression models were established without incorporation of any interaction terms. Since structural information concerning the TMVP was not available, a 3D-QSAR model could not be developed. The numbers of iterations the different models are based on, i.e. the numbers of extracted latent variable vectors, are not reported. The software used by Andersson to apply PLS regression to the measured data were MODDE 4.0 and SIMCA 8.0.

5.3.1 The already performed QSKR modelling

One univariate QSKR model was developed with the association rate constant as response variable and one QSKR model incorporating the logarithmic measurements of the dissociation rate constant as values of the response variable was established. Obviously, in Andersson et. al. (2001), information about the energetic processes was meant to be obtained instead of knowledge about the kinetic process of the dissociation. In the QSKR models relating descriptor variables of the physico-chemical properties of amino acids to the binding parameters measured in the standard buffer, 12 descriptor variables were involved, namely the three ZZ-scales and the HFT-scale (see table **??**) respecting the amino acids at the three mutation sites.

The univariate QSKR models for the association rate constant and the logarithmic dissociation rate constant are specified as follows in Andersson et. al. (2001):

$$\begin{aligned} k_a &= 1175000 - 511000 \cdot HFT_{142} + 82200 \cdot ZZ3_{145} \quad \text{and} \\ log(k_d) &= -1.052 - 1.11 \cdot HFT_{145} - 0.186 \cdot ZZ3_{146}. \end{aligned}$$

Consequently, the QSKR modelling regarding the association rate constant reveals that the value of the HFT-scale of the amino acid at position 142 and the value of the ZZ3-scale of the amino acid at position 145 influence the association of the interaction under study significantly. In detail, the larger the value of the HFT-scale of the amino acid at position 142 is, the smaller

is the value of the association rate constant, whereas the ZZ3-scale of the amino acid at position 145 shows a positive effect on the association.

Furthermore, the QSKR analysis indicates that the logarithmic values of the dissociation rate constant depend significantly on the value of the HFT-scale of the amino acid at position 145. The corresponding QSKR model contains additionally the ZZ3-scale at position 146 as relevant descriptor variable since the incorporation of this variable improved the prediction accuracy. Both of these descriptor variables were determined to have a negative influence on the logarithmic dissociation rate constant.

In terms of the resulting prediction accuracies, the QSKR model referring to the association rate constant leads to a value of 0.49 of the Q_a^2-statistic, whereas the Q_A^2-value of the QSKR model related to the logarithmic dissociation rate constant is reported to be 0.73.

5.3.2 The already realized QBKR modelling

The univariate QBKR models were not presented for the association rate constant since, according to Andersson et. al. (2001), the obtained fingerprints were too noisy to be interpreted and thus, the models could not provide reliable results. The reason for this circumstance is the fact that the correctness of the association rate constant depends on the knowledge of the concentration of the Fab fragment of the antibody 57P in the different perturbation buffers. However, this concentration that might vary in the different buffers could not be determined exactly.

Accordingly, in Andersson et. al. (2001), QBKR models were merely established for the dissociation rate constant as response variable. The QBKR models describing the relationship between the six buffer descriptor variables and the values of the dissociation rate constant measured in the 19 different buffers were developed separately for each of the 17 peptides. Furthermore, if repeated measurements were available, the QBKR models were derived individually for these repetitions.

In Andersson et. al. (2001), the results of the QBKR analysis are summarized by giving the values of the chemical sensitivity fingerprints referring to the various peptides and the different repetitions. These fingerprints are shown in table **??**. However, only one fingerprint is available for each of the mutants RVA, NES and SEA though the measurements were repeated twice for these peptides. Beyond this, merely two fingerprints are given for the

wild-type peptide for which three repeated measurements were performed. The reason why the fingerprints referring to these repetitions were not presented is not explained in Andersson et. al. (2001).

According to Andersson et. al. (2001), the QBKR modelling shows that the buffer components EDTA, KSCN and pH do not influence the dissociation rate constant, whereas the chemical additives urea, DMSO and NaCl have a significant effect on this response variable. The extents of these effects vary with respect to the different peptides. However, information about the prediction accuracy in form of the Q_A^2-values of the obtained QBKR models were not reported.

	fingerprint referring to					
peptide	DMSO	EDTA	NaCl	pH	KSCN	Urea
SES1	0.11000	0.02200	0.05600	-0.03400	0.03500	0.05900
SES2	0.10000	0.02800	0.02300	0.00950	0.01700	0.07700
VQE1	0.18000	-0.02400	-0.00960	0.00031	0.01100	0.08400
MYT1	-0.02500	-0.01300	-0.00710	0.00570	0.01200	0.05000
DYD1	0.17000	-0.01900	-0.05600	-0.02600	0.00580	0.08100
DYD2	0.17000	-0.00910	-0.07800	-0.03100	0.00190	0.05800
GRA1	0.20000	-0.01200	-0.05800	-0.01600	0.02300	0.11000
GRA2	0.17000	-0.02400	-0.02900	-0.02100	0.03400	0.13000
GSQ1	0.14000	-0.00430	-0.01700	-0.02000	0.04100	0.12000
FGR1	0.11000	0.01300	-0.00260	0.01300	0.02000	0.11000
DRK1	0.08500	-0.07000	-0.09200	0.00056	0.02100	0.08600
RVA1	0.14000	0.00037	0.06600	-0.01300	0.02400	0.10000
DSA1	0.13000	-0.00880	-0.05300	0.00620	0.02700	0.08900
RDG1	0.07900	-0.01400	0.04800	-0.01300	0.03700	0.07200
RDG2	0.07000	-0.02400	0.04700	-0.00940	0.03600	0.07000
QDF1	0.12000	-0.00035	0.05100	0.01300	0.05100	0.05000
NES1	0.14000	-0.00930	0.02600	-0.00300	0.01100	0.08900
SEA1	0.12000	-0.02500	0.04100	0.00370	0.01300	0.11000
SAS1	0.18000	-0.00510	0.03500	0.00074	0.01900	0.10000
SAS2	0.17000	0.01500	0.03800	-0.00870	0.02700	0.12000
AES1	0.06000	-0.00330	-0.01400	0.01100	0.01200	0.03900
EES1	0.16000	0.00390	-0.01000	0.00410	0.024000	0.06700

Table 5.7: Results of the already performed QBKR modelling for the different mutants presented in form of the chemical sensitivity fingerprints, i.e. the estimated regression coefficients divided by the corresponding intercept

5.4 Reproduction of the already performed analysis

In order to reproduce the interaction analysis described in Andersson et. al. (2001), the models presented in this publication were developed. In detail, the univariate QSKR models with the association rate constant and the logarithmic dissociation rate constant, respectively, as response variable were established as well as the QBKR models determined separately for the different peptides and repetitions for the dissociation rate constant. The software used to apply the NIPALS-algorithm to the available data was SAS 9.1.

In SAS 9.1, by default, that optimal number A of iterations is determined to be that which results in the minimum prediction residual sum of squares ($PRESS$)-value, since, contrary to the Q_a^2-statistic, small $PRESS$-values indicate a good prediction accuracy. As explained previously, the final regression model is meant to provide accurate predictions on the one hand, and is desired to be based on relatively few iterations, i.e. extracted latent variable vectors, on the other hand. In order to determine that model presenting the best compromise between these two objectives, the test developed by van der Voet can be performed by SAS 9.1. Then, the differences between the minimum $PRESS$-value and those $PRESS$-values obtained after realizing fewer iterations than the optimal number A are tested for significance. Finally, the smallest number A^* showing an insignificantly larger $PRESS$-value compared to the optimal one, i.e resulting in insignificantly larger residuals, can be determined. Consequently, by applying van der Voet's test, the final regression model is based on that number A^* of iterations instead of on the optimal number A of iterations. If all of the $PRESS$-values referring to smaller numbers of iterations than the optimal number A of iterations are significantly larger than the optimal $PRESS$-value, then the model is specified with the help of the optimal number A of iterations. In the following, the regression models were derived by applying this test by van der Voet and both the optimal number A of iterations and the number A^* of iterations smaller than the number A and showing an insignificantly larger $PRESS$-value are reported.

The Q_a^2-statistic is not automatically computed in SAS 9.1. Therefore, the calculation of the $Q_{A^*}^2$-value was programmed, i.e. the Q_a^2-statistic evaluated for A^* iterations to compare the prediction accuracies of the regression models presented in Andersson et. al. (2001) with those of the reproduced models. Further, the $Q_{A^*}^2$-value is preferred to indicate the prediction accu-

racy instead of the $PRESS$-value since the $Q^2_{A^*}$-statistic is a standardized measure in contrast to the $PRESS$-measure.

Beyond this, contrary to the program SIMCA 8.0 applied by Andersson, SAS 9.1 does not perform any significance testing of the estimated original regression coefficients. Instead, the importance of the descriptor variables can be evaluated with the help of the so-called Variable Importance for Projection (VIP)-measure that is also not computed automatically by SAS 9.1.

The VIP-value VIP_j of the j-th descriptor variable is defined as follows:

$$VIP_j = \sqrt{m \sum_{a^*=1}^{A^*} \hat{w}^2_{ja^*} \frac{pctvar_{resp_{a^*}}}{\sum_{a^*=1}^{A^*} pctvar_{resp_{a^*}}}},$$

where the expression $pctvar_{resp_{a^*}}$ denotes the percentage of variation of the response variable being explained in the a^*-th iteration. Accordingly, the sum of these terms over the A^* iterations indicates the total percentage of variation of the response variable accounted for after the performance of A^* iterations of the PLS algorithm. Consequently, the value VIP_j reflects the contribution of the j-th descriptor variable to the fitting of the model, in other words, the importance for the projections performed during the PLS algorithm. According to Chong and Jun (2005), descriptor variables with a VIP-value larger than 1.00 can be considered to be relevant for the explanation of the respective response variable. This threshold value is chosen because the mean of the squared VIP-values equals the value 1.00. Consequently, a VIP-value larger than the value 1.00 indicates that the corresponding descriptor variable contributes to the modelling more than the average importance for the projections among all descriptor variables. Therefore, those descriptor variables showing a VIP-value larger than 1.00 are included in the regression model.

5.4.1 The reproduced QSKR analysis

The results of the reproduction of the QSKR models are presented below. Table **??** lists the VIP-values corresponding to the descriptor variables in the modelling. The VIP-values larger than 1.00 are printed in boldface in this table. Further, the VIP-values for those descriptor variables reported to be significant by Andersson et. al. (2001) are indicated by a star.

For these descriptor variables already determined to be significant, the highest VIP-values were obtained for both the association and the logarithmic

dissociation rate constants. Consequently, the computation of the VIP-values can be considered to provide a reliable evaluation of the contributions of the different descriptor variables to the explanation of the respective response variable. However, the determination of relevant descriptor variables with the help of the VIP-values presents a less strict decision rule compared to the significance test performed by Andersson. This statement can be concluded from the observation that more descriptor variables are found to be important on the basis of the VIP-value criterion using the value 1.00 as threshold value than are reported to be significant by Andersson et. al. (2001).

However, Chong and Jun (2005) report that the proper cutoff value to determine significant descriptor variables may be higher than the value 1.00. Beyond this, comparing the VIP-values reveals that those VIP-values referring to the descriptor variables determined to be significant by Andersson et. al. (2001) are all larger than the value 1.40. Therefore, this value is chosen as the threshold value used to distinguish between descriptor variables presumed to be significant or not significant, though the latter may seeming not to be irrelevant for the response in question. In table **??**, VIP-values larger than the value 1.00 but smaller than 1.40 are written in brackets to indicate these latter descriptor variables.

Using the results of analyzing the computed VIP-values leads to the following specification of the reproduced QSKR models for the association and the logarithmic dissociation rate constants, respectively, as response variable:

$$\begin{aligned} k_a &= 1089483(-44551 \cdot ZZ2_{142}) - 422170 \cdot \mathbf{HFT_{142}} + 53539 \cdot \mathbf{ZZ3_{145}} \\ &\quad (-3391 \cdot ZZ3_{146}), \end{aligned}$$

$$log(k_d) = -2.039(-0.026 \cdot ZZ2_{145}) - 2.960 \cdot \mathbf{HFT_{145}} - 0.315 \cdot \mathbf{ZZ3_{146}}.$$

In these regression models, the descriptor variables with a VIP-value larger than 1.00 are incorporated, where those descriptor variables showing a VIP-value between the values 1.00 and 1.40 are included in brackets. The descriptor variables determined in Andersson et. al. (2001) to be significant are represented by boldface letters and numbers.

Compared to the regression models established in Andersson et. al. (2001), two additional descriptor variables are incorporated in the model referring to the association rate constant and one additional descriptor variable respecting the logarithmic dissociation rate constant. In detail, in the QSKR model

VIP-values referring to			
		response variable	
descriptor variable	position	k_a	$log(k_d)$
ZZ1	142	0.48	0.90
	145	0.47	0.74
	146	0.50	0.38
ZZ2	142	(**1.21**)	0.42
	145	0.90	(**1.19**)
	146	0.79	0.91
ZZ3	142	0.57	0.31
	145	**1.44***	0.61
	146	(**1.34**)	**1.44***
HFT	142	**1.72***	0.49
	145	0.82	**2.19***
	146	0.74	0.69

Table 5.8: The VIP-values for the descriptor variables considered in the reproduced QSKR analysis for the association rate constant and the logarithmic dissociation rate constant (further explanations in the text)

with the association rate constant as response variable, the ZZ2-scale at position 142 and the ZZ3-scale at position 146 are considered to be important in addition to the HFT-scale at position 142 and the ZZ3-scale at position 145. For the QSKR model of the logarithmic dissociation rate constant, the ZZ2-scale at position 145 is determined to be relevant beyond the HFT-scale at position 145 and the ZZ3-scale at position 146.

In table **??**, some measures describing the reproduced QSKR models are summarized. In fact, the following kinds of information are listed:

- the optimal number A of iterations
- the smallest number A^* of iterations resulting in insignificantly larger residuals compared to those ones obtained on the basis of A iterations
- the $Q^2_{A^*}$-value of the prediction accuracy after A^* iterations
- the percentage $pctvar_{descr_{A^*}}$ of variation of the descriptor variables accounted for after A^* iterations and
- the percentage $pctvar_{resp_{A^*}}$ of variation of the respective response variable explained after A^* iterations.

response variable	A	A^*	$Q^2_{A^*}$	$pctvar$ [%]	
				$descr_{A^*}$	$resp_{A^*}$
k_a	3	-	0.77	42.79	84.27
$log(k_d)$	5	3	0.75	48.48	78.71

Table 5.9: Description of the reproduced QSKR models respecting the association rate constant and the logarithmic dissociation rate constant

For the reproduced QSKR model of the association rate constant, the minimum $PRESS$-value is obtained after three iterations, whereas five iterations are required to get the minimum $PRESS$-value for logarithmic dissociation rate constant. For both QSKR models, the smallest number A^* of iterations with an insignificant difference between the respective resulting residuals and those obtained for the optimal model is three. Consequently, for the model for the association rate constant, no adequate smaller number of iterations than the optimal number A could be found. Therefore, both the QSKR model of the association rate constant and the logarithmic dissociation rate constant are established on the basis of the computation of three iterations.

The prediction accuracies determined for the QSKR models approximately coincide with each other. In detail, for the QSKR model of the association rate constant, a $Q^2_{A^*}$-value of 0.77 is obtained, and the $Q^2_{A^*}$-value for the QSKR model of the logarithmic dissociation rate constant is 0.75. Consequently, both QSKR models can be considered to provide sufficiently accurate predictions. The $Q^2_{A^*}$-value computed for the logarithmic dissociation rate constant approximately equals that $Q^2_{A^*}$-value (0.73) reported by Andersson et. al. (2001). However, the $Q^2_{A^*}$-value regarding the association rate constant is much larger than that one (0.49) presented in Andersson et. al. (2001).

Approximately 43% of the variation of the descriptor variables could be explained by the QSKR model of the association rate constant, a result a little bit worse compared to the QSKR model of the logarithmic dissociation rate constant that accounts for approximately 48% of the variation of the descriptor variables. Consequently, the latent variable vectors obtained in the QSKR modelling represent the information inherent in the descriptor variables only to a limited extent. Better results are obtained for both models in terms of the explained percentage of variation of the respective response variable. In particular, approximately 84% of the variation of the measured association rate constants and approximately 79% of variation of

the observed logarithmic dissociation rate constants could be accounted for by the corresponding QSKR models. Since the explanation of the variation of the respective response variable is more important than the explanation of the variation of the descriptor variables to obtain a regression model providing accurate predictions, both QSKR models can be considered to be useful in practical applications. The explained percentages of the total variation of the descriptor variables or the respective response variable could not be compared with those referring to the already derived models as this information is not given in Andersson et. al. (2001).

5.4.2 The reproduced QBKR analysis

According to the performance presented in Andersson et. al. (2001), the QBKR models were derived separately for the diverse peptides and the different repetitions, if repeated measurements were available. However, the QBKR models were developed for each repetition of the measurements of the wild-type and the mutants RVA, NES and SEA since it is not known for which of the available repetitions of these peptides the models were reported in Andersson et. al. (2001). It has to be noted, that the measurement-numbers referring to the peptides are chosen randomly and hence, it cannot be expected that they coincide with the notation used in the context of the already realized analysis.

Those VIP-values larger than the value 1.00, computed for the six buffer variables for the different peptides and repetitions, are listed in table **??**. According to the notation in the previous subsection, the threshold value of 1.40 is assumed to distinguish between relatively important and presumably significant descriptor variables. Therefore, those VIP-values between the values 1.00 and 1.40 are written in brackets in table **??**. The buffer components determined to be significant by Andersson et. al. (2001) are indicated by a star.

Comparing the computed VIP-values reveals that DMSO and urea are chemical additives influencing the dissociation rate constant for many peptides and repetitions. In detail, DMSO corresponds to VIP-values larger than the value 1.00 for 20 of the 26 developed QBKR models, where merely five of these VIP-values are smaller than the value 1.40. In terms of the buffer component urea, a significant effect can be supposed for ten QBKR models and a relative importance in 8 further cases. For both chemical additives DMSO and urea a positive effect on the dissociation rate constant is determined. Consequently, increasing the concentrations of these buffer components results in a faster

dissociation process, i.e. more complexes of bound biomolecules decompose per second. Further, the chemical additive NaCl shows a notably influence on the dissociation rate constant for 8 models, where the effect is probably significant for five of these models. The kind of influence of NaCl depends on the particular peptide referred to.

However, the buffer components EDTA and KSCN as well as the pH-value are not determined to have an important influence on the dissociation rate constant. In particular, EDTA corresponds to VIP-values larger than the value 1.00 for three models but a VIP-value larger than 1.40 is only computed for one of these models. In terms of the pH-value, merely one of the VIP-values is larger than the value 1.00 but is not even larger than the value 1.40. KSCN was not found to be relevant in any of the QBKR models.

The observation that the chemical additives DMSO, urea and NaCl have an effect on the dissociation rate constant in contrast to the buffer components EDTA and KSCN and the pH-value coincides with the statements reported in Andersson et. al. (2001). The different extents of the influences of the relevant chemical additives on the dissociation rate constant for the various peptides can be evaluated not only with the help of the VIP-values but also by comparing the chemical sensitivity fingerprints referred to subsequently.

Further information about the established QBKR models is summarized according to the presentation in table **??**. In fact, the optimal number A of iterations, the smallest number A^* of iterations resulting in insignificantly larger residuals compared to those ones obtained on the basis of A iterations, the $Q^2_{A^*}$-value and the explained percentages $pctvar_{descr_{A^*}}$ and $pctvar_{resp_{A^*}}$ of the total variation of the buffer variables and the measurements of the dissociation rate constant, respectively, are also given in table **??**. The values of the optimal and used numbers of iterations, the obtained prediction accuracies as well as the explained percentages of the total variation of the buffer variables and the measured dissociation rate constants, respectively, could not be compared with those referring to the already established models since this kind of information is not given in Andersson et. al. (2001).

Examining the percentages of explained variation of the descriptor variables reveals that only for four peptides, namely for the mutants DYD, GSQ, RVA and EES, more than half of the total variation could be accounted for. Consequently, the computed latent variable vectors cannot be expected to represent the structure inherent in the buffer variables. Very good results are obtained for the explained percentages of the variation of the measured

dissociation rate constant. In detail, with the exception of five models, the percentage of the variation that could be accounted for is larger than 80%. Two models lead to explained percentages of approximately 83% and the remaining 17 regression models result in more than 90% of variation that could be accounted for.

Further, the $Q^2_{A^*}$-values larger than 0.78, with the exception of five models, indicate sufficiently good prediction accuracies of the obtained QBKR models. The five models for which less accurate predictions are provided are identical to those models explaining less than 80% of the total variation of the measured dissociation rate constants. These models referring to one repetition corresponding to the wild-type and the mutants MYT, AES, NES and VQE can be considered to be relatively useless in practical applications. However, the other 19 developed QBKR models can be expected to provide reliable predictions.

In table **??**, the results of the reproduced QBKR analysis regarding the dissociation rate constant are presented by giving the chemical sensitivity fingerprints obtained for the different peptides and repetitions. The fingerprints referring to a chemical additive showing a VIP-value larger than the value 1.00 for a particular peptide and repetition are written in boldface in this table. Further, the fingerprints corresponding to VIP-values between 1.00 and 1.40 are given in brackets. Conclusions on the reproducibility of the measurements can be drawn by comparing the fingerprints for the repetitions for one peptide. The reproduced fingerprints differ from those presented in Andersson et. al. (2001), most probably for the same reasons explained in the previous subsection.

For the QBKR models for one repetition of the wild-type and the mutant SEA, respectively, the number A^* of iterations leading to an insignificantly larger $PRESS$-value than the optimal one is determined to be zero. This observation can be interpreted as reflecting that these repeated measurements are inappropriate for the model. This might be the reason why the fingerprints for these peptides were not reported in Andersson et. al. (2001). However, it cannot be explained why the second fingerprints referring to the mutants RVA and NES are missing.

mutant	VIP-values referring to						A	A^*	$Q^2_{A^*}$	*pctvar* [%]	
	DMSO*	EDTA	NaCl*	pH	KSCN	Urea*				$descr_{A^*}$	$resp_{A^*}$
SES1	1.86						1	1	0.78	16.67	82.52
SES2	1.85	(1.07)	1.51			1.42	1	1	0.57	16.67	67.16
SES3							4	0	-	-	-
VQE1	1.94					(1.02)	5	1	0.61	18.95	75.68
MYT1	(1.18)					2.04	1	1	0.68	19.95	75.78
DYD1	1.61		(1.26)				6	4	0.97	58.78	98.62
DYD2	1.71						6	3	0.98	42.21	98.92
GRA1	1.79					(1.26)	4	3	0.98	42.24	98.81
GRA2	1.64					1.52	4	2	0.92	32.68	94.90
GSQ1	(1.30)			(1.08)		1.51	5	3	0.96	54.00	98.40
FGR1	(1.35)					1.94	6	2	0.94	31.84	95.77
DRK1		(1.07)	1.51			(1.20)	2	1	0.90	18.79	92.61
RVA1		1.55	1.43				5	1	0.78	19.75	83.02
RVA2	1.63					1.50	3	3	0.92	50.00	96.23
DSA1	1.52					(1.28)	5	2	0.97	33.39	98.30
RDG1	(1.01)		(1.28)			1.57	5	2	0.98	17.74	98.89
RDG2			(1.31)			1.58	6	2	0.95	33.90	96.90
QDF1	1.69		(1.29)				4	1	0.87	17.29	90.68
NES1						(1.21)	3	1	0.88	23.69	91.53
NES2	1.49		(1.11)				5	1	0.70	25.31	77.70
SEA1	1.44					1.77	2	1	0.88	16.78	91.28
SEA2							0	0	-	-	-
SAS1	1.81					(1.31)	6	2	0.95	31.01	95.93
SAS2	1.57					1.54	6	1	0.87	17.46	90.00
AES1	1.72					(1.34)	5	1	0.72	17.64	78.28
EES1	1.94					(1.26)	3	3	0.98	50.06	99.15

Table 5.10: Description of the reproduced QBKR models and VIP-values corresponding to the buffer variables for the different peptides and repetitions

peptide	fingerprint referring to					
	DMSO	EDTA	NaCl	pH	KSCN	Urea
SES1	**0.00071**	0.00187	0.00024	-0.07172	0.00331	0.00028
SES2	**0.00166**	(**0.00612**)	**0.00026**	0.05294	0.00407	**0.00095**
SES3	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VQE1	**0.00387**	0.00173	-0.00091	0.34628	0.02094	(**0.00153**)
MYT1	(**-0.00024**)	-0.00080	-0.00002	0.02638	0.00131	**0.00031**
DYD1	**0.00056**	-0.00108	(**-0.00016**)	-0.07509	0.00025	0.00026
DYD2	**0.00046**	-0.00034	-0.00020	-0.07635	0.00019	0.00016
GRA1	**0.00083**	-0.00087	-0.00022	-0.06026	0.00192	(**0.00044**)
GRA2	**0.00061**	-0.00190	-0.00008	-0.07226	0.00284	**0.00048**
GSQ1	(**0.00059**)	-0.00032	0.00007	(**-0.07589**)	0.00326	**0.00049**
FGR1	(**0.00312**)	0.00457	-0.00019	0.40702	0.00956	**0.00356**
DRK1	0.00061	(**-0.00903**)	**-0.00064**	0.04453	0.00356	(**0.00057**)
RVA1	0.00027	**-0.00588**	**-0.00027**	-0.04031	-0.00132	0.00013
RVA2	**0.00081**	0.00102	0.00030	-0.05522	0.00315	**0.00056**
DSA1	**0.00138**	-0.00217	-0.00046	0.07553	0.00553	(**0.00092**)
RDG1	(**0.00039**)	-0.00093	(**0.00023**)	-0.05263	0.00364	**0.00037**
RDG2	0.00036	-0.00271	(**0.00023**)	-0.02980	0.00353	**0.00044**
QDF1	**0.00172**	0.00309	(**0.00089**)	0.10228	0.01269	0.00062
NES1	0.00038	-0.00347	0.00033	-0.06455	-0.00045	(**0.00032**)
NES2	**0.00035**	-0.00194	(**0.00017**)	-0.05335	-0.00158	0.00021
SEA1	**0.00337**	-0.00545	0.00077	0.35986	0.01974	**0.00309**
SEA2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
SAS1	**0.00160**	-0.00082	0.00029	0.00630	0.00345	(**0.00094**)
SAS2	**0.00096**	0.00596	0.00024	-0.04704	0.00481	**0.00071**
AES1	**0.00050**	-0.00140	-0.00016	0.05793	0.00135	(**0.00030**)
EES1	**0.00163**	0.00076	-0.00010	0.03828	0.00490	(**0.00067**)

Table 5.11: Results of the reproduced QBKR modelling presented in form of the chemical sensitivity fingerprints, i.e. the estimated regression coefficients divided by the corresponding intercept

5.4.3 Conclusions concerning the reproduced regression analysis

The reproduction of the already performed QSKR and QBKR modelling showed that regression models providing relatively reliable predictions could be developed with only a few exceptions. In the following section, novel aspects of the modelling procedure are proposed and evaluated in order to optimize the regression models.

The most important results of the regression analysis, i.e. the determination of the relevant descriptor variables, could be reproduced for both the QSKR and the QBKR models by applying the VIP-value criterion. However, the reproduced estimated model parameters, i.e. the intercepts and the regression coefficients respecting the QSKR models and the chemical sensitivity fingerprints for the QBKR models, respectively, differ from those reported in Andersson et. al. (2001).

The differences between the results of the reproduced QSKR and QBKR modelling and those ones presented in Andersson et. al. (2001) are probably caused by the fact that the regression models were developed using different software, where the exact performance in MODDE 4.0 and SIMCA 8.0 is not described. Therefore, the modelling procedure could not be reproduced in detail. Predominantly, two important aspects about the already realized modelling are not reported in Andersson et. al. (2001).

On the one hand, the number of iterations on which the presented regression models are based are not mentioned. Probably, the models specified in Andersson et. al. (2001) are derived with the help of a different number of iterations than the number of iterations performed to obtain the reproduced models. This distinction is caused by the fact that the useful test by van der Voet was not applied during the development of the already specified models. On the other hand, it is not known which variant of the PLS algorithm was applied to the data, i.e. if also the NIPALS-algorithm was realized.

Furthermore, the significance test for the estimated regression coefficients applied by Andersson et. al. (2001) is not implemented in SAS 9.1. Since the performance of this test is not reported, neither in the literature referred to nor in the manuals of MODDE and SIMCA, it could not be programmed. Consequently, a different criterion, the VIP-value, was used to decide which descriptor variables should be included in the final reproduced regression models.

In terms of the comparison of the obtained prediction accuracies, the values of the Q_a^2-statistic would not coincide with each other even if the reproduced QSKR models would be identical with the already established models. The reason is that SIMCA 7.0, and hence most probably SIMCA 8.0 as well, does not use the standard formula for computing the Q_a^2-values as explained by Freyhult et. al. (2005). However, the standard formula was programmed in SAS 9.1 in order to evaluate the prediction accuracy of the QSKR models conventionally.

Therefore, it is not surprising that the results of the reproduced modelling differ from those presented in Andersson et. al. (2001). The comparison between the reproduced models and the already developed ones illustrates the importance of a detailed documentation of the applied modelling procedure in order to allow for a reproduction of the results. The exact description of the performance of the regression analysis is especially necessary for PLS regression because of this method's complexity.

5.5 Novel aspects of the analysis of the interaction between the TMVP and a Fab fragment of the antibody 57P

Usually, in a biomolecular interaction analysis, univariate regression models respecting the different subgroups of descriptor variables are established without incorporation of interaction terms, though an alternative modelling might be more advantageous. In this section, several novel aspects of the analysis of binding processes are considered. The aim of this research is the determination of that modelling procedure leading to an improved fitting of the data and an increased prediction accuracy of the resulting regression models. By applying this advanced and more comprehensive regression analysis to the data used in Andersson et. al. (2001), the knowledge about the interaction between the tobacco mosaic virus protein and a Fab fragment of the antibody 57P is extended.

To evaluate the benefit of the diverse modelling approaches, PLS regression is applied in modified ways to the available data. Subsequently, the resulting regression models are compared with the help of the achieved prediction accuracy and further measures of interest. In the following, the different modelling approaches are motivated and explained.

In practice, the regression models are specified separately for the different subgroups of descriptor variables, i.e. individual QSKR and QBKR models are developed. However, alternatively, one single unified regression model incorporating both the amino acid and buffer descriptor variables can be established. With the help of a unified model, the effects of all conceivable descriptor variables can be modelled simultaneously. This novel modelling approach can be considered to be preferable to that one of the separate subgroup models since the results of the regression analysis can be presented in a more compact form. Consequently, the unified modelling approach facilitates the biochemical interpretation of the results of the regression analysis.

Beyond this, in a unified model, relevant interaction terms can be taken into account, not only individually among the amino acid descriptor variables or buffer descriptor variables, but even between each amino acid and each buffer descriptor variable. The incorporation of these interaction terms in the unified model is especially relevant since in Andersson et. al. (2001), different extents of the influence of the chemical additives urea, DMSO and NaCl on the binding respecting the diverse mutants are reported. Obviously, interactions between the physico-chemical properties of the amino acids at the mutation sites and the concentrations of these buffer components exist. These supposed interactions can be easily quantified and hence explained by performing the unified modelling approach involving interaction terms.

In practice, it is often the case that QSPR models referring to the chemical additives are established in addition to the QBKR and QSKR models in order to explain the different sensitivities to changes in the concentrations of the buffer components for the diverse mutants. The values of the response variables of these QSPR models are estimated regression coefficients divided by the estimated intercept. Since these models involve estimations instead of measurements as values of the response variables, they cannot be expected to provide very reliable statements concerning the interaction between the amino acid sequence and the chemical environment. Therefore, another advantage of specifying a unified regression model with interaction terms instead of the subgroup models is that this complicated QSPR modelling procedure becomes unnecessary to perform. Consequently, the establishment of a unified regression model including interaction terms replaces not only the numerous subgroup models but also the diverse QSPR models. This fact illustrates the contribution of the unified modelling approach for simplifying the interpretation of the results.

In the following, interaction terms are not only incorporated in the unified models but in all established regression models. Furthermore, the models are

developed as well without interaction terms in order to determine the benefit of taking into account interaction terms. Obtaining knowledge about the importance of interaction terms contributes notably to an improvement of the understanding of the influences on the binding process under investigation.

Usually, univariate models are developed in the analysis of biomolecular interactions, though multivariate modelling might be a useful alternative. In contrast to Ordinary Least Squares, multivariate PLS regression leads to different results compared to the univariate analysis. Therefore, multivariate subgroup models as well as multivariate unified models are established in order to examine whether the multivariate analysis approach can be considered to be advantageous in comparison with the univariate modelling.

Instead of the commonly used three ZZ-scales, the 26 variables considered in Sandberg et. al. (1998) to derive the ZZ-scales (see subsection **??**) can be incorporated as descriptor variables of the physico-chemical properties of amino acids in the corresponding regression models, i.e. the QSKR and unified models. By quantifying the physico-chemical properties of the amino acids at the mutation sites by these 26 variables, a more detailed representation of the possibly relevant features of amino acids can be used in the modelling. Therefore, involving the 26 variables presented in Sandberg et. al. (1998) in the regression analysis permits a more sophisticated modelling of the influences of the physico-chemical characteristics of amino acids on the interaction of interest. Consequently, the use of this quantification of physico-chemical features, that is unusual in the regression analysis of biomolecular interactions so far, can be suspected to lead to an improved prediction accuracy of the resulting models.

In Andersson et. al. (2001), the QBKR models are established separately for the different peptides and repetitions. The QBKR models might be developed as well for each peptide, ignoring the different repetitions. This performance leads to the specification of fewer regression models and would hence contribute to a facilitation of the presentation of the results of the QBKR modelling.

Furthermore, the application of the test from van der Voet already described in the previous section presents a novel modelling aspect for the analysis of the interaction between the tobacco mosaic virus protein and a Fab fragment of the antibody 57P. The reason is that this test is not implemented in the software used to develop the models presented in Andersson et. al. (2001). The application of van der Voet's test leads to the establishment of regres-

sion models being based on less iterations than those models that would have been specified without using this test. The final regression models are hence easier to interpret in terms of the underlying latent variable vectors but show merely an insignificant difference from the optimal obtainable prediction accuracy.

In order to determine the benefit of the different novel modelling approaches, a number of regression models are established and compared. Which model comparisons are performed for which particular objective is explained in the following subsection. Subsequently, the diverse obtained regression models are described by relevant measures, and the novel modelling procedures are evaluated. Finally, the optimal and further useful regression models are established, and the results of these models are used to draw biochemical conclusions concerning the influences of the diverse descriptor variables on the interaction between the tobacco mosaic virus protein and a Fab fragment of the antibody 57P.

5.5.1 Model comparisons

The diverse regression models established for the evaluation of the different modelling approaches are listed and numbered in table **??**. In this table, the incorporated response variable(s) and descriptor variables are indicated by the symbol x for each regression model. Consequently, table **??** can be used to identify whether the respective regression models are univariate or multivariate models or whether they represent subgroup (QBKR or QSKR, respectively) models or unified models.

In the respective subgroup models, the buffer descriptor variables (QBKR), the HFT-scale and the three ZZ-scales or the HFT-scale and the 26 variables presented by Sandberg et. al. (1998) (QSKR) are incorporated as descriptor variables. The unified models involve either the buffer descriptor variables and the HFT-scale and the three ZZ-scales or the buffer descriptor variables, the HFT-scale and the 26 variables used by Sandberg et. al. (1998).

The QSKR models are based on the measurements obtained in the standard buffer, whereas the QBKR models and the unified models are established using the data measured in the perturbation buffers. None of the regression models, except model number 6, refers to the logarithmic measurements since the biochemical interpretation of the results is meant to provide statements concerning the kinetic and not the energetic processes. Model 6 with the logarithmic dissociation rate constant as response variable is included in the investigation because this modelling was described in Andersson et. al. (2001).

The QSKR, QBKR and unified models were developed both under incorporation of all conceivable interaction terms and without interaction terms. Exceptions refer to the models involving the 26 variables considered by Sandberg et. al. (1998). In fact, the QSKR models were only derived without interaction terms since an extremely large number of interaction terms would have been to be taken into account. Further, because of the same reason, the corresponding unified models incorporate only the interaction terms between the buffer variables and the HFT-scale and these 26 variables at the three mutation sites. The respective regression models without and with interaction terms are indicated by one common model number in **??**.

For each conceivable constellation of incorporated descriptor variables, three regression models are established, namely two univariate ones for the association and dissociation rate constant, respectively, and one multivariate one incorporating simultaneously both of these response variables.

For QBKR modelling, 17 models are obtained if the establishment was performed separately for each peptide, and 26 models are specified in case of realizing the regression analysis separately for each peptide and repetition. Consequently, the model numbers in table **??** refer to a number of regression models for the QBKR models.

The application of PLS regression is particularly useful for the establishment of the regression models required for the evaluation of the novel modelling approaches. The reason is that most of the models comprise a large number of descriptor variables. In particular, by taking into account all conceivable interaction terms or even some of them, an enormous number of descriptor variables has to be dealt with additionally. Furthermore, the use of the 26 variables from Sandberg et. al. (1998) instead of the three ZZ-scales of the amino acids at the mutation sites leads to an increase of the number of descriptor variables by 69 in the corresponding regression models. For example, the unified regression models taking into account the 26 variables from Sandberg et. al. (1998) involve 87 descriptor variables, i.e. the HFT-scale and the 26 physico-chemical property variables of the amino acids at the three mutation sites and six buffer descriptor variables, and the corresponding enormous number of interaction terms, not even considering the resulting interaction terms.

The model comparisons performed for the investigation of the usefulness of a specific novel modelling aspect are summarized in table **??** and are described below. In this table, the model comparisons are indicated by the symbol $\leftrightarrow$.

model number and name		response variable		descriptor variables		
				buffer	HFT-scale/	HFT-scale/
		k_a	k_d	variables	ZZ-scales	26 variables
(m1)	QBKRs-ka	x		x		
(m2)	QBKRs-kd		x	x		
(m3)	QBKRs-ka,kd	x	x	x		
(m4)	QSKR-ka	x			x	
(m5)	QSKR-kd		x		x	
(m6)	QSKR-logkd		log(kd)		x	
(m7)	QSKR-ka,kd	x	x		x	
(m8)	QSKR-ka-26vars	x				x
(m9)	QSKR-kd-26vars		x			x
(m10)	QSKR-logkd-26vars		x			x
(m11)	QSKR-ka,kd-26vars	x	x			x
(m12)	unified-ka	x		x	x	
(m13)	unified-kd		x	x	x	
(m14)	unified-ka,kd	x	x	x	x	
(m15)	unified-ka-26vars	x		x		x
(m16)	unified-kd-26vars		x	x		x
(m17)	unified-ka,kd-26vars	x	x	x		x

Table 5.12: The regression models taken into consideration

In order to investigate the usefulness of performing a multivariate regression analysis instead of a univariate one, each of the multivariate regression models is compared to the two corresponding univariate models referring either to the association or the dissociation rate constant. If the resulting prediction accuracies and other measures of the multivariate models are better than the prediction accuracies for the corresponding univariate models, the multivariate modelling procedure should be preferred to the univariate analysis for the interaction in question.

In general, the incorporation of interaction terms in the regression analysis is advised in order to permit the determination of their relevance. The usefulness of involving interaction terms in the modelling for the interaction under study is evaluated by examining whether interaction terms are stated to be important in the regression models developed including them. Further, characteristics of all of the regression models with interaction terms are compared to those without them. In particular, the regression models established

novel aspect	model comparisons
multivariate modelling	m5 ↔ m1/m2, m6 ↔ m3/4, m9 ↔ m7/8, m13 ↔ m11/12, m16 ↔ m14/15
use of interaction terms	m1-m7, m12-17m with interaction terms ↔ m1-m7, m12-17m without interaction terms
use of 26 variables from Sandberg et. al. (2001)	m8 ↔ m4, m9 ↔ m5, m11 ↔ m7, m15 ↔ m12, m16 ↔ m13, m17 ↔ m14, m6 ↔ m10
unified modelling	m12 ↔ m1/m4, m13 ↔ m2/m5, m14 ↔ m3/m7, m15 ↔ m1/m8, m16 ↔ m2/m9, m17 ↔ m3/m11
QBKRs per peptide	m1-m3 per peptide ↔ m1-m3 per peptide and repetition

Table 5.13:

by Andersson et. al. (2001) without interaction terms (see section **??**) are reproduced with interaction terms (m2, m4 and m6), and it is investigated whether their inclusion improves them.

In order to determine whether the more sophisticated representation of the physico-chemical properties of the amino acids at the mutation sites by the 26 variables used by Sandberg et. al. (1998) is advantageous compared to the quantification of these features with the help of the three ZZ-scales, the QSKR and unified models incorporating the 26 variables are compared to the corresponding models involving the three ZZ-scales.

The unified modelling procedure should be preferred to the performance of establishing different subgroup models because of the compactness of the resulting model and the fact that the influences of the numerous descriptor variables can be modelled simultaneously. However, it has to be investigated whether the respective unified regression models show at least the same prediction accuracy obtained by the separate descriptor subgroup models. Therefore, the conceivable unified regression models are compared to the corresponding QBKR and QSKR models.

The usefulness of establishing the QBKR models separately for each peptide in contrast to the performance presented in Andersson et. al. (2001) is evaluated by comparing these models with the corresponding QBKR models developed separately for each peptide and repetition.

5.5.2 Description of the regression models

In order to evaluate the novel modelling approaches for the interaction under study, several measures of interest were determined and summarized in tables for the diverse regression models. In fact, the number A of optimal iterations, the smallest number A^* of iterations resulting in insignificantly larger residuals than those obtained by the optimal model are given as well as the $Q^2_{A^*}$-values indicating the respective prediction accuracies obtained after A^* iterations. Further, the percentage $pctvar_{descr}$ of variation of the descriptor variables accounted for after A^* iterations and the percentage $pctvar_{resp}$ of variation of the considered response variable explained after A^* iterations are listed. In case of a multivariate regression model, the percentages of explained variation of both the association and dissociation rate constant were determined separately beyond the total variation of the response variables accounted for.

In table **??**, these measures are summarized for the QSKR and unified models incorporating the ZZ-scales or the 26 variables used by Sandberg et. al. (1998), respectively. In terms of the QBKR models established per peptide or per peptide and repetition, the corresponding measures are given in the appendix in table **??** and table **??**, respectively.

model		response	inter-action				*pctvar* [%]			
							$descr_{A^*}$	$resp_{A^*}$		
number	type	variable(s)	terms	A	A^*	$Q^2_{A^*}$		total	k_a	k_d
m4		k_a		3	-	0.77	42.79		84.27	
m4		k_a	x	7	3	0.81	44.56		87.84	
m5	Q	k_d		3	1	0.82	23.74			85.17
m5	S	k_d	x	15	2	0.93	26.63			95.23
m6	K	$log(k_d)$		5	3	0.75	48.48			78.71
m6	R	$log(k_d)$	x	12	10	0.97	88.01			98.20
m7		k_a, k_d		6	4	0.70	60.20	86.83	80.07	93.58
m7		k_a, k_d	x	7	5	0.80	66.69	90.16	90.06	90.26
m8	QS	k_a		1	0					
m9	KR-	k_d		4	3	0.92	37.32			95.22
m10	26	$log(k_d)$		12	4	0.94	45.51			94.81
m11	vars	k_a, k_d		4	0					
m12		k_a		8	4	0.62	31.74		62.89	
m12		k_a	x	15	-	0.78	76.76		79.81	
m13	uni-	k_d		12	9	0.93	62.70			93.32
m13	fied	k_d	x	15	13	0.99	73.25			99.30
m14		k_a, k_d		14	9	0.62	63.71	77.70	62.94	92.45
m14		k_a, k_d	x	15	-	0.77	78.52	88.42	78.26	98.57
m15		k_a		15	12	0.73	87.13		74.23	
m15	uni-	k_a	x	15	-	0.75	72.47		76.59	
m16	fied-	k_d		15	13	0.98	90.48			97.93
m16	26	k_d	x	15	-	0.99	73.39			99.62
m17	vars	k_a, k_d		15	-	0.73	93.36	85.96	74.21	97.70
m17		k_a, k_d	x	15	8	0.68	44.66	82.35	68.51	96.20

Table 5.14: Description of the QSKR and unified models involving the ZZ-scales or the 26 variables from Sandberg et. al. (1998) (further explanations in the text)

5.5.3 Evaluation of the novel modelling approaches

The evaluation of the novel modelling aspects is predominantly based on the comparison of the prediction accuracies represented by the $Q^2_{A^*}$-values of the different regression models. The reason for this comparison is the importance of the characteristic of a regression model to provide exact predictions. Though the $Q^2_{A^*}$-values can be expected to present too optimistic an impression of the obtained prediction accuracies, they can be used to compare the different regression models for this feature. The $Q^2_{A^*}$-values for the univariate QSKR and unified models are summarized in table **??**, where the prediction accuracies obtained by the modelling procedure applied in Andersson et. al. (2001) are in bold.

model type	QSKR		QSKR-26vars	unified		unified-26vars	
interaction terms		x			x		x
response variable							
k_a	**0.77**	0.81	-	0.62	0.78	0.73	0.75
k_d	0.82	0.93	0.92	0.93	0.99	0.98	0.99
logkd	**0.75**	0.97	0.94	-	-	-	-

Table 5.15: Summary of the $Q^2_{A^*}$-values obtained for the univariate QSKR and unified models

Further, the percentages of explained variation of the descriptor variables are taken into account. For the multivariate models, the percentages of explained variation of both the association and dissociation rate constant are considered in addition to the total prediction accuracy because these percentages also indicate the extent of exactness of the predictions that can be expected from the models.

Comparing the multivariate QSKR and unified regression models with the corresponding univariate models reveals that in most cases the multivariate modelling leads to worse results than the univariate. In detail, the percentages of variation accounted for by the multivariate models for the association and dissociation rate constant, respectively, are lower than those explained by the univariate models with two exceptions where these percentages are slightly larger. These exceptions refer to the QSKR model without interaction terms for the dissociation rate constant and the QSKR model with interaction terms respecting the association rate constant. The explained percentages of variation of the association and dissociation rate constant co-

incide with each other for the unified model incorporating the 26 variables by Sandberg et. al. (1998) without interaction terms. All of the total $Q^2_{A^*}$-values referring to the multivariate QSKR and unified regression models are lower than the respective $Q^2_{A^*}$-values of the univariate models. A multivariate QSKR model involving the 26 variables from Sandberg et. al. (1998) could not be developed because the number A^* of iterations resulting in insignificantly larger residuals compared to those obtained by the model based on the optimal number A of iterations was determined to be zero. For the multivariate QBKR models, the total $Q^2_{A^*}$-values are more often lower than 0.70 than those of the univariate QBKR models. Consequently, the multivariate modelling approach cannot be considered to be a useful alternative to the univariate modelling procedure for the interaction under investigation since it does not lead to more accurate predictions. This is the reason why the multivariate models are not taken into account in the model comparisons described below.

Many of the interaction terms included in the regression models show VIP-values larger than the value 1.00 and some of them even refer to VIP-values larger than the value 1.40. Consequently, interactions between the descriptor variables contribute relevantly to the binding behaviour of the interaction under study and should hence be incorporated in the regression analysis to allow for an explanation of the interferences between the various variables. A comparison of the prediction accuracies of the regression models involving interaction terms with those of the regression models without interaction terms stresses the fact that it might be useful to take interaction terms into account. In detail, the obtained prediction accuracies improved considerably by including interaction terms in the QSKR models in terms of the dissociation rate and logarithmic dissociation rate constants as well as in the unified model for the association rate constant. The prediction accuracies of the other models incorporating interaction terms are only slightly better than those of the models without interaction terms. In these cases, the interaction terms need not necessary be included in the models, and regression models without interaction terms are preferred as they are easier to interpret. Another interesting observation is that the inclusion of interaction terms in the QSKR and unified models involving the ZZ-scales as descriptor variables results in higher percentages of explained variation of the descriptor variables than is achieved by extracting more latent variable vectors. However, for the unified models using the 26 variables by Sandberg et. al. (1998) as descriptor variables, the percentages of variation accounted for of the descriptor variables are lower with interaction terms than without. Obviously, the fitting of the data suffers from taking into account too many descriptor variables

showing no relevance to the binding behaviour, and the latent variable vectors represent only a part of the data of the descriptor variables. Therefore, in case of a large number of descriptor variables, it is advisable not to include all conceivable interaction terms in the regression models but rather a subset of them thought to be relevant by biochemists.

The incorporation of the more detailed representation of the physico-chemical properties by the 26 variables mentioned in Sandberg et. al. (1998) seems to be advantageous to the inclusion of the ZZ-scales summarizing these variables since the prediction accuracies could be improved for almost all of the models by using these 26 variables instead of the ZZ-scales. The only exceptions refer to the unified model of the association and dissociation rate constant both with interaction terms where the use of these 26 variables results in the same or a slightly worse prediction accuracy, respectively. The QSKR model of the association rate constant could not be established under incorporation of the 26 variables from Sandberg et. al. (1998). The reason for this observation might be the fact that the experimental design was based on the values of the ZZ-scales and not on the values of these 26 variables. Accordingly, it can be expected that this model could be developed as well if a statistical design plan were derived with the help of the 26 physico-chemical properties. If the prediction accuracy of a model involving the 26 variables from Sandberg et. al. (1998) at least equals that of the corresponding model using the ZZ-scales, the model with the 26 variables should be preferred for interpretation and reporting since it provides more detailed information about the influencing physico-chemical properties of the amino acids at the mutation sites.

The unified modelling approach can be considered to be a useful alternative to the commonly separately performed QSKR and QBKR modelling. Comparing the prediction accuracies of the unified models with the corresponding QSKR models shows that the QSKR models in terms of the association rate constant are better than the unified models with one exception. However, the $Q^2_{A^*}$-values of the unified models incorporating interaction terms only differ slightly from the $Q^2_{A^*}$-value of the QSKR model without interaction terms. In comparison with the QSKR models of the dissociation rate constant, the unified models result in better or at least equal prediction accuracies. It has to be noted that for these unified models, $Q^2_{A^*}$-values of 0.93 and 0.98 without interaction terms and 0.99 with interaction terms were computed. Thus, these models can be expected to provide very reliable predictions in practice. The unified model of the association rate constant incorporating the 26 variables from Sandberg et. al. (1998) could be developed both with and without interaction terms in contrast to the corresponding QSKR model

and shows a sufficiently large value of 0.73 and 0.75, respectively, of the $Q^2_{A^*}$-statistic. Many of the QBKR models respecting the association rate constant both with or without interaction terms and both established per peptide or peptide and repetition show $Q^2_{A^*}$-values smaller than the value 0.70 or cannot be developed. Therefore, an alternative modelling procedure like the unified modelling approach is required for the association rate constant in order to obtain knowledge about the influence of the buffer variables. The results of the QBKR models in terms of the dissociation rate constant are better. Since the unified models in terms of the dissociation rate constant provide very exact predictions, it is no problem using these models instead of the QBKR models for the determination of the effect of the considered buffer components. The unified modelling procedure should be preferred to the descriptor variables subgroup models, not only if the resulting models lead to more accurate predictions but also if the prediction accuracy is slightly worse. This fact can be accepted because of the more compact presentation of the results in form of one single model instead of a QSKR model and 17 or 26, respectively, QBKR models and the resulting facility of the interpretation. Beyond this, the unified modelling allows for the quantification of interaction terms between the amino acid and buffer descriptor variables that might be suspected to be relevant.

The question of whether the QBKR models should be established per peptide or per peptide and repetition cannot be answered in general. If the QBKR models developed per peptide and repetition show approximately the same high $Q^2_{A^*}$-values for a particular peptide, the specification of a single QBKR model referring to this peptide is useful. Otherwise, the QBKR models per peptide result in bad prediction accuracies or even cannot be specified and hence, the models should be developed per peptide and repetition. This complication respecting the decision on the exact performance of establishing QBKR models also leads to the suggestion to prefer the unified modelling procedure to the descriptor variables subgroup modelling.

Summarizing, the model comparisons revealed that some of the novel modelling approaches can lead to an improvement of the exactness of the predictions compared to that one of the predictions calculated with the help of the models established according to the common performance presented by Andersson et. al. (2001). In particular, the incorporation of descriptor variables reflecting the physico-chemical properties of amino acids more detailed than the commonly used ZZ-scales as well as the inclusion of interaction terms improved the prediction accuracies of the developed regression models. Beyond this, the unified modelling procedure that should be preferred

to the subgroup modelling because of the interpretability resulted in models showing sufficiently large prediction accuracies regarding the association rate constant and very well prediction accuracies referring to the dissociation rate constant. Therefore, this modelling approach can be proposed as a standard procedure of performing a regression analysis with respect to biomolecular interactions.

The multivariate modelling cannot be considered advantageous to the univariate one for the interaction under study. However, for other interactions, multivariate models might be more appropriate than univariate ones. In contrast to this aspect that might be of different use for different interactions, the statements concerning the incorporation of interaction terms and more detailed descriptor variables as well as the unified modelling approach can be expected to be valid in general for the analysis of biomolecular interactions.

5.5.4 Specification of the optimal regression models

In the following, the regression models which are optimal for the resulting prediction accuracies are presented for the association rate and dissociation rate constants as well as for the logarithmic dissociation rate constant. In the established regression models, the symbol $*$ is used to indicate interaction terms. Since some of the 26 variables from Sandberg et. al. (1998) are indicator variables, the different adjustments of these variables are referred to by giving the corresponding number in the brackets [] in the regression models.

For both the association rate constant and the logarithmic dissociation rate constant, the most accurate predictions are provided by a QSKR model involving interaction terms, where the $Q^2_{A^*}$-values are 0.81 and 0.97, respectively. Consequently, the prediction accuracies of the models established according to the performance described in Andersson et. al. (2001) could be improved partially remarkably by merely including interaction terms in the regression models. In terms of the dissociation rate constant, the best prediction accuracy with a $Q^2_{A^*}$-value of 0.99 was obtained by both the unified model including the 26 variables from Sandberg et. al. (2001) and interaction terms and the unified model with the ZZ-scales and interaction terms, where the latter model is chosen to be presented in section **??** in the appendix.

$$
\begin{aligned}
ka \;=\;& 838227 - 111847 \cdot HFT_{142}(-7907 \cdot ZZ2_{142})(+7538 \cdot ZZ3_{145})(+883 \cdot ZZ3_{146}) \\
& -5820 \cdot ZZ1_{142} * ZZ2_{142} - 97370 \cdot HFT_{142} * HFT_{146} - 4119 \cdot ZZ1_{145} * ZZ2_{145} \\
& +5836 \cdot ZZ3_{145} * ZZ1_{146}(+820 \cdot ZZ1_{142} * ZZ3_{145})(-1245 \cdot ZZ1_{142} * ZZ3_{146}) \\
& (-7586 \cdot ZZ2_{142} * HFT_{145})(-5955 \cdot ZZ2_{142} * HFT_{146})(-3820 \cdot ZZ3_{142} * ZZ2_{145}) \\
& (+6627 \cdot ZZ3_{142} * HFT_{145})(+1228 \cdot ZZ3_{142} * ZZ3_{146})(+8962 \cdot HFT_{142} * ZZ3_{145}) \\
& (-15292 \cdot HFT_{142} * ZZ1_{146})(-1880 \cdot HFT_{142} * ZZ3_{146})(-1014 \cdot ZZ1_{145} * ZZ3_{146}) \\
& (-10529 \cdot ZZ2_{145} * HFT_{145})(-3918 \cdot ZZ2_{145} * ZZ1_{146})(-2921 \cdot ZZ2_{145} * ZZ3_{146}) \\
& (+7362 \cdot ZZ3_{145} * HFT_{145})(-307 \cdot ZZ3_{145} * ZZ3_{146})(+4778 \cdot ZZ3_{145} \cdot HFT_{146}) \\
& (-7687 \cdot HFT_{145} * ZZ2_{146})(+3694 \cdot HFT_{145} * ZZ3_{146})(-769 \cdot ZZ1_{146} * ZZ3_{146}) \\
& (-627 \cdot ZZ3_{146} * HFT_{146})
\end{aligned}
$$

$$
\begin{aligned}
logkd \;=\;& -4.534328 - 0.672850 \cdot HFT_{145} - 0.106923 \cdot ZZ3_{146}(+0.038284 \cdot ZZ1_{142}) \\
& (-0.075566 \cdot ZZ2_{145})(+0.006403 \cdot ZZ2_{146})0.242483 \cdot HFT_{142} * HFT_{145} \\
& -0.005436 \cdot HFT_{142} * ZZ3_{146} - 0.102806 \cdot ZZ1_{145} * HFT_{145} \\
& -0.028224 \cdot ZZ1_{145} * ZZ3_{146} - 0.101335 \cdot HFT_{145} * ZZ3_{146} \\
& -0.091297 \cdot ZZ3_{146} * HFT_{146}(+0.059251 \cdot ZZ1_{142} * HFT_{142}) \\
& (+0.016928 \cdot ZZ1_{142} * ZZ1_{145})(+0.018373 \cdot ZZ1_{142} * HFT_{145}) \\
& (+0.128222 \cdot ZZ2_{142} * ZZ3_{146})(-0.044738 \cdot ZZ3_{142} * ZZ1_{145}) \\
& (-0.067007 \cdot ZZ3_{142} * ZZ2_{145})(-0.087835 \cdot ZZ3_{142} * ZZ3_{146}) \\
& (-0.046019 \cdot HFT_{142} * ZZ2_{145})(-0.136094 \cdot ZZ2_{145} * ZZ3_{145}) \\
& (-0.135864 \cdot ZZ2_{145} * HFT_{145})(-0.040037 \cdot ZZ2_{145} * ZZ1_{146}) \\
& (+0.000955 \cdot ZZ2_{145} * ZZ3_{146})(-0.000476 \cdot ZZ2_{145} * HFT_{146}) \\
& (-0.041420 \cdot HFT_{145} * ZZ1_{146})(+0.071831 \cdot HFT_{145} * ZZ2_{146}) \\
& (-0.545765 \cdot HFT_{145} * HFT_{146})(-0.003467 \cdot ZZ1_{146} * ZZ2_{146}) \\
& (-0.029517 \cdot ZZ1_{146} * ZZ3_{146})
\end{aligned}
$$

Further models are specified in section **??** in the appendix in order to provide information about the relevance of the 26 variables from Sandberg et. al. (1998) and to determine the influence of the buffer variables and the interactions between the amino acid and buffer variables. These models can be considered to be useful in practice because of their potential of biochemical conclusions that might be drawn from them though they show slightly worse prediction accuracies compared to the optimal ones. The corresponding prediction accuracies can be found in table **??**.

In detail, the QSKR model involving the 26 variables from Sandberg et. al. (1998) is shown for the logarithmic dissociation rate constant. Furthermore, the unified model incorporating these 26 variables without interaction terms is presented for the dissociation rate constant. For the association rate constant, the unified model including the ZZ-scales with interaction terms and the unified model using the 26 variables from Sandberg et. al. (1998) without interaction terms are established. Though the ZZ-scales do not represent the physico-chemical properties of amino acids as detailed as these 26 variables do, the model with the ZZ-scales is specified in order to permit a comparison with the statements obtained with the help of the QSKR models.

The determination of the descriptor variables that have to be included in the regression models is based on the VIP-value criterion, i.e. those variables are considered to be relevant that show VIP-values larger than the value 1.00. The descriptor variables referring to the VIP-values between 1.00 and 1.40 are judged to be less important than those belonging to VIP-values larger than the value 1.40 and are hence given in brackets.

In order to present the relevance of the various descriptor variables for the different regression models, the VIP-values were summarized in tables in the appendix, where those values between 1.00 and 1.40 are given in brackets. Because of the enormous number of interaction terms referring to VIP-values larger than the value 1.00, these descriptor variables were not listed in these tables.

In table **??** and table **??**, the VIP-values referring to the QSKR models involving the ZZ-scales or the 26 variables from Sandberg et. al. (1998), respectively, are presented. Those VIP-values belonging to the unified models using the ZZ-scales or the 26 variables mentioned by Sandberg et. al. (1998) are listed in table **??** and table **??**, respectively. Furthermore, the VIP-values referring to the QBKR models established per peptide or per peptide and repetition, respectively, are summarized in table **??** and in table **??**, respectively.

5.5.5 Biochemical conclusions

In the following, biochemical conclusions concerning the association and dissociation rate constants as well as the logarithmic dissociation rate constant are drawn from the regression models specified in the previous subsection and in section **??** in the appendix. This task is of special importance in practice in order to improve knowledge about the binding process under study and to understand the predictions of the kinetic parameters for a particular mutant.

With the help of the established regression models, it is possible to determine the relevant descriptor variables with respect to the interaction between the TMVP and the antibody 57P as well as the nature of their influence. However, further biochemical interpretations, in particular in case of which constellation of the values of the association and dissociation rate constants caused by a specific mutant with determined physico-chemical properties the existing antibody can still be expected to be effective, requires the judgement of a biochemist.

Statements concerning the affinity of the interaction under examination are not presented since they can be concluded directly from those ones respecting the association and dissociation rate constant that are more relevant.

Conclusions concerning the association rate constant

In the QSKR model involving the ZZ-scales with interaction terms that shows the best prediction accuracy in terms of the association rate constant, the HFT-scale at position 142 is determined to be relevant in accordance with the statement in Andersson et. al. (2001). The ZZ3-scale at position 145 included in the model presented in Andersson et. al. (2001) though it was not evaluated to be significant, is incorporated in brackets like the ZZ2-scale at position 142 and the ZZ3-scale at position 146 in the model specified in the previous subsection. Furthermore, four interaction terms showing a negative effect with one exception are incorporated in the model as well as a number of additional interaction terms given in brackets since they refer to VIP-values between 1.00 and 1.40.

Consequently, a mutant characterized by a larger value of the HFT-scale at position 142 than that one of the amino acid serine, i.e. has a stronger tendency to adopt a helical structure at this location than it is present in the wild-type peptide, the association rate constant of the interaction with the antibody 57P is reduced. Since the process of association between the existing antibody and this particular mutant would be limited compared to the association between the antibody and the wild-type peptide, the antibody might be not effective any more depending on the extent of reduction of the association rate constant. However, the influences of the other possibly relevant physico-chemical properties at the specific positions as well as of the interaction terms determined to be important should also be taken into account in these considerations to improve understanding of the interaction under study.

A unified model involving the ZZ-scales with interaction terms is specified in the appendix in order to present additional information about the influence of the buffer variables as well as of the interactions between the amino acid and buffer variables on the association of the interaction of interest. Further, a unified model incorporating the 26 variables from Sandberg et. al. (1998) without interaction terms is established that can be used to obtain more detailed knowledge of the relevant physico-chemical properties of the amino acids at the mutation sites.

Both models reveal that the chemical additives NaCl and urea influence the association process negatively, i.e. the higher the concentrations of these buffer components, the smaller is the association rate constant. The observation that NaCl has an effect on the association rate constant leads to the suggestion that electrostatic forces contribute to the binding process under investigation. According to Andersson et. al. (2001), the fact that both the pH-value and the chemical additive EDTA do not influence the association rate constant indicates that Fab histidines as well as metal ions are probably not incorporated in the binding process, but hydrogen bonds or ionic interactions are relevant.

In the unified model with the ZZ-scales and the interaction terms, a number of interaction terms refer to VIP-values larger than the value 1.40. In detail, six buffer interaction terms were determined to be relevant, most of them with a positive influence. All of the seven important interaction terms between the buffer variables and the physico-chemical properties of the amino acids at the mutation sites show a negative effect on the association rate constant.

For the 26 variables from Sandberg et. al. (1998) whose influences are quantified in the unified model without interaction terms, it can be stated that the physico-chemical properties of the amino acid at position 145 are predominantly important for the association rate constant. In particular, the molecular weight, the van der Waals volume of the side chain as well as the total surface area of the amino acid at this position have a negative effect on the association of the interaction under study. Beyond this, one of the nuclear magnetic resonance shift variables at this mutation site influences the association positively.

Consequently, the three regression models established for the association rate constant contribute relevantly to a better understanding of the binding process under investigation by quantifying the influences of the physico-chemical

properties of the amino acids at the mutation sites, of the buffer variables as well as of the interactions between these variables. In order to obtain reliable predictions of the association rate constant, the QSKR model involving the ZZ-scales and interaction terms should be applied.

Conclusions concerning the dissociation rate constant

The optimal model concerning the dissociation rate constant, i.e. the unified model incorporating the ZZ-scales and interaction terms, reveals that different physico-chemical properties of the amino acids at all of the mutation sites influence the dissociation process. In detail, the ZZ-scale and HFT-scale at position 145 and the ZZ3-scale at position 146 have a relevant negative effect on the dissociation rate constant, whereas the ZZ1-scale at position 142 shows an important positive effect.

Accordingly, mutants characterized by larger values of the relevant physico-chemical properties, compared to those for the wild-type peptide, at the corresponding mutation sites with a negative influence show a smaller value of the dissociation rate constant in comparison with that measured for the wild-type peptide. In detail, mutants with amino acids at position 145 being larger and more polarizable and having a stronger tendency to adopt a helical structure compared to the features of the amino acid glutamic acid of the wild-type at this position or an amino acid with larger values of the electronic properties at position 146 than serine lead to a reduced dissociation process compared to that one of the interaction involving the wild-type peptide. Furthermore, mutants with larger values of the ZZ1-scale at position 142, i.e. with a more hydrophobic amino acid than serine, lead to an increased value of the dissociation rate constant compared to that one referring to the wild-type peptide.

A number of interaction terms are incorporated additionally in the unified regression model, where many of them refer to VIP-values larger than the value 1.40. The interaction terms respecting the physico-chemical properties of the amino acids at the mutation sites show both a negative and a positive effect, whereas the interactions between the buffer and amino acid variables have a negative influence on the dissociation rate constant with one exception. The interaction terms between the buffer and amino acid variables can be used to explain the differences of the influences of the chemical additives between the various peptides observed in the QBKR models.

Buffer variables were not determined to be important in the unified model though the VIP-values concerning the corresponding QBKR models indicate

that DMSO, NaCl as well as urea influence the dissociation rate constant relevantly. The reason for this observation that the buffer variables are not included in the model might be the fact that the incorporated physico-chemical properties of the amino acids already explain the respective influences. In particular, the contribution of DMSO is known to reflect the importance of hydrophobicity being represented by the ZZ1-scale determined to be relevant. Further, the influence of NaCl reveals the contribution of electrostatic forces. These forces are reflected as well by the ZZ3-scale incorporated in the model. However, a corresponding relationship concerning the chemical additive urea is not known but is expected to be existent with respect to either the ZZ2-scale or the HFT-scale.

In order to obtain a more detailed identification of the physico-chemical properties influencing the dissociation rate constant, the unified model including the 26 variables from Sandberg et. al. (1998) without interaction terms, given in the appendix, is considered. According to the unified model using the ZZ-scales, the HFT-scale at position 145 is determined to be relevant. Beyond this, several additional physico-chemical properties at the position 142 and 146 refer to VIP-values larger than the value 1.40. This fact that the model involving the 26 variables from Sandberg et. al. (1998) might be relatively complex could be suspected from the observation that all of the three ZZ-scales were determined to be relevant in the corresponding unified model. In detail, the retention values TL4 and TL7 as well as the semi-empirical molecular orbital index ELUMO of the amino acid at position 142 and the logP-value, the polar surface area and the hydrogen bond donor property of the amino acid at position 146 influence the dissociation rate constant relevantly. With the exception of the variable ELUMO, the effects are positive. Obviously, large values of these physico-chemical properties might correspond partially to small values of the ZZ-scales for which a negative effect was determined.

Conclusions concerning the logarithmic dissociation rate constant

The optimal model respecting the logarithmic dissociation rate constant, i.e. the QSKR model involving the ZZ-scales with interaction terms, shows that the HFT-scale at position 145 and the ZZ3-scale at position 146 influence the logarithmic dissociation rate constant in a negative way. Consequently, amino acids with a higher value of the helix-forming tendency at position 145 lead to a reduction in the logarithmic dissociation rate constant as amino acids with higher values of certain electronic properties like charge, polarity, electrophilicity and electronegativity. This observation coincides with the

results reported in Andersson et. al. (2001). However, the model specified in the previous subsection can be expected to provide more exact predictions than that one presented in Andersson et. al. (2001) as it is indicated by a higher $Q^2_{A^*}$-value.

Beyond the HFT-scale at position 145 and the ZZ3-scale at position 146, a few interaction terms are determined to be relevant. The six interaction terms corresponding to VIP-values larger than the value 1.40 have a negative effect on the logarithmic dissociation rate constant with one exception. In addition to the HFT-scale at position 145 and the ZZ3-scale at position 146, the HFT-scale at the other two mutation sites as well as the ZZ1-scale at position 145 contribute to the most relevant interaction terms. Further interaction terms referring to VIP-values between 1.00 and 1.40 that are consequently of minor importance were included in brackets in the model.

The statement that the ZZ3-scale at position 146 influences the logarithmic dissociation rate constant relevantly gives first hints that electronic properties play an important role. In order to determine the contributing properties more exactly, the QSKR model involving the 26 variables from Sandberg et. al. (1998) without interaction terms were established in the appendix. This model permits the detailed identification of those physico-chemical properties influencing the logarithmic dissociation rate constant.

In detail, a negative side chain at position 145 results in a lower value of the logarithmic dissociation rate constant, whereas the factors of the indicator variable representing the hydrogen bond acceptor properties show both a negative or a positive effect on the logarithmic dissociation rate constant in dependence on the position of the amino acid. A number of further descriptor variables are included in the model that correspond to VIP-values between 1.00 and 1.40 and are consequently relatively important as well.

Summarizing, the QSKR model involving the ZZ-scales with interaction terms can be used to obtain accurate predictions of the logarithmic dissociation rate constant and knowledge of the relevance of the considered physico-chemical properties of the amino acids at the mutation sites. For more detailed information about the physico-chemical properties of importance, the QSKR model including the 26 variables from Sandberg et. al. (1998) without interaction terms can be taken into account.

Summary of the biochemical conclusions

With the help of the established regression models, the relevant descriptor variables could be determined and their effects on the interaction under study could be quantified. In order to present a summary of the descriptor variables influencing the association and dissociation rate constant predominantly, the variables determined to be relevant in terms of these response variables are listed in table **??**, where the kind of the respective effect is indicated by a plus or a minus. Interaction terms are not incorporated in this table.

The regression analysis revealed that different physico-chemical properties of amino acids at different mutation sites influence the association and dissociation. Obviously, different intermolecular forces contribute to these kinetic processes. Table **??** illustrates that most of the descriptor variables lead to undesired kinetic characteristics when they attain larger values compared to the values referring to the wild-type. This statement is based on the fact that almost all of the relevant descriptor variables with respect to the association rate constant have a negative influence, whereas most of the descriptor variables determined to be relevant regarding the dissociation rate constant show a positive effect. Consequently, larger values of the respective descriptor variables compared to those of serine or glutamic acid of the wild-type peptide result in a reduced association process or an increased dissociation process, respectively, in comparison with those regarding the wild-type peptide.

response variable		k_a	k_d
descriptor variable	position		
HFT	142	-	
	145		-
MW	145	-	
vdW	145	-	
Stot	145	-	
NM12	145	+	
TL4	142		+
TL7	142		+
ELUMO	142		-
logP	146		+
Spol	146		+
[4] HDONR	146		+

Table 5.16: Summary of the influences of the most relevant descriptor variables on the association and dissociation rate constant

By determining the mutation sites where amino acid replacements lead predominantly to changes in the binding behaviour, functional domains of the TMVP could be identified. In detail, position 145 can be considered to be crucial for the association process, whereas positions 142 and 146 play an important role in terms of the dissociation process.

Based on the results of the regression analysis, it is possible to judge the efficacy of the antibody 57P for a particular occurred mutant. For mutants presumably leading to an increased association process or a decreased dissociation process compared to those ones referring to the wild-type peptide, it can be expected that the existing antibody 57P is still effective. However, for mutants resulting in a decreased association process or an increased dissociation process in comparison with that ones regarding the wild-type peptide, the available antibody might not be efficient anymore, depending on the extent of change in the kinetic parameters. Whether a particular mutant will show an increased or decreased association or dissociation process, respectively, can be determined by taking into account the effects of the relevant descriptor variables in the established regression models.

For any potential occurring mutant, the values of the important physicochemical properties at the relevant positions can be determined and used in the established regression model to obtain a prediction of the association rate constant of the interaction between this mutant and the existing antibody. By comparing this prediction with that referring to the wild-type peptide, conclusions can be drawn on whether the available antibody might be still effective for the particular mutant. In this context, the additional knowledge of biochemists is required.

Chapter 6

Summary and outlook

This dissertation provides a comprehensive presentation of both the methodology of Partial Least Squares (PLS) regression and its application to the analysis of biomolecular interactions. The explanations and derivations referring to the performance of PLS regression contribute to the completeness of a statistically exactly described methodology of this procedure. Furthermore, the general realization of biomolecular interaction studies is presented in detail. Beyond this, the modelling procedure with respect to a particular binding process, the interaction between an antigen of the tobacco mosaic virus protein (TMVP) and the corresponding antibody 57P, is optimized by applying novel modelling approaches. Consequently, the knowledge of influences on this particular interaction are extended, providing the possibility of determining functional domains of the TMVP and predicting the kinetic parameters with respect to an occurred mutant of this virus.

PLS regression is a method that can be used in cases when the Ordinary Least Squares procedure is not applicable to data showing collinearity among the descriptor variables, e.g. if the number of descriptor variables exceeds the number of objects in the sample. By applying PLS regression, the specification of the corresponding regression model can nevertheless be performed. In the course of the algorithm, latent variable vectors, which compress the information of the descriptor variables using the information of the response variables, are computed. Further terms obtained in the iterations of the algorithm can be used to estimate the unknown model parameters of the regression model.

In biomolecular interaction studies, the aim is to gain knowledge of the factors a certain interaction depends upon. Therefore, regression models relating measured kinetic parameters of the binding process to a number of potential

descriptor variables are established. Since often only a limited number of measurements can be performed but a multitude of possibly relevant factors is included in the regression models, PLS regression is particularly useful in this context.

The case-study concerns the interaction between an antigen of the tobacco mosaic virus protein (TMVP) and a Fab fragment of the antibody 57P. In particular, the objective was to establish useful regression models with the help of the available data to draw biochemical conclusions concerning the interaction of interest and to provide the possibility of judging the efficacy of the existing antibody with respect to a specific mutation of the tobacco mosaic virus. This kind of information can also be used to elucidate the binding behaviour of similar viruses such as Orthomyxovirus causing influenza.

By repeating the analysis of the data obtained for investigating the interaction of interest, the most important conclusions on the relevant descriptor variables could be reproduced from the VIP-value criterion. However, the new model parameter estimates, i.e. the intercepts and regression coefficients, differed from those found earlier. Explanations for these differences are given. The comparison between the new and old models illustrates the importance of a detailed documentation of the application of PLS regression to permit a reproduction of the regression analysis results.

In general, in biomolecular interaction studies, univariate regression models are presented that are specified individually with respect to the different subgroups of descriptor variables. In detail, the following types of regression models are developed: quantitative buffer-kinetics relationship, quantitative sequence-kinetics relationship, quantitative structure-activity relationship or 3D-quantitative structure-activity relationship. These are referred to respectively as (QBKR),(QSKR), (QSAR)or (3D-QSAR)-models.

Usually, interaction terms between the different descriptor variables are not considered in the regression analysis. In this dissertation, alternative modelling approaches were proposed and evaluated in terms of the improvement in the resulting prediction accuracy.

Consequently, a number of different types of models were established. In detail, both univariate and multivariate models were obtained by applying either the univariate or the multivariate PLS algorithm. Further, the different subgroup models were derived as well as a unified regression model incorporating simultaneously all of the potential descriptor variables. The

regression models were developed both with and without interaction terms. Additionally, regression models were established using different quantifications of the physico-chemical properties of amino acids, either the helix-forming tendency- (HFT-)scale and the three ZZ-scales summarizing diverse features or the HFT-scale and the 26 variables from Sandberg et. al. (1998).

The model comparisons showed that some of the novel modelling approaches are advantageous compared with the commonly performed modelling procedure with respect to the analysis of the interaction between the TMVP and the antibody 57P. In particular, the prediction accuracies obtained for the models based on the novel approaches could be improved substantially in comparison with those already published.

The multivariate modelling did not improve the prediction accuracies of the established models and hence, the univariate modelling should be preferred to apply to the data available for the interaction under examination. However, it could be possible that the multivariate modelling leads to better results than univariate in studies of other interactions. Consequently, both modelling approaches should be performed and evaluated in future interaction studies.

It is particularly of note that the inclusion of interaction terms led to models providing considerably more accurate predictions than those without them. However, it is advisable to involve only those interactions that are suspected to be relevant by biochemists. In this research, interactions between the amino acid and buffer descriptor variables were assumed to be present and hence modelled. Therefore, with the exception of the QSKR model incorporating the 26 variables from Sandberg et. al. (1998), all potential interactions were modelled. However, further biochemical knowledge on possible interactions between the descriptors was not available.

Further, it was shown that the inclusion of descriptor variables representing the physico-chemical properties of amino acids which are more sophisticated than the commonly used ZZ-scales results in models with a higher prediction accuracy. Consequently, the statistical design should be based on more detailed variables such as the 26 variables from Sandberg et. al. (1998) in order to optimize the results in future applications.

The unified modelling approach was demonstrated to be a useful alternative to the usual method of developing separate regression models for the different subgroups of descriptors. An important advantage of the unified regression

models is the fact that the results of the regression analysis can be presented in one single model. This considerably facilitates the biochemical interpretation. Additionally, it permits the quantification of interactions between the amino acid and buffer variables, a relevant aspect in practice for obtaining comprehensive information on the influences on the interaction of interest.

In summary, the optimal modelling procedure for the analysis of biomolecular interactions is the construction of unified regression models involving descriptors reflecting both the buffer composition and the physico-chemical properties of amino acids in detail, as well as relevant interaction terms which have biochemical justification. This is demonstrated by considering both the resulting prediction accuracies of the derived models as well as their interpretability. Consequently, these kinds of models should be developed in order to draw biochemical conclusions on the interaction under study.

With the help of both the optimal models and further useful specified models referring to the association and dissociation rate constant, biochemical conclusions improving the understanding of the interaction between the TMV and the antibody 57P were drawn. In detail, the descriptor variables influencing the interaction under study as well as the nature of their effect were determined.

In relation to the association rate constant, the following physico-chemical properties were evaluated to have a negative effect on the interaction of interest: the helix-forming tendency of the amino acid at position 142 as well as the molecular weight, the van der Waals volume of the side chain and the total surface area of the amino acid at the mutation site 145. Further, a nuclear magnetic resonance shift variable at position 145 shows a positive effect. Beyond this, the chemical additives NaCl and urea also influence the association process in negatively. Consequently, since changes in the physico-chemical properties of the amino acid at the mutation site 145 lead predominantly to changes in the association rate constant, this position in the sequence of the TMVP can be considered to be crucial for the association process. At position 142, only the helix-forming tendency plays an important role during the association, whereas the mutation site 146 seems not to be incorporated relevantly in the recognition of the biomolecules.

With respect to the dissociation rate constant, the following physico-chemical properties show a negative influence: the helix-forming tendency and a semi-empirical molecular orbital index at position 142, as well as the size and polarizability of the amino acid at the mutation site 145 and diverse electronic

properties summarized by the ZZ3-scale at position 146. Positive effects were determined for the hydrophobicity and two retention values of the amino acid at the mutation site 142 and the logP-value, the polar surface area and the hydrogen bond donor property of the amino acid at position 146. Obviously, in contrast to the association process, the physico-chemical properties of the amino acid at the position 145 seem to play a minor role during the dissociation process. Instead, the mutation sites 142 and 146 are important with respect to the rate of decay of complexes of bound biomolecules.

Summarizing, it can be stated that different physico-chemical properties as well as different mutation sites are relevant with respect to the association and dissociation rate constant of the interaction under investigation. Obviously, different molecular forces at different positions in the sequence of the TMVP contribute to the process of assembly and decomposition of the complexes of studied biomolecules. These statements illustrate the fact that the application of PLS regression to the analysis of biomolecular interactions can be used to identify functional domains which are relevant to the different kinetic processes.

In any case of the occurrence of a new mutant of the TMV, the respective relevant physico-chemical properties at the important mutation sites can be related to those of the amino acids serine or glutamic acid, respectively, i.e. the amino acids of the wild-type peptide. This comparison permits statements concerning the changes in the association and dissociation of the interaction between the mutant and the antibody 57P that can be expected with reference to the interaction between the wild-type virus and this antibody. The available antibody 57P can be expected to be less or even not at all effective in case of a reduced association rate constant or an increased dissociation rate constant in comparison with the kinetic parameters referring to the wild-type of the TMV.

With respect to the association rate constant, mainly negative effects of the relevant descriptor variables were determined, in contrast to the dissociation rate constant. Consequently, it is unfavourable if the physico-chemical properties which negatively influence the association rate constant at the particular mutation sites correspond to larger values for a mutant than those at the respective positions in the wild-type peptide. Further, the larger the concentrations of NaCl and urea in the cells of the organism infected by the mutant are, the more the association rate constant is reduced. In relation to the dissociation rate constant, it is disadvantageous if the physico-chemical properties with a positive effect attain larger values at the specific mutation sites of a mutant compared with the values in the wild-type peptide.

Exact predictions of the kinetic parameters for the mutant can be obtained by applying the established optimal regression models for the association and dissociation rate constant. However, a reliable assessment of the efficacy of the existing antibody for a particular mutant on the basis of the predicted values of the association and dissociation rate constant requires the judgement of a biochemist.

In order to extend the knowledge of the interaction between the TMVP and the antibody 57P, further experiments might be performed. In particular, additional mutation sites and replaced amino acids should be taken into account. Further, the range of the concentrations of the chemical additives in the buffers might be extended.

A key problem is the question of for which other positions in the sequence the statements derived for the mutation sites studied here are also valid. This corresponds to extending the information about functional domains regarding the association and dissociation that can be achieved by taking into account further mutation sites in the experiments. The determination of functional domains is useful for improving the understanding of the binding process. From this type of knowledge, in a case of a new mutant, it can be concluded whether the mutation occurred at a crucial position for the interaction or whether or not it can be expected to influence the association or dissociation. In this context, the additional knowledge of biochemists about the virus under study is required. The possibility of generalizing the results obtained with the help of the established regression models is important in so far as the mutations in a virus might occur at any position of the sequence.

In Andersson et. al. (2001), problems arising during the QBKR modelling of the association rate constant measured in the perturbation buffers are described. In fact, some of the observations of the association rate constant are quite unreliable because of a lack of the required knowledge of the concentration of the Fab fragment biomolecules in the perturbation buffers. Consequently, the measured data could be only used partially. Therefore, in future investigations of the binding characteristics of particular interactions, it is important to perform experiments in which the concentrations of the involved biomolecules are determined exactly in order to obtain more reliable measurements of the association rate constants.

In order to obtain improved results of regression models established in future with respect to interactions respecting transmembrane proteins, variables quantifying the 2D-structure of the transmembrane biomolecules might also

be incorporated in the modelling. Consequently, results from the research of the topology of the transmembrane proteins involved in the binding could be used. For example, the prediction of the topology based on a Bayesian approach applying the Gibbs Sampling algorithm can be considered to provide a reliable descriptor variable of the 2D-structure. This procedure presented by Sousa et. al. (2004) and Kirschbaum (2005) in a more generalized form results in a predicted localization for each amino acid in the protein sequence.

Conventionally, the ability of the resulting regression models to provide correct predictions is evaluated with the help of the Q_a^2-statistic. However, Freyhult et. al. (2005) propose determining the prediction accuracy of regression models by performing a repeated blind cross validation procedure and using the resulting mean of the P^2-values. This average P^2-value can be considered to be a more realistic measure of the prediction accuracy than the Q_{A*}^2-value. Therefore, the repeated blind cross validation method should be applied to the regression models under examination in addition to the calculation of the Q_{A*}^2-value in future biomolecular interaction studies in order to judge the prediction accuracy.

A future aim in the context of the application of PLS regression might be the optimization of the modelling procedure with respect to other diverse interactions of interest. Further, modifications of the performance of PLS regression might be required in future applications of this methodology. For example, another field of application of PLS methodology is the investigation of mass spectrometry data in proteomics. The PLS procedure can be used to identify biomarkers in samples whose protein expression levels are measured by the Matrix-assisted laser desorption / ionization time-of-flight mass spectrometer (MALDI-TOF-MS) as presented by Podwojski et. al. (2006). In this paper, the PLS method is applied as a classification procedure in combination with linear discriminant analysis. By realizing this analysis, those proteins presenting biomarkers for particular disease states can be determined and it is possible to assign new probes to the given classes. Another example for the successful application of the PLS procedure combined with discriminant analysis (PLS-DA) is presented by Lee et. al. (2003). In this article, the PLS method is also used to classify spectrometric data.

Finally, the analysis of dynamic protein expression from difference gel electrophoresis deserves special interest. Jung et. al. (2005) discuss the preprocessing of proteomic data and present a method for the imputation of missing values. Their analysis of time dependent proteome changes resulting from the activation of Tyrosinkinase receptors by their ligand NGF (nerve growth

factor) identifies candidate tumor markers for neuroblastoma. Further new statistical methods and their applications in functional genomics and clinical proteomics can be found in Urfer and Amaral Turkman (2006).

Appendix A

The 26 variables used by Sandberg et. al. (1998)

Amino acid	MW	TL1	TL2	TL3	TL4	TL5	TL6	TL7	vdW	NM1	NM7	NM12	logP
Alanine	89.1	60	24	29	9	52	23	37	13.7	4.13	3.77	3.30	-3.12
Arginine	174.2	1	9	6	1	19	7	1	64.9	4.06	3.76	3.20	-4.79
Asparagine	132.1	39	14	25	5	21	12	12	32.5	4.33	4.00	3.56	-3.63
Aspartic Acid	133.1	66	16	9	2	38	20	19	30.0	4.34	3.89	3.51	-2.43
Cysteine	121.2	35	23	27	40	46	63	81	25.0	4.30	3.97	3.00	-2.35
Glutamine	146.1	53	16	37	7	32	20	23	42.7	4.10	3.76	3.24	-3.46
Glutamic Acid	147.1	74	24	15	4	42	19	28	40.2	4.12	3.74	3.21	-2.72
Glycine	75.1	44	19	22	6	32	22	17	3.5	3.88	3.54	3.15	-3.21
Histidine	155.2	9	6	16	1	16	20	2	45.1	4.29	4.00	3.47	-3.73
Isoleucine	131.2	80	50	49	28	72	53	83	44.4	4.02	3.66	3.08	-1.76
Leucine	131.2	82	52	50	33	73	53	83	44.4	4.05	3.71	3.23	-1.67
Lysine	146.2	1	7	2	1	19	7	1	51.1	4.05	3.74	3.27	-3.42
Methionine	149.2	77	45	50	29	63	49	79	45.0	4.20	3.84	3.30	-1.73
Phenylalanine	165.2	82	50	52	35	66	52	87	56.1	4.31	3.98	3.48	-1.56
Proline	115.1	52	20	50	11	57	31	68	30.7	4.37	4.11	3.51	-2.41
Serine	105.1	44	19	15	6	36	23	13	18.3	4.18	3.87	3.31	-2.74
Threonine	119.1	58	25	27	11	46	39	30	28.5	3.97	3.57	3.08	-2.43
Tryptophan	204.2	86	54	54	40	49	51	88	75.1	4.36	4.05	3.56	-1.57
Tyrosine	181.2	81	50	47	32	68	51	80	61.6	4.30	3.90	3.40	-2.22
Valine	117.2	74	42	43	22	68	43	72	34.1	3.94	3.60	3.03	-2.29

Table A.1: The 26 variables used for the derivation of the ZZ-scales

Amino acid	EHOMO	ELUMO	HOF	POLAR	EN	HA	Stot	Spol	Snp	HDONR	HACCR	Chpos	Chneg
Alanine	-10.09	1.15	-92.6	5.34	4.47	5.62	133	76	58	0	0	0	0
Arginine	-9.64	1.03	-72.3	11.85	4.31	5.33	221	132	90	4	3	1	0
Asparagine	-10.43	0.68	-132.4	7.72	4.87	5.56	160	121	39	2	3	0	0
Aspartic Acid	-10.35	0.71	-182.3	7.25	4.82	5.53	158	118	41	1	4	0	1
Cysteine	-9.51	0.27	-86.5	7.07	4.62	4.89	150	74	76	0	0	0	0
Glutamine	-10.23	1.01	-137.8	8.88	4.61	5.62	184	126	58	2	3	0	0
Glutamic Acid	-10.37	0.82	-189.8	8.52	4.77	5.60	179	121	59	1	4	0	1
Glycine	-10.21	1.08	-89.0	4.18	4.56	5.64	112	79	33	0	0	0	0
Histidine	-9.16	0.79	-37.4	10.46	4.19	4.98	196	98	98	1	1	1	0
Isoleucine	-10.03	1.16	-109.7	8.83	4.44	5.59	189	71	118	0	0	0	0
Leucine	-10.04	1.17	-108.6	8.82	4.44	5.60	188	70	118	0	0	0	0
Lysine	-9.79	1.13	-108	9.74	4.33	5.46	203	96	108	2	1	1	0
Methionine	-8.61	0.68	-96.9	9.68	3.97	4.65	194	71	125	0	0	0	0
Phenylalanine	-9.74	0.10	-68.0	12.76	4.82	4.92	212	73	140	0	0	0	0
Proline	-9.98	0.79	-94.0	7.47	4.60	5.38	159	60	100	0	0	0	0
Serine	-10.26	0.86	-141.5	5.83	4.70	5.56	137	91	47	1	2	0	0
Threonine	-9.99	1.12	-140.4	7.04	4.43	5.56	159	91	69	1	2	0	0
Tryptophan	-8.61	-0.02	-35.2	16.89	4.31	4.29	247	85	163	1	0	0	0
Tyrosine	-9.27	0.05	-112.2	13.64	4.61	4.66	218	95	123	1	2	0	0
Valine	-10.04	1.18	-102.3	7.64	4.43	5.61	168	70	98	0	0	0	0

Appendix B

Description of the regression models

B.1 QBKR models per peptide

B.2 QBKR models per peptide and repetition

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	*pctvar* [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
SES	k_a		3	1	0.64	15.92		66.36	
	k_a	x	2	1	0.54	21.87		58.31	
	k_d		0	-					
	k_d	x	0	-					
	k_a, k_d		1	-	0.64	15.81	36.20	66.43	5.98
	k_a, k_d	x	1	-	0.52	22.43	31.86	55.92	7.80
VQE	k_a		4	0					
	k_a	x	6	0					
	k_d		5	1		18.95			75.68
	k_d	x	1	-	0.53	23.71			73.60
	k_a, k_d		4	0					
	k_a, k_d	x	3	0					
MYT	k_a		1	-	0.65	20.48		75.39	
	k_a	x	1	0					
	k_d		1	-	0.68	19.95			75.78
	k_d	x	6	1	0.62	19.92			69.99
	k_a, k_d		2	-	0.61	40.40	75.59	75.35	75.83
	k_a, k_d	x	2	0					
DYD	k_a		3	2	0.74	32.40		82.07	
	k_a	x	8	2	0.63	41.18		73.61	
	k_d		5	2	0.95	32.49			95.49
	k_d	x	4	-	0.91	63.95			94.21
	k_a, k_d		6	2	0.72	35.80	86.90	80.71	93.09
	k_a, k_d	x	8	3	0.63	55.50	80.71	73.84	87.57
GRA	k_a		0	-					
	k_a	x	0	-					
	k_d		2	1	0.79	18.24			81.45
	k_d	x	3	1	0.70	23.21			73.80
	k_a, k_d		6	2	0.72	35.80	86.90	80.71	93.09
	k_a, k_d	x	1	0					
GSQ	k_a		5	1	0.90	16.67		92.51	
	k_a	x	1	-	0.71	24.36		82.19	
	k_d		5	3	0.96	54.00			98.40
	k_d	x	3	1	0.87	20.27			89.43
	k_a, k_d		5	0					
	k_a, k_d	x	5	0					

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	*pctvar* [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
FGR	k_a		3	1	0.62	18.52		75.64	
	k_a	x	10	1	0.64	19.38		76.47	
	k_d		6	2	0.94	31.84			95.77
	k_d	x	2	1	0.88	18.92			90.94
	k_a, k_d		4	2	0.64	37.34	84.90	78.87	90.94
	k_a, k_d	x	4	1	0.31	20.74	67.13	50.81	83.46
DRK	k_a		3	1	0.73	18.30		85.56	
	k_a	x	10	1	0.40	26.40		71.87	
	k_d		2	1	0.90	18.79			92.61
	k_d	x	6	1	0.93	16.47			94.26
	k_a, k_d		4	2	0.71	37.12	89.19	85.72	92.66
	k_a, k_d	x	11	1	0.44	24.78	40.42	66.83	14.01
RVA	k_a		2	0					
	k_a	x	1	-	0.46	16.31		53.76	
	k_d		1	-	0.42	17.00			47.20
	k_d	x	1	0					
	k_a, k_d		4	1	0.27	17.31	38.03	35.79	40.28
	k_a, k_d	x	1	-	0.33	17.89	39.64	40.80	38.47
DSA	k_a		4	2	0.88	33.52		92.78	
	k_a	x	2	1	0.75	22.87		85.26	
	k_d		5	2	0.97	33.39			98.30
	k_d	x	10	8	0.99	95.85			99.52
	k_a, k_d		6	3	0.84	54.33	94.03	91.46	96.60
	k_a, k_d	x	4	2	0.73	40.43	90.93	86.19	95.67
RDG	k_a		4	1	0.83	17.84		87.37	
	k_a	x	10	2	0.72	41.96		79.22	
	k_d		6	2	0.95	35.86			96.19
	k_d	x	9	2	0.91	47.19			93.65
	k_a, k_d		4	1	0.76	20.41	85.36	80.63	90.09
	k_a, k_d	x	12	10	0.87				
QDF	k_a		3	1	0.68	20.02		85.07	
	k_a	x	10	2	0.60	41.62		83.67	
	k_d		4	1	0.87	17.29			90.68
	k_d	x	2	1	0.84	26.03			87.14
	k_a, k_d		6	1	0.62	22.08	84.93	81.55	88.32
	k_a, k_d	x	2	-	0.56	45.58	87.85	80.75	94.96

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	$pctvar$ [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
NES	k_a		2	1	0.47	17.56		64.68	
	k_a	x	2	1	0.29	22.40		50.53	
	k_d		4	1	0.76	24.21			79.64
	k_d	x	2	1	0.75				
	k_a, k_d		5	0					
	k_a, k_d	x	7	1	0.31	25.39	54.31	54.34	54.28
SEA	k_a		1	-	0.48	17.50		56.61	
	k_a	x	1	-	0.36	24.61		46.80	
	k_d		1	0					
	k_d	x	1	0					
	k_a, k_d		2	1	0.44	17.48	32.21	53.07	11.34
	k_a, k_d	x	2	1	0.31	25.22	28.36	41.66	15.06
SAS	k_a		3	1	0.80	18.87		83.81	
	k_a	x	9	2	0.72	41.51		79.22	
	k_d		6	2	0.93	31.00			93.64
	k_d	x	3	2	0.91	41.50			92.85
	k_a, k_d		6	2	0.78	35.53	87.28	83.59	90.96
	k_a, k_d	x	9	2	0.65	41.30	79.43	73.22	85.64
AES	k_a		3	-	0.85	46.47		91.79	
	k_a	x	2	1	0.78	18.26		85.09	
	k_d		5	1	0.72	17.64			78.28
	k_d	x	2	0					
	k_a, k_d		5	3	0.85	49.46	85.90	91.13	80.68
	k_a, k_d	x	3	2	0.80	38.25	82.56	87.67	77.44
EES	k_a		3	1	0.66	20.01		84.79	
	k_a	x	12	2	0.56	41.78		82.43	
	k_d		3	-	0.98	50.01			99.15
	k_d	x	12	2	0.93	45.98			97.08
	k_a, k_d		6	2	0.63	39.70	91.15	85.04	97.26
	k_a, k_d	x	13	12		99.98	99.95	99.91	99.99

Table B.1: Description of the QBKR models established per peptide (further explanations in the text)

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	*pctvar* [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
	k_a		2	1	0.74	16.67		86.74	
	k_a	x	8	1	0.43	24.05		70.70	
SES	k_d		1	-	0.78	16.67			82.52
1	k_d	x	1	-	0.56	24.57			70.66
	k_a, k_d		3	1	0.62	16.67	74.02	76.48	71.56
	k_a, k_d	x	2	1	0.36	24.73	63.27	62.08	64.46
	k_a		2	1	0.90	16.67		94.12	
	k_a	x	8	1	0.66	23.91		79.79	
SES	k_d		1	-	0.57	16.67			67.16
2	k_d	x	1	-	0.63	24.39			69.01
	k_a, k_d		3	2	0.90	33.33	80.64	94.12	67.16
	k_a, k_d	x	2	-	0.72	38.95	76.60	84.10	69.10
	k_a		4	3	0.88	46.47		94.13	
	k_a	x	2	1	0.81	18.33		87.17	
SES	k_d		4	0					
3	k_d	x	1	-	0.52	20.91			62.26
	k_a, k_d		3	-	0.88	46.46	77.77	94.85	60.69
	k_a, k_d	x	2	-	0.84	35.99	76.33	90.76	61.89
	k_a		4	0					
	k_a	x	6	0					
VQE	k_d		5	1	0.61	18.95			75.68
1	k_d	x	1	-	0.53	23.71			73.60
	k_a, k_d		4	0					
	k_a, k_d	x	3	0					
	k_a		1	-	0.65	20.48		75.39	
	k_a	x	1	0					
MYT	k_d		1	-	0.68	19.95			75.78
1	k_d	x	6	1	0.62	19.92			69.99
	k_a, k_d		2	-	0.61	40.40	75.59	75.35	75.83
	k_a, k_d	x	2	0					
	k_a		3	1	0.78	20.51		87.30	
	k_a	x	3	1	0.45	27.87		70.60	
DYD	k_d		6	4	0.97	58.78			98.62
1	k_d	x	12	10	0.99	97.47			99.49
	k_a, k_d		5	-	0.75	79.39	94.84	91.07	98.62
	k_a, k_d	x	5	2	0.49	44.93	80.98	74.80	87.16

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	$pctvar$ [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
DYD 2	k_a		2	0					
	k_a	x	9	0					
	k_d		6	3	0.98	42.21			98.92
	k_d	x	10	4	0.92	63.41			97.50
	k_a, k_d		4	2	0.58	37.16	86.86	81.15	92.58
	k_a, k_d	x	9	0					
GRA 1	k_a		2	1	0.75	19.08		87.43	
	k_a	x	9	1	0.43	27.61		72.71	
	k_d		4	3	0.98	42.24			98.81
	k_d	x	12	3	0.95	54.79			97.83
	k_a, k_d		6	1	0.45	18.65	64.67	63.32	66.01
	k_a, k_d	x	9	1	0.38	26.90	56.54	59.68	53.40
GRA 2	k_a		2	1	0.82	16.67		86.50	
	k_a	x	10	9	0.93	95.07		97.21	
	k_d		4	2	0.92	32.68			94.90
	k_d	x	12	3	0.95	54.79			97.83
	k_a, k_d		6	1	0.45	18.65	64.67	63.32	66.01
	k_a, k_d	x	9	1	0.58	25.28	70.22	70.82	69.63
GSQ 1	k_a		5	1	0.90	16.67		92.51	
	k_a	x	1	-	0.71	24.36		82.19	
	k_d		5	3	0.96	54.00			98.40
	k_d	x	3	1	0.87	20.27			89.43
	k_a, k_d		5	0					
	k_a, k_d	x	5	0					
FGR 1	k_a		3	1	0.62	18.52		75.64	
	k_a	x	10	1	0.64	19.38		76.47	
	k_d		6	2	0.94	31.84			95.77
	k_d	x	2	1	0.88	18.92			90.94
	k_a, k_d		4	2	0.64	37.34	84.90	78.87	90.94
	k_a, k_d	x	4	1	0.31	20.74	67.13	50.81	83.46
DRK 1	k_a		3	1	0.73	18.56		85.56	
	k_a	x	10	1	0.40	26.40		71.87	
	k_d		2	1	0.90	18.79			92.61
	k_d	x	6	1	0.93	16.47			94.26
	k_a, k_d		4	2	0.71	37.12	89.19	85.72	92.66
	k_a, k_d	x	11	1	0.44	24.78	40.42	66.83	14.01

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	*pctvar* [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
RVA 1	k_a		2	-	0.93	31.56		95.42	
	k_a	x	3	1	0.89	16.29		92.03	
	k_d		5	1	0.78	19.75			83.02
	k_d	x	2	1	0.75	20.21			80.07
	k_a, k_d		6	3	0.92	47.77	90.86	95.83	85.89
	k_a, k_d	x	3	2	0.89	37.52	87.06	91.92	82.21
RVA 2	k_a		6	1	0.87	14.80		93.18	
	k_a	x	3	1	0.78	16.42		87.50	
	k_d		3	-	0.92	50.00			96.23
	k_d	x	4	-	0.97	57.58			98.92
	k_a, k_d		4	2	0.88	33.79	93.58	94.02	93.13
	k_a, k_d	x	4	1	0.68	16.69	86.55	80.29	92.82
DSA 1	k_a		4	2	0.88	33.52		92.78	
	k_a	x	2	1	0.75	22.87		85.26	
	k_d		5	2	0.97	33.39			98.30
	k_d	x	10	8	0.99	95.85			99.52
	k_a, k_d		6	3	0.84	54.33	94.03	91.46	96.60
	k_a, k_d	x	4	2	0.73	40.43	90.93	86.19	95.67
RDG 1	k_a		2	1	0.68	17.85		84.28	
	k_a	x	1	-	0.39	25.40		69.30	
	k_d		5	2	0.98	17.74			98.89
	k_d	x	9	4	0.97	65.27			99.04
	k_a, k_d		3	1	0.58	20.92	92.70	84.49	76.29
	k_a, k_d	x	3	0					
RDG 2	k_a		2	1	0.93	17.82		95.90	
	k_a	x	2	1	0.71	25.02		84.84	
	k_d		6	2	0.95	33.90			96.90
	k_d	x	12	1	0.73	26.60			86.36
	k_a, k_d		5	1	0.84	20.04	89.58	89.34	89.82
	k_a, k_d	x	2	1	0.69	27.56	80.89	82.23	79.55
QDF 1	k_a		3	1	0.68	20.02		85.07	
	k_a	x	10	2	0.60	41.62		83.67	
	k_d		4	1	0.87	17.29			90.68
	k_d	x	2	1	0.84	26.03			87.14
	k_a, k_d		6	1	0.62	22.08	84.93	81.55	88.32
	k_a, k_d	x	2	-	0.56	45.58	87.85	80.75	94.96

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	*pctvar* [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
NES 1	k_a		1	0					
	k_a	x	1	0					
	k_d		3	1	0.88	23.69			91.53
	k_d	x	10	2	0.80	48.27			90.44
	k_a, k_d		3	0					
	k_a, k_d	x	2	0					
NES 2	k_a		4	2	0.90	30.61		94.29	
	k_a	x	2	1	0.74	17.90		84.35	
	k_d		5	1	0.70	25.31			77.70
	k_d	x	9	1	0.88	22.71			91.19
	k_a, k_d		5	1	0.93	22.95	58.51	95.47	21.56
	k_a, k_d	x	5	3	0.89	62.90	94.67	96.18	93.16
SEA 1	k_a		3	1	0.72	17.54		83.22	
	k_a	x	2	-	0.57	39.78		77.50	
	k_d		2	1	0.88	16.78			91.28
	k_d	x	3	2	0.90	40.78			94.04
	k_a, k_d		5	2	0.69	34.48	87.49	83.12	91.85
	k_a, k_d	x	3	1	0.33	24.74	64.99	57.69	72.28
SEA 2	k_a		4	1	0.79	17.43		87.23	
	k_a	x	7	1	0.51	24.66		73.29	
	k_d		0	-					
	k_d	x	0	-					
	k_a, k_d		2	1	0.79	17.41	43.81	87.24	0.39
	k_a, k_d	x	1	-	0.49	24.89	36.47	72.08	0.85
SAS 1	k_a		3	1	0.85	18.90		90.95	
	k_a	x	2	-	0.69	41.80		86.31	
	k_d		6	2	0.95	31.01			95.93
	k_d	x	3	2	0.91	41.53			93.89
	k_a, k_d		6	1	0.74	18.15	75.42	82.92	67.93
	k_a, k_d	x	4	-	0.67	70.15	91.11	86.71	95.50
SAS 2	k_a		2	1	0.77	18.84		86.99	
	k_a	x	2	-	0.62	41.27		82.24	
	k_d		6	1	0.87	17.46			90.00
	k_d	x	4	2	0.93	41.44			95.77
	k_a, k_d		4	1	0.65	18.16	74.90	78.51	71.30
	k_a, k_d	x	4	2	0.55	40.57	82.20	78.77	85.62

peptide	response variable(s)	inter-action terms	A	A^*	$Q^2_{A^*}$	*pctvar* [%]			
						$descr_{A^*}$	$resp_{A^*}$		
							total	k_a	k_d
AES 1	k_a		3	-	0.85	46.47		91.79	
	k_a	x	2	1	0.78	18.26		85.09	
	k_d		5	1	0.72	17.64			78.28
	k_d	x	2	0					
	k_a, k_d		5	3	0.85	49.46	85.90	91.13	80.68
	k_a, k_d	x	3	2	0.80	38.25	82.56	87.67	77.44
EES 1	k_a		3	1	0.66	20.01		84.79	
	k_a	x	12	2	0.56	41.78		82.43	
	k_d		3	-	0.98	50.06			99.15
	k_d	x	12	2	0.93	45.98			97.08
	k_a, k_d		6	2	0.63	39.70	91.15	85.04	97.26
	k_a, k_d	x	13	12		99.98	99.95	99.91	99.99

Table B.2: Description of the QBKR models established per peptide and repetition (further explanations in the text)

Appendix C

Specification of the regression models

C.1 Established regression models

C.1.1 Unified model with interaction terms respecting the association rate constant

$$
\begin{aligned}
k_a \;=\;& 913593(-6098 \cdot ZZ2_{142})(-45320 \cdot HFT_{142})(-10818 \cdot ZZ2_{145})(+5226 \cdot ZZ3_{145}) \\
& (-19281 \cdot HFT_{146}) - 131 \cdot NaCl - 155 \cdot urea(-39457 \cdot pH) + 0.86 \cdot NaCl * KSCN \\
& +0.02 \cdot NaCl * DMSO - 15 \cdot NaCl * pH + 0.41 \cdot NaCl * urea + 4 \cdot KSCN * urea \\
& -18 \cdot pH * urea(+2 \cdot NaCl * EDTA)(+0.42 \cdot DMSO * urea)(+7 \cdot EDTA * urea) \\
& +6959 \cdot ZZ2_{142} * ZZ2_{145} - 5515 \cdot ZZ3_{142} * ZZ2_{145} - 55561 \cdot ZZ1_{145} * ZZ2_{145} \\
& -23109 \cdot ZZ2_{145} * HFT_{145} + 6445 \cdot ZZ3_{145} * HFT_{145} + 9900 \cdot ZZ3_{145} * ZZ1_{146} \\
& (+282 \cdot ZZ1_{142} * ZZ2_{142})(-3981 \cdot ZZ1_{142} * ZZ2_{145})(+1067 \cdot ZZ1_{142} * ZZ3_{145}) \\
& (-7557 \cdot ZZ1_{142} * ZZ3_{146})(+5800 \cdot ZZ2_{142} * ZZ3_{146})(-1875 \cdot ZZ3_{142} * ZZ3_{146}) \\
& (+6925 \cdot HFT_{142} * ZZ3_{145})(-5983 \cdot ZZ2_{145} * ZZ1_{146})(+482 \cdot ZZ3_{145} * HFT_{146}) \\
& -10 \cdot NaCl * HFT_{142} - 50 \cdot NaCl * ZZ3_{145} - 103 \cdot NaCl * HFT_{145} \\
& -17 \cdot NaCl * HFT_{146} - 5736 \cdot pH * HFT_{142} - 54 \cdot urea * HFT_{142} \\
& -84 \cdot urea * HFT_{146}(-1 \cdot NaCl * ZZ1_{145})(-24 \cdot NaCl * ZZ1_{146}) \\
& (-255 \cdot KSCN * ZZ3_{145})(-1137 \cdot KSCN * HFT_{146})(+7 \cdot DMSO * HFT_{142}) \\
& (-50 \cdot DMSO * ZZ3_{145})(-78 \cdot DMSO * HFT_{146})(-1277 \cdot pH * ZZ2_{145}) \\
& (+479 \cdot pH * ZZ3_{145})(-2287 \cdot pH * HFT_{146})(-295 \cdot EDTA * ZZ3_{145}) \\
& (-43 \cdot urea * ZZ3_{145})(-73 \cdot urea * HFT_{145})
\end{aligned}
$$

C.1.2 Unified model involving the 26 variables from Sandberg et. al. (1998) without interaction terms respecting the association rate constant

$$
\begin{aligned}
k_a \;=\; & -740094 - 357 \cdot NaCl - 273 \cdot urea(-240 \cdot DMSO)(+18350 \cdot HFT_{142}) \\
& (+64293 \cdot HFT_{145})(-20420 \cdot HFT_{146})(-314 \cdot MW_{142})(-515 \cdot vdW_{142}) \\
& (-4077 \cdot POLAR_{142})(-279 \cdot Stot_{142})(+10510 \cdot [2]HACCR_{142}) - 424 \cdot MW_{145} \\
& -760 \cdot vdW_{145} + 241701 \cdot NM12_{145} - 381 \cdot Stot_{145}(+198310 \cdot NM1_{145}) \\
& (+213728 \cdot NM7_{145})(+225 \cdot HOF_{145})(-2877 \cdot POLAR_{145})(-531 \cdot Spol_{145}) \\
& (+50182 \cdot [0]HDONR_{145})(+50182 \cdot [0]HACCR_{145})(-4620 \cdot [4]HACCR_{145}) \\
& (+4620 \cdot [0]Chneg_{145})(-4620 \cdot [1]Chneg_{145})(-14722 \cdot EHOMO_{146})(-168 \cdot Snp_{146})
\end{aligned}
$$

C.1.3 Unified model involving the 26 variables from Sandberg et. al. (1998) without interaction terms respecting the dissociation rate constant

$$
\begin{aligned}
k_d \;=\; & 0.127912 - 0.005101 \cdot HFT_{145} + 0.000091 \cdot TL4_{142} \\
& +0.000010 \cdot TL7_{142} - 0.009015 \cdot ELUMO_{142}(+0.001278 \cdot [0]HDONR_{142}) \\
& (+0.001278 \cdot [0]HACCR_{142})(+0.000067 \cdot TL2_{142})(+0.000034 \cdot TL3_{142}) \\
& (-0.000047 \cdot TL5_{142})(+0.000041 \cdot TL6_{142})(+0.000272 \cdot POLAR_{142}) \\
& (-0.003915 \cdot HA_{142})(+0.000013 \cdot Snp_{142})(-0.000018 \cdot MW_{145})(-0.000039 \cdot vdW_{145}) \\
& (-0.000023 \cdot Stot_{145})(-0.000027 \cdot Spol_{145}) - 0.003182 \cdot logP_{146} + 0.000102 \cdot Spol_{146} \\
& +0.012610 \cdot [4]HDONR_{146} + 0.012757 \cdot [3]HACCR_{146}(+0.000027 \cdot MW_{146}) \\
& (+0.000039 \cdot vdW_{146})(+0.001641 \cdot EHOMO_{146})(+0.000286 \cdot POLAR_{146}) \\
& (+0.000016 \cdot Stot_{146})(-0.000516 \cdot [0]Chpos_{146})(+0.000516 \cdot [1]Chpos_{146})
\end{aligned}
$$

C.1.4 Unified model with interaction terms respecting the dissociation rate constant

$$
\begin{aligned}
k_d \;=\;& 0.021526 + 0.000037 \cdot ZZ1_{142} - 0.000359 \cdot ZZ2_{145} - 0.004013 \cdot HFT_{145} \\
& -0.000860 \cdot ZZ3_{146}(+0.000278 \cdot ZZ2_{146})(-0.000212 \cdot ZZ1_{142} * ZZ2_{142}) \\
& +0.000036 \cdot ZZ1_{142} * HFT_{142}(+0.000014 \cdot ZZ1_{142} * ZZ1_{145}) \\
& (+0.000077 \cdot ZZ1_{142} * ZZ2_{145}) + 0.000066 \cdot ZZ1_{142} * HFT_{145} \\
& (+0.000022 \cdot ZZ1_{142} * ZZ1_{146}) + 0.000138 \cdot ZZ1_{142} * ZZ3_{146} \\
& +0.000055 \cdot ZZ1_{142} * HFT_{146}(+0.000347 \cdot ZZ2_{142} * ZZ1_{146}) \\
& (-0.000166 \cdot ZZ3_{142} * ZZ2_{145}) - 0.000383 \cdot HFT_{142} * ZZ2_{145} \\
& -0.000298 \cdot HFT_{142} * HFT_{145}(+0.000154 \cdot HFT_{142} * ZZ2_{146}) \\
& -0.000387 \cdot HFT_{142} * ZZ3_{146}(-0.000171 \cdot ZZ1_{145} * ZZ2_{145}) \\
& (-0.000133 \cdot ZZ1_{145} * HFT_{145})(0.000017 \cdot ZZ1_{145} * ZZ2_{146}) \\
& -0.000192 \cdot ZZ1_{145} * ZZ3_{146}(-0.000198 \cdot ZZ2_{145} * HFT_{145}) \\
& -0.000273 \cdot ZZ2_{145} * ZZ1_{146} - 0.000354 \cdot ZZ2_{145} * ZZ2_{146} \\
& +0.000390 \cdot ZZ2_{145} * ZZ3_{146} - 0.000208 \cdot ZZ2_{145} * HFT_{146} \\
& (+0.000566 \cdot ZZ3_{145} * ZZ1_{146})(+0.000214 \cdot HFT_{145} * ZZ2_{146}) \\
& -0.000531 \cdot HFT_{145} * ZZ3_{146} - 0.002349 \cdot HFT_{145} * HFT_{146} \\
& +0.000212 \cdot ZZ1_{146} * ZZ2_{146} + -0.000191 \cdot ZZ1_{146} * ZZ3_{146} \\
& -0.000402 \cdot ZZ2_{146} * ZZ3_{146}(+0.000196 \cdot ZZ2_{146} * HFT_{146}) \\
& -0.000818 \cdot ZZ3_{146} * HFT_{146} - 0.0000008 \cdot NaCl * ZZ3_{146} \\
& (+0.0000004 \cdot NaCl * ZZ1_{142})(-0.00000005 \cdot NaCl * ZZ2_{145}) \\
& (-0.0000005 \cdot NaCl * ZZ2_{146}) - 0.000023 \cdot KSCN * ZZ3_{146} \\
& (+0.000012 \cdot KSCN * ZZ1_{142})(-0.000012 \cdot KSCN * ZZ2_{145}) \\
& (+0.000002 \cdot KSCN * ZZ2_{146}) - 0.000004 \cdot DMSO * ZZ3_{146} \\
& (+0.0000002 \cdot DMSO * ZZ1_{142})(-0.000003 \cdot DMSO * ZZ2_{145}) \\
& +0.0000005 \cdot pH * ZZ1_{142} - 0.000048 \cdot pH * ZZ2_{145} \\
& -0.000504 \cdot pH * HFT_{145} - 0.000119 \cdot pH * ZZ3_{146} \\
& (+0.000037 \cdot pH * ZZ2_{146}) - 0.000009 \cdot EDTA * ZZ3_{146} \\
& (+0.000012 \cdot EDTA * ZZ1_{142})(-0.000014 \cdot EDTA * ZZ2_{145}) \\
& -0.000005 \cdot urea * ZZ3_{146}(-0.0000005 \cdot urea * ZZ1_{142}) \\
& (-0.000003 \cdot urea * ZZ2_{145})(+0.000001 \cdot urea * ZZ2_{146})
\end{aligned}
$$

C.1.5 QSKR model involving the 26 variables from Sandberg et. al. (1998) without interaction terms respecting the logarithmic dissociation rate constant

$$
\begin{aligned}
logk_d \;=\; & -2.883252 - 0.391150 \cdot HFT_{145} - 0.370017 \cdot [2]HACCR_{142} \\
& (+0.114110 \cdot [0]HDONR_{142})(-0.163044 \cdot [1]HDONR_{142}) \\
& (+0.114110 \cdot [0]HACCR_{142}) + 0.269987 \cdot [2]HACCR_{145} \\
& -0.191431 \cdot [4]HACCR_{145} + 0.191431 \cdot [0]Chneg_{145} - 0.191431 \cdot [1]Chneg_{145} \\
& (-0.002961 \cdot MW_{145})(+0.872085 \cdot NM1_{145})(+1.217089 \cdot NM7_{145}) \\
& (+1.378180 \cdot NM12_{145})(+0.001778 \cdot HOF_{145})(-0.004229 \cdot Spol_{145}) \\
& +0.277160 \cdot [3]HACCR_{146}(+0.000976 \cdot MW_{146})(+0.001258 \cdot vdW_{146}) \\
& (-0.814360 \cdot NM7_{146})(-0.103571 \cdot logP_{146})(+0.010552 \cdot POLAR_{146}) \\
& (+0.000572 \cdot Stot_{146})(+0.003761 \cdot Spol_{146})(-0.097428 \cdot [1]HDONR_{146}) \\
& (+0.067619 \cdot [2]HDONR_{146})(+0.245382 \cdot [4]HDONR_{146}) \\
& (-0.110593 \cdot [2]HACCR_{146})
\end{aligned}
$$

C.2 VIP-values

C.2.1 QSKR models involving the ZZ-scales

C.2.2 QSKR models involving the 26 variables from Sandberg et. al. (1998)

C.2.3 Unified models involving the ZZ-scales

C.2.4 Unified models involving the 26 variables from Sandberg et. al. (1998)

C.2.5 QBKR models per peptide

C.2.6 QBKR models per peptide and repetition

model type		QSKR							
response variable		k_a		k_d		$\log(k_d)$		k_a,k_d	
interaction terms			x		x		x		x
descriptor variable	position								
ZZ1	142			(1.34)	1.46		(1.07)		
	145								
	146								
ZZ2	142	(1.21)	(1.18)						(1.05)
	145			(1.28)	(1.37)	(1.19)	(1.27)	(1.11)	(1.04)
	146			(1.24)	(1.34)		(1.00)		
ZZ3	142								
	145	1.44	(1.31)					(1.35)	(1.17)
	146	(1.34)	(1.20)	1.72	1.84	1.44	1.55	(1.20)	(1.39)
HFT	142	1.72	1.84					(1.25)	1.40
	145			1.75	1.93	2.19	2.41	1.73	2.08
	146								

Table C.1: VIP-values referring to the descriptor variables included in the QSKR models

model type		QSKR-26vars			
response variable		k_a	k_d	logk_d	k_a,k_d
interaction terms					
descriptor variable	position				
HFT	142 145 146		1.76	2.10	
TL1	142 145 146				
TL2	142 145 146		(1.32)		
TL3	142 145 146		(1.21)		
TL4	142 145 146		1.45		
TL5	142 145 146		(1.02)		
TL6	142 145 146		(1.30)		
TL7	142 145 146		(1.35)		
NM1	142 145 146			(1.01)	
NM7	142 145 146			(1.20) (1.06)	
NM12	142 145 146			(1.34)	

model type		QSKR-26vars			
response variable		k_a	k_d	logk_d	k_a,k_d
interaction terms					
descriptor variable	position				
EHOMO	142 145 146		 (1.04)		
ELUMO	142 145 146		(1.11)		
HOF	142 145 146			(1.29)	
POLAR	142 145 146		 1.42	 (1.22)	
EN	142 145 146				
HA	142 145 146		(1.17)		
HDONR	142 145 146 146		[0] (1.22) [4] 1.71	[0] (1.10) [1] (1.02) [1] (1.09), [2] (1.01), [4] (1.08)	
HACCR	142 145 145 146		[0] (1.22) [2] (1.14), [3] 1.98, [4] (1.04)	[0] (1.10), [2] 1.66 [2] 1.62, [4] 1.63 [2] (1.33), [3] 1.53	
Chpos	142 145 146		 [0] (1.16), [1] (1.16)		
Chneg	142 145 146		[0] (1.04), [1] (1.04)	[0] 1.63, [1] 1.63	
Stot	142 145 146		 (1.30) 1.46	 (1.29)	

model type		QSKR-26vars			
response variable		k_a	k_d	logk_d	k_a,k_d
interaction terms					
descriptor variable	position				
Spol	142				
	145		(1.16)	(1.30)	
	146		1.62	(1.26)	
Snp	142		(1.06)		
	145				
	146				
MW	142				
	145		(1.28)	(1.01)	
	146		1.44	(1.20)	
logP	142				
	145				
	146		1.62	(1.18)	
vdW	142				
	145		(1.27)		
	146		1.50	(1.26)	

Table C.2: VIP-values referring to the QSKR models involving the 26 variables from Sandberg et. al. (1998)

model type		unified					
response variable		k_a		k_d		k_a,k_d	
interaction terms			x		x		x
descriptor variable	position						
ZZ1	142			1.53	1.65	(1.20)	(1.28)
	145						
	146						
ZZ2	142		(1.01)				
	145		(1.09)	1.48	1.54	(1.33)	(1.30)
	146			(1.28)	(1.30)	(1.06)	(1.06)
ZZ3	142						
	145	(1.36)	(1.37)			(1.05)	(1.04)
	146			1.92	2.00	1.55	1.60
HFT	142	(1.27)	(1.37)			(1.06)	(1.14)
	145			1.99	2.14	1.59	1.72
	146	(1.15)	(1.16)				
DMSO						(1.28)	
EDTA							
NaCl		2.07	2.02				(1.37)
pH			(1.10)				
KSCN							
urea		1.46	1.41				(1.02)

Table C.3: VIP-values referring to the descriptor variables included in the unified models

model type		unified-26vars					
response variable		k_a		k_d		k_a,k_d	
interaction terms			x		x		x
descriptor variable	position						
HFT	142	(1.20)	(1.27)		(1.01)	(1.06)	(1.15)
	145	(1.02)	(1.37)	1.80	2.44	1.45	1.95
	146	(1.05)					
TL1	142				(1.18)		
	145						
	146				(1.28)		
TL2	142			(1.37)	1.88	(1.05)	1.46
	145						
	146				(1.11)		
TL3	142			(1.15)	1.58		(1.27)
	145						
	146						
TL4	142			1.52	2.08	(1.14)	1.61
	145						
	146						
TL5	142			(1.14)	1.55		(1.21)
	145						
	146				(1.34)		(1.14)
TL6	142			(1.30)	1.78		(1.37)
	145						
	146				(1.01)		
TL7	142			1.44	1.97	(1.14)	1.57
	145						
	146						
NM1	142						
	145	(1.23)	1.91				1.48
	146						
NM7	142						
	145	(1.33)	1.99				1.52
	146						
NM12	142						
	145	1.49	2.33		(1.06)	(1.15)	1.80
	146						

model type		unified-26vars					
response variable		k_a		k_d		k_a,k_d	
interaction terms			x		x		x
descriptor variable	position						
EHOMO	142				(1.18)		
	145						
	146	(1.04)		(1.07)	1.49	(1.01)	(1.29)
ELUMO	142			1.48	2.05	(1.15)	1.59
	145						
	146						
HOF	142				(1.13)		
	145	(1.07)			(1.23)		(1.07)
	146						
POLAR	142	(1.02)	(1.11)	(1.00)	1.41	(1.06)	(1.35)
	145	(1.23)	(1.20)		(1.03)	(1.04)	(1.08)
	146			(1.24)	1.71	(1.04)	1.40
EN	142						
	145						
	146				(1.15)		(1.00)
HA	142			(1.39)	1.91	(1.11)	1.51
	145						
	146						
HDONR	142			[0] (1.10)	[0] 1.49		[0] (1.16)
	145	[0] (1.09)	[0] (1.09)		[0] (1.10),		[0] (1.09)
	145				[1] (1.04),		
	145				[2] (1.11)		
	146			[4] 2.07	[1] (1.01),	[4] (1.59)	[4] 2.24
	146				[4] 2.83		
HACCR	142	[2] (1.37)	[2] (1.35)	[0] (1.10)	[0] 1.49,	[2] (1.10)	[0] (1.16),
	142				[2] (1.07)		[2] (1.23)
	145	[0] (1.09),	[0] (1.09)		[0] (1.11),		[0] (1.09),
	145	[4] (1.13)			[2] (1.32),		[2] (1.14),
	145				[3] (1.01),		[4] (1.19)
	145				[4] (1.36)		
	146			[3] 2.28	[1] (1.01),	[3] 1.79	[3] 2.39
	146				[2] (1.05),		
	146				[3] 3.04		

model type		unified-26vars					
response variable		k_a		k_d		k_a,k_d	
interaction terms			x		x		x
descriptor variable	position						
Chpos	142						
	145						
	146			[0] (1.34),	[0] 1.81,	[0] (1.06),	[0] 1.45,
	146			[1] (1.34)	[1] 1.81	[1] (1.06)	[1] 1.45
Chneg	142						
	145	[0] (1.13),			[0] (1.36),		[0] (1.19),
	145	[1] (1.13)			[1] (1.36)		[1] (1.19)
	146						
Stot	142	(1.06)	(1.12)		(1.17)	(1.01)	(1.21)
	145	1.41	1.49	(1.12)	1.49	(1.30)	1.49
	146			(1.33)	1.83	(1.09)	1.47
Spol	142						
	145	(1.33)	(1.35)	(1.02)	1.40	(1.20)	1.41
	146			1.53	2.10	(1.20)	1.63
Snp	142			(1.35)	1.86	(1.16)	1.55
	145						
	146	(1.04)					
MW	142	(1.04)	(1.10)		(1.04)		(1.12)
	145	1.47		(1.04)	1.41	(1.29)	1.46
	146			(1.29)	1.78	(1.02)	1.42
logP	142				(1.18)		
	145						
	146			1.54	2.09	(1.22)	1.70
vdW	142	(1.04)	(1.10)		(1.08)		(1.15)
	145	1.44	1.52	(1.09)	1.44	(1.30)	1.48
	146			(1.38)	1.89	(1.10)	1.51
DMSO		(1.12)					
EDTA							
NaCl		2.73	1.57			1.82	(1.06)
pH							
KSCN							
urea		1.83	(1.13)			(1.26)	

Table C.4: VIP-values referring to the unified models involving the 26 variables from Sandberg et. al. (1998)

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
SES	k_a				2.00			(1.22)
	k_a	x			1.87			(1.14)
	k_d							
	k_d	x						
	k_a,k_d				1.90			(1.27)
	k_a,k_d	x			1.72			(1.17)
VQE	k_a							
	k_a	x						
	k_d		1.94					(1.02)
	k_d	x	1.83					
	k_a,k_d							
	k_a,k_d	x						
MYT	k_a				1.61			(1.25)
	k_a	x						
	k_d		(1.18)					2.04
	k_d	x	(1.25)					2.15
	k_a,k_d				(1.16)			(1.69)
	k_a,k_d	x						
DYD	k_a				1.75			(1.01)
	k_a	x			1.68			
	k_d		1.75		(1.15)			
	k_d	x	1.79		(1.24)	(1.16)		
	k_a,k_d		(1.37)		1.46			
	k_a,k_d	x	(1.39)		1.60	(1.00)		
GRA	k_a							
	k_a	x						
	k_d		1.76					1.40
	k_d	x	1.69					(1.35)
	k_a,k_d							
	k_a,k_d	x						
GSQ	k_a		(1.12)		1.51			1.44
	k_a	x	(1.34)					
	k_d		(1.30)			(1.08)		1.51
	k_d	x					(1.04)	1.55
	k_a,k_d							
	k_a,k_d	x						

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
FGR	k_a		(1.02)		1.57			1.55
	k_a	x			1.51			1.49
	k_d		(1.35)					1.94
	k_d	x	(1.32)					1.90
	k_a,k_d		(1.28)		(1.12)			1.75
	k_a,k_d	x	(1.33)					1.84
DRK	k_a				1.40			1.56
	k_a	x			(1.29)			1.44
	k_d			(1.07)	1.51			(1.20)
	k_d	x	(1.03)	(1.14)	1.60			(1.27)
	k_a,k_d				1.46			(1.39)
	k_a,k_d	x	(1.14)					1.76
RVA	k_a							
	k_a	x	(1.05)					1.74
	k_d		1.70					(1.38)
	k_d	x						
	k_a,k_d		1.67					1.67
	k_a,k_d	x	1.56					1.59
DSA	k_a				1.79			1.48
	k_a	x			1.70			1.41
	k_d		1.52					(1.28)
	k_d	x	1.48		(1.06)			(1.24)
	k_a,k_d		(1.18)		(1.38)			1.42
	k_a,k_d	x	(1.10)		1.53			(1.27)
RDG	k_a		(1.16)		1.64			(1.27)
	k_a	x	(1.05)		1.51			(1.16)
	k_d		(1.30)					1.56
	k_d	x			(1.17)			1.43
	k_a,k_d		(1.13)		1.52			1.44
	k_a,k_d	x	(1.04)		(1.31)			(1.26)
QDF	k_a		(1.08)		1.62			
	k_a	x						
	k_d		1.69		(1.29)			
	k_d	x	1.41		(1.08)			
	k_a,k_d		(1.18)		1.41			
	k_a,k_d	x						

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
NES	k_a				1.82	(1.04)		
	k_a	x			1.88	(1.07)		(1.02)
	k_d		(1.31)		(1.35)			(1.27)
	k_d	x	(1.30)		(1.35)			(1.27)
	k_a,k_d							
	k_a,k_d	x	(1.07)		1.63			(1.24)
SEA	k_a				1.82			(1.14)
	k_a	x			1.74			(1.10)
	k_d							
	k_d	x						
	k_a,k_d				1.48			(1.20)
	k_a,k_d	x			(1.26)			(1.09)
SAS	k_a				1.93			
	k_a	x			1.84			
	k_d		1.68					1.42
	k_d	x	1.40					(1.18)
	k_a,k_d		(1.30)		1.44			(1.21)
	k_a,k_d	x	(1.16)		1.40			(1.04)
AES	k_a				2.15			(1.01)
	k_a	x			2.13			
	k_d		1.72					(1.34)
	k_d	x						
	k_a,k_d		(1.18)		1.61			(1.22)
	k_a,k_d	x	(1.15)		1.65			(1.12)
EES	k_a				1.85			
	k_a	x			1.83			
	k_d		1.94					(1.26)
	k_d	x	1.74					(1.12)
	k_a,k_d		1.57		(1.29)			(1.01)
	k_a,k_d	x	(1.37)		(1.33)			

Table C.5: VIP-values referring to the QBKR models established per peptide

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
SES1	k_a							
	k_a	x			1.63			(1.25)
	k_d		1.86					
	k_d	x	1.68					
	k_a,k_d		1.41		1.43			(1.23)
	k_a,k_d	x	(1.29)		(1.28)			(1.10)
SES2	k_a				1.72			(1.35)
	k_a	x			1.60			(1.25)
	k_d		1.85					1.42
	k_d	x	1.53					(1.17)
	k_a,k_d		(1.34)		(1.34)			(1.38)
	k_a,k_d	x	(1.18)		(1.28)			(1.18)
SES3	k_a				2.17			
	k_a	x			2.15			
	k_d							
	k_d	x	1.59					(1.19)
	k_a,k_d		(1.15)		1.73			1.14
	k_a,k_d	x	(1.02)		1.54			(1.09)
VQE1	k_a							
	k_a	x						
	k_d		1.94					(1.01)
	k_d	x	1.83					
	k_a,k_d							
	k_a,k_d	x						
MYT1	k_a				1.61			(1.24)
	k_a	x						
	k_d		(1.18)					2.04
	k_d	x	(1.25)					2.15
	k_a,k_d				(1.16)			1.69
	k_a,k_d	x						
DYD1	k_a				1.99			
	k_a	x			1.80			
	k_d		1.61		(1.26)			
	k_d	x	1.65		(1.35)	(1.01)		
	k_a,k_d		(1.20)		1.51			
	k_a,k_d	x			1.64			

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
DYD2	k_a							
	k_a	x						
	k_d		1.71					
	k_d	x	1.76		(1.06)	(1.21)		
	k_a,k_d		1.45		(1.34)			
	k_a,k_d	x						
GRA1	k_a				1.55			(1.36)
	k_a	x			1.44			(1.26)
	k_d		1.79					(1.26)
	k_d	x	1.67					(1.17)
	k_a,k_d		1.67					1.56
	k_a,k_d	x	1.49					1.42
GRA2	k_a		(1.22)		(1.02)			1.79
	k_a	x	(1.04)			(1.00)		1.54
	k_d		1.64					1.52
	k_d	x	1.51					1.40
	k_a,k_d		1.53					1.72
	k_a,k_d	x	1.42					1.60
GSQ1	k_a		(1.12)		1.51			1.44
	k_a	x			(1.35)			(1.28)
	k_d		(1.30)			(1.08)		1.51
	k_d	x	(1.34)			(1.05)		1.55
	k_a,k_d							
	k_a,k_d	x						
FGR1	k_a		(1.02)		1.57			1.55
	k_a	x			1.51			1.49
	k_d		(1.35)					1.94
	k_d	x	(1.32)					1.90
	k_a,k_d		(1.28)		(1.12)			1.75
	k_a,k_d	x	(1.33)					1.84
DRK1	k_a				1.40			1.56
	k_a	x			(1.29)			1.44
	k_d			(1.07)	1.51			(1.20)
	k_d	x	(1.03)	(1.14)	1.60			(1.27)
	k_a,k_d				1.46			(1.39)
	k_a,k_d	x	(1.14)					1.76

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
RVA1	k_a		(1.15)		(1.06)			1.69
	k_a	x	(1.08)					1.60
	k_d			1.55	1.43			
	k_d	x		1.60	1.48			
	k_a,k_d		(1.26)	(1.07)	(1.28)			(1.15)
	k_a,k_d	x	1.50					(1.29)
RVA2	k_a							2.03
	k_a	x						1.97
	k_d		1.63					1.50
	k_d	x	(1.36)					(1.24)
	k_a,k_d		(1.39)					1.78
	k_a,k_d	x	(1.31)					1.68
DSA1	k_a				1.79			1.48
	k_a	x			1.70			1.41
	k_d		1.52					(1.28)
	k_d	x	1.48		(1.06)			(1.24)
	k_a,k_d		(1.18)		(1.38)			1.42
	k_a,k_d	x	(1.30)		(1.02)			1.50
RDG1	k_a		(1.15)		1.65			(1.20)
	k_a	x	(1.08)		1.55			(1.13)
	k_d		(1.01)		(1.28)			1.57
	k_d	x			(1.15)			1.42
	k_a,k_d		(1.11)		1.52			1.45
	k_a,k_d	x						
RDG2	k_a		(1.16)		1.62			(1.35)
	k_a	x	(1.06)		1.48			(1.23)
	k_d				(1.31)			1.58
	k_d	x			(1.22)			1.49
	k_a,k_d		(1.14)		1.51			1.45
	k_a,k_d	x	(1.04)		(1.38)			(1.33)
QDF1	k_a		(1.08)		1.62			
	k_a	x	(1.00)		1.60	(1.02)		
	k_d		1.69		(1.29)			
	k_d	x	1.41		(1.08)			
	k_a,k_d		(1.36)		1.52			
	k_a,k_d	x	(1.18)					1.41

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
NES1	k_a							
	k_a	x						
	k_d							
	k_d	x	(1.16)		1.48			(1.27)
	k_a,k_d							
	k_a,k_d	x						
NES2	k_a				2.15			
	k_a	x			2.17			
	k_d		1.49		(1.11)			(1.21)
	k_d	x	1.42		(1.05)			(1.14)
	k_a,k_d				1.98			(1.28)
	k_a,k_d	x						
SEA1	k_a				1.86			(1.05)
	k_a	x			1.78			
	k_d		1.44					1.77
	k_d	x	(1.24)					1.52
	k_a,k_d		(1.09)		(1.34)			1.46
	k_a,k_d	x			(1.12)			1.41
SEA2	k_a				1.74			(1.29)
	k_a	x			1.65			(1.22)
	k_d							
	k_d	x						
	k_a,k_d				1.78			(1.30)
	k_a,k_d	x			1.58			(1.17)
SAS1	k_a				1.99			
	k_a	x			1.89			
	k_d		1.81					(1.31)
	k_d	x	1.53					(1.09)
	k_a,k_d		(1.28)		1.48			(1.18)
	k_a,k_d	x						
SAS2	k_a				1.85			
	k_a	x			1.77			
	k_d		1.57					1.54
	k_d	x	(1.26)					(1.24)
	k_a,k_d		(1.24)		1.34			(1.35)
	k_a,k_d	x	(1.06)		(1.35)			(1.10)

peptide	response variable	interaction terms	buffer variable					
			DMSO	EDTA	NaCl	pH	KSCN	Urea
AES1	k_a				2.15			(1.01)
	k_a	x			2.13			
	k_d		1.72					(1.34)
	k_d	x						
	k_a,k_d		(1.18)		1.61			(1.22)
	k_a,k_d	x						
EES1	k_a				1.85			
	k_a	x			1.83			
	k_d		1.94					(1.26)
	k_d	x	1.74					(1.12)
	k_a,k_d		1.57		(1.29)			(1.01)
	k_a,k_d	x	(1.06)		(1.20)			(1.39)

Table C.6: VIP-values referring to the QBKR models established per peptide and repetition

Acknowledgements

First of all, I am indebted to Dr. Karl Andersson from the Biacore AB in Uppsala who made me familiar with proteo-chemometric approaches. He placed the data he collected during his research of biomolecular interactions at my disposal and answered arising questions concerning his analysis. Thereby, he enabled me to investigate the benefit of the novel modelling approaches with respect to the interaction under study.

Furthermore, I would like to thank Prof. Dr. Rolf Kinne, PD Dr. Jürgen Kuhlmann and Dr. Barbara Wimmer from the Max Planck Institute of Molecular Physiology in Dortmund. Professor Kinne gave me insight in the biological background and made it possible for me to get in touch with Dr. Wimmer. The explanations from Dr. Kuhlmann and Dr. Wimmer improved my understanding of the technical knowledge in what concerns the Surface Plasmon Resonance Biosensors.

I am also indebted to Dr. Nick Fieller from the university of Sheffield for correcting my dissertation concerning the English language in detail and giving further helpful comments.

Last but not least, I would like to thank the German Research Foundation (Deutsche Forschungsgemeinschaft) for their financial support within the scope of the Research Training Group (Graduiertenkolleg) "Statistical Modelling" that enabled me to study at this topic of biological relevance.

Bibliography

[1] Andersson, K. (2004): Characterization of biomolecular interactions using a multivariate approach. *Acta universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine*, **1363**.

[2] Andersson, K., Choulier, L., Hämäläinen, M. D., van Regenmortel, M. H. V., Altschuh, D. and Malmqvist, M. (2001): Predicting the kinetics of peptide-antibody interactions using a multivariate experimental design of sequence and chemical space. *Journal of Molecular Recognition*, **14**, 62-71.

[3] Andersson, K., Gülich, S., Hämäläinen, M. D., Nygren, P-Å., Hober, S. and Malmqvist, M. (1999): Kinetic characterization of the interaction of the Z-fragment of protein A with mouse-IgG3 in a volume in chemical space. *Proteins: Structure, Function and Genetics*, **37**, 494-498.

[4] Chong, I.-G. and Jun, C.-H. (2005): Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, **78**, 103-112.

[5] Choulier, L., Andersson, K., Hämäläinen, M. D., van Regenmortel, M. H. V., Malmqvist, M. and Altschuh, D. (2002): QSAR studies applied to the prediction of antigen-antibody interaction kinetics as measured by Biacore. *Protein Engineering*, **15**, 373-382.

[6] De Genst, E., Areskoug, D., Decanniere, K., Muyldermans, S. and Andersson, K. (2002): Kinetic and affinity predictions of a protein-protein interaction using multivariate experimental design. *Journal of Biological Chemistry*, **277**, 29897-29907.

[7] Freyhult, E. K., Andersson, K. and Gustafsson, M. G. (2003): Structural modeling extends QSAR analysis of antibody-lysozyme interactions to 3D-QSAR. *Biophysical Journal*, **84**, 2264-2272.

[8] Denham, M. C. (1997): Prediction intervals in Partial Least Squares. *Journal of Chemometrics*, **11**, 39-52.

[9] Freyhult, E., Prusis, P., Lapinsh, M., Wikberg, J., Moulton, V., Gustafsson, M. (2005): Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. *BMC Bioinformatics*, **6**, Article number 50.

[10] Geladi, P., Kowalski, B. R. (1986): Partial Least-Squares Regression: a tutorial. *Analytica Chimica Acta*, **185**, 1-17.

[11] Gustafsson, G. (2001): A probabilistic derivation of the Partial Least-Squares algorithm. *Journal of Chemical Information and Computer Sciences*, **41**, 288-294.

[12] Helland, I.S. (1988): On the structure of Partial Least Squares regression. *Communications in Statistics - Simulation and Computation*, **17**, 581-607.

[13] Höskuldsson, A. (1988): PLS regression methods. *Journal of Chemometrics*, **2**, 211-228.

[14] Jung, K., Gannoun, A., Sitek, B., Meyer, H. E., Stühler, K. and Urfer, W. (2005): Analysis of dynamic protein expression data. *REVSTAT-Statistical Journal*. **3**, 99-111.

[15] Kirschbaum, N. (2005): Prediction of protein's topology by a Bayesian method using the Gibbs sampling algorithm. *Diploma thesis, Department of Statistics, University of Dortmund*.

[16] Lee, K. R., Lin, X., Park, D., Eslava, S. (2003): Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics*, **3**, 1680-1686.

[17] Manne, R. (1987): Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, **2**, 187-197.

[18] Martens, H., Naes, T. (1989): Multivariate calibration. *Wiley, Chichester*.

[19] Podwojski, K., Kracker, H., Büscher, G., Jung, K., Schleif, F.-M., Decker, J., Fieller, N. and Urfer, W. (2006): Outlier detection and classification methods for protein expression data from MALDI-TOF mass spectrometry. *to be published*.

[20] Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S. (1998): New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry*, **41**, 2481-2491.

[21] Serneels, S., Lemberge, P. and Van Espen, P.J. (2004): Calculation of PLS prediction intervals using efficient recursive relations for the Jacobian matrix. *Journal of Chemometrics*, **18**, 76-80.

[22] Sousa, L., Althoff, T., Raja, M., Kinne, R.K.H. and Kirschbaum, N. (2007): Prediction of nonhelical intramembrane segments in membrane proteins by Bayesian methodology. *In preparation.*

[23] Urfer, W. and Amaral Turkman, M. A. (2006): "Proceedings of the Workshop on Statistics in Genomics and Proteomics", Monte Estoril, 5-8 October 2005. *International Centre of Mathematics*, **Volume 27**, Coimbra: Portugal.

[24] Wimmer, B. (2005): Structure-function relationship of loop 13 of Na^+/Glucose Cotransporter SGLT1 investigated by Atomic Force Microscopy and Surface Plasmon Resonance Spectroscopy. *Dissertation, Department of Biophysics, Johannes Kepler-university of Linz, Austria.*