Proteochemometrics modeling of carbonic anhydrase-ligand interactions using rule-based and linear methods

Min Jia

Student project in Molecular Bioinformatics and Biotechnology
Research Training, 5 points, Nov 2005
Supervisor:
Torgeir R. Hvidsten
The Linnaeus Centre for Bioinformatics, Uppsala University

Abstract

Proteochemometrics is a novel technology for the analysis of interaction of series of receptors with series of ligands. In this study, two different proteochemometric approaches, rough sets and partial least squares (PLS), have been utilized to model the interaction of Carbonic anhydrases (CA I, CA II, CA V) and their ligands. Both approaches analyzed the dataset which correlated the experimentally determined binding affinity (pKi value) with the specific positions of CAs amino acid sequence and ligands. The CAs and ligands are described by vectors of numerical descriptors which are associated with their physico-chemical properties. Rough sets generate If-Then rules using Boolean reasoning. Different partitions of the dataset were evaluated in order to find an optimal partition into rough set decision classes. The performance of the rough set classifier was assessed by 10 fold cross validation (CV) and we reported *accuracy mean, Area Under Curve* (*AUC) mean* and Standard deviation (SD). The results show that a highly valid model was obtained with *accuracy mean* of 0.87 (SD=0.06) and an *AUC mean* of 0.92 (SD=0.05). PLS regression is a recent technique that generalizes and combines features from principal component analysis (PCA) and multiple linear regression (MLR). It is particularly useful when we need to predict a set of response variables (*e.g. pKi*) from a very large set of predictor variables (*e.g. descriptors of the CAs and ligands*).The goal of the PLS regression is to predict *pKi* from *descriptors of the CAs and ligands* and to describe their common structure. PLS also could rank all the attributes and cross terms from most influential to least influential for binding affinity. The PLS model yielded a significant PLS component, the R2 and Q2 being 0.759, and 0.722 respectively. Rough sets and PLS have different strength to construct a valid model, and they could be complementary in some respects, thus we might achieve a more objective and valid prediction result when combination of the two approaches.

1 Introduction

1.1 Motivation

Carbonic anhydrases (CAs) are wide-spread zinc enzymes, present in mammals in at least 14 different isoforms. Some of these isozymes are cytosolic (CA I, CA II, CA III, CA VII), others are membrane-bound (CA IV, CA IX, CA XII and CA XIV), CA V is mitochondrial and CA VI is secreted in the saliva. Three acatalytic forms are also known (CARP VIII, CARP X and CARP XI). These enzymes catalyze a very simple physiological reaction, the interconversion

between carbon dioxide and the bicarbonate ion. They are important targets for the design of inhibitors with clinical applications. Therefore, understanding the CAs-ligands interaction might greatly help the design of potent inhibitors.

1.2 Proteochemometric approach

In proteochemometrics, one analyses the experimentally determined interaction strength of series of biopolymer-molecular. It is based on quantitative descriptions derived from structure and physico-chemical properties of interacting moleculars, which are correlated to interaction affinity using mathematical modeling [1]. Various linear and nonlinear correlation methods can be used.

PLS is the prime choice for a linear approach, which has already modeled peptide interaction with chimeric and wild-type melanocortin GPCRs successfully [2]. When using linear modeling, these descriptions reveal the contribution of linear combination of properties to the interaction. However in reality, complexes nonlinear processes govern interactions, thus *cross terms* are calculated for investigating the contribution of the nonlinear combination of descriptors.

Nonlinear methods have been used to only a limited extent. Rough set theory [3, 4], developed in Poland in the early 1980s, is a relatively new and promising machine learning technique for data mining and knowledge discovery from databases. Rough set is a Boolean method, which is suited to investigate nonlinear phenomena.

In this study, we applied both the linear method PLS and the rule-based nonlinear method rough sets to model the CAs-ligands interactions.

2 Materials and methods

2.1 Datasets

2.1.1 Interaction Data

Data for 191 CAs interaction with their corresponding ligands were obtained from (http://kibank.iis.u-tokyo.ac.jp/) [5]. The CAs represented three CA families (CA I, CA II, CA V), and included 54, 91, 46 subtypes respectively.

2.1.2 Descriptor of CAs and ligands

2.1.2.1 Mutiple sequences alignment

Amino acid sequences were retrieved from the ENZYME database (http://www.expasy.org/enzyme/), and aligned according to the conserved amino acid positions (http://bioinformatics.albany.edu/~cemc/). From the result in alignment (see supplementary data for detail), 8 non-conserved residues were selected, at position 63, 66, 68, 92, 132, 133, 205, 209 (Table 1).

Table 1 The amino acids at site 63, 66, 68, 92, 132, 133, 205, 209 of CA I, CA II, CA V

	63	66	68	92	132	133	205	209
1HCB	Val (V)	Ser (S)	His (H)	Phe (F)	Leu (L)	Ala (A)	His (H)	Tyr (Y)
1A42	Asn (N)	Ala (A)	Asn (N)	Ile (I)	Phe (F)	Gly (G)	Thr (T)	Leu (L)
1DMY	Thr (T)	Phe (F)	Gln (Q)	Lys (K)	Tyr (Y)	Lys (K)	Thr (T)	Ala (A)

2.1.2.2 Descriptor of CAs and ligands

Description of CAs. The non-conserved residues were subsequently coded using the three z-scale descriptors (Table 2) derived by Sandberg et al.[6]. For example, the amino acid at site 63 should be described as [63_z1, 63_z2, 63_z3], for CA I, the residue at the 63rd site is "V" (Val), so CA I is described by [-2.69, -2.53, -1.29], thus the physicochemical differences in the ligand-binding region of CAs were accordingly encoded by a total of 24 descriptors. Therefore the composition of each CA in the dataset is described by a vector whose elements contain 24 numerical values, as [63_z1, 63_z2, 63_z3, 66_z1, 66_z2, 66_z3, 68_z1, 68_z2, 68_z3, 92_z1, 92_z2, 92_z3, 132_z1, 132_z2, 132_z3, 133_z1, 133_z2, 133_z3, 205_z1, 205_z2, 205_z3, 209_z1, 209_z2, 209_z3].

Table 2 Descriptor scales for the coded amino acids

*z*1	hydrophobicity/hydrophilicity
*z*2	side-chain bulk volume
*z*3	polarizability and charge

Amino acid	*z*1	*z*2	*z*3
Phe (F)	–4.92	1.3	0.45
Trp (W)	–4.75	3.66	0.85
Ile (I)	–4.44	–1.68	–1.03
Leu (L)	–4.19	–1.03	–0.98
Val (V)	–2.69	–2.53	–1.29
Met (M)	–2.49	–0.27	–0.41
Tyr (Y)	–1.39	2.32	0.01
Pro (P)	–1.22	0.88	2.23
Ala (A)	0.07	–1.73	0.09
Cys (C)	0.71	–0.97	4.13
Thr (T)	0.92	–2.09	–1.40
Ser (S)	1.96	–1.63	0.57
Gln (Q)	2.19	0.53	–1.14
Gly (G)	2.23	–5.36	0.3
His (H)	2.41	1.74	1.11
Lys (K)	2.84	1.41	–3.14
Arg (R)	2.88	2.52	–3.44
Glu (E)	3.08	0.039	–0.07
Asn (N)	3.22	1.45	0.84
Asp (D)	3.64	1.13	2.36

Description of ligands. Compounds were characterized by 34 descriptors, which were calculated by the Dragon software (Talete S.r.l., Milano, Italy). The descriptors represented different physicochemical properties as well as the numbers of functional groups and structural fragments in the molecule (Table 3).

Table 3 The molecular descriptors annotation

Symbol	Definition
MW	molecular weight
AMW	average molecular weight
Sv	sum of atomic van der Waals volumes (scaled on Carbon atom)
Se	sum of atomic Sanderson electronegativities (scaled on Carbon atom)
Sp	sum of atomic polarizabilities (scaled on Carbon atom)
Ss	sum of Kier-Hall electrotopological states
Mv	mean atomic van der Waals volume (scaled on Carbon atom)
Me	mean atomic Sanderson electronegativity (scaled on Carbon

	atom)
Mp	mean atomic polarizability (scaled on Carbon atom)
Ms	mean electrotopological state
nAT	number of atoms
nSK	number of non-H atoms
nBT	number of bonds
nBO	number of non-H bonds
nBM	number of multiple bonds
SCBO	sum of conventional bond orders (H-depleted)
ARR	aromatic ratio
nCIC	number of rings
nCIR	number of circuits
RBN	number of rotatable bonds
RBF	rotatable bond fraction
nDB	number of double bonds
nAB	number of aromatic bonds
nH	number of Hydrogen atoms
nC	number of Carbon atoms
nN	number of Nitrogen atoms
nO	number of Oxygen atoms
nS	number of Sulfur atoms
nHDon	number of donor atoms for H-bonds (N and O)
nHAcc	number of acceptor atoms for H-bonds (N,O,F)
Hy	hydrophilic factor
MLOGP	Moriguchi octanol-water partition coeff. (logP)
AMR	Ghose-Crippen molar refractivity
PSA	Fragment-based polar surface area

Consequently, the dataset contains 191 CAs-ligands complexes, which are described by a vector of 59 numerical values, where the first 24 values represent the CAs, the following 34 values represent the ligand, and the last one is the pKi value.

2.2 Approach to proteochemometrics

2.2.1 Rough sets model

First we introduce some basic definition and theory of rough sets [7]. The dataset is represented as a table, where each row represents a case, an event, or simply an object. Each column represents an attribute, a variable, an observation, or a property etc that can be measured for each object. This table is called an *information system*. More formally, an

information system is a pair $\mathsf{A} = (U, A)$, where U is a non-empty finite set of *objects* called the *universe* and A is a non-empty finite set of functions $a : U \to V_a$, called *attributes*; for each $a \in A$ the set V_a is called the *value set* of *a*. If there is a known outcome of classification, this a posteriori knowledge is expressed as one distinguished attribute called the *decision attribute*. The process is called supervised learning. An information system of this kind is called a *decision system*. Thus a decision system is any information system of the form $\mathsf{A} = (U, A \cup \{d\})$, where $d \notin A$ is the decision attribute. The element of A is called conditional attributes or simply conditions. The decision attribute may take several values though binary outcomes are rather frequent. The output of the rough set algorithms is a set of minimal *decision rules* of the form $\alpha \to \beta$ .Here α is a Boolean function $U \to \{true, false\}$ built up of the logical connectives $\wedge$ and atom statements of the form $a(\cdot) = v$ where $a \in A, v \in V_a$, Similarly $\beta : U \to \{true, false\}$ is built up from logical connectives $\vee$ and atom statements of the form $d(\cdot) = v$ where $v \in V_d$.

The datasets created is section 2.1 are represented by decision table where the CAs-ligands complexes are objects, the descriptors of the CAs and ligands are condition attributes, the pKi value are the decision attribute. Decision attributes are common for binary outcomes, however the original binding affinity values in the dataset are continuous numbers, therefore we sorted the decision table by the median value of the binding affinity values. The objects whose pKi value are larger than median value are assigned "high", the remaining ones are assigned "low".

Using rough sets to model the receptor-ligand interaction usually start with randomly split ing all the objects into two disjoint sets: A training set from which could induce minimal IF-THEN rules, and a test set which is used to verify how good the rules are at classifying new cases. However, the evaluation of the reliability of the performance using only one test set may not be so accurate, because it might be affected by the size of the set. So 10-fold cross validation (CV) are used to estimate the performance of the rules by computing the *accuracy mean* and *Area Under Curve (AUC) mean* and standard deviations (SD). The AUC is the area under the Receiver Operating Characteristics (ROC) curve and it is a measurement of the discriminatory power of a classifier [8]. The ROC curve results from plotting *sensitivity* against 1-*specificity* while letting the threshold value τ vary. For a binary classifier an AUC of 1.0 means that the discriminatory power is optimal while an AUC of 0.5 means that the classifier does not perform better than a random classification of objects. All the calculations were performed by the Rosetta software [9].

2.2.2 PLS model

PLS [10] is a multivariate analysis method that finds the relationship between a matrix of predictor variables, X and a matrix of dependent variables, Y. (In our case, the Y corresponded to pKi values.) The PLS analysis has the objective of approximating the X and Y spaces and maximizing the correlation between them. A reduction of dimensionality is

accomplished by simultaneously projecting the X and Y matrices on lower dimensionality hyper-planes that are termed PLS components.

Model Creation Prior to PLS, the data in the matrix X and Y should be first mean centered and scaled to unit variance (divided by the standard deviation of the variable). The goodness of fit between X and Y data of the PLS models was characterized by the fraction of explained variation of **Y** (R^2). The predictive ability was characterized by the fraction of the predicted **Y**-variation (Q^2), assessed by cross-validation, as previously described [11]. A model is considered to be acceptable for biological data when $R^2 > 0.7$ and $Q^2 > 0.4$ [12]. Moreover, Q^2 is used to determine how many PLS components should be used for the description of the model.

Cross term Ligand-receptor recognition can evidently only partially be explained by linear combinations of ligand and receptor descriptors. *E.g.*, if the ligands by virtue of some feature (property) interact with non-varied receptor residues, a simple assumption would be that the binding affinity relates linearly with the intensity of this given property. In reality, however, binding is governed by complex processes that depend on the complementarity of the properties of the interacting entities. In proteochemometrics this may be accounted for by computation of ligand-CAs cross-terms. Cross-terms were here formed by multiplying the ordinary descriptors. In this way one additional descriptor block was obtained comprising 1653 descriptors. Thus the total number of descriptors obtained, became 1712.

Coefficients The significance of X variables was assessed by the PLS regression coefficients. As the predictor variables are projected onto the Y by means of the PLS regression equation, a regression coefficient is a measure of the relevance of a chemical property descriptor for explaining the variation in the activity under study. In order to obtain a normalised measure for the significance of a primary term, the absolute value of its regression coefficient was multiplied by the standard deviation of the corresponding descriptor in the data set. All the calculations were performed by using SIMCA 10.0 software (Umetrics, Umea, Sweden).

3 Results

3.1 Rough sets model

3.1.1 Model evaluation

As referred in section 2.2.1, performances of the decision rules are measured by parameter *accuracy mean* and *AUC mean* and Standard deviation (SD). The accuracy mean is the average proportion of correctly predicted objects computed for the k block during CV. The *AUC mean* is the average AUC for the models induced by the k blocks during CV. For the decision table of CAs–ligands complexes, 10-fold CV resulted in an *accuracy mean* of 0.87 (SD=0.06) and an *AUC mean* of 0.92 (SD=0.05). The result of CV shows high validity of the rough set model to classify the objects using induced decision rules.

3.1.2 Interpretation of rules

A set of decision rules were induced which correlated the minimal number of CAs and ligands descriptors with "high" or "low" binding affinity and provided a large number of patterns to determine the binding affinity of CAs and ligands. For instance, one decision rule is "66_z3 ([*, 0.27]) ^ Me ([1.05, *)) ^ nAT ([30, 39)) → pKi (high)" , associating "polarizability and charge" of the residue at site 66 of CAs, "mean atomic Sanderson electronenagativity" and "number of atom" of the ligands with the binding affinity (pKi), and the interval of these 3 descriptors value determing the strength of the binding.

There are several numerical factors associated with decision rules. Most of these are derived from the *support* of a rule, which is the number of objects in the decision system that possess both properties α and β. The factor *coverage*, which is defined as coverage $(\alpha \circledcirc \beta) = \text{support}(\alpha \wedge \beta)/\text{support}(\beta)$, reflects the strength of a rule and gives a measure of how well α describes the decision class(-es) given by β. Thus, we select the rules of highest coverage to explain the "high" and "low" binding affinity of CAs-ligands interaction. The rules with highest coverage for "high" have the patterns:
CAs + Ss ([62.58, *)) + PSA ([113.38, *))
CAs + nCIR ([3, *)) or CAs + CAs + nCIR ([3, *))
CAs + nS ([3, *)) or CAs + CAs + nS ([3, *))
CAs + Ss ([62.58, *)) + nS([3, *))
CAs + nSK ([24, *)) + PSA ([113.38, *))
The rules with highest coverage of the rules for "low" have the patterns:
CAs + MW ([269.32, 362.35)) or CAs + CAs + MW ([269.32, 362.35))
CAs + Se ([*, 30.74)) + nSK ([*, 19))
CAs + Sp ([*, 30.74)) + nSK ([*, 19))
CAs + Sv ([*, 18.96)) + nSK ([*, 19))
CAs + MW ([*, 269.32)) + ARR ([0.151, *))
CAs are more likely to be 63-z1, z2, z3, 66-z1, z2, z3, 132-z1, z2, z3, 133-z1, z2, z3, 92-z2, z3. Thus, the "high binding" decision rules might be associated with ligands descriptors Ss, PSA, nCIR, nS, nSk Within a receptor, the "low binding" decision rules might be associated with ligands descriptors MW, Se, nSk, Sp, Sv, ARR .

3.2 PLS model

Creation of the Proteochemometric Model

PLS modeling of the data set using only descriptors of CAs and ligands resulted in a model explaining R2= 0.755 of the variance of ligands affinities and having a predictive ability of Q2=0.724 (Fig 1). Ligand-CAs cross-terms were then included, allowing us to account for the nonlinearity of the ligand and CAs-affinity profiles. This resulted in a further improved model was obtained with R2=0.810 at Q2=0.737 (Fig 2).

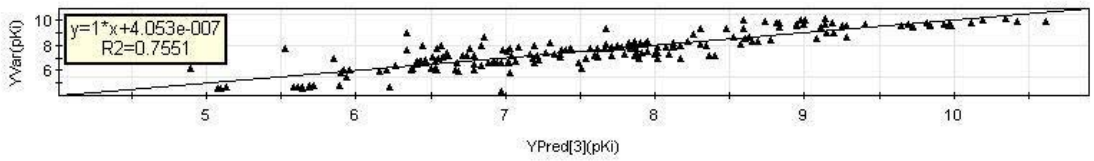

Fig 1 Correlation of calculated versus observed pKi values derived from PLS modelling of protein-ligand interactions without cross terms.

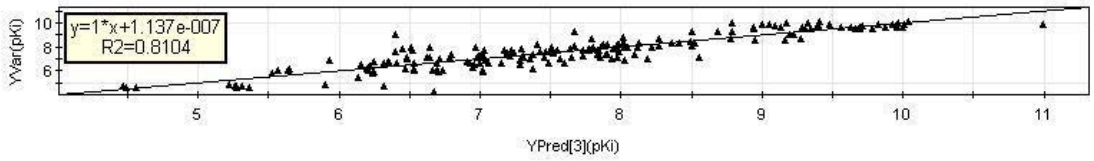

Fig 2 Correlation of calculated versus observed pKi values (i.e., 3log(Ki)) derived from PLS modeling of protein- ligand interactions without cross terms with cross terms.

Interpretation of the Model

Analysis of ligands Properties of Importance for CAs Binding.

To analyze the influence of different properties of the ligands on their overall affinities to CAs, we used the PLS regression equation of the model. The PLS coefficients for ligands descriptors are shown in Fig. 3. As can be seen, the regression coefficients for nS, PSA, nDB, nN, nCIR, nO, AMW, nCIC, MW, Mp attained the largest positive values.
MLOGP, ARR, nHDon, Hy gave the largest negative impact. The nAB, RBF also correlated negatively to the affinity, whereas the remaining descriptors of the ligands correlated positively. And only minor positive correlation to the affinity was associated to the nC and nBM.

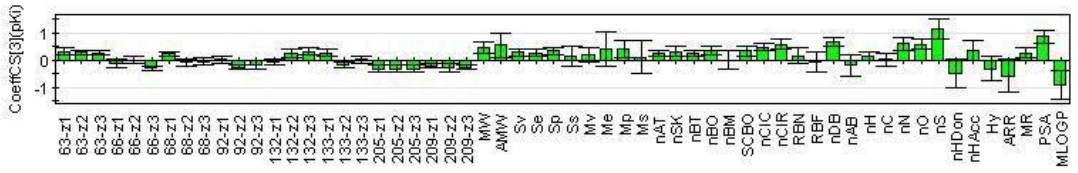

Fig3 PLS coefficients of the descriptors of enzyme and ligand derived from the final proteochemometric model.

The sign and magnitude of the PLS coefficient of the descriptor of the compounds reflects the impact of the underlying property of the ligands to the affinity to the receptor series.

However, depending on the actual descriptor value for a particular ligand, the contribution of the described property to the binding would for some ligands be positive, whereas for others it would be negative. Therefore, to reveal the contribution of the properties of particular compounds to their interaction activity, we multiplied each coefficient with the actual descriptor value for each given compound as follows:

$$\Delta pK_{i,l} = \text{coeff}_l \times (x_l - \bar{x}_l)$$

Using this approach, we found that the overall high affinity of liangds **1** (in the first row of the dataset) is associated with high numbers of nCIR, nHDon. Nevertheless, the largest negative influence is afforded by the MW, PSA. Thus, the model suggests that the high average affinity of the structure is not caused by MW and PSA, though PSA has largest coefficient value.

Contribution of CAs residues for Ligands Affinity.

To analyze the influence of different residues of the CAs on their overall affinities to ligands, we also used the PLS coefficients in Fig 3. The CAs descriptors had overall less impaction on the CAs-ligands interaction affinity. Descriptors 63-z1, 63-z2, 63-z3, 68-z1, 132-z2, 132-z3, 133-z1 achieved largest positive coefficients, 66-z3, 92-z2, 205-z1, z2, z3, 209-z1, z2, z3 obtained the largest negative coefficients. The remaining had little influence and can be neglected.

In general, the results of rough set are agree with those of PLS. The binding affinity are mostly associated with Descriptors 63-z1, z2, z3, 66-z3, 92-z2,132-z1, z2, z3, 205-z1, z2 z3 and 209-z1, z2, z3 for CAs and nS, PSA, nDB, nN, nCIR, nO, AMW, nCIC, MW, Mp, MLOGP, ARR, nHDon, Hy for ligands.

4 Discussion

In this study, we applied a novel technique of proteochemometrics to analysis the interactions between CA I, CA II, CA V and ligands. First, we described the problem mathematically, it seemed that the different parts of the CAs and the ligands contributed to the binding through their biochemical properties. Thus the CAs and ligands were encoded by a serial of quantitative descriptors. The next step was to find a relation between the formed chemical descriptors and the binding data. Two different models, rough sets and PLS, were selected to fulfill the talk. The present study shows that both rough sets and PLS modeling served the purpose quite well. The two approaches are to some respect complementary and may be used in combination to receive a better understanding of receptor-ligand interaction

Rough sets is a rule-based methods, in this case, rough sets model is of high quality with *accuracy mean*= 0.87 (SD=0.06) and an *AUC mean*=0.92 (SD=0.05). The output of rough sets are a series of minimal IF-THEN rules induced from dataset, which facilitate to understand and exploit the binding.

PLS model for the binding data gave a good validated model (i.e., the model displayed in

Fig. 1). The R2 parameter of this model was 0.755, indicating that the binding affinity could be explained by the 58 combined descriptors for the CAs and ligands. Moreover, the validity of the model is shown by cross-validation, which is quantified by Q2 measure. As mentioned in Section 2, this parameter gives an estimate of the predictive power of the model. A value above 0.4 is generally considered to be good; in our case it was 0.724.

Despite the fact that this is already a rather good model, the possibility that parts of CAs and ligands could induce interaction effects were not take into account. In order to improve the model, cross terms were introduced to the model. Results of these cross-terms resulted in a major improvement of the model with R=0.81 and Q2=0.737 (Fig. 2).

In order to interpret the model easily, we calculate the coefficients of each descriptors, which represent a measure of the contribution of the descriptors to the model (Fig 3). Within CAs, 63-z1, 63-z2, 63-z3, 68-z1, 132-z2, 132-z3, 133-z1 enhance the binding affinity largely, while 66-z3, 92-z2, 205-z1, z2, z3, 209-z1, z2, z3 lessen the binding. For ligands descriptors are shown in nS, PSA, nDB, nN, nCIR, nO, AMW, nCIC, MW, Mp associated the binding positively, while MLOGP, ARR, nHDon, Hy have the contary effects.

PLS is a linear method which ranks all the attributes and cross terms from most influential to least influential for binding affinity. The coefficients of the PLS model illustrate this clearly (Fig 3). While rough sets modeling can deal with the nonlinear problem, and it does not produce a ranking of attributes. Instead it selects minimal groups of essential attributes that have the same classification power. In this case, decision rules focus on combination of attributes important for binding. Being a linear model, PLS has to compute cross terms, which represent the non-linear terms, but only two attributes can be combined in each cross term, thus PLS could not describe the nonlinear issues very well. For rough sets, we need not take account of cross terms, each decision rule is a combination of attributes without the number restriction. Moreover, for the rough set approach, decision rules associate high and low binding to certain attribute values while the PLS ranking is not relating to attribute value. To sum up, rough sets provide an explicit model with a serial of classifer to determine the binding, while PLS models are adept at predicting the contribution of descriptors to binding affinity. These two model analysis the binding affinity from different angles, thus combination of the two models will lead to better understand the CAs-ligands interactions.

References

[1] Wikberg, Jarl E S, Lapinsh Maris, Prusis Peteris. Proteochemometrics: A tool for modelling the molecular interaction space. *Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective* (2004) 289-309.
[2] Peteris Prusis , Ruta Muceniece, Per Andersson. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand^receptor interactions. *Biochimica et Biophysica Acta* 1544 (2001) 350^357
[3] Zdzis aw Pawlak. Rough sets. *International Journal of Information and Computer Science*, 11(5):341.356, 1982.
[4] Zdzis aw Pawlak. Rough Sets: Theoretical Aspects of Reasoning about Data, volume 9 of *Series D: System Theory, Knowledge Engineering and Problem Solving*. Kluwer Academic

Publishers, Dordrecht, The Netherlands, 1992.
[5] Zhang, J, Aizawa, M. Amari, S. Iwasawa, Y. Nakano, T. Nakata, K. Development of KiBank, a database supporting structure-based drug design, *Comput Biol Chem*. 5-6 (2054) 401-407.
[6]M. Sandberg, L. Eriksson, J. Jonsson, M. Sjo« stro«m, S. Wold, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, *J. Med. Chem.* 41 (1998) 2481-2491.
[7] Komorowski, J.; Pawlak, Z.; Polkowski, L.; Skowron, A.: Rough sets - a tutorial. In (S. K.Pal, and A. Skowron, Eds.) *Rough-fuzzy hybridization - A new trend in decision making.* Springer Verlag, Singapore, 1999; pp. 3-98.
[8] Bradley,AP The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition* 30(7):(1997) 1149-1157
[9] Øhrn, A.; Komorowski, J.; Skowron, A.; Synak, P.: The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets, The ROSETTA System: In (L. Polkowski, A. Skowron Eds.) *Rough Sets in Knowledge Discovery 1, methodology and applications. Studies in Fuzziness and Soft Computing. Physica-Verlag*, Heidelberg, 1998; Vol. 19, pp. 572-576
[10] P. Geladi, B.R. Kowlaski, Partial least-squares regression: A tutorial, *Anal. Chim. Acta* 185 (1986) 1^17.
[11] Wold S (1995) PLS for multivariate linear modeling, *Chemometric Methods in Molecular Design* (van de Waterbeemd H ed) vol 2, pp 195–218, VCH Verlagsgesellschaft, Weinheim, Germany.
[12] SIMCA 7.0. A New Standard in Multivariate Data Analysis, Manual, Umetrics AB, Umea®, Sweden, 1998.

Key to read JOY format

solvent inaccessible	UPPER CASE	X
solvent accesible	lower case	x
alpha helix	red	x
beta strand	blue	x
3 - 10 helix	maroon	x
hydrogen bond to main chain amide	**bold**	**x**
hydrogen bond to main chain carbonyl	underline	x
disulphide bond	cedilla	ç
positive phi	*italic*	*x*

Multiple Alignment

```
                          10        20        30        40
50
1A42  (    4 )
hwgYgkhnGpehWhkdfpiAkgerQSPVdIdthtAkydpslkpLsVsY
1DMY  (   25 )
gtrgSPInIqwkdSvydpqLapLrVsY
1HCB  (    4 )
dwgYddknGpeqWsklypiAngnnQSPVdIktsetkhdtsLkpIsvsY
                                          bb 333 bb        bbb

                          60        70        80        90
100
1A42  (   52 )
--dqAtSlrIlNnGhaFnVeFddsqdkAvLkgGpLdgtYrLiqFHFHWGs
1DMY  (   52 )
--daasCryLwNtGyFFqVeFddscedSGIsgGPLgnhYrLkQFHFHWGa
```

```
1HCB  (  52 )
--npaTAkeIiNvghSFhVnFednd nrSvLkgGpfsdsYrLfqFHFHWGs
                          bbbbbb     bbbbb        bbb        bbbbbbbbbb

                                110        120        130        140
150
1A42  ( 100 )    ldgqGSEHtvdkkkyAAELHLVHWNtk-ygdfgkAvqqpdGLAVLGIFLk
1DMY  ( 100 )    tdewGSEHAvdghtypAELHLVHWNstkyenykkAsvgenGLAVIGVFLk
1HCB  ( 100 )    tnehGSEHtvdgvkySAELHVAHWNSakysslaeAaskadGLAVIGVLMk
                             bb       bbbbbbbb        aaaa
bbbbbbbbb

                                160        170        180        190
200
1A42  ( 150 )    vg-sakpgLqkVVdvLdsIktkgksadftnFdPrgLlPe---sldYWTYp
1DMY  ( 150 )    lg-ahhqaLqkLVdvLpeVrhkdtqvaMgpFdPscLMpa---crdYwTYp
1HCB  ( 150 )    vg-eanpkLqkVLdaLqaIktkgkrapftnfdPstLLPs---sldFWTYp
                  b         aaaaa 3333      bbb      3333
bbbbb

                                210        220        230        240
250
1A42  ( 196 )    GSLTTPpLleCVtWIVLkepIsVsseQVlkFrk-LNfngegepeelMvdN
1DMY  ( 196 )    GSLTtppIaeSVtWIVQktpveVspsQLsmFrt-LLfsgrgeeedvMvnN
1HCB  ( 196 )    GSlThPpLyeSVtWIICkesisVsseqLaqFrs-Llsnvegdnavpmqhn
                  b           bbbbbb    bbb aaaaaaa

                            260
1A42  ( 245 )     wRpaqplkmrqIkasf
```

```
1DMY  ( 245 )    yrplqplrdrkLrSsfr
1HCB  ( 245 )    nRptqplkgrtVrAsf
                                 bb
```