Proteochemometric Mapping of the Interaction of Organic Compounds with Melanocortin Receptor Subtypes

Maris Lapinsh, Santa Veiksina, Staffan Uhlén, Ramona Petrovska, Ilze Mutule, Felikss Mutulis, Sviatlana Yahorava, Peteris Pruisis and Jarl E. S. Wikberg*

Department of Pharmaceutical Biosciences, Uppsala University, Box 591 BMC, SE-75124 Uppsala, Sweden
Running Title Page

Running title: Proteochemometric Mapping of Melanocortin Receptors

Corresponding author: Jarl Wikberg, Pharmaceutical Pharmacology, Box 591, BMC, SE751 24 Uppsala, Sweden. Phone: +46-18-471 42 38. Fax +46-18-55 97 18. Email: Jarl.Wikberg@farmbio.uu.se

Number of text pages: 30

Number of tables: 2

Number of figures: 7

Number of references: 31

Number of words in Abstract: 247

Number of words in Introduction: 522

Number of words in Discussion: 1082

Abbreviations: PLS, partial least-squares projections to latent structures; MCR, melanocortin receptor; GPCR, G protein-coupled receptor; MSH, melanocyte stimulating hormone; ACTH, adrenocorticotropic hormone; TM, transmembrane region; PCA, principal component analysis; SMD, statistical molecular design.
Abstract

Proteochemometrics was applied in analysis of the binding of organic compounds to wild-type and chimeric melanocortin receptors. Thirteen chimeric melanocortin receptors were designed based on statistical molecular design; each chimera containing parts from three of the MC1, 3-5 receptors. The binding affinities of 18 compounds were determined for these chimeric melanocortin receptors and the four wild-type melanocortin receptors. The data for 14 of these compounds were correlated to the physico-chemical and structural descriptors of compounds, binary descriptors of receptor sequences, and cross-terms derived from ligands and receptors descriptors, in order to obtain a proteochemometric model. (Correlation was performed using partial least-squares projections to latent structures, PLS). A well-fitted mathematical model ($R^2=0.92$) with high predictive ability ($Q^2=0.79$) was obtained. In a further validation of the model, the predictive ability for ligands ($Q^2_{lig}=0.68$) and the predictive ability for receptors ($Q^2_{rec}=0.76$) were estimated. The model was moreover validated by external prediction by using the data for the 4 additional compounds that had not at all been included in the proteochemometric model; the analysis yielding a $Q^2_{ext}=0.73$. Interpretation of the results using PLS coefficients revealed the influence of particular properties of organic compounds on their affinity to melanocortin receptors. Three-dimensional models of melanocortin receptors were also created, and physico-chemical properties of the amino acids inside the receptors transmembrane cavity correlated to the PLS modeling results. The importance of particular amino acids for selective binding of organic compounds was estimated and used to outline the ligand recognition site in the melanocortin receptors.
Introduction

Melanocortin receptors (MCRs) are members of the seven transmembrane-spanning G protein-coupled receptor (GPCR) superfamily. To date, five MCR subtypes, MC$_{1-5}$, are recognized in mammals, all of which stimulate cAMP signal transduction pathways. An endogenous group of peptides, the melanocyte stimulating hormones (MSH) and adrenocorticotropic hormone (ACTH), and the proteins agouti and the agouti-related protein bind to the MCRs with agonistic and antagonistic properties, respectively. However, an exception is the MC$_2$R, which binds only ACTH (Schioth et al., 1996a).

The MCRs have a wide range of physiological roles. The MC$_1$R regulates melanin pigment formation in the skin, and it has a regulating role in the immune system. The MC$_2$R regulates corticosteroid production of the adrenals. The MC$_3$ and MC$_4$Rs play important roles in control of feeding and sexual behaviours, and the MC$_5$R is involved in the regulation of exocrine glands (Wikberg et al., 2000; 2002; Wikberg, 2001). The potential of using the MCRs as targets for drugs treating important medical conditions such as obesity/diabetes, inflammatory conditions and sexual dysfunctions prompts the need for compounds showing high specificity for particular MCR subtypes. However, the design of selective drugs for subtly different receptor subtypes is a difficult task that would become simplified if one had in hand detailed knowledge on the determinants for ligand-receptor interaction.

We have recently invented the proteochemometric modeling technology for the analysis of protein-ligand interactions (Wikberg et al., 2003). This technology is particularly suited to gain knowledge in the differences in the ligand-recognition of different targets. Contrary to traditional QSAR methods that aim to correlate a description of ligands with their affinity for one particular target, proteochemometrics considers many targets and ligands simultaneously. Proteochemometrics thus analyses the experimentally determined interaction activity data of a series of ligands for a series of proteins by correlating the data to descriptors of ligands, proteins and cross-terms derived from ligand and protein
descriptors. The differences in the properties of both interaction partners are accordingly exploited to explain ligand-receptor affinity.

In a previous study we applied proteochemometrics to study the interactions of wild-type and chimeric and point mutated α-adrenergic receptors with piperidyl oxazole derivatives (Lapinsh et al., 2001). In another study we applied it onto twenty-one native amine GPCRs interacting with a set of diverse organic compounds (Lapinsh et al., 2002). We have also modeled the interactions of chimeric melanocortin receptors with natural and synthetic analogues of melanocortin peptides (α-MSH, NDP-MSH and I-NDP-MSH) (Prusis et al., 2001, 2002), and of native MC1,3,5Rs with a large series of organic compounds (Lapinsh et al., 2003). In all cases statistically highly valid models were obtained, which were interpretable in a chemical sense. For example, the most recent of our studies (Lapinsh et al., 2003) allowed us to reveal compound properties influencing the binding of organic compounds to different MCR subtypes. However, using only four receptors in the modeling did not allow us to map the receptors’ ligand binding pockets. In the present study we have undertaken a further analysis to model the recognition site for organic compounds in the MCRs by using four native and thirteen multipart chimeric MCRs, which were designed using the principles of statistical molecular design (SMD).
Materials and Methods

Materials. Primers were from TAG Copenhagen A/S, Denmark. Restriction endonucleases HindIII, XhoI and XbaI were from Promega, USA. \[^{125}I\]-NDP-MSH (\[^{125}I\]-[Nle\(^4\), -Phe\(^7\)]\(\alpha\)-MSH) was custom synthesized by EuroDiagnostica AB, Sweden. The genes of the MC\(_1\) and MC\(_3\)Rs had been cloned earlier in our laboratory (Chhaljani et al., 1992, 1993). The MC\(_3\) and MC\(_4\)R genes were gifts of Dr. Ira Gantz (Gantz et al., 1993a, 1993b).

Experimental Design of Chimeric Receptors. The chimeric receptors contained parts from the human MC\(_1\), MC\(_3\), MC\(_4\) and MC\(_5\)Rs. The native receptors were divided into three parts (A, B and C), which could then maximally be combined into \(4^3\cdot4=60\) three-part chimeras containing portions from three or two melanocortin receptor subtypes. However, we wanted to maximize the information content in the data gained from chimeric receptors, at a minimal experimental work. Therefore, we applied statistical molecular design (SMD) using D-optimal design (Eriksson et al., 1996) as generated by the MODDE 6.0 software (Umetrics AB, Umeå, Sweden). According to the design 12 chimeras were selected, which, together with the four native receptors provided the best representation of all possible combinations of receptor portions in linear terms. The design thus contained all \(4\cdot4\) possible pairs of different sequence combinations in the A/B, A/C and B/C part of the receptors, the 12 selected chimeras being 1-3-4, 1-4-5, 1-5-3, 3-1-5, 3-4-1, 3-5-4, 4-1-3, 4-3-5, 4-5-1, 5-1-4, 5-3-1, 5-4-3 (the numbers representing the respective MCR subtype).

In order to obtain chimeric receptors with unaltered folding and functionality we elected to recombine receptors in the sequence stretches showing the highest conservation among all four melanocortin receptors. We identified four such highly conserved stretches, making it possible to divide the receptors into five segments. For technical reasons the receptors were created as two sets of chimeras. For the first set (F-set) the combinations took place at the beginning of the second of the seven transmembrane
regions (i.e., TM2) (residues 73-81 in the MC$_1$ receptor*) and in the middle of TM6 (residues 253-259).

In the second set (S-set) the combination sites were located at the end of TM3 (residues 138-143) and TM5 (205-210).

**Primer Design and Manufacture of DNA Constructs by PCR.** Part A of the F-set had been built by using vector-specific forward primers, and receptor specific reverse primers (GTA CCT GTC CAC TGC GAT GGC for MC$_1$, MC$_3$ and MC$_5$Rs, and GTA CCT GTC CAC TGC GAT TGA for MC$_4$R) recognizing the sequence encoding the IAVDRY sequence stretch at the end of the third transmembrane helix, which is present in all the four melanocortin receptors. Part B of the F-set was built by using forward primers (ATC GCA GTG GAC AGG TAC ATC TCC A for MC$_1$R and ATC GCA GTG GAC AGG TAC for MC$_3$-5Rs) recognizing IAVDRY, and reverse primers (CAT GTG GAC GTA CAG CAC for MC$_1$R, CAT GTG GAC GTA CAG GGT for MC$_3$R and CAT GTG GAC GTA CAG AGA for MC$_4$-5Rs) recognizing LYVHM at the end of the fifth transmembrane helix of the MC$_1$, MC$_3$, and MC$_4$Rs. (For the MC$_3$R the primer will induce a change in the sequence from LYIHM to LYVHM). Part C had been built using forward primers (CTG TAC GTC CAC ATG CTG for MC$_1$R and CTG TAC GTC CAC ATG TTC CT for MC$_3$-5Rs) recognizing LYVHM, and vector-specific reverse primers. The three sets of PCR amplifications resulted in DNA products for all the four melanocortin receptors with an 18 bp identical overlap between the end of the first and the beginning of the second segment, as well as a 15 bp identical overlap between the end of the second segment and the beginning of the third segment. When segments were combined their overlaps acted as primers for each other. For example, combination of the MC$_1$R part A of F-set and the MC$_3$R part B, and the T7 forward primer and the LYVHM reverse primer resulted in amplification of the MC$_1$/MC$_3$R chimeric part A/B segment. When this A/B segment was combined with part C of the MC$_4$R, together with the T7 primer (forward), and a reverse primer recognizing the 3’-UT region (e.g. the T3 primer), the full length

* Amino acid numbering herein always refer to the position (or corresponding position) in the MC$_1$R
MC₁/MC₃/MC₄ chimeric receptor gene was produced. Combination of parts A, B, and C from different melanocortin receptors could thus produce anyone of the multiple chimeric receptor constructs of the F-set.

The S-set had been built by taking advantage of the almost perfect overlap between the melanocortin receptors at the beginning of the second transmembrane helix (MYFFICSSL; exact match in MC₄R) and at the sixth transmembrane helix (CWAPFFL; exact matches for the MC₃, MC₄, and MC₅Rs). For the annealing site between parts A and B forward primer ATG TAC TTC TTC ATC TGC AGC CTG GC and reverse primer GC CAG GCT GCA GAT GAA GAA GTA CAT had been used. For the annealing site between parts B and C forward primer TGC TGG GCC CCC TTC TTC CT and reverse primer AG GAA GAA GGG GGC CCA GCA had been used. The S-set of PCR amplified bands resulted in a 26 bp overlap between parts A and B, and a 20 bp overlap between parts B and C.

PCR was performed at 95°C for 2 min, followed by 26 cycles of 95°C for 1 min, 47°C for 40 sec, and 68°C for 1 min, using a Technne apparatus, UK. PCR products were isolated by agarose gel electrophoresis. Bands were recovered using a standard NaI/silica-based method (Vogelstein et al., 1979), and combined to produce chimeras by adding the outer primers and then running 31 PCR cycles.

**Cloning, Sequencing and Expression of Chimeric Receptors.** The chimeras where the A segment had been taken from the MC₁ or MC₅R genes had a vector-specific HindIII-site before the start codon, while those taken from the MC₃ or MC₄R genes had a XhoI-site before the start codon. All the chimeras had a vector-specific XbaI-site after the stop codon. The chimeras starting with MC₁ or MC₅R sequences had been cloned with HindIII and XbaI into the pcDNA.3 expression vector and chimeras starting with MC₃ or MC₄R sequences had been cloned with XhoI and XbaI into the pCiNeo vector. To assure that the chimeras were correct they were sequenced, using an ABI Prizm sequencer. (Full accounts on the manufacture of the DNA constructs will be given elsewhere).
For receptor expression, COS-7 cells were grown in Dulbecco’s modified Eagle’s medium with 10% foetal calf serum. Eighty percent confluent cultures were transfected on 100 mm dishes with the expression constructs of chimeric or wild-type melanocortin receptors (10 µg DNA per dish, mixed with liposomes, as described (Schioth et al., 1996b)). 12-16 hours after transfection the serum-free medium was replaced with growth medium and the cells were cultivated for about 48 hours, then scraped off, centrifuged and used for radioligand binding.

**Data Set for Proteochemometric Modeling.** Unfortunately we failed to obtain full length constructs for some chimeras, while some other chimeras showed very low levels of expression making their use unfeasible. In order to obtain a working-set of receptors we therefore combined the F- and S-sets so that the final set included four native and 13 chimeras (eight from the F-set and five from the S-set), as is schematically depicted in the Figure 1. As seen, each of the receptors can be considered as consisting of five segments. Parts of all four native receptors are well represented in the chimeras, with the exception of the third segment of MC4R and the fifth segment of MC5R, which are present only in the native receptors and one of the chimeric receptors. In the following the number system given in Figure 1 will be used to denote these multiple chimeric receptors.

Eighteen organic compounds showing binding activity for MCRs were synthesized in our laboratory (Figure 2). Of these one (1) had been designed earlier by others (Sebhat et al., 2002) while the rest were our original designs. Compounds 2 and 4 were reported earlier by us (Mutulis et al., 2002a, 2002b). Full accounts on the synthesis of the other compounds will be given elsewhere.

Interaction affinities (expressed as the negative logarithm of dissociation constants, pKᵦs) were determined using a competition binding assays with the radioligand [¹²⁵I]-NDP-MSH. Dissociation constants of the radioligand for each receptor were estimated by saturation assays and the dissociation constants of competing compounds were then determined by competition assays. All calculations were based on non-linear curve-fitting assuming that ligands bind to site according to the law of mass action, essentially using the approach described earlier (Schioth et al, 1995, 1996b). The radioligand binding
was performed in Dulbecco’s MEM containing 2g/l albumin and 0.2 g/l phenanthroline without foetal calf serum. Cells were washed with binding buffer, scraped off, and distributed into 96-well non-culture-coated plates, which were centrifuged and the binding buffer removed. The cells were then immediately incubated with radioligand and organic compound for 1 h at 37 °C in 50 µl binding buffer/well. In the saturation experiments the concentration of [125I]-NDP-MSH was varied by dilution in 2-fold intervals covering a range of about 6 pM to 12 nM. Non-specific binding was defined in the presence of 3 µM NDP-MSH. In the competition experiments about 1 nM of [125I]-NDP-MSH and various concentrations of competing ligand were added to the assays. After incubation, the cells were washed with 0.2 ml ice-cold binding buffer and then detached from the plates with 0.2 ml of 0.1 N NaOH. The binding assays were performed in duplicates and repeated at least three times. Curve fitting for computing of dissociation constants was performed using Prism 4 software (GraphPad Software, Inc., San Diego, CA). The dissociation constants (pK) for [125I]-NDP-MSH obtained from the saturation studies of the wild-type and chimeric melanocortin receptors were as follows; MC1=10.34, MC3=9.31, MC4=8.73, MC5=8.58, 11533=10.38, 33544=8.66, 44133=8.64, 44355=8.81, 44511=8.54, 55144=8.45, 55311=10.85, 55433=9.51, 13334=9.98, 35554=8.62, 45551=9.35, 51114=10.04 and 53331=9.18. (Results from the competition studies are given in Table 1).

Our data set obtained on 18 organic compounds and 17 receptors thus included 18*17=306 interaction affinity values (Table 1). In a few cases competition binding was not observed up to a concentration of one millimolar (pKi<3). In these cases we arbitrarily assigned pKi=3.

The large number of observations allowed us to divide the data into a work set, comprising the receptor-affinities of 14 compounds which were used for model creation, and a test set comprising the receptor-affinities of four compounds, which was set aside and used after the creation of the proteochemometric model to assess the model’s predictive ability.

**Description of Organic Compounds.** Compounds were characterized by 62 descriptors, calculated by the Dragon 2.1 software (Talete S.r.l., Milano, Italy). Descriptors represented different physico-
chemical properties (molecular weight, van der Waals volume, electronegativity, polarizability, molar refractivity, polar surface area, logP etc.), and the numbers of functional groups and structural fragments in the molecule. Before their use the descriptors were checked for mutual correlation. For each pair of descriptors with mutual correlation higher than 0.95, the one showing the highest correlation with any other descriptor was excluded. A fair number of the remaining descriptors showed invariant (constant) values for more than 4/5 of the compounds. For reasons suggested elsewhere (Q2 manual, Multivariate Infometric Analysis S.r.l., Perugia, Italy) these descriptors were also discarded. After these procedures 31 descriptors remained in the data set (see list in the legend to Figure 4).

**Binary Description of Receptors.** As discussed above the receptors can be considered to consist of five segments. We described each of these segments separately by using four binary descriptors. The first descriptor was equal to one when the segment was taken from MC$_1$R; otherwise it was set to minus one. The second descriptor was equal to one when the segment was taken from MC$_3$R; otherwise it was set to minus one, and so forth. In this way each receptor was represented by $5 \times 4 = 20$ descriptors.

**Ligand-Receptor Cross-Terms.** Ligand-receptor recognition depends on the complementarity of properties of two interacting entities. Such complementarity cannot be explained by linear combinations of ligand and receptor descriptors as complex non-linear processes govern it. In proteochemometrics the non-linearity may be accounted for by computing ligand-receptor cross-terms (Wikberg et al., 2003; Lapinsht et al., 2001, 2002; Prusis et al., 2001, 2002). Cross-terms were here obtained by multiplying mean centered descriptors of compounds and receptors. In this way an additional descriptor block was obtained, comprising $31 \times 20 = 620$ descriptors. Cross terms were also calculated between mean centered descriptors of receptor, representing different sequence segments. This block included $(20 \times 16)/2 = 160$ descriptors.

**Pre-Processing of Data.** All descriptors were first mean centered and scaled to unit variance. Since the data set comprised descriptors of different types (i.e., descriptors of ligands, receptors and cross-terms) block scaling was applied. While the variables of the same type kept equal variance, scaling
weights between blocks were systematically varied until an optimal (i.e., the most predictive) model was obtained. Prior to use in the computations the response variable (pK$_i$) was also mean centered.

**Partial Least-Squares Projections to Latent Structures.** Descriptors were correlated to the affinity data by partial least-squares projection to latent structures (PLS). PLS is a multivariate analysis method that finds the relationship between predictor variables (X matrix) and response variables (Y matrix or vector; in our case the Y corresponded to pK$_i$). The PLS analysis has the objective to approximate X and Y by simultaneous projecting to latent variables (components), with an additional constraint to maximize the covariance between projections of X and Y. For each response, PLS derives a regression equation, where regression coefficients reveal the direction and magnitude of the influence of X-variables on the response (for a detailed description of PLS algorithms see Geladi and Kowalski, 1986; Wold, 1995).

For a proteochemometric model comprising descriptors of receptors, ligands, ligand-receptor cross-terms and intra-receptor cross-terms the regression equation can be expressed as:

$$y = \bar{y} + \sum \left(\text{coeff}_r \ast (x_r - \bar{x}_r)\right) + \sum \left(\text{coeff}_i \ast (x_i - \bar{x}_i)\right) + \sum \left(\text{coeff}_{r,j} \ast (x_r - \bar{x}_r) \ast (x_j - \bar{x}_j)\right) + \sum \left(\text{coeff}_{r_1,r_2} \ast (x_{r_1} - \bar{x}_{r_1}) \ast (x_{r_2} - \bar{x}_{r_2})\right)$$

(eq. 1)

PLS analysis was carried out using the Q2 software (Multivariate Infometric Analysis S.r.l., Perugia, Italy).

The goodness of fit of the PLS models was characterized by the fraction of explained variation of Y (R$^2$Y). The predictive capability was characterized by the fraction of the predicted Y-variation (Q$^2$), assessed by cross-validation, as previously described (Baroni et al., 1993, Ericsson, 1996). R$^2$Y may vary between 0 and 1; the value increases by each extracted PLS component. Q$^2$ values usually vary between 0 and 1; however, negative values can also be encountered, indicating non-predictive models. A model of biological data is generally considered acceptable if R$^2$Y > 0.7 and Q$^2$>0.4 (Lundstedt et al., 1998). In the current study cross-validation was performed using 5 randomly formed groups using 100
repeats. The Q² estimates were used to adjust block-scaling weights and to determine the optimal number of PLS components.

**Three-Dimensional Modeling and Physico-Chemical Characterization of the Melanocortin Receptors’ Ligand Binding Pocket.** We constructed three-dimensional models of transmembrane regions of MCRs using the crystal structure of bovine Rhodopsin as template (Palczewski et al., 2000). Sequence alignments of human MC₁,₃-₅Rs and of bovine Rhodopsin were taken from the GPCR database (Horn et al., 2003). Alignment of transmembrane regions of MCRs showed that over 40% (78 out of 178) amino acids were conserved among all the four MCR subtypes used herein. Using the 3D models we selected the residues that varied between the receptors, and which faced the inside of the ligand-binding cavity. Thirty-seven residues were chosen and subsequently coded by z-scale descriptors (Sandberg et al., 1998). These z-scales encapsulate 26 measured and computed physico-chemical properties of amino acids and are obtained by principal component analysis from the original properties and are accordingly i) orthogonal to each other and ii) scaled in the way that the same numerical difference in each z-scale corresponds to the same physico-chemical difference between amino acids. Furthermore, z-scales are interpretable and represent essentially hydrophobicity (z₁), steric bulk properties and polarizability (z₂), polarity (z₃) and electronic effects (z₄, z₅) of amino acids. Moreover these 5 scales represent more than 95% of the original measured and computed properties of the amino acids. In this way, the differences in presumed binding pockets of MCRs were encoded by 37*5=185 descriptors.
Results

Results of Radioligand Binding

The affinities of the 18 compounds for the 4 wild type MC1,3-5, and the 13 multiple chimeric melanocortin receptors, determined by radioligand binding, are shown in Table 1. As seen the affinities covered a range of more than four logarithmic units. Most of the compounds were MC1R selective, while three compounds showed their highest affinities for the MC4R. As seen from the table the data was divided into a work set comprising 14 compounds and a test set comprising 4 compounds. For the subsequent modeling only the work set was used, whereas the test set was used to validate the model, by using so called external prediction (see below).

Creation of the Proteochemometric Model

PLS modeling of the work set (Table 1) using only descriptors of receptors and organic compounds resulted in a five component model explaining $R^2_Y = 0.77$ of the variance of compound affinities, and having a predictive ability $Q^2 = 0.70$. Ligand-receptor cross-terms were then included, allowing us to account for the non-linearity of the ligand and receptor affinity profiles. This resulted in a five-dimensional model explaining $R^2_Y = 0.86$ at $Q^2 = 0.73$. Further improvement was obtained by including intra-receptor cross-terms. This was a reasonable measure, as the creation of chimeras would be expected to result in receptors with altered folding having a negative influence on the ligand binding. Indeed, PLS modeling showed that some of the intra-receptor cross-terms attained large negative coefficients, indicating that some particular combinations of receptor segments diminishes the ligands’ affinities. However, the majority of the intra-receptor cross-terms were insignificantly small. The final model was therefore created by including only the 16 intra-receptor cross-terms showing the largest negative coefficients. The performances of the models after extracting different numbers of PLS components are summarized in Table 2. As seen extracting five to seven PLS components led to models with the same predictive abilities, $Q^2 = 0.79$. Cross-validation was further performed so that all 17 observations of each compound were included in the same cross-validation group. In this way the
capacity of the model to predict the affinity of novel ligands (here termed Q²lig) was assessed. Similarly we assessed Q²rec by including all 14 observations of each receptor in the same cross-validation group. As seen from Table 2 very high Q²rec values, of about the same magnitudes, were obtained after extracting 5 to 7 PLS components. However, the Q²lig value reached its largest value after extracting 7 components. Closer inspection of the cross-validation results showed that one of the compounds (4; the only structure containing two guanidine groups) was systematically predicted with a too low affinity. Without including the predictions for this compound Q²lig would have reached 0.73. Results for the final model (i.e., the model with seven extracted components) are illustrated graphically in Figure 3. External predictions for the four compounds, not included in the data set during centering of descriptors, calculations of cross-terms and model elaboration, confirmed the high predictive capacities of the model (Q²ext = 0.73). The results are presented graphically in Figure 3, where the observed versus predicted pKi values are plotted.

In the following, the seven-component model will be referred to as “the model” and is, unless otherwise stated, the one used in all subsequent analysis.

**Interpretation of the Model**

**Analysis of Compound Properties of Importance for Melanocortin Receptor Binding.** In order to analyze the influence of different properties of the compounds on their over all affinities to MCRs we used the PLS regression equation of the model. The PLS coefficients for compound descriptors are shown in Figure 4. As can be seen, the regression coefficients for the numbers of nitrogen atoms, secondary aliphatic amines and amides, unsaturation index, number of unsubstituted aromatic sp2 carbon atoms, molecular weight and number of circuits attained the largest positive values. The presence of phenol in the molecule gave a large negative impact. The numbers of tertiary aliphatic amines, oxygen atoms, rotatable bond fraction and mean electrotopological state correlated also negatively to the affinity, while the mean atomic van der Waals volume and logP correlated positively. A negative coefficient was also assigned to the number of halogen atoms in the molecule; however
halogen atoms attached to the aromatic ring were assessed positively. Surprisingly, only minor positive correlation to the affinity was associated to the number of 6-, 9- and 10-membered rings.

The sign and magnitude of the PLS coefficient of the descriptor of the compounds reflects the impact of the underlying property of the compounds to the affinity to the receptor series. However, depending on the actual descriptor value for a particular compound the contribution of the described property to the binding would for some compounds be positive, whereas for others it would be negative. Accordingly, in order to reveal 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.

$$\Delta pK_i = \text{coeff}_i * (x_i - \bar{x}_i)$$ (eq. 2)

Using this approach we, e.g., found that the overall high affinity of compound 1 is associated with high numbers of nitrogen atoms and 6-membered rings, a low rotatable bond fraction (i.e., lack of long alkyl chains) and the presence of secondary and tertiary amides in the scaffold of the structure. Nevertheless, a negative influence is afforded by the presence of chlorine (and halogen, although positively assessed is attachment of halogen to the aromatic ring). Thus, the model suggests that the high average affinity of the structure is not due to the presence of chlorine, and would not be lost by replacement of chlorobenzene group by, e.g., naphthalene (such a modification, however, would essentially change the selectivity profile of the compound).

In compound 2 the most positive influence is afforded by a high number of unsubstituted aromatic carbon atoms and a high value of unsaturation index, i.e., properties that in the present case indicate presence of two naphthalene moieties in the structure. Positively assessed are also properties that indicate the presence of guanidine; thus, we may conclude that further increase in affinity of 2 could be sought by modifying the scaffold of the structure. Similar analysis reveals that the affinity of compound
4 could be significantly improved by including more aromatic groups and by increasing the hydrophobicity (logP) of the structure.

**Contribution of Receptor Segments for Binding of Organic Compounds.** We used the PLS regression equation of the model to calculate the change in pKi for each compound for 20 hypothetical receptors, in which one of the sequence segments is taken from MC1, 3-5, while the descriptor values for the other four segments are replaced by the mean value for the data set (i.e., zero after centering). Thus, 20 parameters were obtained for each compound, here termed $\Delta p\text{Ki}(1,MC_1)$ … $\Delta p\text{Ki}(5,MC_5)$, allowing comparisons of contribution of different receptor sequence segments to the compounds’ affinities.

In order to assess the importance of sequence segments for binding of the whole compound series we calculated the averages of the $\Delta p\text{Ki}(1,MC_1)$ to $\Delta p\text{Ki}(5,MC_5)$ values for the 14 compounds ($\Delta p\text{Ki}(1,MC_1)$ to $\Delta p\text{Ki}(5,MC_5)$, as is depicted in Figure 5 Panel A). However, our data set included compounds with differing selectivity. Thus, two compounds (1 and 3) were MC4R selective, while the majority of compounds showed their highest affinities for the MC1R. Hence, we plotted separately the $\Delta p\text{Ki}(1,MC_1)$ to $\Delta p\text{Ki}(5,MC_5)$ for the 12 MC1R selective and the two MC4R selective compounds (Figure 4, Panel B and C, respectively).

As seen in Panel A of Figure 5, the first, second and third segments give a major influence to the affinity of the series of compounds. For example, the model predicts that exchange of the first segment in the MC3R with the corresponding in the MC4R would result in an increase in the average affinity of the compounds by 0.15 pKi units. The second and third segments of MC4R also show positive influence. The higher affinity of the compounds for MC1R versus MC3R and MC5R is explained mainly by differences in the third segment. E.g., exchange of the third segment in the MC1R with the corresponding in the MC5R is estimated to reduce the affinity by about 0.2 pKi units. Moreover, as seen from Panels B and C, exchange of the third segment in the MC1R with the corresponding in the MC3R is estimated to reduce the affinity of MC1R selective compounds, whereas it would give an increase in
affinity of the two MC₄R preferring compounds. (In fact, the pattern in Panel C could suggest that the high affinity binding of the latter compounds is accomplished by interactions with those residues in the third segment that are identical or similar in MC₃R and MC₄R).

By contrast the fourth and fifth segments have only minor influence on the average affinity of the compound series. However, analyzing the MC₁R and MC₄R selective compounds separately (Figure 5, Panels B and C) shows that the third and fifth segments are responsible for differentiating the compounds into MC₁R and MC₄R selective ones. As seen from panel C, the affinity of the MC₄R selective compounds for some chimeras is actually expected to be even higher than for the native receptor. This is in agreement with the observed high affinity of compounds 1 and 15 to, e.g., chimera 44355 (see Table 1).

**Contribution of Single Sequence Residues for Ligand Affinity.** As described above, the four \( \Delta pK_i \) values for each sequence segment characterize the changes in ligands affinity once the segment is exchanged between MC₁₃₅Rs. Needless to say, these affinity changes originate from the differences in physico-chemical properties of non conserved residues of the given segment in the MC₁₃₅Rs. In a further analysis of the model we wanted to find out which particular amino acids and amino acid properties that are involved in creating selectivity for the melanocortin receptor ligands. This was done by correlating z-scale descriptors of the non-conserved amino acids located inside of the TM cavity of the MCRs to the \( \Delta pK_i(1,MC_1) \ldots \Delta pK_i(5,MC_5) \) values calculated from the proteochemometric model.

The analysis was performed separately for each of the five receptor segments by applying PLS. Thus, in each PLS model the Y vector (a single column with four rows) comprised the four \( \Delta pK_i \) (segment Nr, MC₁₃₅) values, while the X matrix (5*n columns, four rows) comprised the z-scales of n amino acids in the respective segment of the four (MC₁₃₅) receptors. Since the numerical differences in z-scales reflect physico-chemical differences between amino acids, only centering of descriptors, but not re-scaling to unit variance, was performed prior to the PLS modeling.
The analysis revealed that the $\Delta pKi(1,MC_{1,3-5})s$ for the first segment essentially correlate to the physicochemical properties of amino acid position 41 (Ser, Lys, Ser, Ala in $MC_{1,3-5}$, respectively); the most important being the z-scales 3 and 4 for this amino acid (the results are presented graphically in Figure 6, Panel A; as seen, the third z-scale shows the largest positive, whereas z4 shows the largest negative coefficient). Inspection of z-scale values for particular amino acids suggests that it is the differences in polarity between nucleophilic Ser ($z3=1.15$, $z4=-1.39$) and electrophilic Lys ($z3=-2.49$, $z4=1.49$) that are responsible for the compounds’ preference for $MC_4R$ and $MC_1R$ versus $MC_3R$, and to a lesser extent versus $MC_5R$ (z-scale values for Ala being $z3=0.60$, $z4=-0.14$). Position 41, however, is occupied by the same amino acid in $MC_1$ and $MC_4$, and can thus not explain selectivity between these two receptors. Explanation is instead found by the large negative coefficient for $z1$ of residue 38 (Val, Val, Leu, Met) which assess positively the more hydrophobic Leu ($z1=-4.28$) in this position of the $MC_4R$.

For the second segment (Figure 6, Panel B) several sequence residues seem to jointly explain the differences in selectivity of the compounds for $MC_4R$ versus $MC_3R$ and $MC_5R$. The largest negative coefficient is assigned to $z1$ of sequence position 114 (Gln, Gln, Val, Arg). This z-scale differentiates hydrophobic Val ($z1=-2.59$) in the $MC_4R$ from the hydrophilic Arg ($z1=3.52$) and (to a somewhat lesser extent) Gln ($z1=1.75$). High negative coefficients are also assigned to $z2$ and $z3$ at sequence position 120 (Ile, Phe, Ile, Phe) (i.e., a binary variation) preferring Ile in $MC_1$ and $MC_4$ over Phe in $MC_3$ and $MC_5$Rs. It is noticeable, that also the other residues located at the extracellular end of TM2 (residue 99) and TM3 (116) are assigned higher absolute values of coefficients than the residues located deeper in the TM cavity (positions 83, 91, 124, 128, 129, 132, 136).

For the third segment the affinities are strongly increased if this segment is taken from the $MC_1R$ or $MC_4R$, while they are decreased if it is taken from the $MC_5R$ (vide ante, Figure 5 Panel A). Moreover, the third segment (together with the fifth segment; vide infra) shows major importance for the creation
of the selectivities of the two MC₄R selective compounds. Therefore, the data corresponding to Fig. 5 Panels B and C were interpreted rather than those corresponding to panel A. Accordingly two separate PLS models were created. The results from this analysis are depicted in Fig 6, Panel C. As can be seen several residues, namely 171 (Ala, Cys, Ala, Phe), 175 (Phe, Cys, Ser, Cys) and 187 (Ala, Met, Ala, Tyr), were assessed to be important by both models, although the coefficients for particular z-scales are different or even opposite.

Thus, from the model assessing the binding of MC₄R selective compounds it can be inferred that the Phe in position 171 is affording a negative influence due to the hydrophobicity (z₁= -4.22) and high polarizability (z₂=1.94) of this residue. The third and fifth z-scales of this position, separating Cys from Ala, show less importance. By contrast, from the model for MC₁R selective compounds it can be inferred that the Cys (z₃=3.71, z₅= -2.65) of the MC₃R at this position has a role to diminish the MC₁R selective compounds’ ability to bind. Taken together, Ala is a preferable amino acid at this position for the binding of both MC₁R and MC₄R selective compounds. A corresponding analysis for position 175 shows that MC₄R selective compounds prefer Ser instead of Phe, while MC₁R selective compounds prefer Ser instead of Cys.

In addition to the three above-mentioned residues, a few additional sequence positions appear to have some importance in the third segment, namely for MC₁R selective compounds residues 200 and 184, and for MC₄R selective compounds residue 203.

The fourth segment has only a marginal influence on selectivity. However, the fifth segment gives an opposite effect on the MC₁R and MC₄R selective compounds. The analysis shows that this effect is caused primarily by the amino acid in position 264 (Ile, Ile, Tyr, Met) (Figure 6, Panel D). Thus, according to the analysis the Met (z₄=1.94, z₃=0.47) in the MC₄R has a larger positive impact to the binding of MC₄R selective compounds than the Tyr (z₄=0.04, z₃=0.43) in the MC₄R. The presence of an Ile (z₄= -0.84, z₃= -1.71) at this position is estimated to diminish the affinity of the MC₄R selective compounds, while it increases the affinity of all the other compounds. Coefficients with opposite signs
are also assigned to z1 and z2 of residue 285(Ala, Val, Ile, Ile) in the two models. For residues 265 (Val, Ile, Ile, Leu) and 292 (Ile, Val, Ile, Val) (i.e., positions occupied by physico-chemically similar aliphatic amino acid) the signs are also opposite, but the coefficients of the z-scales are rather small, making it unlikely that the latter amino acids have any larger impact on the compounds’ selectivity.

**Three-Dimensional Model of the Binding Pocket in Melanocortin Receptors.** A three-dimensional model of the MC1R is shown in Figure 7 (see methods for details on 3D modeling), with the described above residues being important for receptor subtype selectivity marked in color. As seen, most of the marked amino acids are located close to the extracellular border of the receptor transmembrane cavity, but residues 171, 175, 200 and 203 form a cluster located deeper inside the transmembrane bundle. Moreover, two other clusters of significant residues can be recognized, suggesting binding regions of functional groups of organic compounds. One of these regions is outlined by residues 38, 41, 114, 120 and 285, and the other by residues 184, 187 and 264. Thus, although the proteochemometric model cannot assess the involvement of non-varying residues to the ligand binding, it reveals amino acids with dominant role for creation of ligand selectivity, thereby tentatively mapping out the recognition site for the organic compounds in the MCRs.
Discussion

We recently applied proteochemometrics to analyze the binding of natural melanocortin receptor ligands, namely the α-MSH peptide and some of its synthetic analogues to chimeric MC1/MC3 receptors (Prusis et al., 2001, 2002). However, these studies could exploit only a binary description of the receptors as they were chimeras of two receptors, MC1/MC3, and therefore these studies allowed only a rough mapping of the ligand binding site for the melanocortin peptides. In yet another study we applied proteochemometrics to study the interactions of a series of 54 organic compounds to native melanocortin receptors (Lapinsh et al., 2003). In this study we could separately reveal chemical properties of the organic compounds that are important for their affinity and receptor subtype selectivity. However, as this study did not include any systematic variations of the receptor sequences, beyond that which is present in the wild type receptors, it was not possible to reveal any information about the receptor properties that are responsible for creation of affinity and of those that are involved in discriminating selective from non-selective compounds.

Accordingly, in the present study we evaluated a larger number of organic compounds on 4 native and 13 multiple chimeric melanocortin receptors, where the latter had been created by use of statistical design methods; the purpose being to represent all possible combinations of 3 parts in 4 receptors as good as possible in a minimal set. The binding data was then analyzed with the major aim of mapping out the receptors binding pocket for the organic compounds. Applying proteochemometrics onto the data produced a model with high predictive ability. The standard deviation of errors of prediction corresponding to the obtained $Q^2 = 0.79$ of the model is 0.32 pKi units. In view of the intrinsic statistical error of the biological measurements the modeling accuracy is thus very high.

Alongside with the conventional $Q^2$ parameter, we introduced two additional estimates of model predictive ability, namely the $Q^2_{rec}$ and $Q^2_{lig}$. This seemed rational since one purpose of a proteochemometric models is to make predictions for novel ligands and/or receptors, rather than to
merely predict affinity of untested combinations of moieties already present in the data set (i.e., filling gaps in a data table). The high values of $Q^2_{\text{rec}}=0.76$ and $Q^2_{\text{lig}}=0.68$, as well as the ability of proteochemometrics to perform accurate predictions for the four compounds not included during data pre-processing (i.e., centering of variables prior to calculation of cross-terms) and model creation (i.e., adjustment of block-scaling weights), affirms a high reliability of the proteochemometric modeling approach. Thus, these results indicate the usefulness of proteochemometrics for *a priori* drug design.

Although the present model was based on binary descriptors of receptors, we succeeded in a further analysis to reveal particular sequence residues that are the ones most likely contributing to ligand selectivity. For this purpose 3-dimensional models of the receptors were first created and properties of non-conserved residues that could form ligand binding pocket(s) were characterized by physico-chemical descriptors. By PLS analysis the values of these descriptors in the MC1,3,5Rs were correlated to the contributions of binary descriptors ($\Delta pK_i$(segment,MC1,3,5)) for each of the five receptor segments. As shown in Figure 7, twelve residues were found that potentially influence ligand selectivity. An interesting finding of these investigations was that for the MC1R selective compounds only residues from the third sequence segment showed up to be of large importance. Three-dimensional modeling further indicates that only the two clusters shown on the left side in Figure 7 are important for these compounds. By contrast, for the two MC4R selective compounds also the region between residues 38, 41, 114, 120 and 285 appears important, suggesting a more complex binding mechanism for these compounds.

Several studies were previously performed using site directed mutagenesis in attempts to identify determinants for the MC3R selectivity of melanocortin peptides. Thus, Nickolls et al., 2003, elucidated the effects of point mutations in MC4R on the binding of eleven natural and synthetic peptides. It was found that an Ile125Phe mutation (corresponding to residue 120 (Ile, Phe, Ile, Phe) in MC1,3,5; in our model Ile is preferred to Phe) results in a 2- to 5-fold decrease in the affinity of MC4 selective ligands,
while the affinities of α-MSH and NDP-MSH are not significantly affected. The naturally occurring mutation of Ile137 (here corresponding to 132 (Leu, Ile, Ile, Met)) to Thr significantly decreased the binding of most ligands. Since Thr shows higher values of z1 to z4 and lower of z5 scale compared to Ile, a drop of affinity by such a mutation is in agreement with our results. In a study by Haskell-Luevano et al., 2001) a number of mutations in the MC₄R were evaluated. It was found that mutation of a Ser to Phe [herein position 175 (Phe, Cys, Ser, Cys); according to our model, MC₄R selective compounds prefer Ser instead of Phe] resulted in a 4 to 6-fold decrease in the affinity of two cyclic peptides, while the affinity of NDP-MSH remained unchanged. Mutation of Met to Phe [herein position 195 (Phe, Met, Met, Met); unimportant according to our model] was reported to have no effect to agonist binding or potency. However, the mutation resulted in a constitutively active receptor. Mutation of the same residue to Ala in the study by Yang et al., 2000) was reported to decrease affinity of α-MSH, but not that of NDP-MSH.

Interpretation of mutagenesis data is straightforward when the change is one-dimensional and the effect it causes is a simple “additive” one (e.g., involving only one alteration and/or causing only a direct effect). When changes cause many simultaneous effects (e.g. by multiple interactions with the ligand and/or inside the receptor) the relations of the changes in activity to the changes in structure may become difficult or even impossible to reveal from just a few scattered observations. Using instead a set of mutated proteins, which are designed to cover as much as possible of a selected region of structural variation, in conjunction with mathematical multivariate analysis, as applied herein, constitutes then a solution. The data of the present study indicate indeed that the proteochemometrics modeling applied on data derived from the interactions of organic molecules with statistically designed multiple chimeric proteins is useful to map ligand recognition. Moreover, the models are quantitative and reveal the underlying chemical properties that determine ligand recognition, which is information that is highly desired in ligand design. As the proteochemometrics approach is general it could be applied to analyze the molecular recognition processes of any set of proteins.
References


Footnotes

Financial support was obtained from the Swedish VR (04X-045959 and 621-2002-4711).

Address correspondence to: Jarl Wikberg, Pharmaceutical Pharmacology, Box 591, BMC, SE751 24 Uppsala, Sweden. Phone: +46-18-471 42 38. Fax +46-18-55 97 18. Email: Jarl.Wikberg@farmbio.uu.se
Legends for Figures

Figure 1. Schematic representation of the 4 native and 13 chimeric melanocortin receptor variants used.

Figure 2. Structures of organic compounds used herein.

Figure 3. Correlation of calculated and predicted pKi versus measured pKi values derived by the PLS modeling of receptor-ligand interactions.

Figure 4. PLS coefficients of the descriptors of organic compounds derived from the final proteochemometric model. Abbreviations correspond to descriptors as follows: AMW: average molecular weight; Sv: sum of atomic van der Waals volumes (scaled on carbon atom); Mv: mean atomic van der Waals volume (scaled on carbon atom); Ms: mean electrotopological state; nCIR: number of circuits; RBN: number of rotatable bonds; RBF: rotatable bond fraction; nAB: number of aromatic bonds; nN: number of nitrogen atoms; nO: number of oxygen atoms; nCL: number of chlorine atoms; nX: number of halogen atoms; nR06: number of 6-membered rings; nR09: number of 9-membered rings; nR10: number of 10-membered rings; nCs: number of total secondary C(sp3); nCt: number of total tertiary C(sp3); nCrH2: number of ring secondary C(sp3); nCrHR: number of ring tertiary C(sp3); nCaH: number of unsubstituted aromatic C(sp2); nCaR: number of substituted aromatic C(sp2); nCONHR: number of secondary amides (aliphatic); nCONR2: number of tertiary amides (aliphatic); nNH2: number of primary amines (aliphatic); nNHR: number of secondary amines (aliphatic); nNR2: number of tertiary amines (aliphatic); nOHPh: number of phenols; nPhX: number of X-C on aromatic ring; Ui: unsaturation index; ARR: aromatic ratio; MLOGP: Moriguchi octanol-water partition coeff. (logP).

Figure 5. Contribution of receptor segments for binding of organic compounds derived from the proteochemometric model. Panel A shows $\Delta pK_i(1-5,MC_{1,3-5})$ for all fourteen compounds, while in
Panels B and C the $\Delta pK_i(1-5,MC_{1,3.5})$ is plotted separately for the twelve MC$_1$ and the two MC$_4$R selective compounds. (See text for further details).

**Figure 6.** Regression coefficients of PLS models correlating the physico-chemical properties of varying sequence residues to the $\Delta pK_i$(segment Nr, MC$_{1,3.5}$) parameters computed from the proteochemometric model. (For details see text).

**Figure 7.** Three-dimensional model of the MC$_1$ receptor with amino acids computed from the proteochemometric model to be important for creation of selective binding of organic amines. Top panel shows the transmembrane domains viewed from the extracellular side. Indicated are the side chains of the sequence residues found to be important for selective binding of organic compounds (also discussed in detail in the text). Bottom panel shows a side view of transmembrane domains, the extracellular portion facing up and the intracellular down, with the same residues indicated.
Table 1

<table>
<thead>
<tr>
<th>Comp. Nr</th>
<th>Receptor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Work set</td>
</tr>
<tr>
<td></td>
<td>MC1</td>
</tr>
<tr>
<td>1</td>
<td>6.1</td>
</tr>
<tr>
<td>2</td>
<td>5.8</td>
</tr>
<tr>
<td>3</td>
<td>4.9</td>
</tr>
<tr>
<td>4</td>
<td>5.3</td>
</tr>
<tr>
<td>5</td>
<td>5.4</td>
</tr>
<tr>
<td>6</td>
<td>5.2</td>
</tr>
<tr>
<td>7</td>
<td>5.2</td>
</tr>
<tr>
<td>8</td>
<td>5.2</td>
</tr>
<tr>
<td>9</td>
<td>5.4</td>
</tr>
<tr>
<td>10</td>
<td>5.1</td>
</tr>
<tr>
<td>11</td>
<td>4.6</td>
</tr>
<tr>
<td>12</td>
<td>4.4</td>
</tr>
<tr>
<td>13</td>
<td>4.7</td>
</tr>
<tr>
<td>14</td>
<td>3.9</td>
</tr>
<tr>
<td>15</td>
<td>6.5</td>
</tr>
<tr>
<td>16</td>
<td>5.4</td>
</tr>
<tr>
<td>17</td>
<td>5.1</td>
</tr>
<tr>
<td>18</td>
<td>5.5</td>
</tr>
</tbody>
</table>

**Table 1.** Binding affinities (pKi) of organic compounds for 4 native and 13 multiple chimeric melanocortin receptors (pKi values represent the mean of at least three separate determinations). The number coding of chimeric receptors represents the presence of the corresponding sequence of the MC₁, MC₃, MC₄, or MC₃R in each of the five varying sequence stretches.
**Table 2.** Results of the PLS modeling after extraction of different number (A=1 to 7) of PLS components. Shown are the explained X and Y variance ($R^2_X$ and $R^2_Y$) and predictive ability ($Q^2$) of the models. $Q^2$ was assessed by cross validation with five randomly formed groups, repeated 100 times; $Q^2_{lig}$ and $Q^2_{rec}$ were assessed by cross validation leaving out one ligand or receptor at a time, respectively; $Q^2_{ext}$ was calculated from the predictions for the four compounds of the test set.

<table>
<thead>
<tr>
<th>A</th>
<th>$R^2_X$</th>
<th>$R^2_Y$</th>
<th>$Q^2$</th>
<th>$Q^2_{lig}$</th>
<th>$Q^2_{rec}$</th>
<th>$Q^2_{ext}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.15</td>
<td>0.57</td>
<td>0.54</td>
<td>0.42</td>
<td>0.52</td>
<td>0.31</td>
</tr>
<tr>
<td>2</td>
<td>0.23</td>
<td>0.70</td>
<td>0.64</td>
<td>0.48</td>
<td>0.63</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>0.27</td>
<td>0.84</td>
<td>0.75</td>
<td>0.36</td>
<td>0.71</td>
<td>0.64</td>
</tr>
<tr>
<td>4</td>
<td>0.34</td>
<td>0.88</td>
<td>0.78</td>
<td>0.41</td>
<td>0.74</td>
<td>0.66</td>
</tr>
<tr>
<td>5</td>
<td>0.42</td>
<td>0.89</td>
<td>0.79</td>
<td>0.57</td>
<td>0.75</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>0.48</td>
<td>0.91</td>
<td>0.79</td>
<td>0.65</td>
<td>0.76</td>
<td>0.71</td>
</tr>
<tr>
<td>7</td>
<td>0.52</td>
<td>0.92</td>
<td>0.79</td>
<td>0.68</td>
<td>0.76</td>
<td>0.73</td>
</tr>
</tbody>
</table>
Fig. 1
Fig. 2

Workset

1 

2 

3 

4 

5 

6 

7 

8 

9 

10 

11 

12 

13 

14 

Test set

15 

16 

17 

18
Fig. 3

This article has not been copyedited and formatted. The final version may differ from this version.
Fig. 4
Fig. 5

A

Average

0.2

-0.2

0.0

0.2

MC1 selective compounds

B

0.2

-0.2

0.0

0.2

MC4 selective compounds

C

0.4

0.2

-0.2

MC1  MC4

MC3  MC5

Segment:  1  2  3  4  5
Fig. 6

A  First segment

B  Second segment

C  Third segment
MC1 selective compounds

MC4 selective compounds

D  Fifth segment
MC1 selective compounds

MC4 selective compounds