Articles

QSAR and Proteo-chemometric Analysis of the Interaction of a Series of Organic Compounds with Melanocortin Receptor Subtypes

Maris Lapinsh, Peteris Prusis, Ilze Mutule, Felikss Mutulis, and Jarl E. S. Wikberg*

Department of Pharmaceutical Biosciences, Uppsala University, Box 591 BMC, SE751 24 Uppsala, Sweden

Received May 27, 2002

We have created quantitative structure−activity relationship (QSAR) models describing the interaction of a series of 54 organic compounds with four melanocortin (MC) receptor subtypes, MC_1, MC_3, MC_4, and MC_5. In addition to traditional QSAR analysis, we applied our recently developed proteo-chemometrics approach. Proteo-chemometrics is based on the combined analysis of series of receptors and ligands, wherein descriptions of ligands, proteins, and so-called ligand−protein cross-terms are correlated with interaction activities. The compounds were characterized by structural descriptors, including three-dimensional grid-independent descriptors (GRINDs), topological descriptors, and geometrical descriptors. Description of receptors was obtained by computing the receptors' amino acid sequence identities. Both the QSAR and proteo-chemometrics approaches resulted in models with essentially the same statistical significance: the cross-validated correlation coefficient q^2 for the proteo-chemometric model being 0.71, while for the QSAR models the q^2s were 0.75, 0.68, 0.63, and 0.71 for the MC_1, MC_{3-5} receptor, respectively. However, the proteo-chemometrics modeling provided more detailed information about receptor−ligand interactions and determinants for receptor subtype selectivity than did QSAR.

Introduction

Melanocortin (MC) receptors are members of the G-protein-coupled receptor (GPCR) superfamily. Five different subtypes of the MC receptors, MC_{1-5}, have been cloned to date in mammals.[1] Physiological roles for melanocortin receptors include pigment regulation (MC_1), adrenal gland control (MC_2), regulation of sexual and feeding behaviors (MC_3 and MC_4), exocrine gland control (MC_5), and regulation of the immune system (MC_1 and MC_3).[1−3] The natural ligands for melanocortin receptors include melanocyte stimulating hormones (MSH), adrenocorticotropic hormone (ACTH), and agouti and agouti-related peptide, all of which are peptides with various chain lengths. These natural ligands show different MC receptor selectivity. For example, while the MC_2 receptors bind only ACTH,[4] the MC_1 and MC_{3-5} receptors interact with both ACTH and MSHs.[1−3]

The potential of using the MC receptors as targets for novel drugs[5] has prompted the need of compounds with high specificity for particular MC receptor subtypes. Unfortunately, natural melanocortin peptides do not show unique MC receptor subtype selectivity, and they are not suited for therapeutic applications due to their large molecular weight and susceptibility to enzymatic degradation. Large efforts are currently concentrated on the design of organic compounds showing high selectivity and affinity for MC receptor subtypes. However, quite a few successes have so far been reported.[5]

In two preceding papers we reported a series of 55 *N*-alkylamino acids and other organic compounds, some of which showed sub-micromolar affinities for the four MSH peptide binding MC receptors (i.e., MC_1, MC_{3-5}).[6,7] Moreover, recently we developed proteo-chemometrics, which is a new technology suited for analysis of drug−receptor interactions.[8,9] Contrary to the traditional QSAR approaches that aim to correlate description of ligands with affinity toward one particular target receptor, proteo-chemometrics considers many targets and ligands simultaneously. Thus, proteo-chemometrics take advantage of the properties of both the receptors and the ligands, allowing one to explain ligand−receptor binding and selectivity, and analyze the ligands' interactions with sets of receptors.

Our previous proteo-chemometrics studies were performed on chimeric melanocortin, α_1-adrenergic receptors,[8−10] and on a wide set of different G-protein-coupled receptors for amines.[11] In all cases we succeeded in creating very robust models that gave insights into the ligand−receptor interactions. Moreover, these models were capable to delineate determinants of importance for creation of the ligands' affinities for series of related receptors and those of importance for creation of the ligands' selectivites.

The goal of present study was to develop predictive structure−activity models for small molecular weight compounds active on melanocortin receptors. To this end we compared models that used (1) uniresponse QSAR (one response variable, i.e., separate analysis of binding to each receptor), (2) multiresponse QSAR (several

* To whom correspondence should be addressed. Tel. +46-18-471 42 38, Fax +46-18-55 97 18. E-mail: Jarl.Wikberg@farmbio.uu.se.

10.1021/jm020945m CCC: $25.00
Published on Web 05/23/2003

response variables), and (3) proteo-chemometrics approaches. Despite the fact that the present data set included only four receptors (contrasting our previous proteo-chemometrics studies that included large sets of receptors), the best modeling approach turned out to be proteo-chemometrics. We therefore used it for interpreting properties, required for ligand recognition of each of the four MC receptor subtypes.

Results and Discussion

Data Set. K_i values for a series of 55 organic compounds for the MC_1, MC_3, MC_4, and MC_5 receptor subtypes were obtained from two of our preceding papers.[6,7] The K_is had been estimated using radioligand binding on recombinant human receptors. The substances contained moieties that mimicked amino acids of the core sequence of melanocortin peptides[12] (Phe-Arg-Trp). Thus, all compounds included at least two of the benzene/naphthalene, amino/guanidine, and indole groups. Some of the compounds contained, in addition, imidazole, dimethoxyphenyl, chlorodimethoxyphenyl, methylindole, or other moieties. Affinities (expressed as the negative logarithm of the K_i values, pK_is) for the four MC receptors covered a range of more than three logarithmic units; the most active compounds showing sub-micromolar K_is. In some cases no binding had been observed up to a 0.5 mM concentration of the compound, or the K_i values had been reported to be >1 mM. In these cases we assigned p$K_i = 3$ (i.e., the K_i being set to 1 mM) for the sake of the modeling. The structures and pK_is of the most active compound of the series are shown in Figure 1.

Description of Organic Compounds. 3D structures of compounds were created and characterized by three descriptor blocks, namely: (1) GRid INdependent Descriptors (GRINDs)[13] calculated by Almond 2.0,[14] (2) topological, and (3) geometrical descriptors calculated by Dragon 1.11.[15]

GRINDs are novel 3D descriptors that characterize the ability of a molecule to form energetically favorable interactions with selected pharmacophore pairs. A major advantage of GRINDs is that they do not require spatial alignment of compounds.

Besides the GRINDs, we computed two additional blocks of descriptor variables that represented, respectively, topological and geometrical descriptors. The topological descriptor block comprised 68 conformationally independent 2D descriptors related to topology, connectivity, atomic composition, molecular size, and shape. The geometrical descriptor block comprised 18 descriptors that depended on molecular geometry (see Todeschini and Consonni[16]).

Removal of Correlated Descriptors of Organic Compounds. Correlation coefficients for all pairs of descriptor variables for the organic compounds were evaluated in order to identify highly correlated descriptors, i.e., to detect redundancy in the data set. Such redundancy might lead to overexploitation of a chemical property in the explanation of the dependent variable. Hence, the removal of some highly correlated variables, or scaling them down relative to the other variables, might be helpful in development of a model.[17,18] Inspection of the descriptor data identified strong correlation among a large set of topological (and a few geometrical)

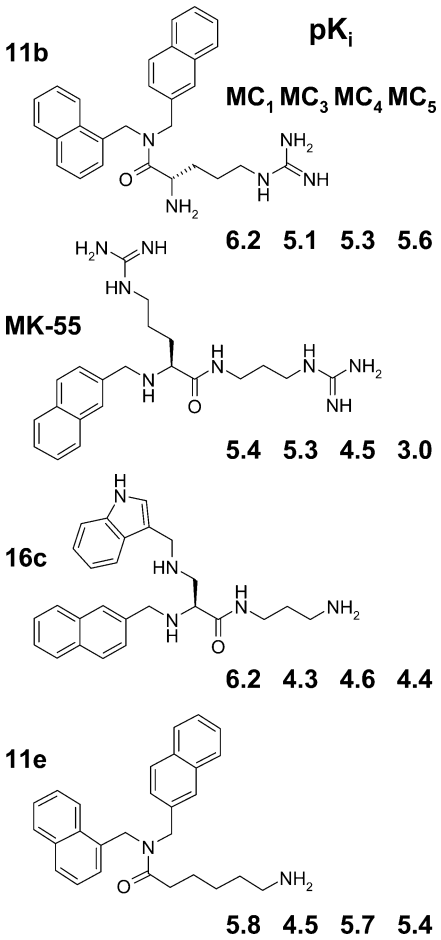

Figure 1. Structures and binding affinities of the most active compounds for each MC receptor subtype. The highest binding affinities were shown for the MC_1 receptor by compounds **11b**[6] and **16c**;[6] **11b** was the most active compound of the series on the MC_5 receptor. MK-55[7] was the highest affinity binder to the MC_3 receptor, while **11e**[6] was the highest affinity binder to the MC_4 receptor.

descriptors. For 41 of these descriptors, the squared correlation coefficients (r^2) to ISIZ (molecular size index) were over 0.8. This thus indicates that a large portion of these topological descriptors merely describes the size of a molecule while quite little information is present on other properties. A similar examination of GRINDs showed a presence of strong correlation of descriptors representing the same molecular interaction at neighboring distance ranges. These findings prompted us to apply block scaling prior to any further analysis (see Experimental Section for details).

Description of Receptors. Proteo-chemometrics is based on the use of a single affinity variable, although the binding to several targets is studied. This is possible by utilizing not only descriptors characterizing the ligands, but also by using a physicochemical description of receptors, and by deriving ligand−receptor cross-terms. In the present study, four receptor subtypes had been studied: MC_1, MC_{3-5}; the sequence identities of their transmembrane regions are shown in Table 1. (The location of the TM regions was as in ref 19). The small number of receptors of the present study, when compared to our earlier studies, suggested that little benefit could be gained by using an extensive set of descriptors based on the receptor sequence physicochemical properties. We here therefore elected to describe the MC

Table 1. Sequence Identity of the Transmembrane Regions of Human Melanocortin Receptor Subtypes (location of TM regions is as in ref 19)

	MC_1	MC_3	MC_4	MC_5
MC_1	1	0.56	0.57	0.53
MC_3	0.56	1	0.68	0.72
MC_4	0.57	0.68	1	0.72
MC_5	0.53	0.72	0.72	1

receptors with only four variables based on the receptor sequence identity (SI) of the four MC receptor subtypes (Table 1).

Ligand−Receptor Cross-Terms. Ligand−receptor recognition depends on the complimentarity of the properties of two interacting moieties. This is something that cannot be described by a linear combination of ligand and receptor descriptors. In proteo-chemometrics the nonlinearity is accounted for by computing ligand−receptor cross-terms.[8−11] Accordingly, we here obtained cross-terms by multiplying mean centered descriptors of organic compounds and receptors, giving one additional descriptor block comprising 4 × 343= 1372 variables.

Conventional QSAR Modeling. Descriptors were correlated to the binding data using partial least-squares projection to latent structures (PLS).[17] PLS allows a simultaneous analysis of several *Y*s in a single model ('multiresponse model'), as well as creation of separate models for each of the *Y*s. (In our case, *Y* variables corresponded to pK_is of compounds' binding to MC_1, MC_3, MC_4, and MC_5 receptors). In the case *Y*s are strongly correlated (squared correlation coefficient R^2 being >0.5), multiresponse modeling may result in a more stable model.[18,20] On the other hand, independent modeling of each *Y* might allow one to explain better the uniqueness of each response, and it allows more flexibility in the model building. In the current data set, the r^2 values for the correlation of the binding affinity (pK_i) variables for the four MC receptor subtypes ranged from 0.45 (MC_1 to MC_5) to 0.69 (MC_1 to MC_3). We therefore created models using both approaches. Simultaneous PLS modeling using four *Y*s resulted in a five component model with the fraction of explained variation of *X* and *Y* (r^2X and r^2Y) of 0.79 and 0.78, respectively. The predictive ability was assessed by cross-validation,[21,22] the fraction of the predicted *Y*-variation (q^2) being 0.60. For the sake of information, r^2X, r^2Y, and q^2 may vary between 0 and 1, but negative q^2 values can also be encountered, indicating nonpredictive models.[20] In QSAR modeling comprising biological data it is generally considered acceptable if q^2 is higher than 0.4.[23] For the sake of model interpretability the margin $r^2Y - q^2$ should be as small as possible, preferably not exceeding 0.2.[20]

r^2Y and q^2 values for particular *Y*s are shown in Table 2 alongside with results of separate modeling of binding to each receptor. When separate models were built, four-dimensional models were obtained in all cases, with r^2Xs ranging 0.756 to 0.764. As can be seen from Table 2, similar q^2 values were obtained for the multiresponse model and the uniresponse models, with the exception for the MC_5 receptor, where uniresponse modeling seemed to produce a more predictive model.

To improve models, we applied variable selection by removing variables with unstable regression coefficients

Table 2. Goodness of Fit (r^2Y) and Predictive Ability (q^2) of QSAR Models

	multiresponse model				uniresponse models			
	r^2Y	q^2	r^2Y [a]	q^2 [a]	r^2Y	q^2	r^2Y [a]	q^2 [a]
MC_1	0.83	0.66	0.84	0.75	0.83	0.67	0.85	0.75
MC_3	0.78	0.58	0.76	0.62	0.81	0.61	0.81	0.68
MC_4	0.73	0.52	0.70	0.55	0.78	0.56	0.80	0.63
MC_5	0.74	0.56	0.72	0.57	0.85	0.65	0.83	0.71

[a] Obtained after variable selection.

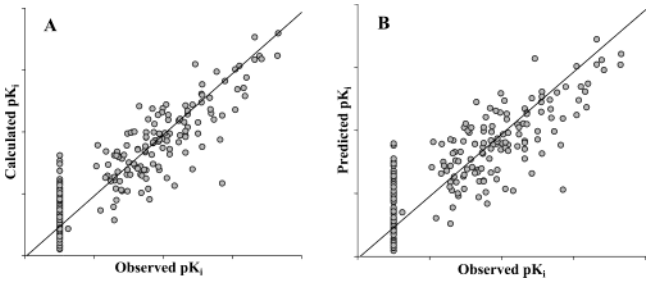

Figure 2. Correlation of calculated versus observed pK_i values (panel A; $r^2Y = 0.80$) and predicted versus observed pK_i values (panel B; $q^2 = 0.71$) derived from the proteo-chemometric model.

(see Experimental Section for details). Comparison of q^2 values after variable selection shows an advantage for uniresponse modeling, with an exception of MC_1, where both approaches performed equally well.

Models were also validated by response permutations, as described.[22] Validation gave negative q^2 intercepts for all cases. These results thus indicate that predictive models could not be obtained with the data randomized and give further validity to the models and the selected descriptors of compounds.

Proteo-chemometric Modeling. Proteo-chemometric models were built by PLS modeling using the descriptor blocks for compounds, and the receptor descriptor and receptor-compound cross descriptor blocks. The initial model described $r^2Y = 0.76$ of the variance in *Y* with a predictive capacity $q^2 = 0.66$. After variable selection the improved models had $r^2Y = 0.80$ and $q^2 = 0.75$. The final model was based on 230 X variables, from which 22 were topological, 12 geometrical, 117 GRIND, 77 ligand−receptor cross-descriptors, as well as SI_{MC1} and SI_{MC3}.

Cross-validation of the final model was also performed so that all four observations of each ligand were included in the same cross-validation group. Thus, when the affinity of a compound was predicted for a particular receptor, no information was available for the binding of the compound to any other of the MC receptors. Despite this the cross-validation gave the very high q^2 value 0.71. The small difference between the thus estimated $q^2 = 0.71$ and the $r^2Y = 0.80$ of the model is noteworthy (cf. Figure 2 panels A and B).

Comparison of Proteo-chemometric and Conventional QSAR Models. The proteo-chemometric and conventional QSAR models were compared by calculating the standard deviation of error of calculation (SDEC) and standard deviation of error of prediction (SDEP).[24] By contrast to r^2Y and q^2, which are measures relative to the initial sums of squares, SDEC and SDEP are expressed in pK_i units and can therefore be used to compare the accuracy of the different modeling ap-

Table 3. Comparison of Standard Deviation of Errors of Calculation (SDEC) and Standard Deviation of Errors of Prediction (SDEP) of QSAR and Proteo-chemometric Models

	QSAR models		proteo-chemometric model		
	SDEC	SDEP	SDEC	SDEP	SDEP[a]
MC_1	0.44	0.57	0.49	0.53	0.56
MC_3	0.31	0.41	0.36	0.40	0.44
MC_4	0.36	0.48	0.38	0.42	0.45
MC_5	0.33	0.43	0.35	0.40	0.44

[a] Rearranged cross-validation groups.

proaches. Comparisons of proteo-chemometric and traditional QSAR models are shown in Table 3. As can be seen from the table, both the conventional and proteo-chemometric approaches resulted in models with good predictive ability. The slightly lower SDECs for QSAR models, and the larger margin between SDEC and SDEP compared to the proteo-chemometric model, may indicate that more of the variance was explained by chance correlations in the traditional QSAR. The proteo-chemometric model is therefore more reliable, e.g., for interpretations. Moreover, the proteo-chemometric model was useful for explaining the selectivity of the compounds, which proved to be based on a relatively small number of cross-descriptors (see below), while in the QSAR models the affinity differences became accumulated from all descriptors. Accordingly the proteo-chemometric model was straightforward to interpret in terms of selectivity, compared to the QSAR model, where distinct factors responsible for selectivity were difficult to isolate.

Interpretation of the Proteo-chemometric Model. The PLS coefficients of proteo-chemometric model were used to assess the properties of the compounds, which were important for their affinity to each studied MC receptor subtype. We performed this analysis separately for GRIND and topological/geometrical descriptors.

Interpretation of GRINDs. In Figure 3, panel A, the PLS coefficients of the GRIND descriptors are plotted versus the distance between the molecular interaction field (MIF) nodes. As seen, each of the six types of MIF probe-pairs are plotted separately. Panels B−E show the contribution for each MC receptor accumulated from GRINDs and their cross-terms with SI descriptors. Shown is the change in calculated pK_i value for the particular receptor if the value of GRIND descriptor of a compound is increased by 1 unit. In Figure 3 positive values represent properties that contribute favorably for creating the compounds' affinities, whereas negative ones identify properties having a negative influence.

In Figure 3 the PLS coefficients that correspond to interactions over small distances describe the importance of the presence or absence of particular pharmacophores. As can be seen from Figure 3A, hydrophobic groups are required for high affinity binding. This was found to be due to extremely high values for short distance DRY−DRY descriptors in molecules containing one or several naphthalene moieties. By contrast, groups favorably interacting with the N1 field seem rather to have a negative influence. High values of N1−N1 descriptors could be attributed to the presence of an amide group, while low values were due to the lack of H-bond acceptors in a molecule. Thus, according to the model, the presence of an amide group is not required for binding to MC receptors.

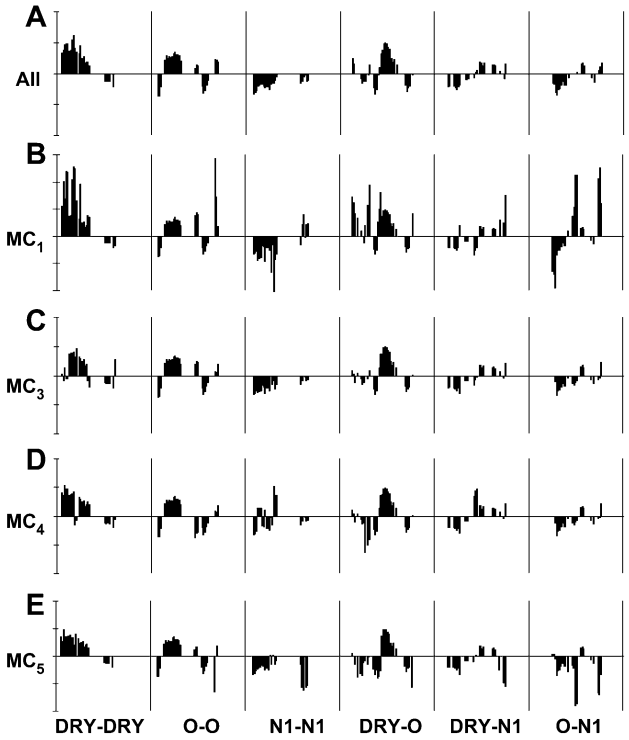

Figure 3. Contribution of GRIND descriptors for explaining ligand−receptor affinity in the proteo-chemometric model for the interaction of 54 organic compounds with melanocortin receptors MC_1, MC_{3-5}. Panel A represent the PLS regression coefficients of GRIND descriptors. Panels B−E show the sum of the PLS coefficients of GRIND descriptors and their cross-terms with SI descriptors multiplied with the actual values of the SI descriptor for the indicated MC receptor. The vertical separators between MIF pairs represent a distance of between 0 and 24.8 Å. Increments on the *Y*-axes represent the change of affinity by 0.02 pK_i units, when descriptor variables are changed by one standard deviation (see text for further details).

Coefficients for descriptors derived from MIFs separated by a larger distance were useful to judge the spatial arrangement of pharmacophores. Very large positive values were seen for DRY-DRY fields at distances <9 Å, for O−O fields at a distance from 5 to 8 Å, and for DRY−O fields at a distance from 10 to 13 Å. Further clues to the interpretation of the importance of these fields could be obtained by tracing the GRINDs to the MIFs surrounding particular molecules. For example, shown in Figure 4 are two compounds **11b**[6] and **16c**[6] that show sub-micromolar affinities for the MC_1 receptor. The high affinity of **11b** can mainly be explained by hydrophobic interactions, while the high affinity of **16c** is to a much larger extent explained by the two major clusters of descriptors in the O−O and DRY−O correlograms that show high positive coefficients.

Further inspection of compound **11b** (Figure 4) reveals that interactions corresponding to ranges of positive coefficients in the DRY−DRY coefficient plots represents the hydrophobic fields around its two naphthalene groups, which according to the model contributes positively to affinity. Interactions at distances over 14−16 Å in **11b** involve a weaker field next to the guanidine. For compound **16c** DRY−DRY interactions

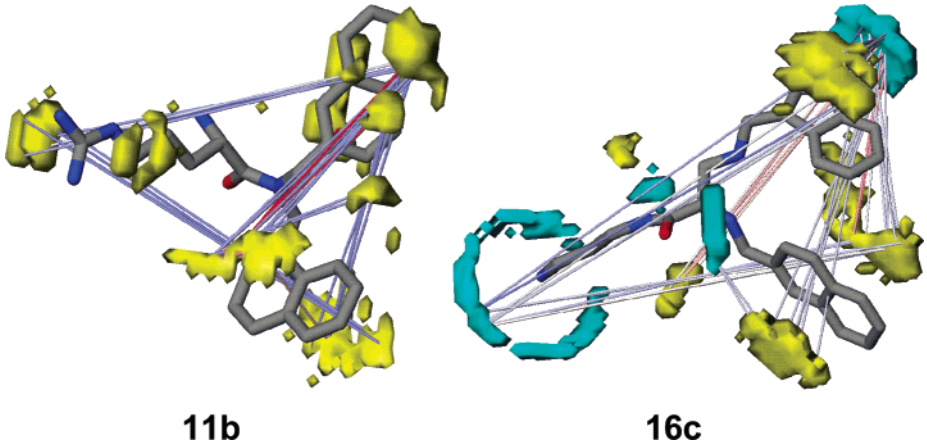

Figure 4. DRY molecular interaction field of compound **11b**, and DRY (yellow) and O (green) molecular interaction fields of compound **16c**. Lines connect the node couples used for calculation of GRINDs.

over long distances are completely absent, while they are present in several other compounds with moderate affinity that share distantly situated aromatic moieties. From these observations it can be concluded that a too distant location of aromatic rings does not contribute positively to affinity.

In the current data set higher than average values of the shortest distance O−O field descriptors were due to the presence of an indole or guanidine group. For several compounds interactions of the indole with the O probe is augmented by spatially closely situated amide nitrogen. According to the model, such a combination shows some negative influence. Moreover, guanidine is lacking in several active compounds, while it is present in all nonbinders in the data set. Instead, a positive effect on affinity is indicated by the interaction of a molecule with O nodes situated 5−8 Å from each other. For **16c** this corresponds to the O field next to the indole together with the O field shown in front of the molecule (Figure 4). A very positive effect to the affinity of **16c** is also related to the O field next to the indole group that is linked with the hydrophobic field below the naphthalene group. (This field is also important for MC_1 selectivity of the compound). Similarly, a link between the O and N1 fields next to the indole group, with the O field around the very distantly located (16−18 Å) amino function appears to be important for both the affinity and MC_1 receptor selectivity of **16c**.

Careful comparisons of the coefficient plots for particular receptors (panels B−E of Figure 3) reveal clear determinants for MC_1/MC_5 selectivity. Thus, MC_1/MC_5 selectivity is associated with large MIF values at the distantly situated O and N1 probe node couples (one of them, as a rule, being located next to the indole). Moreover, MC_1/MC_3 selectivity is mainly explained by the DRY−DRY field descriptors at distances <4 Å. Thus, the mere presence of a naphthalene group is not sufficient for creating affinity for the MC_3 receptor, although it strongly contributes to MC_1 receptor binding. A fairly large number of descriptors explain binding selectivity for the MC_4 receptor. For example, interactions with the N1 probe give no negative influence on binding, as it does for binding to the MC_1 receptor, whereas a negative factor is a close location of hydrophobic and H-bond donor groups (interaction with DRY−O field nodes at a distance <8 Å).

Interpretation of Topological/Geometrical and SI Descriptors. PLS coefficients for the geometrical and topological descriptors of the proteo-chemometric model are shown in Figure 5. The largest negative coefficients were computed for such geometrical descriptors as molecular eccentricity and electrotopological variation, while the largest positive coefficients occurred for E-state topological parameter, mass weighted radius of gyration and span (these latter two are highly correlated parameters). For topological descriptors, the largest negative coefficient was obtained for Randic shape, a parameter that characterizes the size of the molecule and its degree of skeletal branching. Thus, small and unbranched molecules are predicted to show lower affinity. A large negative coefficient was also assigned to average atomic composition, a parameter related to molecular complexity in terms of atom types. Thus, a high fraction of heteroatoms in a molecule is unfavorable. Instead, benzene likeliness of a molecule (BLI index; Figure 5) is a favorable property.

The majority of topological descriptors with moderate positive coefficients, however, contribute jointly in the model, as indicated by their clustering in the loading plots (data not shown). As expected, these parameters (which relate to molecular size) show some positive correlation with affinity.

As can be seen from Figure 5, only a few significant cross-descriptor terms involving topological/geometrical descriptors were retained in the model. This can actually be an advantage as these descriptors are less intuitive than the GRIND descriptors, when new structures are to be designed.

The final proteo-chemometric model retained SI_{MC1} and SI_{MC3} descriptors, while SI_{MC4} and SI_{MC5} were discarded. It therefore seems that the somewhat higher than average affinity of the series of studied compounds for MC_1 receptor and lower for MC_3 is partially explained by specific ligand−receptor interactions and partially by receptor properties alone.

Use of the Proteo-chemometric Model for Prediction of New Compounds. To demonstrate the use of the proteo-chemometric model for design of new compounds, we created a virtual compounds library by including, replacing, and modifying selected substituents on the most active compounds in the data set. The affinities of the resulting structures were subsequently

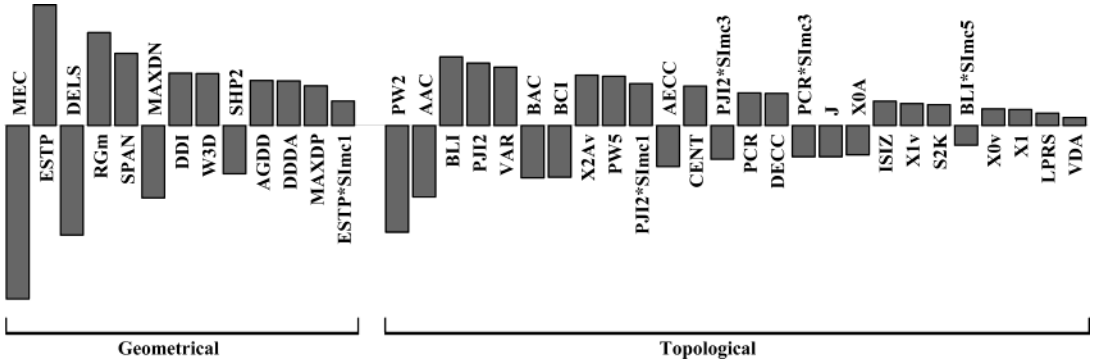

Figure 5. PLS coefficients of topological and geometrical descriptors obtained from the proteo-chemometric model for the interaction of 54 organic compounds with melanocortin receptors MC_1, MC_{3-5}. Abbreviations shown above and below each bar correspond to descriptors as follows: MEC: molecular eccentricity; ESTP: E-state topological parameter; DELS: molecular electrotopological variation; RGm: radius of gyration (mass weighted); SPAN: span radius; MAXDN: maximal electrotopological negative variation; DDI: distance−distance index; W3D: 3D-Wiener index; SHP2: average shape profile index of order 2; AGDD: average geometric distance degree; DDDA: distance−distance degree average; MAXDP: maximal electrotopological positive variation; ESTP*SImc1: E-state topological parameter*Sequence identity to MC_1; PW2: path/walk 2 − Randic shape; AAC: average atomic composition index; BLI: Kier benzene-likeliness index; PJI2: 2D Petitjean shape index; VAR: variation; BAC: Balaban centric index; BCI: Bertz molecular complexity index; x2Av: average valence connectivity index chi-2; PW5: path/walk 5 − Randic shape; PJI2*SImc1: 2D Petitjean shape index*Sequence identity to MC_1; AECC: average eccentricity; CENT: centralization; PJI2*SImc3: 2D Petitjean shape index*Sequence identity to MC_3; PCR: ratio of multiple path counts to path counts; DECC: eccentricity deviation; PCR*SImc3: ratio of multiple path counts to path counts*Sequence identity to MC_3; J: Balaban J index; x0A: average connectivity index chi-0; ISIZ: molecular size index; x1v: valence connectivity index chi-1; S2K: 2-path Kier shape index; BLI*SImc5: Kier benzene-likeliness index*Sequence identity to MC_5; x0v: valence connectivity index chi-0; x1: connectivity index chi-1; LPRS: log of product row sums (PRS); VDA: average vertex distance degree.

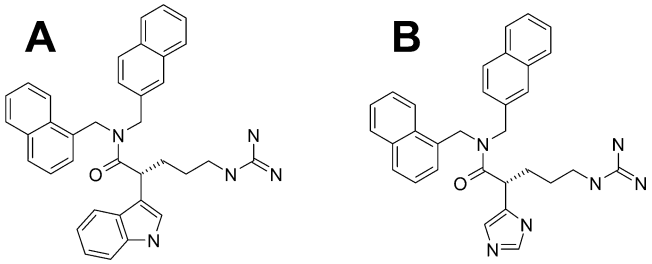

Figure 6. Modified structures of **11b**, designed by including 3-indole (A) or 5-imidazole (B).

predicted using the proteo-chemometric model. Using this approach many structures were found for which the model predicted increased affinity. Here just a few such examples are given. Thus, for example, inspection of GRIND fields and PLS coefficients showed that the molecular interactions, which could be associated with the α-amino group of the arginine residue of compound **11b** contributed little to affinity. Thus, it seems unlikely that this group is involved in hydrogen bonding with the MC receptors. This prompted us to make predictions for structures where the amino function was located farther from the aromatic moieties and closer to the guanidine. Indeed, according to the model, replacement of the amino group with aminoethyl led to an increase in the predicted affinity for all the four MC receptor subtypes by approximately 0.2 pK_i units. Replacement of the amino group by guanidine improved the predicted affinity for the MC_1 receptor further, the increase amounting to 0.4 pK_i units, whereas the predicted affinities for the other receptor subtypes remained unchanged. Replacing the amino group by 3-indole and 5-imidazole (Figure 6) led to even larger predicted increases of the affinity for the MC_1 receptor (up to 1 pK_i unit), due to augmented DRY molecular interaction field and changes in molecular geometry.

Conclusion

In this study we evaluated conventional QSAR and proteo-chemometric approaches for the analysis of the binding of organic compounds to melanocortin receptor subtypes. Both approaches provided models showing good correlations and good predictive abilities. It shall be noted that the modeling accuracy, with SDEP values ranging from 0.3 to 0.5 pK_i units, was close to the accuracy of the biological measurements.

For the QSAR modeling, we found that separate models for each MC receptor performed better than when all receptors were included in a single model. The difference in predictive ability of these two variants of QSAR models became obvious after the improvement of models by variable selection. The better performances of separate models was not surprising as the binding affinity profiles for receptor subtypes were significantly different. However, the proteo-chemometrics approach allowed us to create a model where all the receptors were included in a single model, and this model showed at least as good predictive ability as the separate QSAR models. Moreover, the proteo-chemometric model revealed some determinants that were important for the differences in the MC receptors' recognition of compounds that could not have been revealed from the QSAR model. It seems thus quite promising that good proteo-chemometric models can be created, despite the fact that we had here included only four different wild type receptor subtypes. Increasing the number of receptors would of course allow a more detailed description of the receptor properties and might lead to further improvements in the models. Studies in this direction are highly warranted (cf. refs 10, 11).

The present study includes a combined use of three-dimensional and two-dimensional descriptors and a novel method for variable selection. A well-known disadvantage of 3-D descriptors is their dependence on the selected conformation. The uncertainties that could

arise from the flexibility of the compounds in the current data set were partially avoided by using a rule-based generator of 3-D structures. However, the inclusion of topological and geometrical descriptors allowed us to further improve modeling quality. This could be expected, as the use of different types of descriptors would allow one to capture a broader amount of information on the properties of the compounds. (For a further discussion on advantages of the combined use of different types of descriptors see, e.g., ref 25) Still the inspection of PLS coefficients indicates the dominating role of GRINDs in the models.

Experimental Section

A. Generation of the 3D Structures. Structures of investigated compounds were created using ISIS/Draw and converted to 3D by using the Corina unit of the Tsar 3.3 software package.[26] Corina generates low energy conformations that are close to the X-ray determined structures, as has been shown by the evaluation on a series of 639 compounds.[27] Optimization of the Corina structures were then performed by energy minimization using the Cosmic utility of Tsar 3.3 (COSMIC force field), after first having derived partial charges using Charge2 unit of Tsar 3.3.

B. GRIND Descriptors. Computing GRINDs was performed by Almond 2.0[14] and involved several steps. First, molecular interaction fields (MIFs) were calculated by placing a probe group on grid points surrounding the molecule (this was performed by program GRID[28,29]). We used the three Almond default probes DRY (hydrophobic), O (carbonyl oxygen, i.e., H-bond acceptor), and N1 (amide nitrogen, i.e., H-bond donor), as well as the default 0.5 Å spacing between grid points. A number of grid nodes were then selected for each probe meeting two requirements, namely, showing highly favorable interactions with the molecule and being situated as far as possible from each other. The products of the energy values for all node pairs were then calculated. Finally, the maxima of the thus obtained values falling within specified distance ranges (windows) for node pairs representing the same type of MIF (i.e., DRY-DRY, O−O, and N1−N1), and different MIFs (DRY−O, DRY−N1, O−N1) were used as descriptors of the compounds. Two hundred grid nodes were extracted for each probe, and a window width of 0.4 Å was used for the generation of GRINDs. Thus, six blocks of GRINDs (also termed auto-correlograms, cross-correlograms, or simply correlograms) were obtained, the number of descriptors in each correlogram being equal to the largest distance between nodes divided by the window's width. This resulted in total in $6 \times 62 = 372$ descriptors. However, for a fair number of GRINDs only one or two compounds in the data set obtained nonzero values. These descriptors were discarded, leaving 257 GRINDs in the data set.

C. Preprocessing of Data. All descriptors were first mean centered and scaled to unit variance. Moreover, to account for differences in the number (and mutual correlation) of descriptors of each type, block scaling was then applied. Thus, each of the four variable blocks (GRIND, topological, geometrical, and cross-descriptors) was scaled to equal variance by computing block weight as 1/sqrt(N), where N is number of variables in the given block. Accordingly, the standard deviation of each GRIND descriptor was set to 1/sqrt(257), the standard deviation of each topological descriptor to 1/sqrt(68), the standard deviation of each geometrical descriptor to 1/sqrt(18), and the standard deviation of each cross-descriptor to 1/sqrt(1372). When we elaborated models by variable selection (see below), block-scaling weights were recalculated in order to account for the actual number of remaining variables in each descriptor block. Since we only had four receptor sequence identity (SI) descriptors, the standard deviation of each of these was set to 0.2, rather than 1/sqrt(4), and were not changed during the model elaboration. The response variables (i.e., the negative logarithm of K_i values) were also mean centered prior to applying any further calculations.

D. Principal Component Analysis. Principal component analysis (PCA) was performed in order to reveal and remove eventual outliers in the series of compounds. PCA approximates multivariate data by projecting it onto a lower dimensionality variable space, called latent variables or principal components.[30] Components are computed iteratively and are orthogonal to each other.

PCA was performed using SIMCA-P 9.0 software.[21] The number of components (dimensions) was determined using the eigenvalue criterion, as suggested in the SIMCA manual.[21] Thus, only components with a normalized eigenvalue larger than 2 were considered significant. Using these criterion we obtained a four-component model on our data with the explained sums of squares $r^2X = 0.78$.

To identify outliers we used the distance to model (DModX) parameter.[21] One compound (**16a**[6]) significantly surpassed the critical distance (DModX being 1.65 versus $\mathrm{DModXcritical}_{(5\%\ \text{membership probability})} = 1.22$) and was excluded from further analysis.

E. Partial Least-Squares Projection to Latent Structures. PLS can be regarded as an extension of PCA, which correlates two data matrixes. Thus in PLS a matrix of predictor variables X and a matrix of responses Y are simultaneously projected to latent variables (components), with an additional constraint to maximize the covariance between projections of X and Y.

PLS is insensitive to collinearity among the predictor variables and allows one to handle data sets where the number of variables is larger than number of observations (for an account on the PLS method, see Wold[17]). The PLS analysis was carried out using SIMCA-P 9.0.[21]

F. Validation of PLS Models. The goodness of fit of PLS models was assessed by calculating the fraction of explained variation of X and Y (r^2X and r^2Y). The predictive ability was assessed by calculating the fraction of the predicted Y-variation (q^2), according to cross-validation.[21,22] If not mentioned otherwise, we used seven cross-validation groups.

Models were also validated by response permutation. In short, models were recalculated 20 times for randomly reordered Y data. r^2Y and q^2 were plotted as a function of correlation coefficient between the original Y and permuted Y. The intercepts of the regression lines (correlation coefficient being zero) indicate the degree of overfit. A negative q^2 intercept would thus indicate that original q^2 values are not obtained by chance.[20]

G. Improvements of Model by Variable Selection. The PLS models were improved by removing descriptors deemed to correlate with the responses by chance. To estimate possibility of chance correlation, we calculated standard deviation of PLS coefficients (σ_{coeff}) from the seven cross-validation rounds, and the absolute value of the ratio between PLS coefficient (coeff) and its standard deviation: $|\mathrm{Coeff}|/\sigma_{\mathrm{coeff}}$. The descriptors having the lowest $|\mathrm{Coeff}|/\sigma_{\mathrm{coeff}}$ values were considered least relevant and accordingly removed. A new PLS models were elaborated excluding these variables, and in all cases improvement of q^2 were achieved. In the current study, we iteratively excluded 20% of the remaining variables until the q^2 reached a plateau or started to decline. For models containing several response variables instead of $|\mathrm{Coeff}|/\sigma_{\mathrm{coeff}}$ value the $\Sigma|\mathrm{Coeff}|/\Sigma\sigma_{\mathrm{coeff}}$ was used for variable selection. Block-scaling weights were recalculated during the process in order to account for the change of the number of remaining variables in each descriptor block.

Acknowledgment. Supported by Melacure Therapeutics AB, the Swedish VR (04X-05957 and 230-2000-291).

References

(1) Wikberg, J. E. S. Melanocortin receptors: new opportunities in drug discovery. *Exp. Opin. Ther. Pat.* **2001**, *11*, 61−76.
(2) Wikberg, J. E. S.; Muceniece, R.; Mandrika, I.; Prusis, P.; Lindblom, J.; Post, C.; Skottner, A. New aspects on the melanocortins and their receptors. *Pharmacol. Res.* **2000**, *42*, 393−420.

(3) Abdel-Malek, Z. A. Melanocortin receptors: their functions and regulation by physiological agonists and antagonists. *Cell. Mol. Life Sci.* **2001**, *58*, 434−441.
(4) Schioth, H. B.; Chhajlani, V.; Muceniece, R.; Klusa, V.; Wikberg, J. E. S. Major pharmacological distinction of the ACTH receptor from other melanocortin receptors. *Life Sci.* **1996**, *59*, 797−801.
(5) Andersson, P. M.; Boman, A.; Seifert, E.; Skottner, A.; Lundstedt, T. Ligands to the melanocortin receptors. *Expert Opin. Ther. Pat.* **2001**, *11*, 1583−1592.
(6) Mutulis, F.; Mutule, I.; Lapinsh, M.; Wikberg, J. E. S. Reductive amination products containing naphthalene and indole moieties bind to melanocortin receptors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1035−1038.
(7) Mutulis, F.; Mutule, I.; Wikberg, J. E. S. N−Alkylamino acids and their derivatives interact with melanocortin receptors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1039−1042.
(8) Prusis, P.; Muceniece, R.; Andersson, P.; Post, C.; Lundstedt, T.; Wikberg, J. E. S. PLS modeling of chimeric MS04/MSH-peptide and MC_1/MC_3-receptor interactions reveals a novel method for the analysis of ligand−receptor interactions. *Biochim. Biophys. Acta* **2001**, *1544*, 350−357.
(9) Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. *Protein Eng.* **2002**, *15*, 305−311.
(10) Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. S. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1525*, 180−190.
(11) Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteo-chemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharm.* **2002**, *61*, 1465−1475.
(12) Eberle, A. N. *The melanotropins. Chemistry, physiology and mechanisms of action;* Karger: Basel, 1988.
(13) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.
(14) Almond 2.0, Multivariate Infometric Analysis S.r.l., Perugia, Italy, http://miasrl.com.
(15) Dragon 1.11, Talete S.r.l., Milano, Italy, http://www.disat.unimib.it/chm.
(16) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors.* Wiley − VCH: Weinheim, Germany, 2000.
(17) Wold, S. PLS for multivariate linear modeling. In *Chemometric Methods in Molecular Design*; van de Waterbeemd H., Ed; VCH: Weinheim, Germany, 1995; pp 195−218.
(18) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Introduction to multi- and megavariate data analysis using projection methods (PCA/PLS)*; Umetrics AB: Umeå, Sweden, 1999.
(19) Prusis, P.; Schioth, H. B.; Muceniece, R.; Herzyk, P.; Afshar, M.; Hubbard, R. E.; Wikberg, J. E. S. Modeling of the three-dimensional structure of the human melanocortin 1 receptor, using an automated method and docking of a rigid cyclic melanocyte-stimulating hormone core peptide. *J. Mol. Graphics Modell.* **1997**, *15*, 307−317.
(20) Eriksson, L.; Johansson, E. Multivariate design and modeling in QSAR. *Chemom. Intell. Lab. Syst.* **1996**, *34*, 1−19.
(21) SIMCA-P 9.0. A new standard in multivariate data analysis, Manual, Umetrics AB: Umeå, Sweden, 2001. http://www.umetrics.com.
(22) Eriksson, L.; Johansson, E.; Wold, S. Quantitative structure−activity relationship model validation. In *Quantitative Structure−Activity Relationships in Environmental Sciences − VII*; Chen F., Schuurmann G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 381−397.
(23) Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, Å.; Pettersen, J.; Bergman, R. Experimental design and optimization. *Chemometr. Intell. Lab. Syst.* **1998**, *42*, 3−40.
(24) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(25) Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693−705.
(26) Tsar 3.3, Accelrys Inc., http://www.accelrys.com.
(27) Sadowski, J.; Gasteiger, J. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(28) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(29) GRID v.19; Molecular Discovery Ltd., http://www.moldiscovery-.com.
(30) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

JM020945M