Rough Set-Based Proteochemometrics Modeling of G-protein-coupled Receptor-Ligand Interactions

Helena Strömbergsson,[1] Peteris Prusis,[1,2] Herman Midelfart,[3] Maris Lapinsh,[2] Jarl E.S. Wikberg,[2*] and Jan Komorowski[1*]

[1]*Uppsala University, The Linnaeus Centre for Bioinformatics, Uppsala, Sweden*
[2]*Uppsala University, Department of Pharmaceutical Biosciences, Uppsala, Sweden*
[3]*Norwegian University of Science and Technology, Department of Biology, Trondheim, Norway*

ABSTRACT: **G-Protein-coupled receptors (GPCRs) are among the most important drug targets. Because of a shortage of 3D crystal structures, most of the drug design for GPCRs has been ligand-based. We propose a novel, rough set-based proteochemometric approach to the study of receptor and ligand recognition. The approach is validated on three datasets containing GPCRs. In proteochemometrics, properties of receptors and ligands are used in conjunction and modeled to predict binding affinity. The rough set (RS) rule-based models presented herein consist of minimal decision rules that associate properties of receptors and ligands with high or low binding affinity. The information provided by the rules is then used to develop a mechanistic interpretation of interactions between the ligands and receptors included in the datasets. The first two datasets contained descriptors of melanocortin receptors and peptide ligands. The third set contained descriptors of adrenergic receptors and ligands. All the rule models induced from these datasets have a high predictive quality. An example of a decision rule is "**If **R1_ligand(Ethyl)** and **TM helix 2 position 27(Methionine)** then **Binding(High)."** **The easily interpretable rule sets are able to identify determinative receptor and ligand parts. For instance, all three models suggest that transmembrane helix 2 is determinative for high and low binding affinity. RS models show that it is possible to use rule-based models to predict ligand-binding affinities. The models may be used to gain a deeper biological understanding of the combinatorial nature of receptor-ligand interactions. Proteins 2006; 63:24–34.** **© 2006 Wiley-Liss, Inc.**

Key words: drug design; QSAR; GPCRs; machine learning; rough sets; partial least squares

INTRODUCTION

The G-protein-coupled receptors (GPCRs) share a conserved structure composed of seven transmembrane (TM) helices. About 50% of all recently launched drugs[1] are targeted towards GPCRs, which makes the receptor family a drug target of paramount importance. However, drug discovery is hampered by the fact that GPCRs, similarly to other membrane embedded proteins, have characteristics that make their 3D structure extremely difficult to determine experimentally. For this reason, computational methods for drug design have relied primarily on ligand-based techniques such as 2D substructure similarity searching and quantitative structure activity relationship (QSAR) modeling.[2] Homology based modeling of the 3D-structure of the GPCRs' and ligand docking have also been in use for some time (see e.g., Cavasotto et al.[3] and Varady et al.[4]). However, such methods are often too inaccurate to be highly useful because of the lack of solved structures of GPCR's with close enough homology (at this time, only the crystal structure of rhodopsin is available). Recently, we introduced proteochemometrics (PCM),[5] which is a new approach to the study of molecular recognition. In PCM the receptor-ligand interaction space is modeled using descriptors of both receptors and ligands. These descriptors are combined and associated with experimentally measured binding affinity data. Associations between receptor-ligand properties and binding affinity have previously been modeled[5–7] by the linear method partial least squares (PLS).[8] PLS was able to successfully predict the binding affinity, expressed as a real value, but was dependent on a predefinition of binary combination of attributes, the so-called cross terms, which indicates that there are nonlinear factors influencing the model. It cannot be excluded that even higher order cross terms (such as triplets or quadruples) will be necessary for modeling. Obtaining models that use combinations of several attributes may, in addition, provide a deeper understanding of the mechanisms that govern the receptor-ligand interactions. To this end we introduce a new methodology for obtaining PCM models. It is based on a nonlinear rule-based technique known as rough sets (RS).[9,10] RS are a new approach to use machine learning in genomic and molecular applications. RS approach is particularly well

Grant sponsor: This research was supported by grants from the Knut and Alice Wallenberg Foundation, Swedish Foundation for Strategic Research and the Swedish Research Council; Grant number: (VR 621-2002-4711, 04X-05957)

*Correspondence to: Jan Komorowski, The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, SE-751 24 Uppsala, Sweden or Jarl Wikberg, Department of Pharmaceutical Pharmacology, Uppsala University, Box 591, SE-751 24 Uppsala, Sweden. E-mail: jan.komorowski@lcb.uu.se or jarl.wikberg@farmbio.uu.se

Received 10 May 2005; Accepted 9 August 2005

Published online 24 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20777

suited to model from uncertain, approximate, and inconsistent data and have been proven to be successful in, for example, fold recognition,[11] prediction of gene function from time series expression profiles,[12] and the discovery of combinatorial elements in gene regulation.[13] The generated rule-models are combinatorial in their nature and easy to interpret by nonexperts. They are also minimal in the sense of not using redundant attributes.

In order to validate our methodology in modeling receptor-ligand interactions, we applied it to three distinctive PCM datasets that represent a superfamily of membrane receptors. The first two datasets consist of melanocortin receptors and peptide ligands[14–16] and the third dataset contain adrenergic receptors and ligands.[17] Melanocortin[18] and adrenergic[19] receptors belong to the superfamily of GPCRs. The melanocortin receptor family participates in a large number of physiological functions such as pigmentation, cardiovascular regulation, and neuromuscular regeneration, whereas the adrenergic receptors are involved in sympathetic nervous system responses. The descriptors used in this study are independent of the 3D structures of the receptors and ligands.

It appears that RS-induced models are robust and enable a classification of high and low receptor-ligand binding affinities. Furthermore, the rules obtained from the RS models provide an insight about combinations of properties important for receptor-ligand interactions and affinity. Although RS models do not produce continuous output parameters (e.g., the binding affinity value), they nevertheless were able to easily identify determinative receptor and ligand fragments that separate high and low binding affinities. For the first time to our knowledge, it is possible to identify higher-order combinations of fragments that play a significant role in these interactions.

METHODS

Rough Sets

Proteochemometrics receptor-ligand interaction data is represented in a tabular form. Disjunctive receptor-ligand complexes are in the rows and descriptors of the receptors and ligands are in the columns. The very last column denotes a measurement of receptor-ligand binding affinity.

Here, we give a brief introduction to rough sets. For a complete overview of the method we refer to Komorowski et al.[20] All computations were performed using the ROSETTA system[21] (http://rosetta.lcb.uu.se). ROSETTA implements inductive learning using the mathematical framework of RS.[9,10] This is a relatively new approach to representing and reasoning with incomplete or uncertain knowledge. It deals with the classificatory analysis of data tables. A dataset is represented as a table, where each row represents a case and each column represents an attribute. This table is called an *information system*. More formally, an information system is a pair $\mathscr{A} = (U, A)$, where U is a nonempty finite set of *objects* called the *universe* and A is a nonempty finite set of functions a: $U \rightarrow V_a$, called *attributes*; for each $A \in a$ the set V_a is called the *value set* of a.

If there is a known outcome or classification, this *a posteriori* knowledge is expressed as one distinguished attribute called the *decision attribute*. An information system of this kind is called a *decision system*. Thus, a decision system is any information system of the form $\mathscr{A} = (U, A \cup \{d\})$, where $d \notin A$ is the decision attribute. Two objects x, y are said to belong to the same *decision class* if $d(x) = d(y)$. A subset of attributes, $E \subset A$ that separates U to the same degree as A is called a *full reduct* whereas a subset of attributes that separates $u \in U$ from the other objects in U is called an *object related reduct*. Intuitively, a reduct represents a minimal subset of attributes whose values are necessary to discover a given object from the other objects in the universe. A variant of full reducts called *approximate reducts* is often found to be valuable for classification. Approximate reducts separate a given subset of U to approximately the same degree as the full set of attributes. ROSETTA implements a number of reduct computing algorithms such as the genetic,[22] Johnson's,[23] and Hacid's[24] algorithm.

From the reducts, the RS algorithms derive a set of *decision rules* of the form $\alpha \rightarrow \beta$. Here α is a Boolean function $U \rightarrow \{true, false\}$ built up of the logical connectives $(\wedge, \vee, \neg)$ and atom statements of the form $a(\cdot) = v$, where $a \in A$, $v \in V_a$. Similarly, $\beta: U \rightarrow \{true, false\}$ is built up from atom statements of the form $d(\cdot) = v$, where $v \in V_d$.

There are several numerical factors associated with decision rules. Most of these are derived from the *support* of a rule, which is the number of objects in the decision system that possess the properties of α and β. The factor *coverage*, which is defined as coverage$(\alpha \rightarrow \beta)$ = support$(\alpha \wedge \beta)$/support(β), reflects the strength of a rule and gives a measure of how well α describes the decision class(es) given by β. An example of a decision rule will be presented in the results section.

Within the RS framework, the three datasets included in this study are decision systems in which the descriptors of receptors and ligands are attributes and the binding affinity values (given as pK_i) are decision attribute values. In this study we are interested in inducing models separating high and low binding receptor-ligand complexes. RS are dependent on discrete decision classes. Therefore, each dataset in this study was sorted by *pKi* value and divided into one “high-affinity binding” and one “low-affinity binding” decision class of equal size.

To validate the models the Johnson[23] algorithm was applied to compute object related reducts. These reducts were formulated into decision rules, which were used for validation and interpretation of the induced models. For an implicit ranking of attributes, approximate reducts were computed by the genetic algorithm.[22]

Model Validation

RS modeling is an inductive learning process. Such processes begin with a division of the selected dataset into two subsets: a training set, which is used to train the system, and a test set, which provides a means to evaluate the induced model.

TABLE I. A Summary of the Three Datasets Included in This Study[†]

Dataset	No. of receptor attributes	No. of ligand attributes	No. of objects training set	pK_i range	Cutoff value	No. of objects test set
I (Melanocortin)	4	2	32	4.5–9.0	6.8	8
II (Melanocortin)	4	4	48	6.6–10.3	9.0	12
III (Adrenergic)	52	3	105	5.8–9.9	8.0	26

[†] The number of receptor and ligand attributes and the number of objects in the training and test set is given for each dataset. The range of pK_i values of each training set is given and the cutoff point used for discretization into high and low binding affinity decision classes.

To estimate the robustness of RS model induction, a k-fold cross-validation (CV) on each training set was performed. In a CV, the data is randomly divided into k blocks. Each block is left out once and a model is induced on the remaining $k - 1$ blocks. From a ROSETTA k-fold CV, the performance estimates *accuracy mean* and *area under curve* (*AUC*) *mean* are calculated. The accuracy mean is the average proportion of correctly predicted objects computed for the k blocks during CV. The accuracy mean value ranges between 0 and 1. If the accuracy mean is 1, all objects are classified correctly and if the accuracy is 0 all objects are misclassified. The AUC is the area under the receiver operating characteristics[25] (ROC) curve and it is a measurement of the discriminatory power of a classifier. The ROC curve results from plotting *sensitivity* against 1-*specificity* while letting the threshold value τ vary. For a binary classifier an AUC of 1.0 means that the discriminatory power is optimal, whereas an AUC of 0.5 means that the classifier does not perform better than a random classification of objects. The AUC mean computed by ROSETTA is the average AUC for the models induced during the k folds of a CV. The standard deviation (SD) is reported for both accuracy and AUC mean. In the present study we used a 10-fold CV.

We also performed permutation validations as follows; a randomization algorithm was applied to the training sets. This algorithm randomly shuffled the decision attribute values and then performed a 10-fold CV computing accuracy and AUC mean. This process was repeated in 1000 iterations. The fractions of the iterations resulting in an accuracy and AUC mean higher or equal to the ones obtained from the original training sets are counted and reported as *accuracy mean p-value* and *AUC mean p-value*.

For external test set validation, a model was induced on the training set. The objects in the external test set were classified and accuracy and AUC were computed. The set of decision rules used for classification was minimized by systematically filtering out rules with the lowest support and removing rules with multiple decision values. When the accuracy markedly decreased no further rules were filtered out. The minimal set of decision rules used for classification were analyzed and interpreted for their biological meaning.

In this study each dataset was randomly divided into a training set of 80% and an external test set of 20% of the objects. The properties and number of objects in training- and test sets are summarized in Table I.

Ranking of Attributes

Approximate reducts were computed from the training sets by the Genetic algorithm[22] using a population size of 1000, a keep size of 1000 and a hitting fraction cutoff of 0.5. The *performance filtering algorithm* implemented in the ROSETTA tool kit was applied on the set of reducts. This algorithm filters and ranks a set of reducts by applying those to a test set. In this study, a set of approximate reducts was used for classification of the receptor-ligand complexes present in the test sets. Reducts resulting in an accuracy of 0.7 or higher were used to compute attribute occurrence. The occurrence of each attribute in the high scoring reduct set provides an indirect ranking of attributes since attributes important to classification come up more frequently in reducts.

MATERIALS

Three datasets are used in this study. The first two datasets contain melanocortin (MCs) receptors and the third dataset contains adrenergic receptors (ARs). The MC receptors[18] are involved in a diverse number of physiological functions including pigmentation, steroidogenesis, inflammation, body weight, cardiovascular regulation, and neuromuscular regeneration. The MC system consists of some peptide hormones and the melanocortin receptors. The peptide hormones, termed melanocortins, include melanocyte stimulating hormones and adrenocorticotropic hormone. Five types of MC receptors; MC_1–MC_5; have been identified showing distinct distribution in the body. The ARs[19] are targets for the neurotransmitters and hormones norepinephrine and epinephrine. Both hormones play key roles in sympathetic nervous system responses, especially those involved in cardiovascular homeostasis. ARs are thus targets for many therapeutically important drugs, including those for cardiovascular diseases, asthma, and nasal congestion. Gene cloning studies indicate that three adrenergic α_1 receptor genes exist: α_{1a}, α_{1b}, and α_{1d}.

Dataset I (Melanocortin Receptors)

The first dataset contains 40 MC receptor-ligand complexes. Each object of the dataset is a description of the composition of MC receptor chimeras and peptide ligands together with the receptor-ligand binding affinity. The binding affinities are given in pK_i units, which are the negative logarithm of the inhibition constants K_i. The pK_i values were taken from Muceniece et al.[16]

The receptor chimeras comprised four parts; A, B, C, and D (Fig. 1), each originating either from wild-type MC_1 or

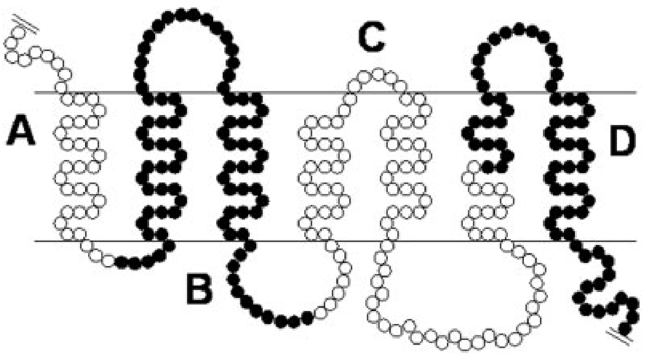


Fig. 1. A schematic overview of the division of chimeric MC-receptors into parts. The A, B, C, and D region originate either from wild type MC_1 or MC_3.

TABLE II. Sequences of the Highly Active Melanocortin α-MSH/MS04 Peptides Used in Dataset 1[†]

	Sequence	
Peptide	R1 R2	Descriptor
α-MSH	SYSMEHFRWGKPV	[MSH, MSH]
MS04	SSIISHFRWGKCN	[MS04, MS04]
MS05	SSIISHFRWGKPV	[MS04, MSH]
MS06	SYSMEHFRWGKCN	[MSH, MS04]

[†] The peptide ligands α-MSH and MS04 are used as building blocks for MS05 and MS06. The middle part of each ligand is constant while the shaded N-terminal (R1) and C-terminal (R2) parts are variable. The two strings of each descriptor represent the origin of the N- and C-terminal part. Underline sequences denote the variable parts of the ligands R and R2.

MC_3. The composition of each receptor was described by a vector of four strings. The string "MC_1" was assigned to parts originating from MC_1 and the string "MC_3" to parts originating from MC_3. For instance, a receptor chimera in which parts A and B originate from MC_1 and parts C and D originate from MC_3 was described by the vector [MC_1, MC_1, MC_3, MC_3]. The wild-type receptors MC_1 or MC_3 were described by the vectors [MC_1, MC_1, MC_1, MC_1] and [MC_3, MC_3, MC_3, MC_3], respectively.

The peptide ligands were α-MSH and MS04 and chimeric analogues of these. The peptide α-MSH is a natural ligand, whereas the MS04 ligand is a result from phage display selections.[15] In the dataset, the middle part of the ligands is nonvariable, whereas the N- and C-terminal parts, denoted R1 and R2, originate either from α-MSH or MS04. The composition of each ligand in the dataset was described by a vector of two strings. The string "MSH" was assigned if the part originates from α-MSH and the string "MS04" if the part originates from MS04 (Table II).

Each receptor-ligand complex was thus described by a vector of six strings and one real number. The first four strings describe the composition of the receptors and the following two strings describe the properties of the ligand. The real number is the binding affinity between the receptor and the ligand given as pK_i. For example, the object [MC_3,MC_3,MC_1,MC_1,MSH,MS04, 6.38] is a receptor-ligand complex, where the A and B parts of the receptor originate from MC_3 and the C and D parts originate from MC_1, the N-terminal part of the ligand is from α-MSH and the C-terminal part is from MS04, and the binding affinity is 6.38.

Dataset II (Melanocortin Receptors)

The second dataset was composed of 60 receptor-ligand complexes, where the receptors were the same as for Dataset I. Each object in the dataset describes the composition of MC receptor chimeras and six linear and cyclic peptide ligands. The receptors were described by the same descriptors as for the first Dataset I. The binding affinity pK_i values were taken from the study made by Schioth et al.[14]

The ligands were various derivatives of α-MSH (Table III) where one or several amino acids had been altered. An amino acid alignment of the ligands showed that there are four variable sites, R3–R6. Some of the ligands had modified leucine (Nle), phenylalanine (dF), and/or tyrosine (I-Y) at the variable sites. Ligands with two cysteine (C) residues at the variable sites were cyclic with a C-C bridge. Vectors of four strings were used to describe the amino acid composition at the variable sites. The binding affinities were described in the same manner as for Dataset I.

A receptor binding complex was hence described by a vector of eight strings and one real number. The first four strings describe the composition of the receptor, the following four strings describe the ligand alterations and the real number is the binding affinity. For instance, the receptor ligand complex [MC_1, MC_1, MC_3, MC_3, Y, M, F, G, 8.27] has part A and part B of the receptor from MC_1 and part C and part D from MC_3, the ligand has the residues tyrosine (Y), methionine (M), phenylalanine (F), and glycine (G) at the four variable sites, and the binding affinity is 8.27.

Dataset III (Adrenergic Receptors)

The third dataset consists of 131 binding complexes. The objects in the dataset were wild type, chimeric, and point-mutated adrenergic receptors with derivatives of 4-piperidyl oxazole antagonists. The binding affinity pK_i values on the interaction were taken from Hamaguchi et al.[17]

The 18 receptor sequences in this study were chimeras and point mutation combinations of the three human wild types α_{1a}, α_{1b}, and α_{1d}. A schematic overview of the receptors is given in Figure 2. For the assignment of descriptors, only the sequences within the receptors' aligned TM regions were taken into account. The alignment of the wild-type receptors was obtained from the GPCR database.[26] The considered TM regions were as follows (using the amino acid numbering of α_{1a}): TM1 29–45, TM2 70–86, TM3 100–122, TM4 145–159, TM5 183–199, TM6 280–292¤ and TM7 310–321. Within these TM regions there are 52 variable sites. The 18 receptors were aligned and the amino acids at the 52 variable sites were used as descriptors. A schematic view on the variable sites is illustrated in Figure 3. For convenience the traditional numbering system based on conserved residues is used.[27] Each receptor was thus described by a vector of 52 strings, where each string is the one letter code of one amino acid.

TABLE III. Amino Acid Sequence of the Peptides Used in Dataset II and Their Descriptors†

	Sequence													
Peptide		R3		R4			R5			R6				Descriptor
α-MSH	S	Y	S	M	E	H	F	R	W	G	K	P	V	[Y, M, F, G]
[^{125}I]-NDP-MSH	S	I-Y	S	Nle	E	H	dF	R	W	G	K	P	V	[I-Y, Nle, dF, G]
NDP-MSH	S	Y	S	Nle	E	H	dF	R	W	G	K	P	V	[Y, Nle, dF, G]
[Nle^4]-α-MSH	S	Y	S	Nle	E	H	F	R	W	G	K	P	V	[Y, Nle, F, G]
cCDC	S	Y	S	C	E	H	dF	R	W	C	K	P	V	[Y, C, dF, C]
cCLC	S	Y	S	C	E	H	F	R	W	C	K	P	V	[Y, C, F, C]

†The underscored positions are the alteration sites R3–R4.

The 12 ligands included herein were derivatives of 4-piperidyl oxazole that had been modified at three positions (Fig. 4), R1, R2, and R3. The chemical formulas were used to describe the substituents at the variable sites. One ligand was thus described by a vector of three strings. For instance, the ligand [H, H, OCF_3] has an H-substitution at R1 and R2 and an OCF_3-substitution at R3.

One receptor-ligand complex was hence described by a vector of 52 + 3 strings and 1 real number. The first 52 strings represent the amino acid composition of the receptor, the following three strings describe the ligand and the real number is the binding affinity expressed as pK_i.

RESULTS AND DISCUSSION

Model Generation and Validation

The results from model induction and predictions of training and test sets are presented in Table IV. The accuracy and AUC means are the outputs from the CVs performed on the training set, and the accuracies and AUCs are the results of the predictions of the external test sets. The CV results show that all datasets have a relatively high stability. The higher standard deviation reported for the first two datasets reflects the fact that these sets have fewer objects than the third dataset. The results from the permutation tests clearly illustrate that there is very little noise in the models and that the probability to obtain a model regardless of the decision attribute value is very small. The test set classifications show that the induced models, consisting of decision rules, are able to successfully separate "low-affinity" and "high-affinity" receptor-ligand objects.

In this study, the set of minimal decision rules associates a minimal number of receptor-ligand descriptors with high or low binding affinity. The results from model validation prove that the rough sets models exhibit a high classification quality. Given high quality rule-based models, biological and biochemical knowledge may be extracted from the induced rules. An example of a decision rule is "**If** A(MC_1) **and** D(MC_1) **and** R4(Nle) **then** Binding-(High)", associating the A part of MC_1 and the D part of MC_1 and a modified Leucine at the variable site R4 of the ligand with high binding affinity. This and other examples

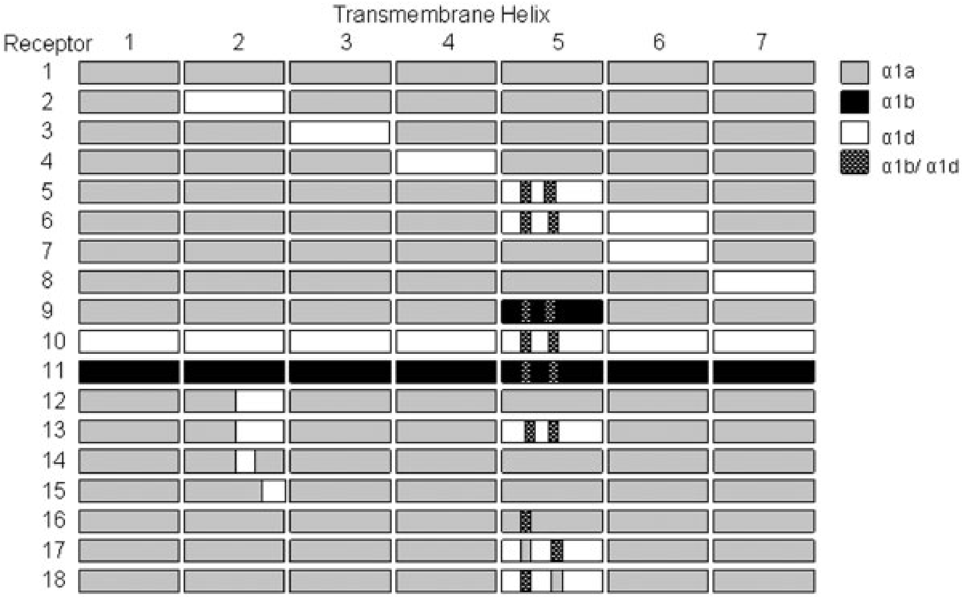


Fig. 2. Schematic overview of the 18 human adrenergic receptors included in this study. Each of the seven transmembrane helices is composed of amino acids from adrenergic receptor subtype α_{1a}, α_{1b} and/or α_{1d}.

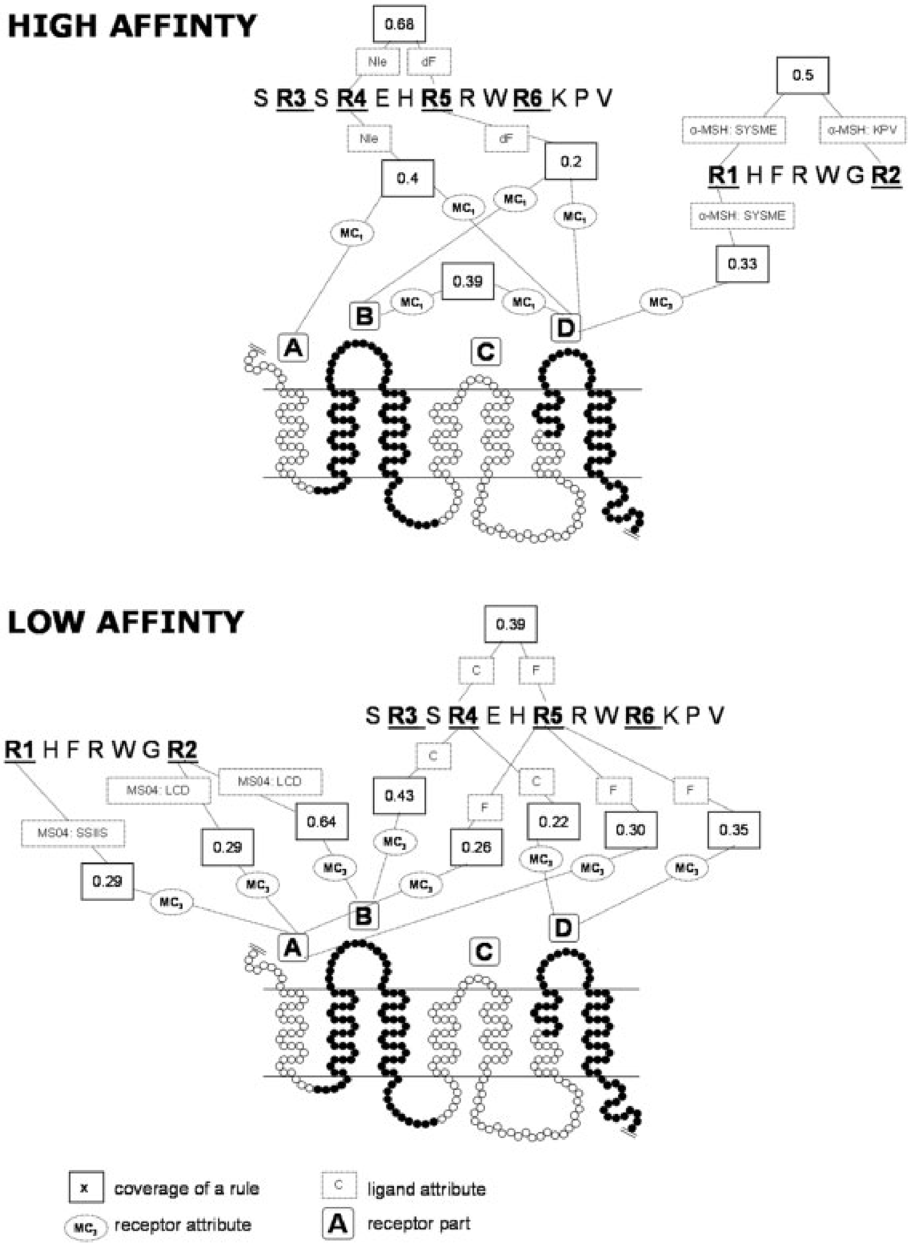


Fig. 3. Illustration of a minimal set of decision rules for Melanocortin Dataset I and Melanocortin Dataset II. The rules are separated into "high affinity binding" and "low affinity binding" rules. In the figure, each rule is centered round a box containing the *coverage* value of the rule (see Methods for details). The "coverage box" connects ligands and/or receptor attributes. The variable parts of the ligands are denoted R1-R6. The variable receptor parts are represented by the letters A, B, C, and D. For instance the 0.4 coverage box in the upper part of the figure connects the ligand attribute Nle with the A and D parts of receptor MC_1, illustrating the rule "**If** A(MC_1) **and** D(MC_1) **and** 2′nd_ligand_position(Nle) **then** Binding(High)".

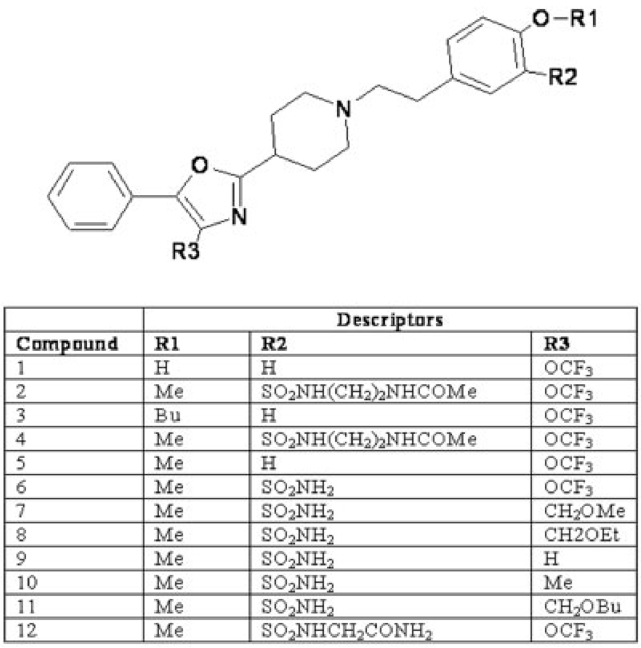
in the datasets show that combinations of three parameters do occur in the models and are determinative to modeling. The complete models can be found at our web site (www.lcb.uu.se/papers/stroembergsson/gpcr/).

Interpretation of Decision Rules for Melanocortin Dataset I and II

A set of minimum decision rules induced by the Johnson algorithm from the training sets of melanocortin Dataset I and Dataset II is illustrated in Figure 3. Because the ligands in the datasets bind to the same type to receptor chimeras, the rules from the two datasets are illustrated in the same figure. The rule sets are divided into "high affinity" and "low affinity" rules. This allows analyzing ligand- and receptor characteristics important to low and high receptor-ligand binding affinity.

Figure 3 illustrates that part C of the receptors and the variable ligand sites R3 and R6 are not included in any of the decision rules. This suggests that part C of the receptors and the ligand sites R3 and R6 are not important for classification of objects into high and low affinity.

The high affinity binding rules include the A, B, and D receptor parts. The A and B parts are exclusively from the MC_1 receptor, whereas the D part can have both a MC_1 and a MC_3 receptor origin. The included variable ligand sites are R1, R2, R4, and R5. Within the receptors, a combination of part B and D, both from the MC_1 receptor quite strongly supports high affinity. Within the ligands, a high affinity is promoted by a combination of R1 and R2 from α-MSH or a combination of modified leucine (Nle) and modified phenylalanine (dF) at R4 and R5. Between ligands and receptor chimeras a high affinity binding is obtained by a combination of part A and D from MC_1 and an Nle at R4. This suggests that Nle interacts with part A and D. Because part A and part D (according to the 3D modeling of the melanocortin receptors[28]) are spatially situated close to each other this interaction is rather likely. High binding affinity is also achieved by a combination of part D from MC_3 and an α-MSH origin at R1. This indicates that only the N-terminal part of the peptide ligand from Dataset I is involved in the interaction with the receptor. Hence, a high binding affinity receptor should have part A and part B from MC_1 and a high binding ligand should have R1 and R2 from α-MSH or Nle at R4 and dF at R5.



	Descriptors		
Compound	R1	R2	R3
1	H	H	OCF_3
2	Me	$SO_2NH(CH_2)_2NHCOMe$	OCF_3
3	Bu	H	OCF_3
4	Me	$SO_2NH(CH_2)_2NHCOMe$	OCF_3
5	Me	H	OCF_3
6	Me	SO_2NH_2	OCF_3
7	Me	SO_2NH_2	CH_2OMe
8	Me	SO_2NH_2	CH2OEt
9	Me	SO_2NH_2	H
10	Me	SO_2NH_2	Me
11	Me	SO_2NH_2	CH_2OBu
12	Me	$SO_2NHCH_2CONH_2$	OCF_3

Fig. 4. Descriptors of derivatives of 4-piperidyl oxazole. The compound has three variable sites. The molecular names of the substituents and their molecular weights were used as descriptors for modeling.

The low binding affinity rules are higher in number resulting in a slightly more complex model. Similarly to the high affinity binding rules, the A, B, and D parts of the receptor, and R1, R2, R4, and R5 of the ligands are included by the rules. The receptors parts are only from the MC_3 receptor, suggesting that this receptor in general has a weaker interaction with its ligands. Within ligands a Cysteine (C) at R4 in combination with a phenylalanine (F) at R5 results in low binding affinity. Between ligands and receptors, one quite strong rule (see Fig. 3) associates the B part from MC_3 with a C ligand attribute at R4 with low binding affinity. The C ligand attribute is exclusively included by the low binding rules indicating that the cyclic nature of the ligands in general causes low binding affinity. Both R1 and R2 occur in low binding rules suggesting that both the N- and C-terminal parts are important to low binding affinity. In summary, a low binding affinity receptor should have the A, B, and D parts from MC_3 and a low binding ligand should have an N- or C-terminal part from MS04 or a C at R4 and an F at R5.

Interpretation of Adrenergic Dataset III Decision Rules

The decision rules induced from the adrenergic Dataset III are illustrated in Figure 5. Similarly to the melanocortin datasets, the rules are divided into high and low binding affinity. The numbering of the 52 receptor descriptors is by helix and based on the location of conserved residues.[27] Only receptor descriptors in the upper part of the helices are included by the rule set. This indicates that the ligands in this dataset are interacting with the nonconserved amino acid residues located towards the extracellular end of the helices upon binding.

The rules associated with low affinity include helix TM1, TM2, and TM5 and the ligand variable sites R1 and R2. There are three strong rules in the rule set. The first rule combines a methyl (Me) group at R1 with a hydrogen (H) atom at R2. The second rule associates a phenylalanine (F) at position 28 in TM2 and a valine (V) at position 2 in TM5 and the substituent $SO_2NH(CH_2)_2NHCOMe$ at R2 with

TABLE IV. Results from Model Validation of the Three Datasets Included in This Study[†]

Dataset	Accuracy mean 10-fold CV	AUC mean 10-fold CV	Accuracy mean p-value	AUC mean p-value	Accuracy test set	AUC test set	No. of Decision Rules
I (Melanocortin)	0.87 (0.17)	0.98 (0.08)	0.000	0.000	0.89	0.92	6
II (Melanocortin)	0.86 (0.19)	0.87 (0.21)	0.000	0.000	0.92	1.0	9
III (Adrenergic)	0.81 (0.09)	0.89 (0.07)	0.000	0.000	0.88	0.91	14

[†]Accuracy and area under curve (AUC) mean are reported for the internal cross validations (CV) performed on the training sets. The standard deviation is given for each calculated mean within parentheses. The accuracy mean and AUC mean p-values are the results from the permutation validations of the training sets. Results from classification of the test sets are given as accuracy, AUC, and number of induced decision rules.

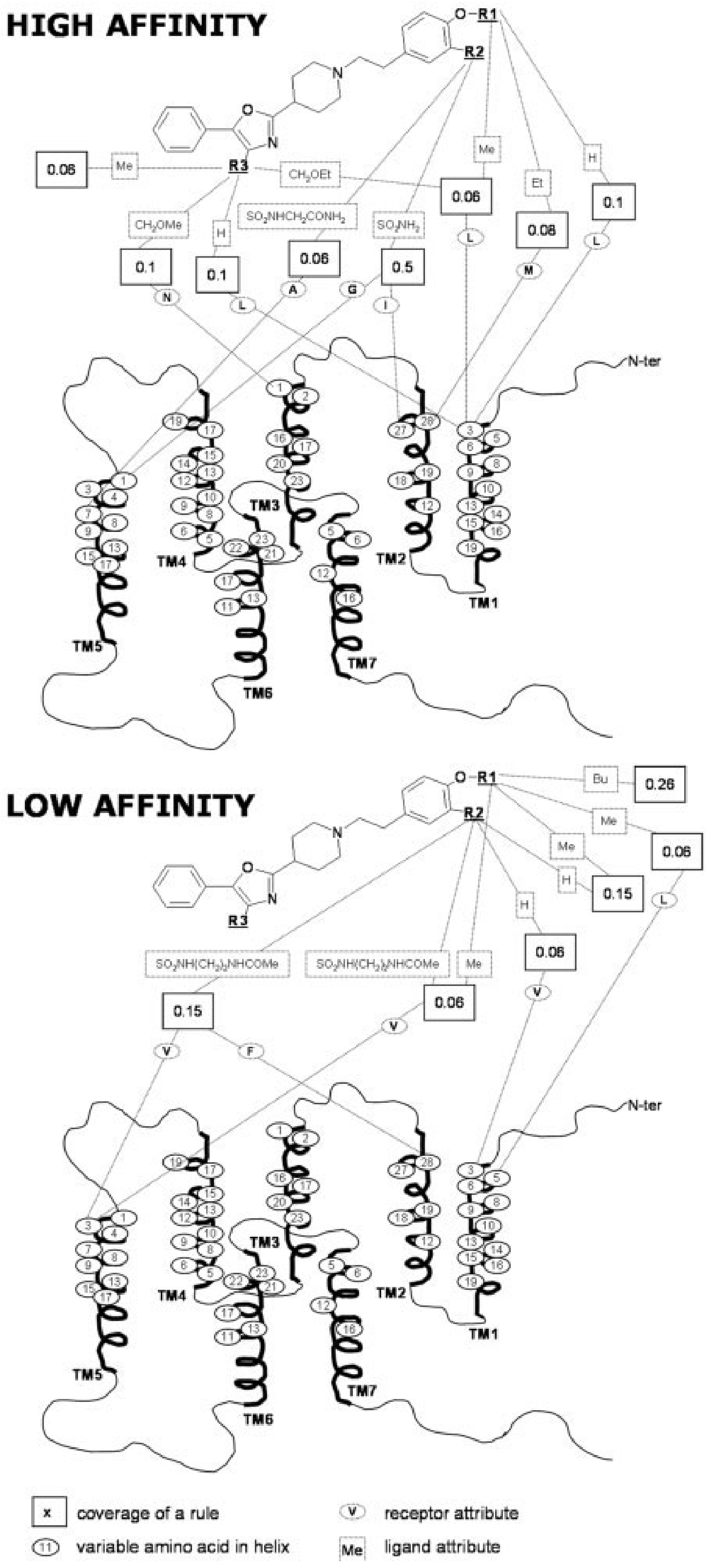


Fig. 5. Schematic representation of a minimal set of decision rules for the adrenergic Dataset III. As in Figure 3, the rules are divided into low affinity binding and high affinity binding rules and each rule is centered round a “coverage box” (c.f. legend to Fig. 3). The variable positions in the ligands are denoted R1–R3. The lower parts of the transmembrane helices, TM1–TM7, are pointing towards the cytosol. The variable amino acids are numbered according to the standard GPCR numbering.[27] For example the 0.5 coverage box in the upper part of the figure is associating the ligand attribute SO_2NH_2 at R2 with an isoleucine at position 27 of TM2 and a glycine at position 1 of TM5 illustrating the rule: “**If** R2(SO_2NH_2) **and** TM2_pos27(I) **and** TM5_pos1(G) **then** Binding(High)”.

low binding. The third rule states that there should be a butyl (Bu) group at R1.

The high affinity binding rules include the receptor helixes TM1, TM2, TM3, and TM5 and the ligand variable sites R2 and R3. The rule set has one strong rule associating an isoleucine (I) at position 27 in TM2, a glycine (G) at position 1 in TM5, and the ligand substituent SO_2NH_2 at R2 with high binding. This indicates that the SO_2NH_2 substituent at R2 interacts both with TM2 and TM5, which are located quite far from each other in the membrane. This could possibly be due to indirect effects of helix-ligand interactions. Thus according to this set of minimal decision rules, a high affinity ligand should have an SO_2NH_2-substitutent at R2 and a R1 substituent that is relatively small in size, and a high affinity binding receptor should have an I at position 27 in TM2 and a G at position 1 in TM5 as supported by the strongest rule in the high affinity rule set.

Ranking of Attributes

RS compute minimal sets of decision rules. To obtain an overview of the influence of each single attribute, a ranking of the attributes was performed by generating a large number of approximate reducts and applying a filtering algorithm to them. The filtering of reducts resulted in 2 reducts from melanocortin Dataset I, 13 reducts from melanocortin Dataset II and 86 reducts from adrenergic Dataset III scoring 0.7 or higher accuracy on the test sets.

The ranking of attributes from the melanocortin Dataset I and II are shown in Figure 6(a,b). Part B of the receptor and the ligand sites R1 and R2 are ranked as the most important attributes, which is in agreement with the previous PLS model.[5] In melanocortin Dataset II, part D and part R5 are ranked as the most influential attributes. This is essentially in agreement with the PLS model[5] that ranks ligand attribute values at position R4 and R5 as the most important but also highly ranks part D. The new results provided by this study are as follows. Receptor part A and part C are lowly ranked in both of the melanocortin datasets, which is a strong indication that these parts are not important to determine binding affinity. Part B of the receptor has a high rank in both datasets, which reflects a general importance of this receptor part. The highly ranked single attributes are well covered by the strong rules illustrated in Figure 3. For instance, the highly ranked B and D parts of the receptors are included by several strong rules.

The results of the ranking of attributes in the adrenergic Dataset III are shown in Figure 6(c). The highest ranked attributes are the ligand variable sites R1 and R2. Site R3 has a significantly lower ranking although it is still among the strongest sites. This indicates that R1 and R2 sites are likely to be more involved in receptor-ligand interaction than the variable site R3. The highest ranked receptor attributes are position 3 in TM5 and position 27 in TM2, which is in accordance with the previous PLS model.[6] This study shows additionally the following: TM2 is included in the highly ranked part B in melanocortin Dataset I and II, and position 27 is one of the most highly ranked receptor attributes in adrenergic Dataset III. This suggests that TM2 is of general importance for the G-protein-coupled receptor-ligand interactions in this study. Position 27 in TM2 is also included by strong decision rules (Fig. 5), which confirms its influence on the model. The ranking of receptor parts in the melanocortin Dataset I and II suggest that part C (TM4 and TM5) is of little importance. In the ranking of the adrenergic Dataset III, the importance of TM5 is supported by position 3 in the helix. This indicates that the ligands included in the adrenergic Dataset III also interact with TM5. The low ranking of receptor attributes in TM4 suggest that this helix is of less importance for receptor-ligand interaction in adrenergic Dataset III.

Rough Sets vs. Partial Least Squares

When comparing the induced RS models to the recently reported PLS models,[5–7] it is possible to conclude that the models provide different benefits. PLS predict real values, whereas RS predict discrete decision classes. The output from previously induced PLS models[5–7] is thus directly comparable with the original data. RS modeling is dependent on distinct decision classes. In this study the real binding affinity values were discretized into high- and low-affinity binding, resulting in a coarser classification. However, the interpretability of RS models is higher than of the PLS models. RS do not require numerical encoding of attribute values, which facilitates the description of receptor-ligand complexes and the interpretation of the models. For example, the amino acids in the adrenergic Dataset III were described by the one-letter amino acid codes, whereas in PLS modeling each amino acid was described by a sequence of real values that to some extent complicates interpretation. Moreover, the RS algorithm does not require the formulation of nonlinear terms (cross-terms). Using PLS, all possible binary cross terms have to be defined before model induction. Normally, not more than two attributes are combined. Because cross terms influence the model to the same degree as single attributes, a large number of cross terms obscures the importance of each single attribute. In the RS approach cross terms do not have to be predefined since a set of minimal decision rules is equivalent to indispensable cross terms and without any restriction on the number of attributes combined. The PLS algorithm ranks the attributes and cross terms from most influential to least influential on binding affinity. RS does not explicitly rank attributes. However, the occurrence of each attribute present in high scoring approximate reducts computed by the genetic algorithm provides an indirect ranking of attributes. Although the ranking of each single attribute is very useful for model validation and comparison with PLS, it should be emphasized that the rule sets provide a deeper understanding of receptor-ligand interaction. The main reason for this is that the rules define minimal combinations of attributes. The rules are ordered by the coverage value, which gives an assessment of their importance. Moreover, the rule sets are divided into high and low affinity binding rules, pointing out receptor and ligand attributes important to the determination of binding affinity. Perhaps the most important result of this work is that we are able to obtain combinations of biologically important receptor-ligand attributes. These are by far more interesting to

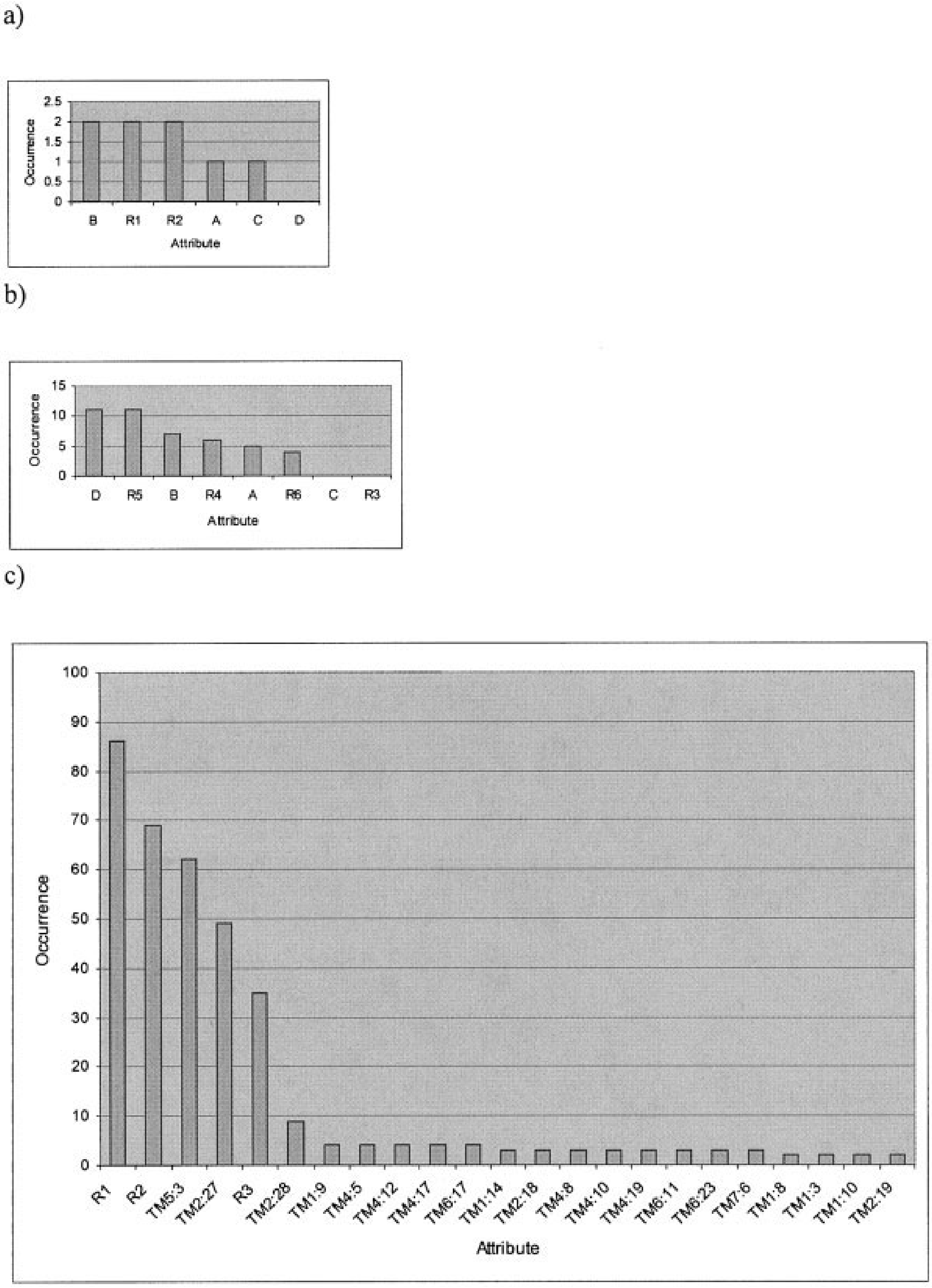


Fig. 6. Ranking of attributes in melanocortin Dataset I (a), melanocortin Dataset II (b), and adrenergic Dataset III (c). Reducts resulting in an accuracy of more than 0.7 are included in the ranking. Occurrence is the number of times a certain attribute occurs in the set of approximate reducts reported for each dataset.

investigate than each attribute alone, because they provide a more holistic approach to understanding the biological nature of G-protein-coupled receptor-ligand interactions.

CONCLUSIONS

A new methodology for modeling a class of receptor-ligand interactions has been introduced. It is based on rough sets and has been validated on three datasets containing G-protein-coupled receptor-ligand complexes and their binding affinities. The models are minimal in terms of used combinations of attributes, robust and of high quality. Sets of minimal decision rules allow a more holistic approach to understanding and exploring the biological nature of the interactions important to binding affinity. A comparison between PLS and RS shows that the two approaches have different strengths and weaknesses. PLS models, being multivariate regression models, are suitable for predicting the numerical values of binding affinities. On the other hand, the RS models are highly interpretable and provide a combinatorial view of receptor-ligand properties important to binding interactions. Moreover, because the overall interpretation to a high degree corresponds with the PLS models, the RS discretization does not seem to affect the modeling. Thus the two approaches are in several respects complementary and may be used in combination so that a better understanding of receptor-ligand interaction may be obtained. We believe that the RS methodology described here is likely to be applicable to model and interpret receptor-ligand interactions taking place in other protein families.

ACKNOWLEDGMENTS

We thank Dr. Andreas Strömbergsson at the Department of Mathematics at Uppsala University for discussions of rough sets and Robin Andersson, Jakub Orzechowski Westholm, Dr. Torgeir Hvidsten and Hugo de Boer at the Linnaeus Centre for Bioinformatics for help with ROSETTA and proofreading of the manuscript. A preliminary account of this work was presented at The German Conference on Bioinformatics, 2004.[29]

REFERENCES

1. Klabunde T, Hessler G. Drug design strategies for targeting G-protein-coupled receptors. Chembiochem 2002;3(10):928–944.
2. Bajorath J. Integration of virtual and high-throughput screening. Nat Rev Drug Discov 2002;1:882–894.
3. Cavasotto CN, Orry AJ, Abagyan RA. Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. Proteins 2003;51(3):423–433.
4. Varady J, Wu X, Fang X, Min J, Hu Z, Levant B, Wang S. Molecular modeling of the three-dimensional structure of dopamine 3 (D3) subtype receptor: discovery of novel and potent D3 ligands through a hybrid pharmacophore- and structure-based database searching approach. J Med Chem 2003;46(21):4377–4392.
5. Prusis P, Mucaniece R, Andersson P, Post C, Lundstedt T, Wikberg JES. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. Biochim Biophys Acta 2001;1544(1-2):350–357.
6. Lapinsh M, Prusis P, Gutcaits A, Lundstedt T, Wikberg JES. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. Biochim Biophys Acta 2001;1525:180–190.
7. Prusis P, Lundstedt T, Wikberg JES. Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. Protein Eng 2002;4:305–311.
8. Geladi P, Kowalski BR. Partial least-square regression: a tutorial. Anal Chim Acta 1986;185:1–17.
9. Pawlak Z. Rough sets. Int J Comp Inf Sci 1982;11:341–356.
10. Pawlak Z. Rough sets—theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic Publishers; 1991.
11. Hvidsten TR, Kryshtafovych A, Komorowski J, Fidelis K. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. Bioinformatics 2003;19:1116–1123.
12. Lægreid A, Hvidsten T, Midelfart H, Komorowski J, Sandvik AK. Predicting gene ontology biological process from temporal gene expression patterns. Genome Res 2003;13(5):965–979.
13. Hvidsten TR, Wilczynski B, Kryshtafovych A, Tiuryn J, Komorowski J, Fidelis K. Discovering regulatory binding site modules using rule-based learning. Genome Res 2005;15(6):856–866.
14. Schioth HB, Yook R, Muceniece R, Wikberg JES, Szardenings M. Chimeric melanocortin MC1 and MC3 receptors: identification of domains participating in binding of melanocyte-stimulating hormone peptides. Mol Pharmacol 1998;54(1):154–161.
15. Szardenings M, Törnroth R, Mucaniece R, Keinänen A, Kuusinen A, Wikberg JES. Phage display selection on whole cells yields a peptide specific for melanocortin receptor 1. J Biol Chem 1997;272: 27943–27948.
16. Muceniece R, Mutule I, Mutulis F, Prusis P, Szardenings M, Wikberg JES. Detection of regions in the MC1 receptor of importance for the selectivity of the MC1 receptor super-selective MS04/MS05 peptides. Biochim Biophys Acta 2001;1544(278–282).
17. Hamaguchi N, True TA, Goetz AS, Stouffer MJ, Lybrand TP, Jeffs PW. α_1-Adrenergic receptor subtype determinants for 4-piperidyl oxazole antagonists. Biochem 1998;37(16):5730–5737.
18. Wikberg JES. Melanocortin receptors: perspectives for novel drugs. Eur J Pharmacology 1999;375:295–310.
19. Piascik MT, Perez DM. Alpha1-adrenergic receptors: new insights and directions. J Pharmacol Exp Ther 2001;298(2):403–410.
20. Komorowski J, Pawlak Z, Polkowski L, Skowron A. Rough sets—a tutorial. In: Pal SK, Skowron A, editors. Rough-fuzzy hybridization—A new trend in decision making. Singapore: Springer-Verlag; 1999. p 3–98.
21. Øhrn A, Komorowski J, Skowron A, Synak P. The design and implementation of a knowledge discovery toolkit based on rough sets: The ROSETTA system. In: Polkowski L, Skowron A, editors. Rough sets in knowledge discovery 1: methodology and applications. Volume 18, Studies in Fuzziness and Soft Computing. Heidelberg: Physica-Verlag; 1998. p 376–399.
22. Holland JH. Outline for a logical theory of adaptive systems. J ACM 1962;9:297–314.
23. Johnson DS. Approximation algorithms for combinatorial problems. Austin, TX: ACM Press; 1973
24. Hacid MH, Leger A, Rey C, Toumani F. An algorithm and a prototype for the dynamic discovery of e-services. LIMOS Technical Report; 2003.
25. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143: 29-36.
26. Horn F, Weare J, Beukers MW, Horsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G. GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res 1998; 26(1):275–279.
27. Baldwin JM, Schertler GFX, Unger VM. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J Mol Biol 1997;272(1):144–164.
28. Prusis P, Schioth HB, Muceniece R, Herzyk P, Afshar M, Hubbard RE, Wikberg JE. 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 Graph Model 1997;15(5):307–313.
29. Strömbergsson H, Prusis P, Midelfart H, Wikberg JES, Komorowski J. Proteochemometrics modeling of receptor ligand interactions using rough sets. In: Giegerich R, Stoye J, editors. Proceedings of the German conference on Bioinformatics; 2004 October 4–6; Bielefeld, Germany. GI. (Proceedings of the German Conference on Bioinformatics).