Proteochemometrics Modeling of Receptor-Ligand Interactions Using Rough Sets

Helena Strömbergsson (1, 2), Peteris Prusis (1, 2), Herman Midelfart (1)
Jarl E. S. Wikberg (2) and Jan Komorowski(1)

(1) The Linnaeus Centre for Bioinformatics
Uppsala University
Box 598
SE-751 24 Uppsala
SWEDEN
helena.strombergsson@lcb.uu.se
herman.midelfart@lcb.uu.se
jan.komorowski@lcb.uu.se

(2) Department of Pharmaceutical Pharmacology
Uppsala University
Box 591
SE-751 24 Uppsala
SWEDEN
peteris.prusis@farmbio.uu.se
jarl.wikberg@farmbio.uu.se

Abstract: We report on a model for the interaction of chimeric melanocortin G-protein coupled receptors with peptide ligands using the rough set approach. Rough sets generate If-Then rule models using Boolean reasoning. Two separate datasets have been analyzed, for which the binding affinities have previously been measured experimentally. The receptors and ligands are described by vectors of strings. Different partitions of each dataset were evaluated in order to find an optimal partition into rough set decision classes. To obtain a measurement of the accuracy of the rough set classifier generated from each dataset, a 10-fold cross validation (CV) was performed. The Area Under Curve (AUC) was calculated for each iteration during CV. This resulted in an AUC mean of 0.94 (SD 0.12) and 0.93 (SD 0.16) for the first and second dataset respectively. The CV results show that the rough set models exhibit a high classification quality. The decision rules generated from the rough set model inductions are easy to interpret. We apply this information to develop models of the interaction between ligands and receptors.

1 Introduction

The melanocortins and adrenocorticotropins [WI99] are involved in a diverse number of physiological functions. The melanocortin system consists of some peptide hormones and the melanocortin receptors. Melanocortins include melanocyte stimulating hormones (MSHs) and adrenocorticotropic hormone (ACTH). The receptors for these are termed melanocortin receptors and belong to the G-protein-coupled receptor family. Five such receptors (MC_1 through to MC_5) have been identified and most of these show different binding affinities for each of the melanocortin hormones.

Recently, we have developed proteochemometrics to model MC receptor-ligand interactions [PLW02, Pr01]. In proteochemometrics, descriptors of proteins and ligands are combined and correlated with binding affinity data. The linear method partial least squares (PLS) [GK86] was applied very successfully for the correlation. We have here evaluated rough sets [Pa82, Pa91], which is a Boolean method suited to investigate non-linear phenomena to find out whether or not this approach adds information and/or gives new perspectives on MC receptor-ligand interactions.

2 Materials and Methods

2.1 Datasets

Models were induced from two datasets describing the interaction between chimeric melanocortin receptors and peptide ligands. The binding affinities of highly selective melanocortin peptides for chimeras of the MC_1 and MC_3 receptors were reported [Mu01, Sh98]. (The same types of receptor chimeras were used in both datasets.) The chimeras are composed of four parts (**A**, **B**, **C** and **D**) (Fig. 2.1.1), each originating either from wild-type MC_1 or MC_3. The composition of each receptor is described by a vector whose elements are four strings. Thus the string "MC1" is assigned to parts originating from MC_1 and the string "MC3" 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 is described by the vector [MC1, MC1, MC3, MC3]. The wild-type receptors MC_1 or MC_3 are described by the vectors [MC1, MC1, MC1, MC1] and [MC3, MC3, MC3, MC3], respectively.

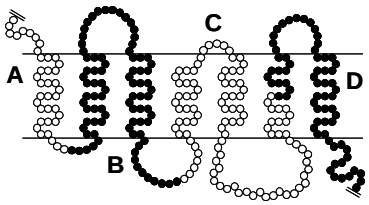

Figure 2.1.1: Schematic overview of the division of MC-receptors into parts.

The first dataset contains 40 receptor-ligand complexes. Each object of the dataset is a description of the composition of MC receptor chimeras and **α-MSH**, **MS04** peptides or chimeric analogues of these peptides together with the receptor-ligand binding affinity. The peptide **α-MSH** is a natural ligand while the **MS04** peptide is a result from phage display selection [Sz97]. In the dataset, the middle part of the ligand is constant while the N-terminal and C-terminal parts originate either from **α-MSH** or **MS04**. The composition of each ligand in the dataset is described by a vector of two strings. The string "MSH" is assigned if the part originates from **α-MSH** and the string "MS04" if the part originates from **MS04** (Table 2.1.1). Each receptor-ligand complex is 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 which is the negative logarithm of the inhibition constant K_i. The pK_i values are taken from [Mu01]. For example, the object [MC3,MC3,MC1,MC1,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.

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

Table 2.1.1: Sequences of the highly active melanocortin α-MSH/MS04 peptides used in the first dataset. 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- and C-terminal parts are variable. The two strings of each descriptor represent the origin of the N- and C-terminal part.

The second dataset contains 60 receptor-ligand complexes. Each dataset describes the composition of MC receptor chimeras and six linear and cyclic peptide ligands. The ligands are various derivatives of α-MSH (Table 2.1.2) where one or several amino acids have been altered. An amino acid alignment of the ligands shows that there are four variable sites. Vectors of four strings have been used to describe the properties the ligand focusing on the amino acid composition at the variable sites. The receptors and the biding affinities are described in the same manner as the first dataset. The pK_i values are taken from [Sh98]. A receptor binding complex is hence described by a vector of eight strings and one real number where the first four strings represent the receptor, the following four strings describe the ligand and the real number is the binding affinity. For instance, the receptor ligand complex [MC1, MC1, MC3, MC3, Tyr, Met, Phe, Gly, 8.27] has part **A** and **B** of the receptor from MC_1 and part **C** and **D** from MC_3, the ligand has the residues **Tyr**, **Met**, **Phe** and **Gly** at the variable sites, and the binding affinity is 8.27.

Peptide	Sequence													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]

Table 2.1.1: Amino acid sequence of the peptides used in the second dataset and their descriptors. The shaded positions are the alteration sites.

2.2 Rough sets

For an overview of the rough set method we refer to [Ko99]. All computations were performed using the Rosetta system [Øh98] (rosetta.lcb.uu.se). Rosetta implements inductive learning using the mathematical framework of rough sets [Pa82, Pa91]. 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 $\mathcal{A} = (U, A)$ where U is a non-empty finite set of *objects* called the *universe* and A is a non-empty finite set of functions $a : U \to V_a$, called *attributes*; for each $a \in A$ the set V_a is called the *value set* of a.

If there is a known outcome 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 $\mathcal{A} = (U, A \cup \{d\})$, where $d \notin A$ is the decision attribute. Two objects x, y are said to belong to the same *decision class* if $d(x) = d(y)$.

The output of the rough set algorithms is a set of minimal *decision rules* of the form $\alpha \to \beta$. Here α is a Boolean function $U \to \{\text{true}, \text{false}\}$ 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 $\beta : U \to \{\text{false}, \text{true}\}$ is built up from atom statements of the form $d(\cdot) = v$ where $v \in V_d$. An example of a decision rule will be given in section 3.3. The extracted set of minimal decision rules is applied to classify new objects.

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

2.2 Rough sets and proteochemometrics data

Using the rough set terminology the two datasets are decision systems where the receptor-binding complexes are objects, the descriptors of the receptor-ligand complexes are attribute values and the binding affinities are decision attribute values. The first dataset will be referred to as Decision System 1 (DS1) and the second dataset will be referred to as Decision System 2 (DS2).

2.3 Model validation

The Rosetta system computes the Area Under Curve (AUC) (Hanley and McNeil, 1982) for an induced model. The AUC is the area under the Receiver Operating Characteristics (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 while an AUC of 0.5 means that the classifier does not perform better than a random classification of objects.

Rough set modeling is an inductive learning process. It begins 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. However for a small dataset, the reliability of the performance of one single partitioning can be questioned. The random division of the dataset could have been particularly "lucky" or "unlucky". In order to deal with this issue, the Rosetta system implements k-fold cross validation (CV). From a Rosetta k-fold CV, the performance estimates *accuracy mean* and *AUC mean* are obtained. The accuracy mean is the average proportion of correctly predicted objects computed for the k blocks during CV. The AUC mean computed by Rosetta is the average AUC for the models induced by the k blocks during CV. The standard deviation (SD) is reported for both accuracy- and AUC mean.

Three types of CVs were performed; 10-fold, leave-one-out and leave-one-receptor-out. In 10-fold CV the objects are randomly divided into 10 blocks. In leave-one-out CV the n objects of a decision system are divided into n blocks (each containing one single object). It is not possible to calculate AUC for each of the iterations. Instead an AUC of the overall performance is computed and the standard error (SE) of the computations is reported. In leave-one-receptor-out CV each block consist of one type of melanocortin receptor chimera. As information about one entire receptor type is left out this is a more rigid test than k-fold CV where the partitioning of blocks is at random.

2.4 Optimization of partitioning into decision classes

Both decision systems have real and continuous decision attribute values without any obvious cutoff value for discretization into decision classes. In order to find an optimal partition of the decision systems into two decision classes, rough set models were systematically induced for a number of partitions. Each decision system was sorted by the decision attribute value. For DS1 in the first iteration, the decision attribute values of

the first two objects were assigned the number “1” (low binding) and the remaining objects were assigned the number “2” (high binding). A rough set model was induced and the model was validated using 10-fold CV, for which the AUC mean was calculated. In the next iteration the following two “high binding” objects were set as “low binding” objects while the remaining objects were unchanged and the 10-fold CV performed. This was repeated until all but the two last objects of DS1 were assigned as “low binding” objects. The same procedure was repeated for DS2 with the sole difference that three objects were reassigned in each of the iterations. In all, models of 19 partitions of each datasets were induced and validated. The AUC means and standard deviations are presented in Figure 2.4.1.

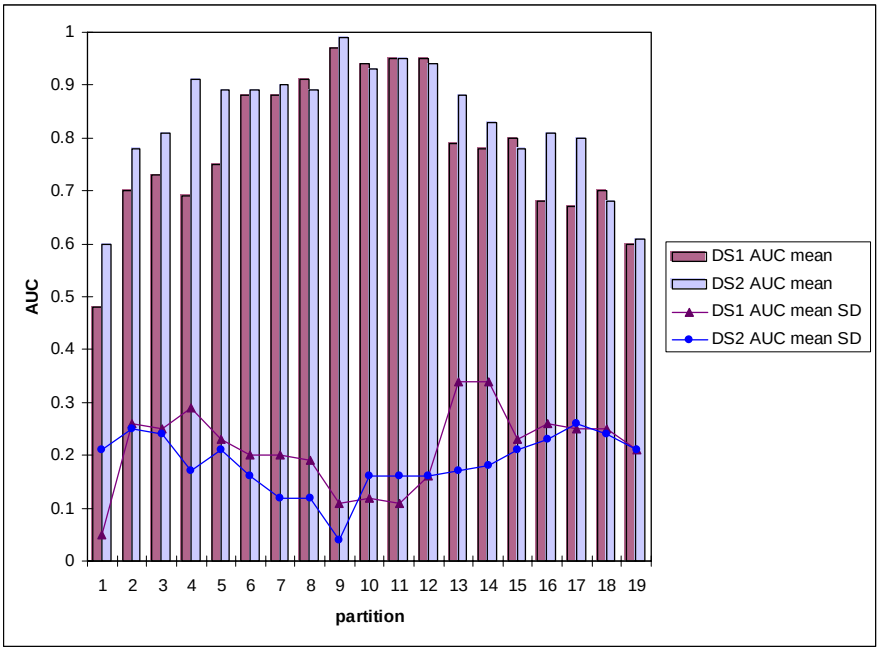

Figure 2.4.1: Optimization of partitioning of decision system 1 (DS1) and decision system 2 (DS2) into two decision classes.

2.5 Export of decision rules

From the 10-fold CV the set of minimal decision rules was exported from each induced model. The coverage of each rule was summarized and mean and standard deviation were calculated. Rules that occurred in less than 7 CV folds or had a coverage mean of less than 0.2 were omitted from interpretation.

3 Results

3.1 Optimization of decision classes

The plot (Fig. 2.4.1) of the AUC means and standard deviations for both decision systems shows that the AUC mean peaks and the standard deviations are at their lowest at partition 9. This suggests that a partitioning of objects into decision classes of equal cardinality is suitable for model induction in this particular case. The median value of each decision system was therefore used to discretize the systems into two decision classes.

3.2 Model validation

For DS1, 10-fold CV resulted in an accuracy mean of 0.9 (SD 0.13) and an AUC mean of 0.94 (SD 0.12). Leave-one-out CV produced an accuracy mean of 0.9 (SD 0.3) and an AUC of 0.94 (SE 0.04). Leave-one-receptor-out CV resulted in an accuracy mean of 0.8 (SD 0.16) and an AUC mean of 0.85 (SD 0.21). For DS2, 10-fold CV resulted in an accuracy mean of 0.94 (SD 0.12) and an AUC mean of 0.91 (SD 0.19). Leave-one-out CV produced an accuracy mean of 0.93 (SD 0.25) and an AUC of 0.98 (SE 0.05). Leave-one-receptor-out CV resulted in an accuracy mean of AUC of 0.85 (SD 0.15) and an AUC mean of 1.0 (SD 0.0).

3.3 Interpretation of rules

The cross validations results show that the rough set models exhibit a high classification quality. Classification produced by a model is a direct consequence of the decision rules. In this study, the set of minimal decision rules associates a minimal number of receptor-ligand descriptors with high or low binding affinity. It is therefore of biological and biochemical interest to study the set of rules exported from cross validations to discover patterns determining binding affinity. An example of a decision rule is "A(MC1) $\wedge$ D(MC1) $\wedge$ pos4_ligand(G) $\rightarrow$ Binding(2)", associating the **A** and **D** part of MC_1 and a glycine (**G**) at position 4 of a ligand with high binding affinity. Illustrations of high and low binding rules are shown in Figure 3.3.1.

The "high binding" decision rules are always associated with receptor parts from MC_1. Within a receptor, a combination of part **B** and **D** results in a high binding. Within ligands, a high binding affinity is promoted by the N- and C-terminal part of **α-MSH** and a combination of the residues **Nle** and **D-Phe**. Between ligands and receptors chimeras, a high binding affinity is obtained by a combination of the N-terminal part of **α-MSH** and part **B**, a combination of a **G** residue and part **A** and **D,** and a combination of a **Nle** residue and part **B** and **D**. Receptor parts from MC_3 are frequently associated with "low binding" decision rules. Within ligands, low binding is affected by a combination of **Cys** at position four and a **F** residue at position three. Between ligands and receptor chimeras, low binding is promoted by a combination of **Cys-Cys** bridge and part **B** and **D**, part **A** in combination with the N- or C-terminal part of **MS04**.

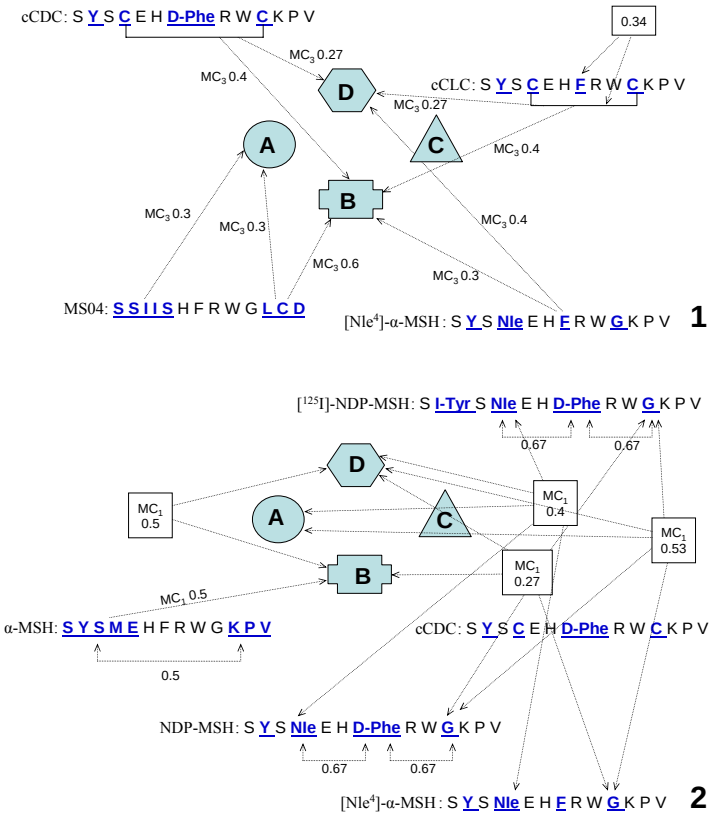

Figure 3.3.1: (1) Interpretation of low affinity decision rules, (2) interpretation of high affinity decision rules. Each arrow or combination of arrows represents one decision rule. For each decision rule we state the wild-type receptor associated with the rule and the RHS coverage mean for the rule.

4 Discussion

Rough set models have been induced and validated from two datasets containing binary descriptors of melanocortin receptor-ligand complexes and their respective binding affinity. In this context, the set of minimal decision rules can be used to understand and explore the biochemical nature of the interactions important for binding affinity. The decision rules suggest that part **A**, **B** and **D** of the receptor are involved in the interactions while part **C** seems to be of little significance. Within the ligands the first variable position of the DS2 ligands does not seem affect binding affinity. Receptor parts from MC_1 are involved in high binding affinity interactions while parts from MC_3 are involved in low binding affinity interactions suggesting that the peptides used in this study in general bind more weakly to MC_3 parts than to MC_1 parts. The rough set decision rules suggest that a high binding affinity may be achieved by an interaction

between receptor parts **B** and **D**. The rules also suggest that the N-terminal part of **α-MSH** interacts with part **B**. A high binding affinity may be promoted by a **Nle** residue in the peptide interacting with part **A** and **D,** and a **G** residue in the peptide interacting with part **A** and **D** or part **B** and **D**. By counting the number of interactions suggested by the rules it is possible to suggest that part **B** and **D** of the receptor are the most important for high binding affinity. Low binding affinity is achieved by the presence of a **Cys-Cys** bridge in the peptide supposedly interacting with parts **B** and **D**. A **F** residue in the peptide interacting with part **B** or **D** may cause low binding. Low binding affinity may also be caused by a possible interaction between the N-terminal part of **MS04** with part **A** and an interaction between the C-terminal part of **MS04** with part **A** and **B**. By counting the number of interactions of each part of the receptor it can be assumed that the receptor **B** and **D** parts are the most influential on "low affinity" interaction.

When comparing the induced rough set models to the recently reported PLS models [PLW02, Pr01], it is possible to conclude that the models have different strengths. PLS ranks all the attributes and cross terms from most influential to least influential for binding affinity. For instance, both PLS models single out part **B** of the receptor as the most important for binding affinity. Rough set modeling does not produce a ranking of attributes. Instead it selects minimal groups of essential attributes that have the same classification power as the full set of attributes. In this case, decision rules focus on combination of attributes important for binding, which from a biological point of view is desirable. One major advantage with the rough set approach is that it is independent of numbers as attribute values which facilitates the description of receptor-ligand complexes and the interpretation of the model. Moreover, the rough set decision rules are more specific as they associate high and low binding to certain attribute values while the PLS ranking is not relating to attribute value. For linear models such as PLS, non-linear terms (cross terms) have to be defined before building a model and usually not more than two attributes can be combined in each cross terms. In the rough set approach the set of minimal decision rules is equivalent to cross terms without any restriction to number of attributes combined. For instance, the rule "B(MC1) $\wedge$ D(MC1) $\wedge$ pos2_ligand(Nle)" combines the attributes **B**, **D** and pos2_ligand thus adding information about the interaction that could not have been generated by PLS. One disadvantage with the rough sets approach is that it is dependent on distinct decision classes which makes it necessary to discretize the binding affinity values into "high" and "low" binding, while PLS can deal with and predict real binding affinity values. However, as the rough set models of receptor-ligand interactions are essentially in agreement with the PLS models the discretization does not appear to be harmful in this particular study. In conclusion, rough sets provide an explicit model with parameters (i.e. attributes) that are easy to interpret. On the other hand PLS models, being multivariate regression models are better at predicting the numerical values of binding affinity. Thus the two approaches are in some respects complementary and may be used in combination to receive a better understanding of receptor-ligand interaction.

To our knowledge these are the first rough set models induced on proteochemometrics data. We have shown that the rough set approach can be used as a modeling tool and that the resulting model agrees with and in some respects complements previously induced

PLS models. Additionally, we have proposed a novel approach to handling continuous decision attributes in rough sets. In the future, we would like to add some features to the rough set approach to select optimal boundaries for decision classes, a feature that will be useful when building models from larger dataset.

Acknowledgements

We would like to thank Dr. Andreas Strömbergsson at the Department of Mathematics at Uppsala University for discussions on rough sets and Torgeir Hvidsten at the Linnaeus Centre for Bioinformatics for help with Rosetta. This research was supported by grants from the Knut and Alice Wallenberg Foundation and the Swedish VR(621-2002-4711).

References

[GK86] Geladi, P.; Kowalski, B. R.: Partial least-square regression, A tutorial. Anal. Chim. Acta. 1986; pp. 1-17.

[HM82] Hanley, J. A.; McNeil, B. J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; pp. 29-36.

[Ko99] Komorowski, J.; Pawlak, Z.; Polkowski, L.; Skowron, A.: Rough sets - a tutorial. In (S. K. Pal, and A. Skowron, Eds.) Rough-fuzzy hybridization - A new trend in decision making. Springer Verlag, Singapore, 1999; pp. 3-98.

[Mu01] Mucaniece, R.; Mutule, I.; Mutulis, F.; Prusis, P.; Szardenings, M.; Wikberg, J. E. S.: 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; pp. 278-282.

[Øh98] Øhrn, A.; Komorowski, J.; Skowron, A.; Synak, P.: The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets, The ROSETTA System: In (L. Polkowski, A. Skowron Eds.) Rough Sets in Knowledge Discovery 1, methodology and applications. Studies in Fuzziness and Soft Computing. Physica-Verlag, Heidelberg, 1998; Vol. 19, pp. 572-576.

[Pa82] Pawlak, Z.: Rough Sets. Int. J. of Comp. Inf. Sci. 1982; pp. 341-356.

[Pa91] Pawlak, Z.: Rough Sets - Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, 1991.

[PLW02] Prusis, P.; Lundstedt, T.; Wikberg, J. E. S.: Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. Protein. Eng. 2002; pp. 305-311.

[Pr01] Prusis, P.; Mucaniece, R.; Andersson, P.; Post, C., Lundstedt, T.; Wikberg, J. E. S.: 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; pp. 350-357.

[Sh98] Shioth, H. B.; Yook, R.; Mucaniece, R.; Wikberg, J. E. S.; Szardenings, M.: Chimeric melanocortin MC1 and MC3 receptors, identification of domains participating in binding of melanocyte-stimulating hormone peptides. Mol. Pharmacol. 1998; pp. 154-161.

[Sz97] Szardenings, M.; Törnroth, R.; Mucaniece, R.; Keinänen, A.; Kuusinen, A.; Wikberg, J. E. S.: Phage display selection on whole cells yields a peptide specific for melanocortin receptor 1. J. Biol. Chem. 1997; pp. 27943-27948.

[WI99] Wikberg, J. E. S.: Melanocortin receptors, perspectives for novel drugs. Eur. J. Pharmacology. 1999; pp. 295-310.