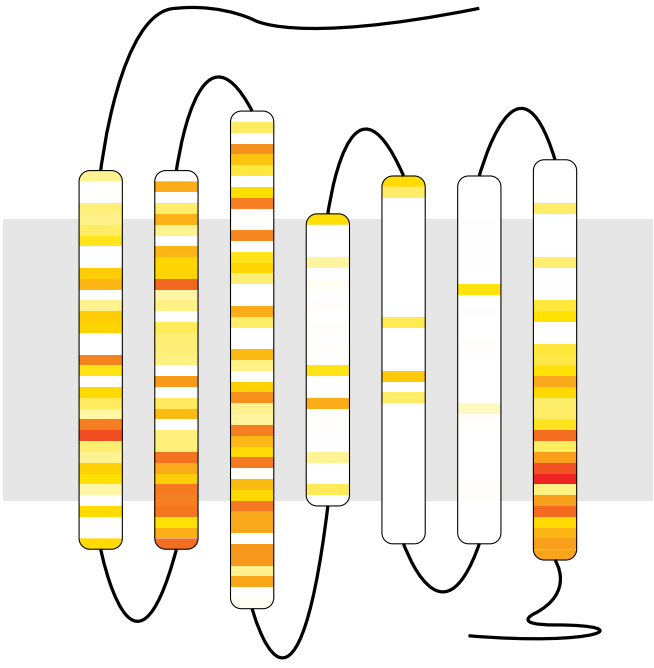
Uncovering the Classification Characteristics of Olfactory G Protein-Coupled Receptors

B.M. Groenendijk

The Leiden Institute of Advanced Computer Science (LIACS)
University of Leiden, The Netherlands

Master's Thesis

March 26, 2007



Mentors: Prof.dr. Th. Bäck, LIACS[1] and Prof.dr. J.N. Kok, LIACS
Tutor: Drs. E.V. Samsonova, LACDR[2]/LIACS.

[1]LIACS: The Leiden Institute of Advanced Computer Science.
[2]LACDR: Leiden / Amsterdam Center for Drug Research.

Abstract

The human sense of smell is made possible through the use of the so-called olfactory receptors. These receptors are proteins which are embedded into the cellular membranes located in the nose. The olfactory receptors are members of the G protein-coupled receptors (GPCRs) class. The function of a GPCR is to transmit a signal into the cell structure when a molecule binds to the GPCR on the outside of the cellular membrane.

The olfactory receptors are mostly found inside the nose. However a number of these receptors are also found on other locations. For this reason the olfactory receptors may also be involved in functions other than making the sense of smell possible.

Generally speaking, GPCRs are already essential in drug discovery. Because of this the pharmacological importance to map the hidden function of the olfactory receptors may introduce new drug targets.

The paper by Samsonova et al. proposed a rule-based characterisation method for classifying olfactory receptors using multiple sequence alignment (MSA). In a MSA the olfactory receptor sequences can be compared with each other. This thesis continues the effort of finding characteristic amino acid combinations in olfactory GPCRs. Two existing rule discovery algorithms, PRISM and Tertius, are used in combination with two optimisation methods, *rule set optimisation* and *rule subset generation.*

The combined methods found highly supported rule sets which classify the olfactory receptor class very successfully. Moreover these rule sets cover multiple highly characteristic features in a non-sequential manner throughout the transmembrane (TM) domains of the olfactory receptors. In contrast with the well known motifs, which uses sequences of conserved amino acids in chronological order to classify a receptor class.

The used characterisation methods in this thesis could be used to examine the characteristics of other GPCRs. Furthermore these method could use other alignment methods such as a structural alignment. This could uncover important characterisations within the 3D structure which may lead to a better understanding of the functions of the olfactory receptor.

Contents

1 Introduction

The human sense of smell is made possible through the use of the so-called olfactory receptors. These receptors are proteins which are embedded into the cellular membranes located in the nose. The olfactory receptors are members of the GPCRs class. The general structure of the olfactory GPCRs are described in Section 2.1.

After a deeper analysis and identification of the olfactory receptors' functions [13, 69], it became clear that the main part of all G protein-coupled receptors (GPCRs) of the human genome consists of olfactory genes. The olfactory GPCRs are mainly used for odorant signal transduction [25], making the sensing of smell possible. However a number of olfactory receptors are also found on other locations, such as the heart [19] or sperm, testis [21] and various brain regions [65]. For this reason, the olfactory receptors may also be involved in functions other than making the sense of smell possible.

Generally speaking, GPCRs are already essential in drug discovery [47]. Because of this the pharmacological importance to map the hidden function of the olfactory receptors may introduce new drug targets. Therefore, we set out to determine the characteristics of the olfactory receptors. These characterisations could lead to a better understanding of the structure which may reveal the inner mechanism of the olfactory receptors.

Today, research using sequence analysis is an increasing source of information. In sequence analysis a sequence of elements is used as input data which is analysed with a particular goal in mind. The sequence analysis using amino acids for example has been a valuable resource on a variety of subjects related to GPCRs [50, 51, 66, 7, 67]. Besides amino acids, sequence analysis using RNA [36] and DNA [73] are also used to classify these sequences into classes.

When analysing the sequence of the olfactory GPCRs, the hidden function may become clear through the finding of its classification characteristics. Classification characteristics are elements which not only characterise the olfactory receptors but are also used to classify the olfactory receptor class. For example these characterisations could be used in specially designed biological experiments or in theoretical computational models to uncover the true structure and function of these receptors.

In the paper by Samsonova et al. [63], sequence analysis of the olfactory GPCRs combined with rule discovery is used to find characteristic amino acid combinations. Rule discovery uses rules to classify the olfactory GPCRs using the aligned sequences as input data. The composition of such rules is described in Section 2.2. This thesis continues the effort of finding characteristic amino acid combinations using two alternative rule discovery algorithms.

Rule discovery is still a very active research area in computer science and is used in

a variety of subjects such as knowledge discovery using cardiovascular and heart disease data [59, 52] or the discovery of transcription factors [58]. Rule discovery is also used in Samsonova et al. [63] where the method is based on fuzzy logic [32]. This method is rather complex for the task performed, therefore two alternative methods will be used in this thesis:

- First a simple algorithm was chosen in an attempt to gain equal results compared to the results found in [63]. The Apriori [1] algorithm is usually well known for rule discovery, related to databases. Nevertheless this association rule algorithm induces rules not only with the classification class but also with other attributes. Variations on association rule algorithms have also been considered to include constraining certain attributes, like the classification class, as described in [29]. However there is a crucial difference between classification and discovery of association rules as discussed in [24]. Therefore the PRISM [8] algorithm which uses a different approach is chosen. PRISM, still being a simple algorithm, is preferred because it is designed as a classifier. The methods and results are described in Section 3.

- Second an algorithm called Tertius [22] is chosen. Tertius uses a different approach, which discovers rules with the best classification quality. The used method is more complex compared to PRISM. The methods and results are described in Section 4.

Besides these algorithms two additional methods where developed, Rule set optimisation and Subset generation, which are used to optimise and enhance the results of PRISM and Tertius. These methods use different approaches for selecting characteristic amino acid combinations to find multiple classification characteristics for the olfactory receptors.

The contents of the rest of the thesis are as follows. After a general approach section the methods and results regarding the two algorithms, PRISM and Tertius, are presented in Section 3 and Section 4 respectively. Next, the findings are discussed in Section 5 including some future thoughts. The conclusion of this thesis can be found in Section 6.

2 General approach

In this section we describe the general approach. The contents of this section are as follows. Rule discovery use a training set as input, this is defined in the next subsection. The rule discovery algorithm returns a set of rules with a certain classification quality. The rule format, rule set and classification method are defined in Section 2.2. In Section 2.3 three quality measures are defined which are used by the rule set optimisation and subset generation methods. In addition, the developed algorithms were tested using cross-validation which is defined in Section 2.4. This section ends with a summary of the developed software and a short discussion.

2.1 Training set

Rule discovery algorithms use a training set as input data to generate rules which can classify items from the training set. The data set used throughout this thesis consists of olfactory and non-olfactory GPCRs and was also the training set used in [63]. The data set will be described in more detail in the next paragraph. The data set is constructed using only the seven transmembrane (TM) regions of the multiple sequence alignment (MSA). A short introduction about G protein-coupled receptor (GPCR) structure and sequence alignment are described in two paragraphs later on followed by a formal definition of the training set. The data set normally uses the individual amino acid model. However in two additional experiments the group model has been used. Both models are described in the last paragraph of this section.

Composition of the data set The data set contains olfactory and non-olfactory GPCRs from GPCRDB, release February 2004 [30]. GPCRDB is Molecular-Specific Information System for GPCRs, a database which contains the sequences of most GPCRs. Table 1 indicates the composition of each class. Orphan GPCRs do not belong to any class and thus the 'Others' group is composed of non-orphan GPCRs. Various classes are added to the non-olfactory class to have approximately an equal amount of GPCRs compared to the olfactory class. This is necessary to find good classification criteria.

Class	Number	GPCR Class
olfactory	391	Class A (olfactory)
non-olfactory	369	Class A (non-olfactory) Class B Class C Others (non-orphan)

Table 1: Composition of olfactory and non-olfactory classes.

GPCR structure All proteins are composed of amino acids, where the sequence determines its structure. Nevertheless the true 3D structure cannot be derived from the amino acid sequence yet and remains unknown. The amino acid names and the abbreviations can be found in Table 2.

The GPCR is embedded in the cellular membrane and functions as a receptor which binds a particular molecule called a ligand. A receptor is a chemical agent that acts like a sensor and response to stimulation. When a ligand binds to a GPCR, it will transmits a signal through the use of the G protein into the cell structure. This complex system is involved in numerous processes such as the sense of smell, behavioural and mood regulation, inter-cellular communication between cells of the immune system and many automatic functions of the body such as digestive processes, heart rate and blood pressure.

It is believed that GPCRs have approximately the same structures; seven α-helices representing the transmembrane (TM) domains. This model is derived from the only known 3D structure of bovine rhodopsin [54, 49] which is shown in Figure 1. The seven spiral loops represent the α-helices located inside the membrane. These helices are connected with loops which are located outside the membrane. Through the use of the rhodopsin model, GPCRs can be compared using their sequences.

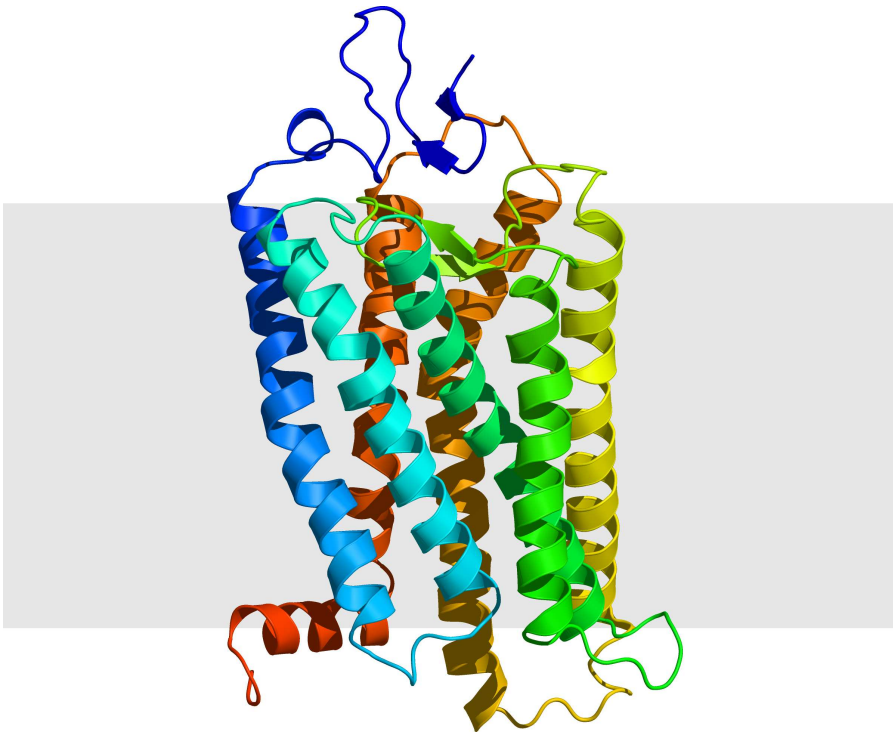


Figure 1: This molecule model represents the structure derived from the crystal structure of bovine rhodopsin [54, 49]. The molecule is embedded in the cellular membrane which is represented by the grey area. The illustration was generated using the PyMOL Molecular Graphics System [17] and POV-Ray [70].

GPCR alignment The GPCRs in GPCRDB [30] are aligned in a multiple sequence alignment (MSA) [9, 71], minimising both misalignments and the number of introduced gaps. In other words the number of amino acids that matches in the MSA is maximised. In Figure 2 part of a MSA is visualised. For example the amino acids like Phenylalanine (F), Tyrosine (Y) or Proline (P) are maximised towards a certain position. The MSA results in equally sized sequences with fixed positions which can be used for comparison.

Only the sequences of the seven transmembrane (TM) regions are used for the alignment of the data set, because these regions are most conserved. In contrast with the loops outside the TM region which are very variable in length and their sequences are not very conserved. For this reason these regions are not suitable for a good alignment.

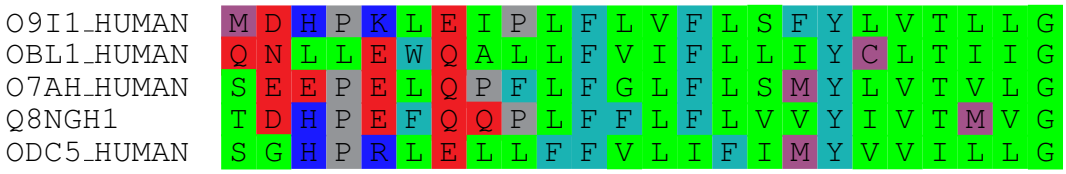


Figure 2: This visualisation represents a part of the non-olfactory/olfactory multiple sequence alignment (MSA). Each row represents a sequence of amino acids with the abbreviation of the receptors name on the left. The letters in the coloured boxes represent the amino acid abbreviations and the boxes are coloured according to the amino acid group model, which are both describe in the last paragraph of this subsection. Although the colouration has no meaning other than giving the amino acids some contrast. The alignment shows the position of the sequences which maximises the number of amino acids on the same aligned location.

Defining the training set A subset of each GPCR denoted G is composed of the seven transmembrane (TM) regions concatenated from I to VII next to each other as defined below.

$$G = \text{TM-I} \ldots \text{TM-VII}$$

After concatenation, each G contains a total of 248 TM region positions. These positions are called 'features' denoted as f. Each feature f is assigned a number according to the Ballesteros and Weinstein [4] numbering system, where the TM region is represented by the first digit x as shown here:

$$f = x.(50 \pm i)$$

The number after the dot is an index which represents the position relative to the most conserved feature. A feature number is computed using the index i as position before or after the most conserved feature with value 50. For example 1.49 is read as helix I and one position before the most conserved feature of helix I.

Training set T in (1) contains 391 olfactory and 369 non-olfactory GPCRs G as specified in Table 1. In total there are $n = 760$ GPCRs in training set T.

$$T = \{G_1, G_2, \ldots, G_n\} \tag{1}$$

Amino acid models The data set uses the individual amino acid model. This model contains the 20 standard amino acids, their names are shown in Table 2. The 'Gap' defines an empty position when aligned in a multiple sequence alignment (MSA).

Alanine	Ala	A	Lysine	Lys	K	Threonine	Thr	T
Cysteine	Cys	C	Leucine	Leu	L	Valine	Val	V
Aspartate	Asp	D	Methionine	Met	M	Tryptophan	Trp	W
Glutamate	Glu	E	Asparagine	Asn	N	Tyrosine	Tyr	Y
Phenylalanine	Phe	F	Proline	Pro	P			
Glycine	Gly	G	Glutamine	Gln	Q			
Histidine	His	H	Arginine	Arg	R			
Isoleucine	Ile	I	Serine	Ser	S	Gap	–	–

Table 2: Amino acids three-letter and one-letter abbreviations [48]. The 'Gap' is used in the multiple sequence alignment (MSA) to indicate an empty position in the alignment.

Besides the individual amino acid model the group model is used as well. In this model, groups of amino acids are used rather than individual amino acids. The individual amino acids are grouped into six sets according to their chemical properties as described by Conn and Stumpf [10]. Table 3 shows the content of the six amino acid groups, where the individual amino acid abbreviations can be found in Table 2.

1	Acidic residues and their amides	DENQ
2	Basic	HKR
3	Sulfur-containing	CM
4	Aromatic	FYW
5	Aliphatic and hydroxyl	GVLIAST
6	Cyclic imino acid	P
7	Gap	-

Table 3: Amino acids groups according to Conn and Stumpf [10]. The 'Gap' is used in the multiple sequence alignment (MSA) to indicate an empty position in the alignment.

2.2 Classification rules

Classification rules are defined as a conjunction of terms, the body, that make a logical implication towards the head as shown in the formula below, where the $\rightarrow$ represents the implication symbol.

$$\text{rule:} \quad \text{body} \quad \rightarrow \text{head}$$

The head contains one term, a GPCR class from Ω as defined below, which is the target for classification.

$$\Omega = \{\text{olfactory}, \text{non-olfactory}\}$$

The terms in the body are composed of a feature f and an amino acid $\phi \in \Phi$ or amino acid group $\psi \in \Psi$. Where Φ represents the set of amino acids (2) as described in Table 2 and Ψ the set of amino acid groups (3) as described in Table 3.

$$\Phi = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, -\} \tag{2}$$

$$\Psi = \{1, 2, 3, 4, 5, 6, 7\} \tag{3}$$

The formula for a rule r is shown below, where x is the number of terms, feature f connected towards an individual amino acid or group and c a classification class from Ω. Notice the difference between the rules with ϕ or ψ which contain individual amino acids or grouped amino acids respectively. These type of rules can be mixed into one rule set. Nevertheless in our experiments the models are strictly separated. The $\wedge$ and $\rightarrow$ represent the conjunction and implication symbol respectively.

$$r = \bigwedge^{x} (f = \phi) \rightarrow c, \qquad x > 0, \phi \in \Phi, c \in \Omega$$

$$r = \bigwedge^{x} (f = \psi) \rightarrow c, \qquad x > 0, \psi \in \Psi, c \in \Omega$$

Classification usually requires a number of rules to classify the complete training set T. Therefore a rule set R is defined below, where n represents the number of rules in the set.

$$R = \{r_1, r_2, \ldots, r_n\}$$

The unclassified training set T which is defined in (1) can be classified using rule set R. Classification is defined in (4) where training set T is classified towards item set I

using rule set R. Item set I contains all correctly classified items. Thus item set I is a subset (5) of training set T. A complete item set I contains all instances from training set T.

Classification can also be done with incomplete item sets. So in general some item set I is classified toward item set I' using rule set R as shown in (6). Notice that item set I' could be empty if rule set R cannot classify any item at all, from item set I.

$$T \xrightarrow{R} I \tag{4}$$

$$I \subseteq T \tag{5}$$

$$I \xrightarrow{R} I' \tag{6}$$

2.3 Quality measures

Rules, as defined in Section 2.2, generated by an algorithm could be very general but also very specific. A general rule classifies many items, while a specific rule classifies a few items or just one. A rule set is overfitted if it contains many very specific rules. Such a rule set could classify the training set fully but new items, which are not in the training set, could remain unclassified or classified incorrectly. A rule set containing mostly general rules tends to overlap the target classification. In other words, items are classified more than once. So to get a robust classification, rules should be more general rather than specific.

To compare rule sets from different algorithms a quality measure is needed, therefore each rule must be qualified. The quality of a rule depends on the features it contains, thus the importance of a feature must be analysed. Important features characterise the item set. As a result, a set of rules which covers many important features will produce a robust classification. To represent the quality measurement, three different measures are defined:

- Feature support
- Feature rank
- Rule support

Feature support A good classification is based on key features which characterise the item set. Feature support is a measurement of feature significance equal to the number of items correctly classified by only the rules containing a given feature f. Such a rule set is denoted R_f. In formula (7), the rule set R_f classifies item set I into a correctly

classified item set I'_f.

$$I \xrightarrow{R_f} I'_f \tag{7}$$

Then the feature support S_f is defined in (8), where $|I|$ translates to the number of items in item set I. Notice that item set I could contain items from all classes.

$$S_f = \frac{\left|I'_f\right|}{|I|} \tag{8}$$

Feature support S_f represents a value between 0 and 1. The feature with the highest feature support corresponds to the most important feature in the rule set. If only one particular class is of interest, the feature support per class S_{fc} can be calculated analogously to (7) and (8). This is illustrated in the formula below, where rule set R_{fc} only contain rules with feature f and class c.

$$I \xrightarrow{R_{fc}} I'_{fc}, \quad S_{fc} = \frac{\left|I'_{fc}\right|}{|I|}$$

Feature rank To rate the importance of a feature f in the rule set it is necessary to know how many items depend on the rule set R_f containing this feature. This is done by removing all rules containing this feature from the rule set R, yielding a rule set $R_{\bar{f}}$. Formula (9) shows the normal classification, where item set I is classified using rule set R. The correctly classified item will be in item set I'. Furthermore, classification error E has a value between 0 and 1, where 0 means that the used rule set classifies the item set without error and 1 if rule set R do not classify any item.

In formula (10) rule set R has been replaced by $R_{\bar{f}}$. Rule set $R_{\bar{f}}$ only contains rules without feature f. After classification with rule set $R_{\bar{f}}$ item set $I'_{\bar{f}}$ only contains correctly classified items which not do solely depend on feature f. Furthermore, classification with $R_{\bar{f}}$ corresponds to the classification error E_f of feature f.

$$I \xrightarrow{R} I' \quad E = \frac{|I| - |I'|}{|I|} \tag{9}$$

$$I \xrightarrow{R_{\bar{f}}} I'_{\bar{f}} \quad E_f = \frac{|I| - |I'_{\bar{f}}|}{|I|} \tag{10}$$

The increase in the classification error is defined by (11). It could be negative, in which case removing all the rules with feature f from the rule set actually improves the classification quality. This situation only applies in case conflicting rules exist in the rule set.

The features are ranked with formula (12) where N_f represents the rank of feature f and K the normalisation value which depends on the situation. Rank N_f has a value between $-\infty$ and 1. Furthermore the $\exists$ and $\forall$ symbols represent respectively the existential quantor and universal quantor, in other words 'there exists at least one' and 'for all'. The formula distinguishes three situations:

$\exists(\Delta E_f > 0)$ In this situation an error must occur if a feature f is removed from the set. In this case the feature which introduces the largest error is ranked last.

$\forall(\Delta E_f = 0)$ This case only occurs if none of the features introduces an error when removed from the rule set. Which means that $K = 0$, which cannot be computed. Therefore all features are equally ranked with value 0.

$\forall(\Delta E_f \leq 0)$ The final case occurs when all features f in the rule set have a $\Delta E_f \leq 0$ error. Which means that if such a rule with a negative error is removed, the classification quality increases. This part of the formula is needed to rank the features with the most negative impact on classification last.

$$\Delta E_f = E_f - E \tag{11}$$

$$N_f = \begin{cases} \frac{\Delta E_f}{K}, & K = \max_f(\Delta E_f), \quad \text{if } \exists(\Delta E_f > 0) \\ 0, & \text{if } \forall(\Delta E_f = 0) \\ \frac{\Delta E_f}{K}, & K = -\min_f(\Delta E_f), \quad \text{if } \forall(\Delta E_f \leq 0) \end{cases} \tag{12}$$

Features with rank ≤ 0 do not play any significant role in classification. Therefore rules with only such features may be removed from the rule set without decreasing the classification quality. However, such rules may overlap with other rules in the rule set, so that these features might still be of interest for the characterisation of the item set. Notice that the current feature rank must be recalculated after rules with a certain feature are removed.

Rule support The quality of a rule can be measured through the number of items it classifies. In formula (13), the rule r_c, which classifies class c, classifies item set I_c into a correctly classified item set I'_c. Item set I_c only includes class c items. Then the rule support S_r is defined in (14), where $|I|$ translates to the number of items in item set I.

$$I_c \xrightarrow{r_c} I'_c \tag{13}$$

$$S_r = \frac{|I'_c|}{|I_c|} \tag{14}$$

Thus rule support S_r is a value between 0 and 1, where the value 1 represents the rule that classifies all items. If the rule support has a value of 0 no items are classified.

2.4 Cross-validation

A classification method is based upon the data from its training set. When the same method is applied on new data from outside the data set a percentage of the new data might remain unclassified.

Using k-fold cross-validation (CV) it is possible to predict a statistical classification quality. In k-fold cross-validation, the original data is partitioned into k equally distributed subsets. The $k-1$ subsets are used for training, the subset which remains is used for testing. The CV process is done for each test/training set combination, thus in k-fold. The k-fold average is the estimated error. Usually with $k = 10$ an accurate error indication is gained.

The CV results can depend greatly on the way the original data is partitioned. Therefore an average of ten 10-fold cross-validations and its standard deviation is given as the final CV result.

2.5 Software

The main program used in this thesis is Weka [74] version 3.4.5, which includes both the PRISM and Tertius algorithms. Besides the use of Weka additional utilities were developed. The additional algorithms as defined in Sections 2.4, 3.1 and 4.1 are implemented into a the so-called Rule set Optimisation Application (RSOA) with is part of the *Rule set Optimisation Package*. A more detailed description of the RSOA can be found in Appendix A. The Rule set Optimisation Package includes:

Rule set Optimisation Application This application contains the implementation of the following algorithms: Classification, optimisation, subsets generation and cross-validation.

Conversion scripts Since the Weka implementation was used, the output format from both the PRISM and Tertius algorithms were converted to a simplified rule format used by the Rule set Optimisation Application.

Additional scripts Some small shell scripts were written to automate certain tests. For example, the generation of results and cross-validation.

2.6 Discussion

There are some discussion points regarding the used classification methods which depend on a given alignment and the quality measures involved for optimisation.

Alignment dependency The classification depends greatly on the alignment of the data set. Generally speaking, if a data set is not very conserved in a multiple sequence alignment (MSA), the classification may not be very robust and might be useless for classifying other instances outside the data set. Moreover, if another MSA method is used, other classification characteristics could apply.

Quality measures Besides the quality measures discussed in Section 2.3 other measures could be used for optimisation. For example, using quality measures with various biological or chemical experiments in mind could improve the feasibility of these experiments using the obtained optimised classification. These special options could also be included in the rule discovery methods.

3 PRISM

PRISM [8] is an algorithm for inducing modular rules from a training set. The composition of such rules is defined in Section 2.2. Furthermore the rules PRISM generates represent a decision tree which leads to classification of the training examples. A simplified version of the PRISM algorithm is described below:

1. Search for a rule which has a maximum information gain on the current training set without conflicting the rules in the rule set.
2. The rule is added to the rule set.
3. The training set is split into a classified and unclassified part.
4. Steps 1 through 4 are repeated with the unclassified training subset until the training set is empty.

The induced set of rules fully classifies the training set. PRISM tends to overlap its rule set a little so it is not a minimal rule set. Furthermore, each iteration in the algorithm will generate more specific rules rather than general ones.

3.1 Methods

Our goal is to analyse the possibility if a simple algorithm like PRISM could generate results which are comparable to the results found by Samsonova et al. [63]. As input the olfactory – non-olfactory GPCR training set as described in Section 2.1 is used. In addition to PRISM an optimisation algorithm is used to optimise the number of features used in the rule set. This optimisation method is described in the second paragraph. Besides the optimisation method two additional experiments are done. The first experiment uses the group amino acid model, as described in Section 2.1 rather than the individual amino acid model. The second experiments uses a selection of features which are found in [63]. The two methods are described in the last two paragraphs of this section.

Implementations The PRISM algorithm has already been implemented in the Weka software package [74] which is used in the experiments. Furthermore the classification and optimisation algorithms described in the next paragraph are implemented in the Rule set Optimisation Application as described in Section 2.5.

Rule set optimisation The rule set optimisation method is divided into two parts.

- The primary goal of this project is to classify and characterise olfactory GPCRs. The non-olfactory class is composed of a variety of non-olfactory GPCR families, see Section 2.1. Since PRISM only generates rules without conflict, rules with the conclusion non-olfactory which are not relevant are discarded from the rule set. This can only be done because all non-olfactory GPCRs are also fully classified. This rule set only classifies the olfactory class of the training set fully, GPCRs which could not be classified are assigned to the non-olfactory class. Thus the same classification quality of the training set is achieved with a smaller rule set.

- The rules in the PRISM rule set show some overlap, so that some rules could be removed without decreasing classification quality. Therefore an optimisation algorithm is created using the rule measurements as defined in Section 2.3. According to these quality measures, rules which cover features with the lowest rank are good candidates for removal. Besides feature rank, feature support may also play a significant role. Rules which cover features with a low support are second best candidates. To optimise the rule set the following steps are performed:

  1. Feature rank and support for all features is calculated.
  2. The feature with the lowest rank and support is selected.
  3. The rules which contain the selected feature are removed from the rule set.
  4. Feature rank and support of the remaining features is recalculated.
  5. Steps 2 through 4 are repeated until a given termination criterion is reached. Termination criteria may be: classification quality, rule set size, feature set size, etc.

Amino acid group model The amino acid group model, as described in Section 2.1, was used by Samsonova et al. [63]. This was done because of the computational complexity involved using the 20 individual amino acid model. Therefore the 6 amino acid group model was chosen to decrease the computational complexity.

In this experiment the PRISM training set which is composed of the individual amino acid model is converted to the six groups used in the amino acid group model. Consequently PRISM uses the new training set to generate a new rule set. This new rule set is constructed from rules which contain groups from the group model and none from the individual amino acid model.

Selected features The results from Samsonova et al. [63] discovered seven important features that are used for classification. The positions of these features are selected not to be removed from the PRISM training set, with the result that all GPCRs in this training set only contain the selected 7 features. Using this training set PRISM generates the rule set which will be optimised, analysed and compared.

Notice that the final phase of [63] uses the individual amino acid model. Therefore this model is also used in the PRISM training set.

3.2 Results

PRISM generated a rule set consisting of 66 rules with 50 unique features. These rules classify the training set fully, both olfactory and non-olfactory GPCRs.

Optimised rule set Using the method described in Section 3.1, the PRISM rule set is reduced to 47 rules with 36 unique features in the first step. The second step reduces the rule set to 25 rules with 16 features without decreasing classification quality.

The error introduced in the optimising method is rated through 10-fold CV as described in Section 2.4. An average classification quality of 99.1% $\pm$ 0.4 is obtained. Without this optimisation, PRISM 10-fold CV rated an average classification quality of 89.6% $\pm$ 0.4. The average classification quality increases significantly using the optimisation method. The diversity in the non-olfactory class, which is removed during the first phase of the optimisation process, is the main reason for this improvement. Thus this optimising method is very powerful for reducing the rule set with equal classification quality. This set can be optimised further, with the same method. However the classification quality will decrease with each reduction.

In Figure 3 all features from the optimised rule set are plotted in a histogram. The feature rank and support are shown next to each other. Notice that feature rank does not always depend on feature support. Due to overlapping, most GPCRs are classified through more than one feature. For example if $S_f \gg N_f$ then this feature classifies many GPCRs that are already classified through more important features.

For instance feature 1.51 has the lowest rank and support. As a result all rules that contain this feature are removed from the rule set. However this decreases the classification quality.

Table 4 shows the optimised rule set, where rules are sorted by rule support. The extended tables indicate the decrease in classification quality, if rules with certain features are removed.

Rules with a high rule support contain features which both have a high feature rank and support. To illustrate this, feature 1.49 and 1.43 both have a high feature

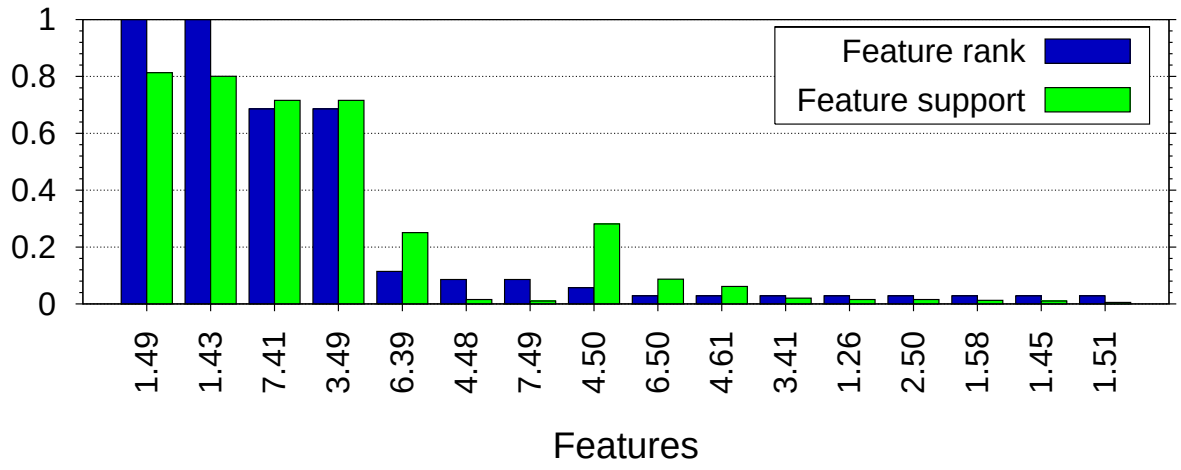


Figure 3: This histogram shows the features from the optimised rule set. Each set of bars represents a feature with its rank N_{fc} and support S_{fc}, where class c is olfactory.

rank and support, as shown in Figure 3. Both features are included in the first rule $[1.43 = \mathrm{Y} \wedge 1.49 = \mathrm{G} \rightarrow \text{olfactory}]$ with the highest support. Nevertheless there are also two rules, $[1.49 = \mathrm{W} \rightarrow \text{olfactory}]$ and $[1.49 = \mathrm{C} \rightarrow \text{olfactory}]$, both contain feature 1.49 with a low rule support. These two rules are not removed during optimisation because the first rule, with the highest support, contains feature 1.49.

	Features															
S_r	1.26	1.43	1.45	1.49	1.51	1.58	2.50	3.41	3.49	4.48	4.50	4.61	6.39	6.50	7.41	7.49
0.801		Y		G												
0.716									D						Y	
0.238													Y			
0.210											L					
0.061												H				
0.054														V		
0.038											M					
0.020								P								
0.020														G		
0.015										P						
0.015	A						A									
0.013													H			
0.013											A					
0.013														I		
0.013						N										
0.010				W												
0.010											S					
0.008			-													
0.005																Q
0.005											G					
0.005					D											
0.005											R					
0.005																R
0.003			K													
0.003				C												

	Classification Quality						
	100.0%	99.5%	98.9%	98.6%	98.2%	97.6%	96.6%
0.801	×	×	×	×	×	×	×
0.716	×	×	×	×	×	×	×
0.238	×	×	×	×	×	×	
0.210	×	×	×	×			
0.061	×	×					
0.054	×	×					
0.038	×	×	×	×			
0.020	×	×					
0.020	×	×					
0.015	×	×	×	×	×		
0.015	×						
0.013	×	×	×	×	×	×	
0.013	×	×	×	×			
0.013	×	×					
0.013	×						
0.010	×	×	×	×	×	×	×
0.010	×	×	×	×			
0.008	×						
0.005	×	×	×				
0.005	×	×	×	×			
0.005	×						
0.005	×	×	×	×			
0.005	×	×	×				
0.003	×						
0.003	×	×	×	×	×	×	×

100.0%	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
99.5%		×		×				×	×	×	×	×	×	×	×	×
98.9%		×		×					×	×	×		×		×	×
98.2%		×		×					×	×	×		×		×	
96.4%		×		×					×	×			×		×	
97.6%		×		×					×				×		×	
96.6%		×		×					×						×	

Table 4: The optimised PRISM rule set. Each row represents a rule with olfactory as conclusion. For example the top row represents the rule $[1.43 = \mathrm{Y} \wedge 1.49 = \mathrm{G} \rightarrow \text{olfactory}]$. On the left the rule support S_r indicates the classification quality of that rule. In the extended lower table the percentage indicates the classification quality of the rule set, which only includes the features marked with '×'. In the extended right table the corresponding rules are marked similarly.

In Figure 4 the optimised rule set is displayed as a decision tree. Starting from the root, each rule can be read by following its branch until a leaf is reached. For example the branch that reads "1.49,G,1.43,Y" represents the rule [$1.49 = \mathrm{G} \wedge 1.43 = \mathrm{Y} \rightarrow$ olfactory]. If a training example is not classified by the tree then it is assigned to class non-olfactory.

The features and amino acid printed in boldface show the decision tree which classifies 96.6% of the training set. Notice that the 96.6% decision tree is significantly reduced, which means that the olfactory class can be classified with less features at the cost of approximately 3% classification quality.

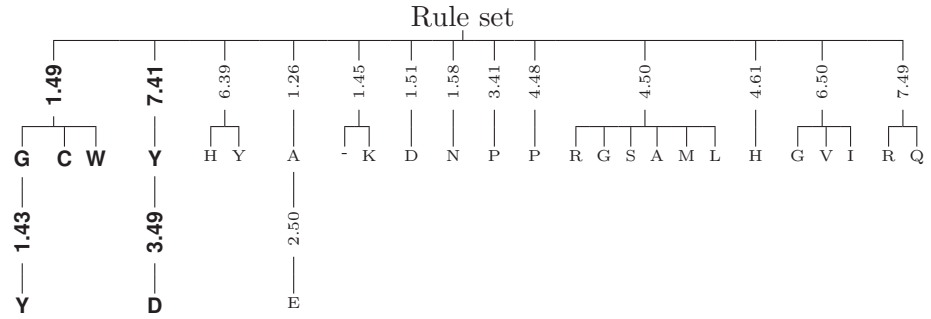


Figure 4: Optimised PRISM rules displayed as a decision tree. Notice that all rules classify towards the olfactory class. The rule set which classifies 96.6% is printed in boldface.

In Figure 5 both the olfactory and non-olfactory sub-alignments[3] are presented as sequence logos [64, 27]. Each character represents an amino acid, where its size represents the degree of conservation. Gaps are excluded from the calculations. The information content is indicated through bits reflecting the relative frequency of occurring amino acids. All the features from the optimised rule set which classifies 100% of the training examples, are marked with boxes.

The results from [63] share only a single feature when compared with the PRISM optimised rule set. Only feature 7.41 is included in both rule sets. Furthermore, both rule sets contain rules with high and low rule support. Almost all selected features from the model in [63] have a high feature support. In contrast with the features PRISM found, where only a few features have a high support.

[3]The amino acid letters are coloured according to the amino acid group model, described in Section 2.1. However in this case the colouration has no meaning other than giving the amino acids some contrast.

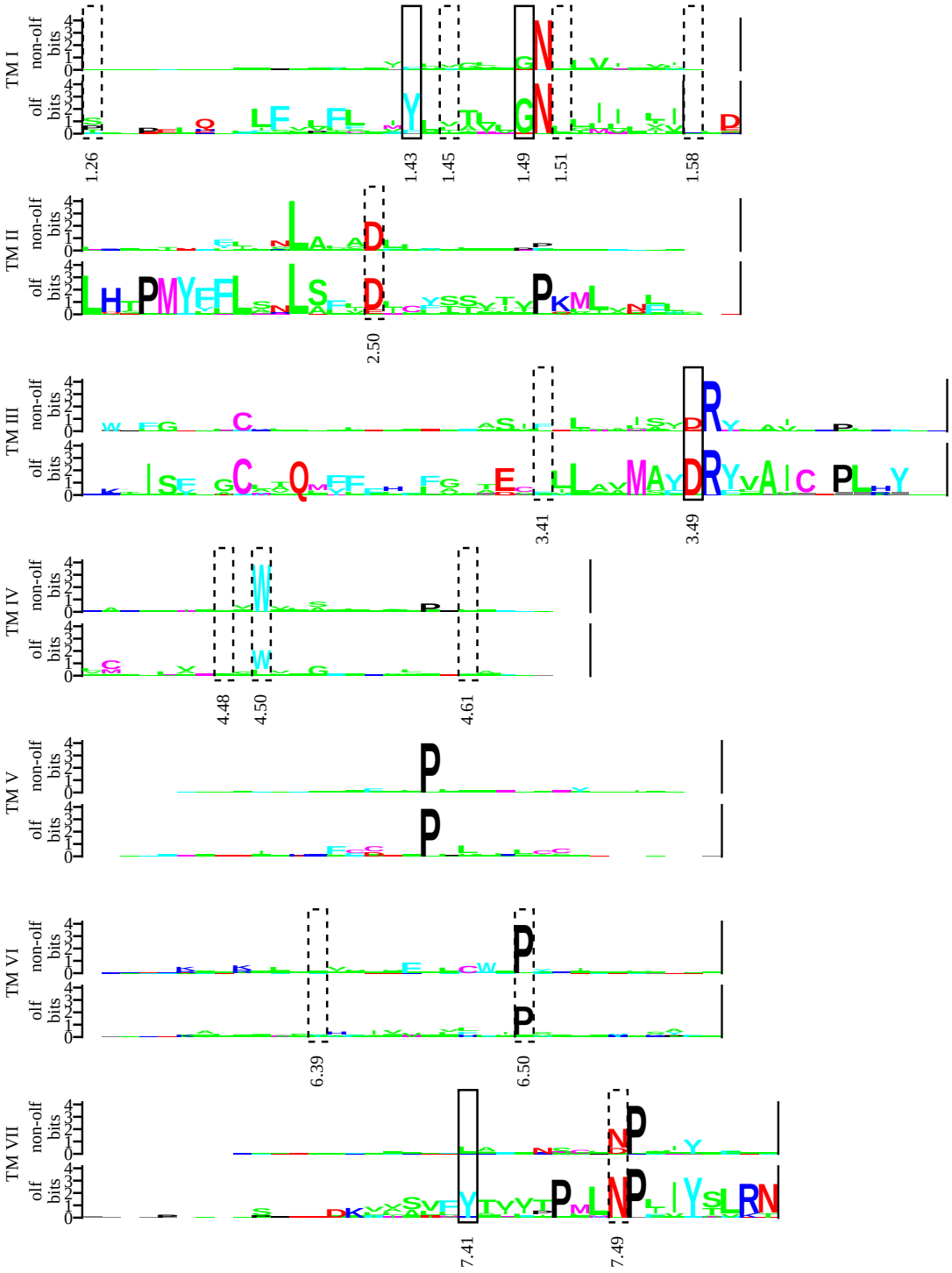


Figure 5: Alignment of the seven transmembrane (TM) regions of the human GPCRs, shown as sequence logos of sub-alignments containing olfactory and non-olfactory receptors respectively as defined in Section 2.1. Features in the optimised PRISM rule set are indicated with boxes. After optimising further to a classification quality of 96.6% the rule set only contains features marked with solid boxes.

Amino acid group rule set PRISM generated a rule set using the amino acid group model as described in Section 3.1. The results are shown in Table 5 where the number of rules and features of both models are compared. This rule set is also optimised using the method from Section 3.1 The results are also shown in the table.

	Normal		Optimised	
	Rules	Features	Rules	Features
Individual amino acid model	47	36	25	16
Amino acid group model	69	61	20	23

Table 5: In this table the usage of the individual amino acids and amino acid groups model are compared. The number of rules and the number of the used features are shown for both the normal and optimised PRISM rule set.

The amino acid group model uses much more rules and different features to fully classify the training set compared to the other model. This applies both to the normal and optimised rule set. Only the optimised rule set using the amino acid group model contains less rules. In Figure 6 the histogram shows the features found in the PRISM optimised rule set.

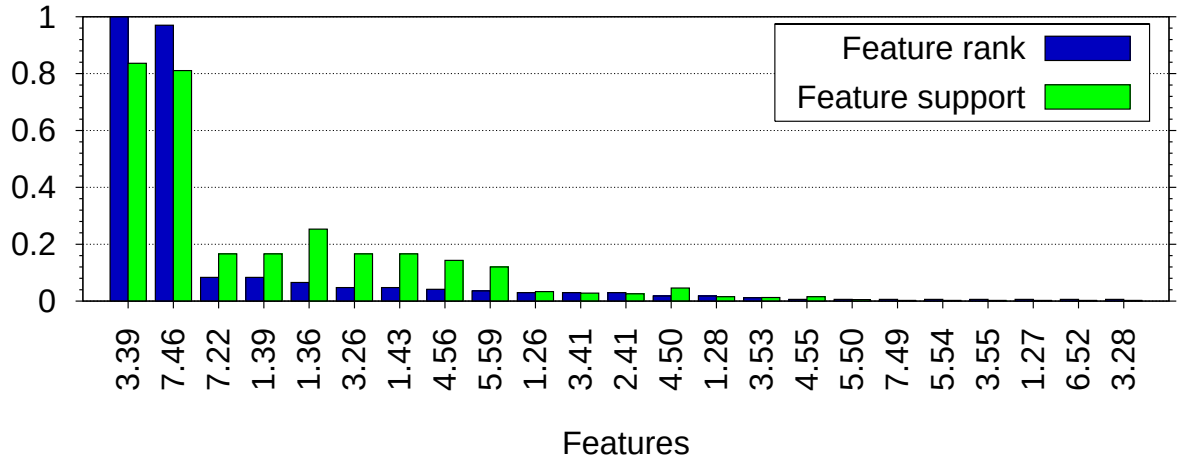


Figure 6: This histogram shows the optimised PRISM rule set using the amino acid group model. Each set of bars represents a feature with its rank N_{fc} and support S_{fc}, where class c is olfactory.

The two particularly highly supported features among the other low supported features classify the most part of the training set. Compared to the individual amino acid model as shown in Figure 3, four features show a high support. This result may be very interesting because only two main features may be needed to classify most of the olfactory GPCRs.

The used method, including the usage of the amino acid group model, is tested using 10-fold cross-validation as described in Section 2.4. Using PRISM without optimisation the average classification quality value is 88.95% ± 0.56, which is almost equal to the individual amino acid model, discussed in Section 3.2. When including the optimisation step the average classification quality increases to 97.59% ± 0.31 which is slightly less than of the individual amino acid model. So it can be concluded that the amino acid group model has no significant effect on the classification quality.

Selected features rule set The results from Samsonova et al. [63] found 70 rules containing only 7 features which classify the training set for 96.2%, where the olfactory class is classified using 32 rules. However, the used model needs both olfactory and non-olfactory rules for classification.

Using only these selected features in the training set, as described in the last paragraph in Section 3.1, PRISM generated 68 rules. This rule set classifies the training set fully, where only 16 rules are needed to classify the olfactory class. As discussed before in Section 3.1, the non-olfactory rules are removed from the set. Therefore, the instances which cannot be classified using the olfactory rule set, are assigned to the non-olfactory class.

In Figure 7 the feature rank and support are plotted in a histogram. Notice that almost all selected features have a high feature support S_{fc}.

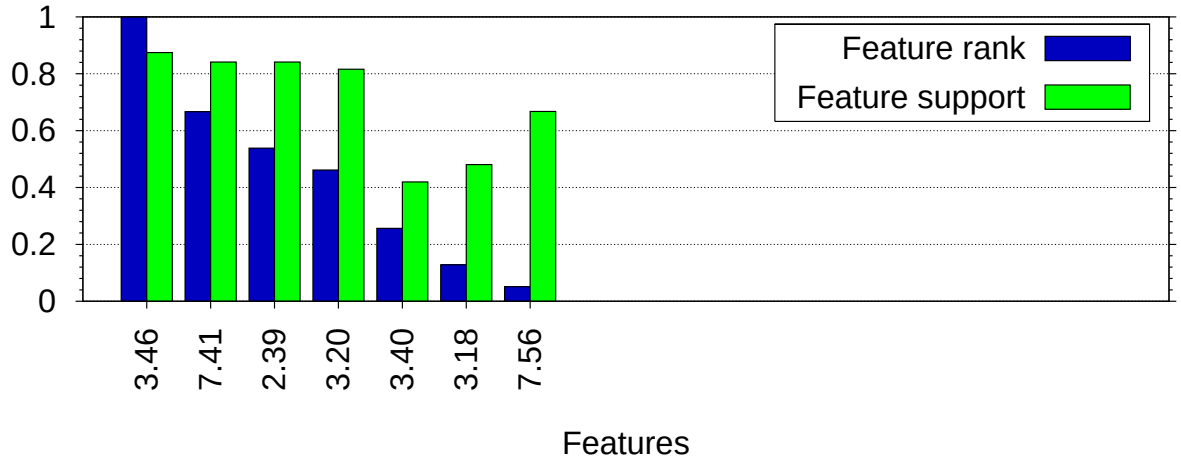


Figure 7: This histogram shows the 7 features from the PRISM rule set, selected from the results in [63]. Each set of bars represents a feature with its rank N_{fc} and support S_{fc}, where class c is olfactory.

The generated rules are shown in Table 6. In the extended tables the optimised rule sets are shown using the optimisation method described in Section 3.1. Notice that the rule set cannot be optimised further without decreasing the classification quality.

	Features							Classification Quality				
S_r	2.39	3.18	3.20	3.40	3.46	7.41	7.56	100.0%	99.7%	98.9%	97.5%	93.6%
0.816			**I**		**M**			×	×	×	×	×
0.716	**M**					**Y**		×	×	×	×	
0.714					**M**	**Y**		×	×	×	×	×
0.668						**Y**	**R**	×				
0.427		**K**			**M**			×	×			
0.325	**M**			**C**				×	×	×		
0.043	**M**			**F**				×	×	×		
0.041				**D**				×	×	×		
0.031	**M**					**H**		×	×	×	×	
0.026	**M**					**N**		×	×	×	×	
0.020	**M**	**P**						×	×			
0.018	**M**	**G**						×	×			
0.015		**F**				**Y**		×	×			
0.010	**M**			**K**				×	×	×		
0.003					**N**			×	×	×	×	×
0.003						**-**		×	×	×	×	×

100.0%	×	×	×	×	×	×	×
99.7%	×	×	×	×	×	×	
98.9%	×		×	×	×	×	
97.5%	×		×		×	×	
93.6%			×		×	×	

Table 6: The PRISM 7-feature rule set. Each row represents a rule, assembled from features and amino acids. All rules have the same conclusion, olfactory. On the left the rule support S_r indicates the classification quality of that rule. In the extended lower table the percentage indicates the classification quality of the rule set, which only includes the features marked with '×'. In the extended right table the corresponding rules are marked similarly.

In addition, a decision tree is shown in Figure 8 which displays the 7-feature PRISM rule set. The rule set, sufficient for the classification quality of 97.5%, is indicated in boldface. With only 4 features, a very high classification quality is reached at the cost of 2.5%.

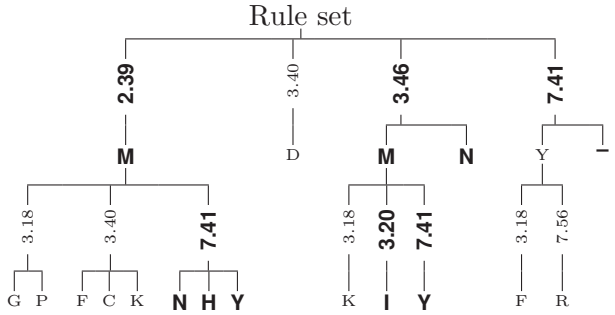


Figure 8: The 7-feature PRISM rule set displayed as a decision tree. All rules in the decision tree classify to the olfactory class. The branches shown in boldfaced represent the branches still used in the rule set with a classification quality of 97.5%.

The PRISM rule set is much smaller compared to the rule set found by Samsonova et al. [63], moreover this rule set fully classifies the training examples. When comparing the two sets, only 10 rules are equal, including 4 olfactory rules.

After optimising the 7-feature set, below the 100% threshold, a rule set with only 4 features remains. Compared to the results from [63] with a rule set which classifies 96.2%, this 4-feature rule set has nearly the same classification quality (97.5%) with much fewer features and rules.

3.3 Conclusion

To conclude this section, PRISM has some interesting results. Classification and characterisation are two important aspects which depend on each other. The strategy PRISM uses to select rules has a totally different approach than the model used by Samsonova et al. [63]. PRISM generates a non-conflicting rule set which can fully classify all training examples and can be visualised as a decision tree which is easy to interpret.

PRISM is a very simple method if it comes to classification compared to the model in [63]. However, this classification is not very robust because features with very low support are included in the rule set. Nevertheless the method can be successful as shown in the selected feature rule set, where PRISM generated a smaller rule set with similar classification quality compared to [63].

When it comes to characterisation, the model used in [63] found 7 features which all have a high feature support, in contrast with PRISM which induced more rules covering features with a low support. Features with a high support characterise the rule set through rules with a high rule support. These features combined with the right amino acid are very good candidates for the characterisation of an olfactory GPCR.

4 Tertius

Tertius [22] is a rule discovery algorithm for deriving rules from a training set where it is unknown which attributes are useful for classification. The format of the classification rules is as defined in Section 2.2.

The bases for rule discovery within Tertius is a heuristic measure of confirmation. The measure of confirmation is a statistical analysis of a term under consideration. This analysis is between the expected and the observed frequencies of counter-instances. A counter-instance classifies a GPCR toward both classes so that the GPCR remains unclassified. The degree of confirmation is used as an estimation to what depth a rule can be improved by refinement.

Refinement occurs through strict ordering of the following refinement operations: Adding a new term, unifying two variables and instantiating a variable with a constant. Where a variable is an undetermined feature or amino acid and likewise a constant is a selected feature or amino acid. The refinement operator *unifying two variables* is not used in our case because unification can only occur using two variables of the same type. In a term (feature = amino acid) no such variables exists. The operation *instantiating a variable with a constant* is done in sorted order, using a constant from its domain. The domain represents a set of amino acids or features which can be obtained from the training set. This is necessary to avoid searching the search space more than once.

The exploring starts from an empty rule by refining it iteratively using the refinement operations described above. Multiple paths from an empty rule to a hypothetical rule could appear in the Tertius search space, therefore strict ordering in refinement operations is required. Thus paths leading to hypothetical rules which are already considered are pruned. Finally, the confirmation function is able to rank the hypotheses, finding the k most confirmed hypotheses.

The Tertius system uses a complete best-first search [62], which means that all possible hypothetical rules are explored, moreover the best confirmed hypothetical rules are explored first. However the search space of all possible hypothetical rules is pruned according to the degree of confirmation. In other words, hypothetical rules with a low degree of confirmation are not refined, as a result their search branches are pruned.

Limitation criteria using various threshold parameters such as the number of terms in rule discovery, relative frequency of counter instances, confirmation threshold, etc. are needed to limit the search space. This may also be useful for the algorithm to avoid the problem of insufficient computer memory.

4.1 Methods

The used methods are based on an existing Tertius implementation which will be described in the next paragraph. Besides Tertius a number of algorithms have been developed to find better classification characteristics for the olfactory GPCRs compared to the characteristics found in Samsonova et al. [63]. Tertius uses the training set, as described in Section 2.1, to generate a list of rules. With these rules various classification characteristics may come forward using the developed algorithms. Furthermore one additional experiment is done using the amino acid group model which is described in the last paragraph.

Tertius implementation In this study the Tertius implementation in Weka [74, 18] is used to generate a list of rules.

The search space Tertius uses is very large, which leads to the problem of insufficient computer memory. To avoid such problems the search has been split up into six smaller problems, each using the parameter which sets the maximum number of terms in a rule with the value 2 through 7. Also the maximum number of confirmation values to be found is limited using $k = 200$ for rules with 2 through 4 terms and $k = 100$ for rules with 5 through 7 terms. Furthermore, the maximum proportion of counter-instances of rules called the 'noise Threshold' parameter is set to zero. This means that only rules without counter-instances are listed in the final list of rules.

Is was necessary to use these limitations to avoid insufficient computer memory[4]. Consequently, these limitation affect the resulting lists of rules. Thus these lists are limited and might miss some less important rules, because the rules with the best confirmation are listed first.

The six generated rule lists were combined into one list. This could only be done because the rule lists do not contain any conflicting rule. As a result a lot of identical rules occur, which were removed. Furthermore the non-olfactory rules were also removed. The resulting list is defined as *The Tertius rule set* which only contains olfactory rules. If a GPCR cannot be classified using this rule set, it is automatically classified as non-olfactory.

The Tertius rule set is composed of more than 5000 rules and should be optimised or split into robust subsets. The goal is to find a robust rule set which is composed of rules with a high rule support. First, the method described in Section 3.1 is used for optimisation. In addition the optimised rule set can be optimised further using the subset construction, which will be defined in the next paragraph. Second, the Tertius rule set is split using only subset construction to find optimised rule subsets. These rule

[4]Using a AMD Athlon 2000+ with 768MB RAM.

subsets should give a robust classification with features that are very characteristics for olfactory GPCRs, because the rules selected for the subsets involve features that are highly supported in the olfactory training set.

Subset construction Finding optimal rule subsets is similar to the knapsack problem [26, 43] which is NP-complete. Its name was derived from the maximisation problem of choosing as much as possible items that could fit into one bag. In other words the search space to examine all rule combinations is too complex to compute. A number of algorithms could be used to find good solutions without exploring all rule combinations.

There are different approaches for choosing rules to classify olfactory GPCRs. Evolutionary algorithms [3] could be used to get near optimal subsets. In particular genetic algorithms are already used for robust feature selection [72, 76], nevertheless our features are already selected by Tertius. At the same time the algorithms could be adapted for robust rule selection. However in our case no real optimum exists, because a subset could be optimised using a variety of criteria such as the number of rules or the number of different features in a subset.

An other approach is to find robust rule sets. A robust rule set as discussed in [31, 40] uses redundant rules for an increasing robustness. The algorithm described in [31] has some similarities with the implemented subset algorithms, although finding the optimal robust rule set is not the main intention.

The goal is to have different kinds of rule subsets which are composed using various rule combinations. So various features should appear in these subsets with the result that different characterisations may come forward. Therefore different approaches are implemented for finding good rule subsets which still can classify all olfactory GPCRs.

The algorithms that are implemented use different criteria for generating the rule subsets. Multiple criteria are combined using a lexicographic approach. A critical review about multi-objective optimisation [23] compared the Weighted-formula, Lexicographic and Pareto approach with each other.

The Weighted-formula approach. This approach combines multiple criteria into one value. However criteria with different measurements cannot be mixed as concluded in the review.

The Lexicographic approach. This approach gives priority to each different criteria, which means that criteria with the highest priority are treated first. Although different measurements can be mixed because each criteria is treated separately. A well known algorithm AQ18 [34] uses this approach.

The Pareto approach. This approach uses a multi object algorithm as discussed in [35, 55] using Evolutionary Search or Genetic Algorithm respectively. The Pareto approach returns a set containing one or more non-dominated solutions (A non-dominated set means that there is no better set of solutions). Where as only one solution is needed, one rule to continue creating a rule subset. This makes the approach unnecessarily complex compared to the two other methods.

In conclusion the lexicographic approach for multi-objective optimisation problems is in our case the best choice. The lexicographic approach has been adapted slightly and is used without the use of a tolerance value in the subset algorithms. This approach is used in all different subset algorithms, which are described below.

Algorithm 1, Greedy rule selection. In this algorithm a greedy search method [11] is used, which selects the best rule considered. The construction of the rule subset uses the following steps:

1. Search for a rule in the Tertius rule set which classifies the maximum number of GPCRs that are not classified yet. This is the rule with the highest rule support.
2. Add this rule to the rule subset and remove it from the rule set. Consequently, the rule support must be recalculated.
3. Continue with steps 1 and 2 until one of the stop criteria is met. These stop criteria are:
   - All GPCRs are classified.
   - There are no rules left in the Tertius rule set which increases the number of classified GPCRs.
   - The Tertius rule set is empty.
4. Multiple rule subsets are constructed by this process, repeating the steps 1 through 3, and terminates by the second or third stop criteria.

Algorithm 2, Minimising the number of features. The idea behind this algorithm is to minimise the number of features in the subsets. This algorithm is almost the same as the first algorithm except an extra criterion is added using lexicographic

ordering as described earlier. The rule selection must not only classify a maximum number of GPCRs but also *minimise the number of new features* added to the rule subset.

The first rule that will be added to the rule subset is found using the first algorithm because the extra criterion needs some features in the rule subset to be successful. Once the first rule is added to the subset, the next rules are tested on both criteria.

Algorithm 3, Worst classified GPCR first. This algorithm is based on classifying the worst classified GPCR first. In other words there is a minimum number of rules in the Tertius rule set that could be used to classify the worst classified GPCR. The algorithm uses the same methods as in algorithm 2, with one adjustment. *Rather than sequentially classifying GPCRs, the worst classified GPCR is selected and classified first.* Each time the worst classified GPCR in the unclassified GPCR set is selected for classification until all GPCR are classified. The goal of this approach is to find other classification characteristics through the use of different feature combinations in the rule subset.

Algorithm 4, Overlapping. This algorithm is an addition to algorithm 1, because sometimes the rule subset produced by algorithm 1 has some site effects. For example, the last rule added to the rule subset using algorithm 1 selects the first rule found in the Tertius rule set that classifies the last unclassified GPCR(s). In our case one GPCR was not classified, therefore the first rule found classifying this particular GPCR will be added to the rule subset. However this rule may only classify the only GPCR left. To make the subset more robust the last rule should be a rule that not only classifies the last GPCR(s) but also maximises the overlap with the GPCRs already classified.

To accomplish this a criterion is added also in lexicographic order, to algorithm 1, which selects rules that not only classify a maximum number of GPCRs but also *maximise the number of already classified GPCRs.* Adding redundant rules which classify an overlapping GPCR set contributes to the robustness of the rule set as described in [31].

Combining algorithms. Algorithm 3 already demonstrates how to combine three criteria into one algorithm using the lexicographic approach. Algorithm 4 can also be combined with algorithm 2 and 3, which is implemented in respectively algorithm 5 and 6. Combining these algorithms could be useful for getting better rule subsets and find other classification characteristics.

Amino acid group model The amino acid group model as described in Section 2.1 was already used in the PRISM training set which resulted in an interesting finding, see Section 3.2. Therefore an additional experiment includes the use of this model in the Tertius training set.

The method as described in Section 4.1 remains unchanged. However due to problem of insufficient computer memory the number of rules and terms are limited through the parameters Tertius uses. The number of terms uses the value 2 through 5 with $k = 200$ for the maximum number of rules. Although $k = 100$ is used for the rule discovery with a maximum of 5 terms in one rule.

The goal for this additional experiment is to find out if this simplified model gains equal or better results compared to the individual amino acid based training data. In Samsonova et al. [63] the group model was used to circumvent computational problems which emerged when using the individual amino acid model. Therefore this experiment tests the hypotheses whether or not this model is useful in classification or only as preprocessing step.

4.2 Results

The result section is divided into three paragraphs. The main part contains results generated with the *individual amino acid model* as training set. Besides the individual amino acid model an additional experiment is done with the *amino acid group model*. Both models are described in Section 2.1. The final paragraph shows the findings about highly supported features which are unsuitable to characterise the olfactory receptors.

Individual amino acid rule set Tertius generates a list of rules as described in Section 4.1. The Tertius rule set is composed of 5411 olfactory rules which cover 159 features.

In Figure 9 a 2D feature support visualisation of a GPCR is shown. Each feature is coloured according to the Tertius rule set feature support. Features that are highly supported play a very active role in classifying olfactory GPCRs. These features are orange-red coloured in the figure. The yellow coloured features are less active in classification yet they play an important role in classification, especially in combination with other features. The features coloured towards white do not play any significant role in classifying olfactory GPCRs.

The classification quality could increase by combining features with each other [28]. Indeed, almost all features are only functional in combination with other features in a rule or rule set. Accordingly the feature support in the figure only applies to a specific rule set, in this case the Tertius rule set.

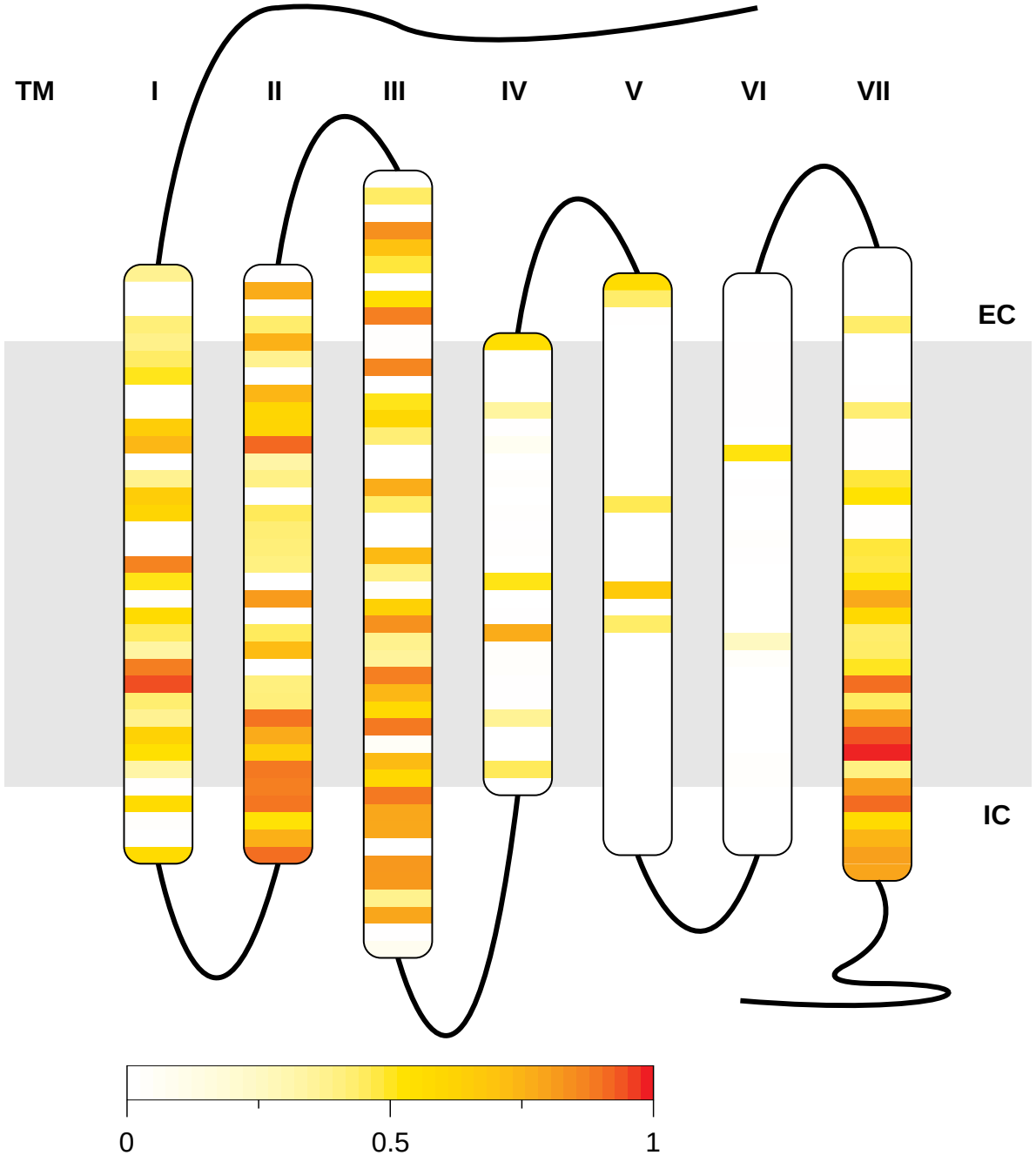


Figure 9: This visualisation shows the 2D feature support of a rule set and is created using Postscript [12]. The visualisation represents a GPCR within a membrane, containing the 7 transmembrane (TM) domains. The gray area represents the membrane and separates the extracellular (EC) and intracellular (IC) domains. Inside these domains the extra- and intracellular loops respectively connects the TM regions. The feature support from the Tertius rule set is used to indicate the classification characteristics. The bars inside each TM domain represent the feature support. The features support is indicated using a spectrum with the colours white-yellow-orange-red representing the value 0 through 1 respectively.

A histogram of all features with a feature support ≥ 0.8 is shown in Figure 10. These 26 features correspond to the orange-red coloured features in Figure 9. Furthermore there are 49 less active features between a feature support of 0.5 and 0.8. Also there are 45 features between a feature support of 0.5 and 0.25. The features below a feature support of 0.25 are unimportant for classification. Notice that from the total of 248 features, 128 features are useless for classifying olfactory GPCRs.

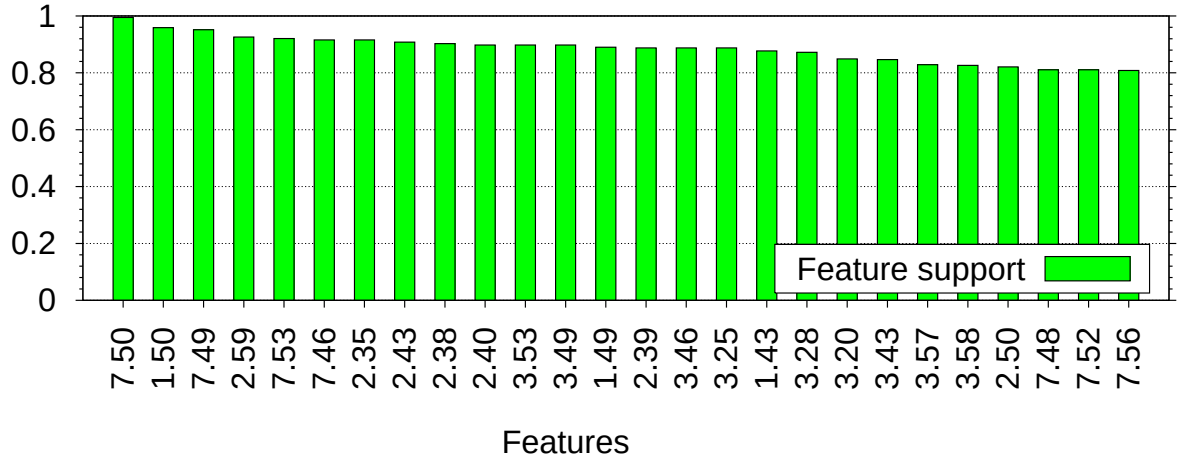
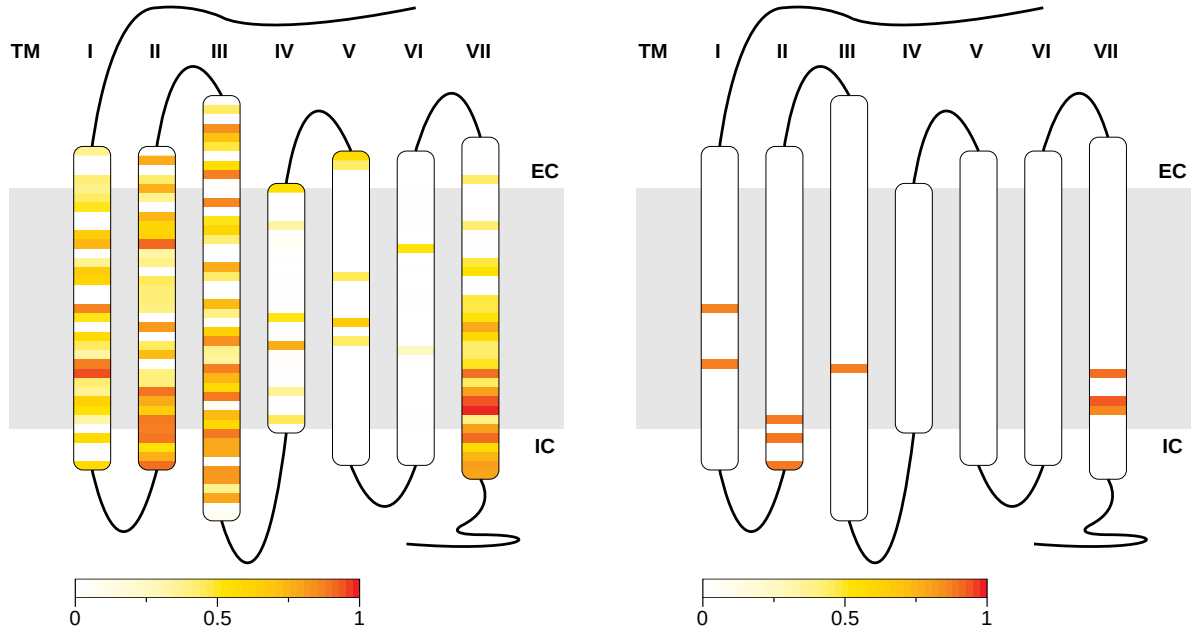


Figure 10: This histogram shows the 26 best features from the Tertius rule set. Each bar represents the feature support S_{fc}, where class c is olfactory. Note that the feature rank N_{fc} is not shown, because all features have 0 as rank.

Optimised rule set After optimising with the same method as described in Section 3.1, 32 rules covering only 9 features where found. This rule set was optimised further using subset construction to remove irrelevant rules. The six subset algorithms described in Section 4.1 found two different rule subsets with 7 and 8 rules all covering the same 9 features. Algorithm $1, 3, 4$ or 6 generated the subset with 7 rules and is chosen to be the double optimised rule set. In Figure 11, a 2D feature support visualisation shows both the Tertius double optimised rule set in Figure 11(b) and the original Tertius rule set in Figure 11(a) (only for comparison, but equal to Figure 9).



(a) Original Tertius rule set

(b) Double optimised Tertius rule set

Figure 11: Comparing the Tertius rule set with the double optimised rule set.

When comparing the two rule sets some interesting differences emerge. The number of remaining features in the double optimised rule set is impressive. With only 9 features left all olfactory GPCRs, present in the training data, could be classified. The position of the features in the double optimised rule set is also very interesting. The features are only located on TM domain I, II, III and VII and none are found on TM domain IV through VI. Most features are also located very close to the intracellular (IC) domain rather than the extracellular (EC) domain. Even more impressive is their feature support. Only a number of very active features remained in the double optimised rule set, indicated by the orange-red colours. In Figure 12 the histogram shows the features rank and support of these features. All features are almost equally highly supported with a feature support of ≥ 0.8 which indicates a very general rule set. The very high feature support is gained through the compact rule set, shown in Table 7, where multiple features share the same rule. Almost all rules have a rule support > 0.75. However the last rule supports only one GPCR, probably because this GPCR differs very much in sequence compared to the main part. This will be discussed later in Section 4.3.

The feature ranks indicate the importance of a feature. However it only indicates the most important feature 7.46 and least important feature 2.35 of the current rule set. This is because the feature rank must be recalculated if all rules containing a certain feature are removed.

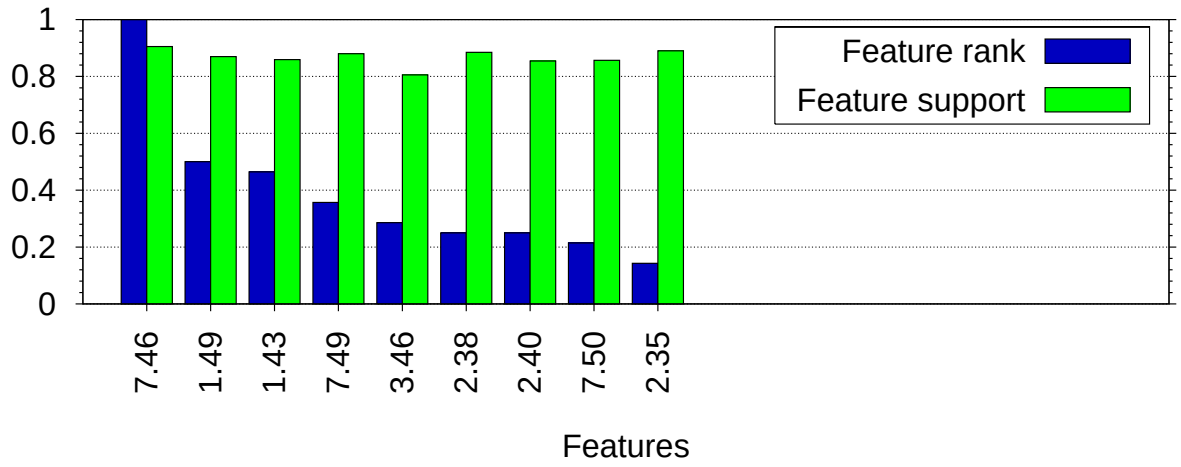


Figure 12: This histogram shows the features from the Tertius double optimised rule set. Each set of bars represents a feature with its rank N_{fc} and support S_{fc}, where class c is olfactory.

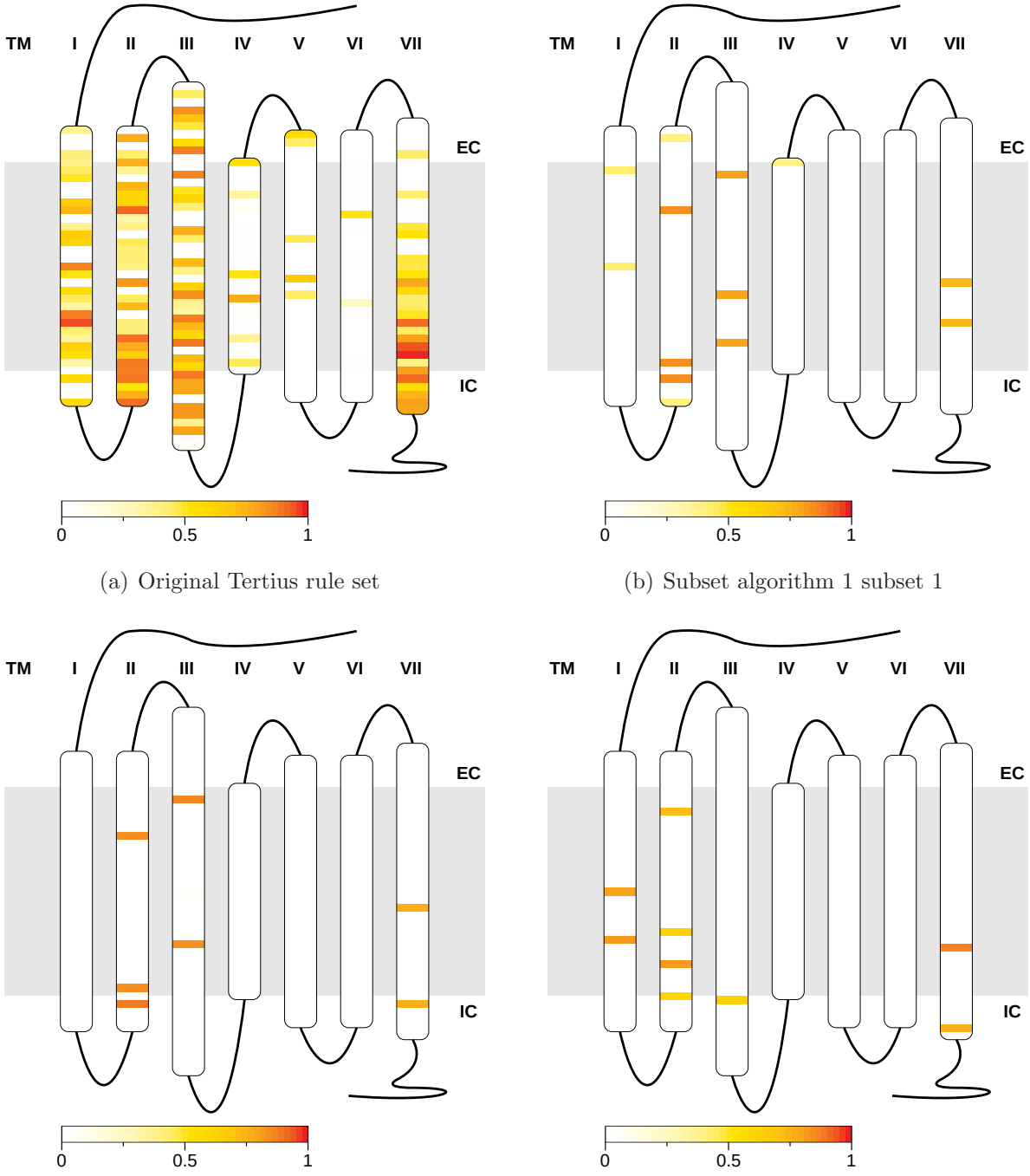
	Features								
S_r	1.43	1.49	2.35	2.38	2.40	3.46	7.46	7.49	7.50
0.824				P			P		P
0.816			L				P	N	P
0.806						M	P	N	
0.801	Y	G							
0.772		G		P	Y				
0.762	Y		L		Y				
0.003							-		

Table 7: The double optimised Tertius rule set. Each row represents a rule with olfactory conclusion. On the left the rule support S_r indicates the classification quality of that rule.

Tertius subsets All subset algorithms created a number of different subsets, where as only the first subset generation of each algorithm is used to compare the algorithm differences and their classification characteristics. This is done because each subset algorithm searches for the best rules according to their algorithm. So the first subsets contain the most optimal rules. The next generation of subsets will contain more specific rules, as described in Section 2.3, because the more general rules are already used in the previously generated subsets.

The first subsets generated by subset algorithm 1 through 3 are shown in Figure 13, where the 2D feature support visualisation of these subsets can be compared with each other and with the Tertius rule set, shown in Figure 13(a).

Subset algorithm 1, shown in Figure 13(b), covers other features with various feature support compared to the double optimised rule set. In this rule subset 13 different



(a) Original Tertius rule set

(b) Subset algorithm 1 subset 1

(c) Subset algorithm 2 subset 1

(d) Subset algorithm 3 subset 1

Figure 13: Comparing the Tertius rule set with the first three subset algorithms.

features were used in contrast with the 9 features used in the double optimised rule set. The features are also located on TM I, II, III and VII, however only TM II, III and VII are most active in this classification. The less active features, coloured in yellow, are still needed to completely classify the training set.

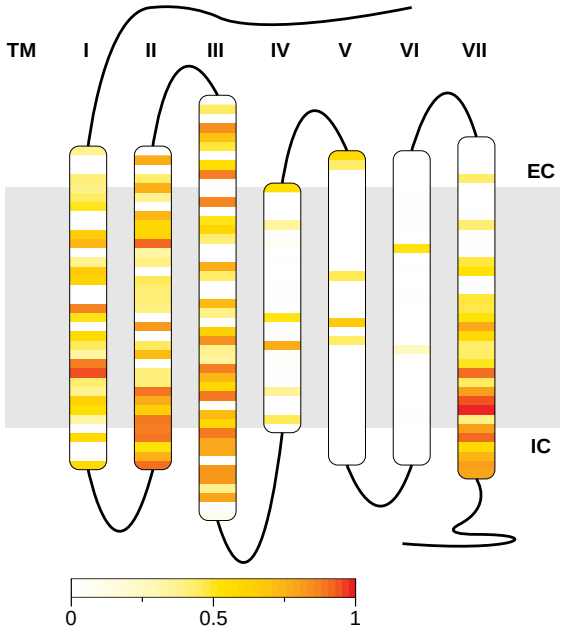
Figure 13(c) shows the results of subset algorithm 2, where only 7 very active features are visible. One invisible feature (3.40) is almost inactive and has been coloured white. This subset is very impressive and uses only 8 different features rather than 9 in the double optimised rule set. The used features should indicate a very compact characterisation to classify olfactory GPCRs. Once more the features only occurs on TM II, III and VII, but not on TM I as in the double optimised rule set.

Figure 13(d) shows subset algorithm 3. The subset covers a totally different feature set than used in the subsets before. Equal to the double optimised rule set there are only 9 different features used in the rule set to fully classify the training set. The location of the features is once again on TM I, II, III and VII. Later on a more detailed analyses of this subset will be discussed.

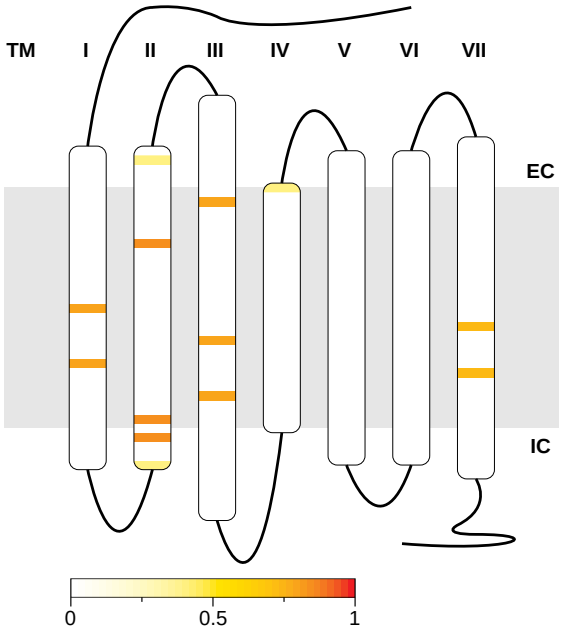
In Figure 14(b) shows algorithm 4 subset 1. The subsets generated by algorithm 1 and 4 are almost identical. When comparing the rule sets (not shown), only one rule differs. Moreover the features in TM I became more active in algorithm 4. So this subset algorithm should generate more robust subsets compared to subset algorithm 1.

After comparing the first three subsets between algorithm 3 and 6, which are equal, subset four is used for comparison. Both subsets are shown in the 2D feature support visualisation in Figure 14(c) and 14(d) respectively. A few very active features are used in both subsets, because the more active features were used in the previous three subsets. The classification of both subsets depend greatly on the less active features, which are coloured yellow. Nevertheless these features combined in a rule set still classify the complete training set. Furthermore, the features are not only located on TM I, II, III and VII but also on TM IV and V. So these two subsets use different classification characteristics.

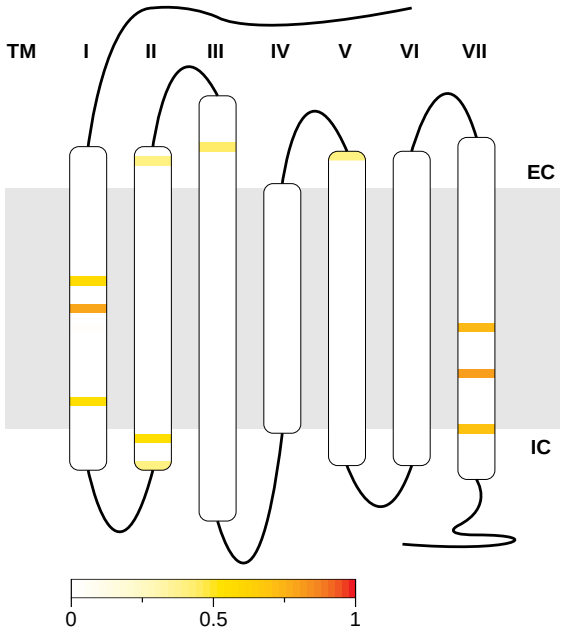
Notice that subset algorithm 5 is missing in Figure 14, because all generated subsets from algorithm 5 are equal to the subsets generated by subset algorithm 2. Apparently when using the lexicographical method, as described in Section 4.1, no rule in the training set could make a better subset.



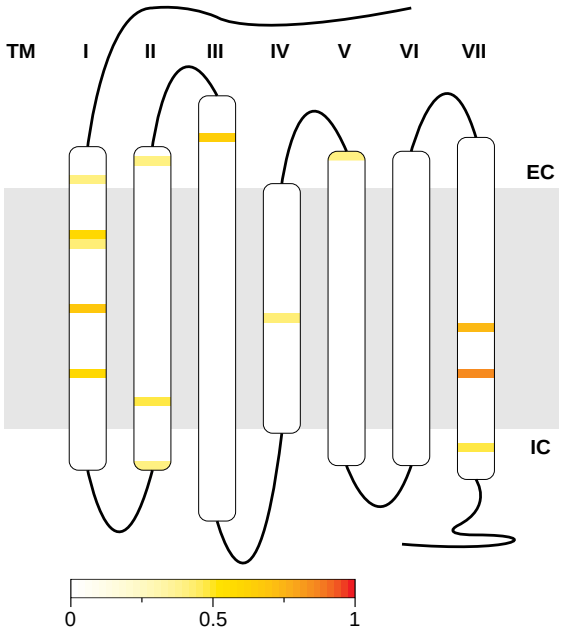
(a) Original Tertius rule set



(b) Subset algorithm 4 subset 1



(c) Subset algorithm 3 subset 4



(d) Subset algorithm 6 subset 4

Figure 14: Comparing the Tertius rule set with algorithm 4 subset 1 and the fourth subset from algorithm 3 and 6.

When comparing the subset more closely certain features return regularly in different subsets. All subsets and their features are shown in Table 8. Features with only one instance found in the subsets are removed from the table, because they cannot be compared.

	Features														
Rule set	1.43	1.49	2.35	2.38	2.40	2.59	2.68	3.28	3.43	3.46	3.49	4.67	5.32	7.41	7.46
Double optimised	×	×	×	×	×					×					×
Algorithm 1 subset 1	×		×	×	×	×	×	×	×		×	×		×	×
Algorithm 2 subset 1				×	×	×		×		×				×	
Algorithm 3 subset 1	×	×													×
Algorithm 4 subset 1	×	×	×	×	×	×	×	×	×		×	×		×	×
Algorithm 3 subset 4	×		×	×			×						×	×	×
Algorithm 6 subset 4	×		×				×						×	×	×

Table 8: This table shows only the features from the subsets which are used more than once. A rule subset contains a particular feature if it is marked with '×'.

When analysing the first 4 subsets, some features are more active than others. The features in the table are almost all very active, as can be seen in Figure 13 and 14 and should give a good characterisation for classifying olfactory GPCRs. Furthermore algorithm 3 subset 1 uses a totally different feature set, as indicated before. In Figure 15 a histogram shows the feature support in more detail. The first 3 features in the histogram are also used in other subsets and listed in Table 8.

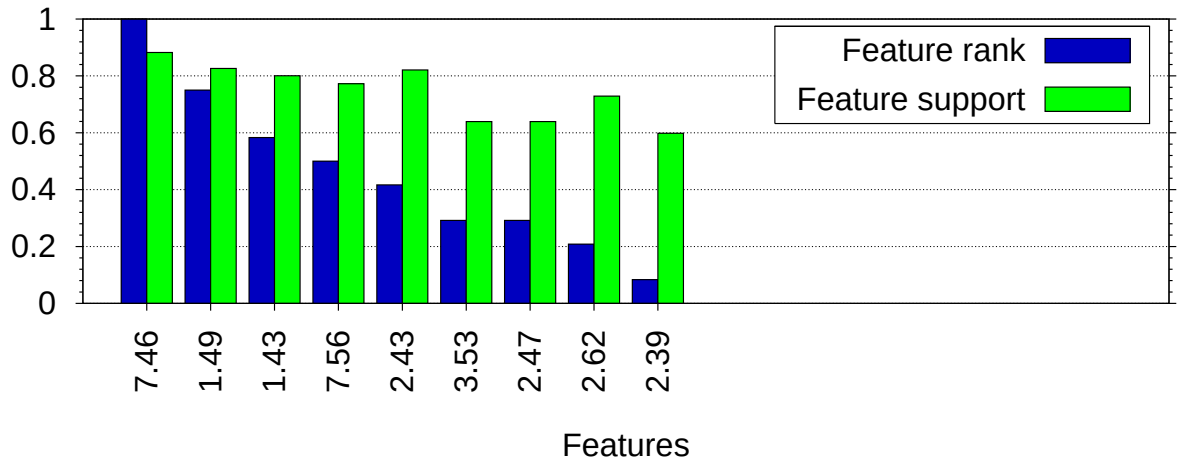


Figure 15: This histogram shows the features from the first subset generated by algorithm 3. Each set of bars represents a feature with its rank N_{fc} and support S_{fc}, where class c is olfactory.

The used features have an overall feature support of ≥ 0.6 and should indicate a robust rule set. The rule set is shown in Table 9 with a rule support of $\gtrsim 0.6$ for each

rule. This rule set contains three rules with only 2 terms in the body in contrast with the double optimised rule set where 3 term body rules are more frequently used. Smaller rules should be more robust in case an error or uncertainty occurs in one of the used features.

	Features								
S_r	1.43	1.49	2.39	2.43	2.47	2.62	3.53	7.46	7.56
0.801	Y	G							
0.772								P	R
0.688						L		P	
0.639				L	S		A		
0.599		G	M	L		L			

Table 9: The first rule subset generated by algorithm 3. Each row represents a rule with olfactory conclusion. On the left the rule support S_r indicates the classification quality of that rule.

Cross-validation subsets The used subset algorithms are tested using CV as described in Section 2.4. The CV is validated using the individual amino acid model as training data to test the subset algorithms. In Table 10 the CV results of all subset algorithms are shown. The classification quality is impressive, the algorithms only introduce a very small error, which is neglectable.

Algorithm	1	2	3	4	5	6
Classification quality	99.82	99.79	99.78	99.84	99.79	99.80
Standard deviation	0.02	0.05	0.03	0.02	0.03	0.02

Table 10: The average classification quality and its standard deviation for each subset algorithm, obtained using 10-fold cross-validation as described in Section 2.4

Amino acid group rule set Using the method as described in Section 4.1, a rule set with 16842 rules covering 224 features was generated and called the Tertius amino acid group rule set. The same optimisation method as in Section 4.2 is used with subset algorithm 3 or 6 as second optimisation. After this optimisation one rule set with 8 rules covering 10 features called the double optimised amino acid group rule set remains. In Figure 16 the found features are shown in a histogram. The features: 7.46, 1.43, 2.35, 2.40 and 3.46 are also used in the individual amino acid Tertius double optimised rule set. The remaining features are used in various subsets as well. This result will be discussed later in Section 4.3.

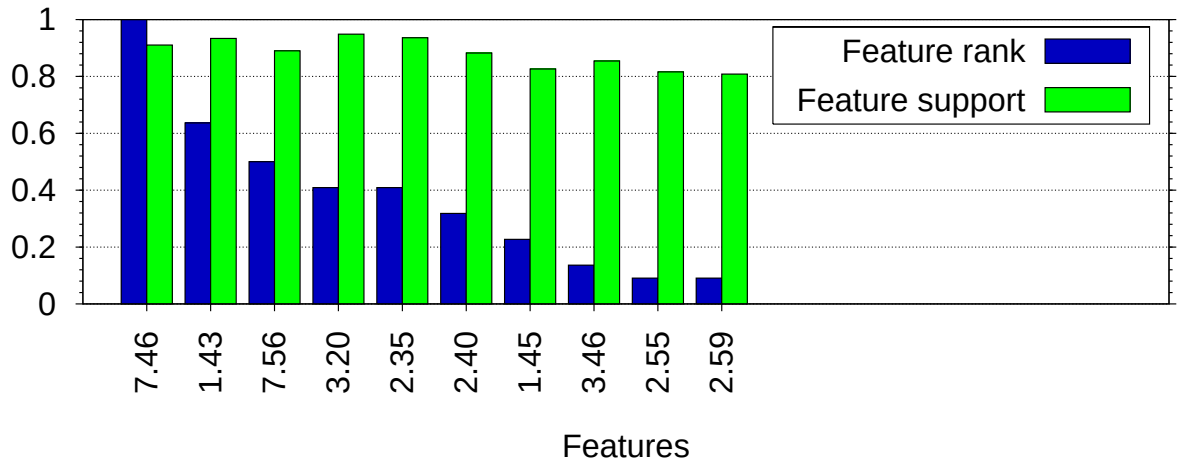


Figure 16: This histogram shows the features from the double optimised Tertius amino acid group rule set. Each set of bars represents a feature with its rank N_{fc} and support S_{fc}, where class c is olfactory.

Conserved amino acids In Figure 17 both the olfactory and non-olfactory sub-alignments[5] are presented as sequence logos [64, 27] similar to Figure 5. The marked features indicate approximately equal conserved amino acids which are both found in non-olfactory and olfactory GPCRs. These amino acids should play an insignificant role in classification because they do not contribute to differentiating the classification characteristics.

However these feature amino acid combinations are frequently chosen in rules generated by Tertius. Tertius uses these features because the classification quality could be increased in combination with other features. While this may be true for features like 2.45, 2.51, 4.50 and 6.50, the other marked feature amino acid combinations are approximately completely conserved over all GPCRs. If such feature amino acid combination is added to a hypothetical rule in Tertius then the rule is unique but the rule support would be approximately equal to the original rule. So the addition of such a feature amino acid combination does not increase the classification quality, although the classification quality is not decreased either.

[5]The amino acid letters are coloured according to the amino acid group model, described in Section 2.1. However in this case the colouration has no meaning other than giving the amino acids some contrast.

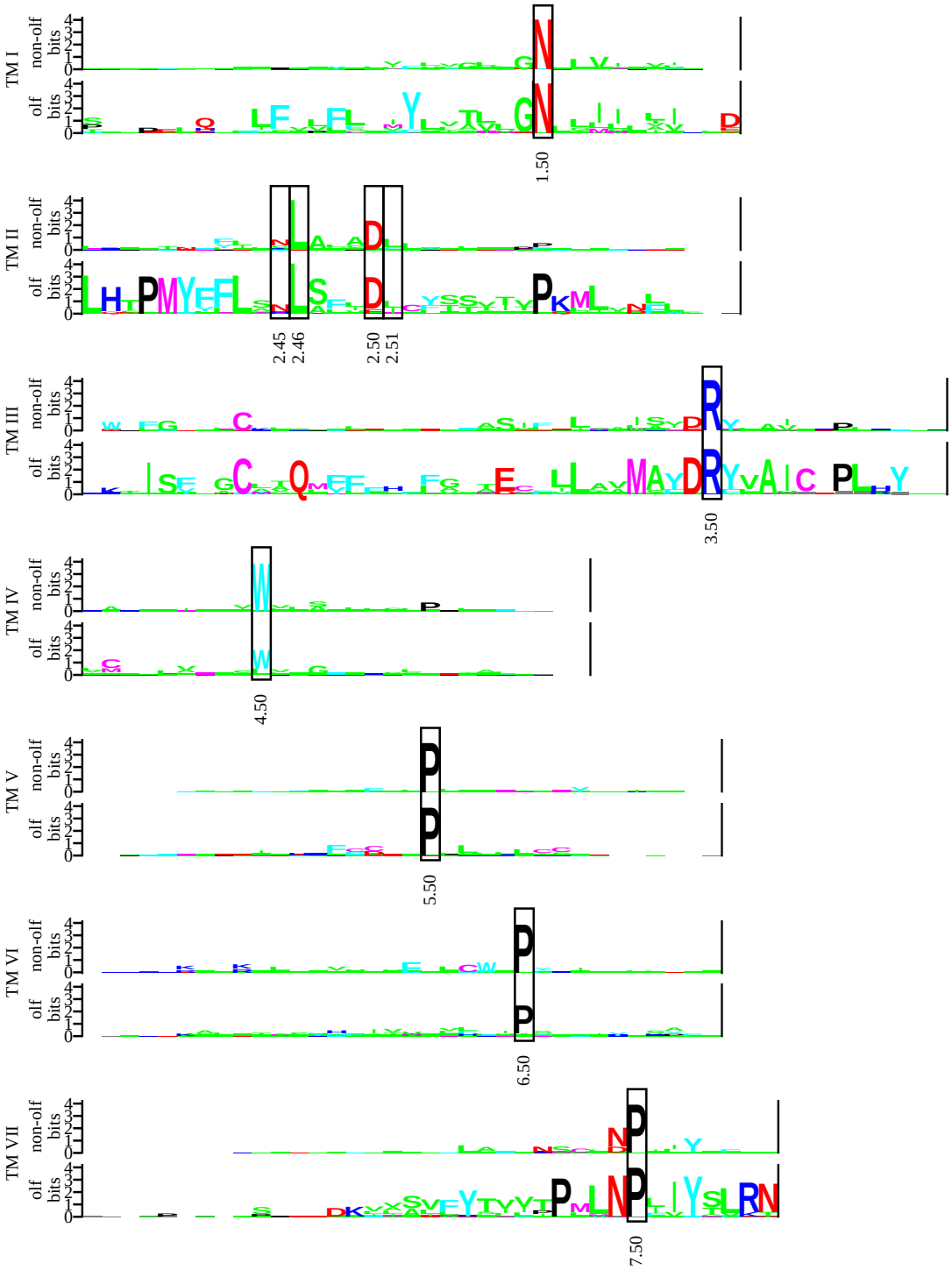


Figure 17: Alignment of the seven transmembrane (TM) regions of the human GPCRs, shown as sequence logos of sub-alignments containing olfactory and non-olfactory receptors respectively as defined in Section 2.1. Features where the same amino acid is approximately equally conserved in non-olfactory and olfactory are marked with solid boxes.

4.3 Discussion

There are some discussion points regarding Tertius and the additional algorithms used in this section.

Improving Tertius rule discovery As discussed in Section 4.2, the conserved amino acids should be pruned before the rule discovery starts. With the result that only rules without these conserved amino acids can be discovered.

Less conserved GPCRs After the alignment most GPCRs have many conserved features. However there are a number of GPCRs that cannot match the induced alignment. As a result these GPCRs have only very few conserved features that could be used for classification. Nevertheless Tertius discovered a number of rules to classify such GPCRs, although these rules only have a small rule support. For example the double optimised rule set contains a rule that only classify one GPCR as discussed in Section 4.2. This kind of GPCRs lack the intermediate GPCR characteristics the alignment induced within the chosen rule set. The intermediate characteristics of a GPCR rule set could be illustrated by the 2D feature support visualisation as can be seen in Figure 9. This intermediate visualisation only shows the classification activity rather than the composition of rule set. In other words the rule set used for classification is invisible in the 2D feature support visualisation. For example the rule set generated by subset algorithm 3, also discussed in Section 4.2, contain only rules with a rule support of $\gtrsim 0.6$ as shown in Table 9. So the 2D feature support visualisation, shown in Figure 13(d), do not show equal values like the rule support, but rather the usage of the features in these rules.

Amino acid group rule set As described in Section 4.1 additional experiments are done using the amino acid group model for Tertius and related algorithms. The results 4.2 show that the features found are not unique. Only other classification characteristics came forward through other feature combinations used in the rule set. So the use for the amino acid group model is not necessary. Moreover Tertius uses more computer memory and computational time for discovering rules using the amino acid group model. Because in the way Tertius works, more hypothetical rule combinations that exist can be explored, resulting in an exponential search space. As a result no computational time is gained when using this model for preprocessing.

4.4 Conclusion

Tertius uses a different approach to discover rules compared to PRISM or the Fuzzy logic method used by Samsonova et al. [63]. This approach resulted in rule sets containing more general rules compared to the very specific rules used in the PRISM rule sets. Moreover Tertius introduces different kinds of characteristics which appear through the various rule sets generated using the optimisation and subset algorithms.

The results in Section 4.2 show a range of classification characteristics in contrast with [63] where only one characterisation is presented. Nevertheless some features are also found in our results. Yet their impact on classification is completely different.

The 2D feature support visualisation as shown in Figure 9 shows the complete set of features that can be used to characterise the olfactory GPCRs. Various subsets are extracted from the Tertius rule set, using the optimisation and subset algorithms. These subsets introduce various intermediate GPCR characteristics that could characterise the olfactory GPCRs.

In conclusion Tertius has an outstanding ability to discover rules compared to PRISM or the method used by Samsonova et al. [63]. The goal to find various characterisations has been reached using the subset and optimisation algorithms. With this result in mind it can be concluded that the olfactory GPCRs can be classified using rule sets which cover various characterisations.

5 Discussion

In this section a number of areas related to GPCRs are discussed. First a comparison between motifs and the results found in the Tertius section are discussed. After that the MSA will be discussed because the results in this thesis depend greatly on the chosen alignment. Continuing with the importance of uncovering the true 3D structure of GPCRs, where structural alignment could be an alternative for the MSA. This section ends with some words regarding the developed visualisation and possible future work.

Motifs The most common characterisation method in literature to classify proteins are motifs. Motifs are based on chronological sequences of conserved amino acids [45, 79] which characterise a certain protein class. Several motifs are characteristic for olfactory GPCRs as already discussed in Samsonova et al. [63]. The features found in this thesis also contain the conserved amino acids used in the motifs [45, 79] located on the transmembrane (TM) domains. However the features are not necessarily in sequential order but found throughout the TM domains, which is also the case in Samsonova et al. [63].

The idea behind motifs is that the conserved sequence represents a key structure which is associated with functions of the receptor, like the binding side for both ligand and/or G-protein [5]. When reviewing the non-sequential features from the Tertius section, no obvious structure can be specified that could associate to a receptors function. Nevertheless the found features should reveal characteristic locations when mapped upon the bovine rhodopsin model. In other words the areas around a very active feature should play a significant role in classifying the olfactory receptor class. These characteristic locations could also play an active role in the receptor's structure and its function, which will be discussed later on.

Both the results in this thesis and the motifs are based on a chosen multiple sequence alignment (MSA), which uses the structural knowledge of bovine rhodopsin. However a MSA could be generated using other specific criteria which may introduce a better alignment for the olfactory receptor.

Alignments Besides the multiple sequence alignment (MSA) used in the GPCRDB [30] other MSA methods could be used to find other desirable GPCRs alignments. For instance, the MSA used in the GPCRDB, release February 2004, uses predefined transmembrane (TM) domains. However in the new release of March 2005, the numbering system has been modified and now concerns helices and not the TM domains [41]. As a result the non-transmembrane helix 8 has been added. At the same time TM domain 1 through 7 are already helices, so helix 8 is only an addition to the model. This illustrates the choice of adopting other models in a multiple sequence alignment (MSA).

There are some methods which use raw sequences rather than aligned ones. For example the alignment-independent [39] or alignment free [53] methods which are used to classify GPCRs suggests that an alignment is not needed. However some sort of intermediate alignments are made inside the used algorithms. For instance, the alignment-independent method makes use of self-organising maps [37, 68]. Other methods, which use unaligned sequences, characterise the sequences using a composition of amino acid n-tuples [15] or pairwise posterior probabilities [61]. Using only the sequence for classification methods like Nearest Neighbour approach (BLAST) [44], Hidden Markov Model (HMM) [20] and Support Vector Machines (SVMs) [33, 14] are frequently used.

As indicated before, GPCRDB uses a predefined set of transmembrane (TM) domains. These TM domains are hydrophobic and should remain inside the membrane due to there hydrophobic character. There are more alignment methods that are focused on structural elements like thermodynamic stability [16], where the stability of the amino acid pairs are taken into account. Just recently some progress has been made to use structural information from known proteins to be used in the effort to get a better structural alignment [57, 2].

3D Structure As already described in Section 2.1 the true three dimensional structure of GPCRs is still not known. However when using the known 3D structure of bovine rhodopsin [54, 49], which is of the same class as olfactory, GPCRs could be modelled using this structure as a template. This can only be done because the majority of the GPCRs share the seven transmembrane (TM) helix topology [78]. However some GPCRs are predicted to have structures which may diverge from that of bovine rhodopsin [60].

It is crucial to comprehend the true 3D structural information so it can be used to unravel its function. For example, the bovine rhodopsin model is a static model yet the GPCRs are dynamic when interacting with a ligand. So the activated structure of rhodopsin may provide important information which could be useful for pharmaceutical drug design [5].

The methods used in this thesis to find characteristic amino acid combinations for classification might also be useful in structure prediction. Using structural alignment [6] rather than multiple sequence alignment (MSA) should uncover important positions within the 3D structure. These structures may reveal key information on the functional behaviour of GPCRs. Proteins of the same class frequently share secondary structures, like helices in GPCRs. The arrangement of these local structures may be in a different order as appears in the sequence. When aligning these local structures, different topologies can be exploited [77, 38]. Structural information is also used to predict G-protein coupling selectivity [75] which is also useful from the viewpoint of drug design.

Furthermore, homology modelling of GPCRs [56] and using the interaction between theory and experiment may provide useful information to simulate the dynamic structure of the receptors in the process when a ligand is binding [46].

The field of structural based methods is very active, because it might be possible to predict the function and metabolism of a drug target directly from its 3D structure. This also could make the requirement for animal testing obsolete in due time [42].

Visualisation The visualisation model used to visualise the feature support of a rule set, as shown in Figure 9, could be used to visualise other data. For example data like entropy, consensus, mutation activity, etc. could be visualised upon the 2D GPCR model. This model can be found in the Rule set Optimisation Package as described in Appendix A.

Future Work As previously discussed, it is interesting to investigate the kind of characteristic compositions a structural alignment or other specialised alignments could introduce.

Furthermore the additional subset algorithm could be enhanced or new algorithms could be inserted only to select special features. These features should be selected on experimental feasibility or features where experimental data already exist. With this in mind, biological experiments could be done to gain specific results.

The characteristic composition of other GPCR classes and or proteins besides the olfactory receptors should be investigated. The methods described in this thesis should also be analysed and tested if they are valuable to pharmaceutical drug design.

6 Conclusions

This thesis provides a method of finding characteristic amino acid combinations in the olfactory G protein-coupled receptors (GPCRs). Based on multiple sequence alignment (MSA) and rule discovery, two methods where used to find these characterisations.

First a very simple method called PRISM was tested and induced a rule set which classifies the olfactory training set. In spite of using an optimisation method the induced rules where too specific and resulted in an overall low feature support. Accordingly almost all features present in the rule set were unusable to be characteristic for the olfactory receptors.

Second a more complex algorithm, Tertius was tested. Together with some additional subset algorithms and the optimisation method multiple characterisations came forward. The best subsets contain highly supported features which indeed are very characteristic for the olfactory receptors. Moreover the rule subsets successfully classify the olfactory GPCRs using these characteristic features.

Overall the Tertius algorithm combined with additional methods introduced a very reliable way to find characteristic feature combinations. Furthermore the resulting characterisatis are non-sequential in contrast with the well known motifs, this may provide new insights on biological experiments. Moreover when combining these methods with structural alignment, it could result in a new understanding in characteristic structural elements within the 3D structure of GPCRs.

7 Acknowledgements

First I want to thank my tutor Elena for her support. After working on a software project, also under Elena's supervision, the opportunity for doing my masters thesis with her support was an easy choice. It was fun working with you, thanks Elena.

I also want to thank my mentors Prof. Dr. Thomas Bäck and Prof. Dr. Joost Kok for giving me advice and comments about my thesis.

In conclusion I want to thank all people whom I worked with during the time I spend on the LACDR faculty under supervision of Prof. IJzerman. In particular I what to thank Eric-Wubbo, Jeroen, Jan-Willem and Kai for helping me with various issues regarding the writing of my thesis and the development of the software.

Last but not least I want to thank my wife Marietje for her support. I also want to thank a friend of mine, Stephan for helping me with the writing of my thesis.

Appendix

A Software

The Rule set Optimisation Package includes the following items, which are described shortly in the next subsections:

- Rule set Optimisation Application (RSOA).
- Additional scripts.
- 2D GPCR visualisation.

A.1 Rule set Optimisation Application

The Rule set Optimisation Application (RSOA) is a command line based program written in C++. The application is still in a development stage and includes the following methods: Classification, Optimisation, Subsets generation and Cross-validation. A short description about the components is described in the paragraphs below.

Classification The classification quality as defined in Section 2.2 classifies the training set using a given rule set.

Optimisation The optimisation method as described in Section 3.1 is implemented and can automatically optimise a rule set toward a given classification quality.

Subsets generation The subset algorithms, which are described in Section 4.1, are implemented to be used individually. Using a training set and a rule set, the subset algorithm generates multiple rule subsets according to the selected subset algorithm.

Cross-validation The cross-validation (CV) method as described in 2.4 has been implemented. Besides the implementation some shell scripts were used to automate the steps involved. The implementation is used internally to measure the introduced error of the optimisation and subset algorithms and externally to measure the error that PRISM introduced.

A.2 Additional scripts

A number of scripts are used in combination with the RSOA. The scripts are involved in converting the Weka [74] PRISM and Tertius output to the space separated value file format used by RSOA, where the values are separated using white spaces. The conversion scripts are written in Perl. A few other scripts are only needed to simplify the removal of features, positions or to generate files needed by the RSOA.

A.3 2D GPCR visualisation

Also included in the Rule set Optimisation Package is the 2D visualisation which visualises the feature support of a rule set as shown in Figure 9. The visualisation was made using Postscript [12] and can be modified to include other data besides the feature support. Currently the visualisation can only be edited manually by editing the postscript file. However this process could be automated using a simple script.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994. ISBN 1-55860-153-8.

[2] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucl. Acids Res.*, 34(suppl 2):W604–608, 2006. doi: 10.1093/nar/gkl092. URL `http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl_2/W604`.

[3] T. Bäck. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, Oxford, UK, 1996. ISBN 0-19-509971-0.

[4] J. A. Ballesteros and H. Weinstein. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in Neurosciences*, 25:366–428, 1995.

[5] L. Bosch, L. Iarriccio, and P. Garriga. New prospects for drug discovery from structural studies of rhodopsin. *Current Pharmaceutical Design*, 11(17):2243–2256, July 2005. doi: 10.2174/1381612054367436.

[6] P. E. Bourne and I. N. Shindyalov. Structure comparison and alignment. In P. E. Bourne and H. Weissig, editors, *Structural Bioinformatics*. Wiley-Liss, Hoboken NJ, February 2003. ISBN 0-471-20200-2. doi: 10.1002/0471721204.ch16. URL `http://dx.doi.org/10.1002/0471721204.ch16`.

[7] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido, and S. Ahmad. A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics*, 19(2):234–240, 2003. doi: 10.1093/bioinformatics/19.2.234. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/2/234`.

[8] J. Cendrowska. Prism: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27(4):349–370, 1987.

[9] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–826, April 1986. ISSN 0261-4189. URL `http://www.pubmedcentral.gov/articlerender.fcgi?tool=pubmed&pubmedid=3709526`.

[10] E. E. Conn and P. K. Stumpf. *Outlines of Biochemistry*. John Wiley and Sons, New York, 1963. ISBN 0471168440.

[11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. stein. *Introduction to Algorithms*. MIT Press, 2nd edition, 2001. ISBN 0262531968.

[12] Corporate Adobe Systems Inc. *PostScript language reference manual*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition, 1990. ISBN 0-201-18127-4.

[13] C. Crasto, M. S. Singer, and G. M. Shepherd. The olfactory receptor family album. *Genome Biology*, 2(10):reviews1027.1–1027.4, 2001. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=11597337`.

[14] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines (and other kernel-based learning methods)*. Addison-Wesley Longman Publishing Co., Inc., 2000. ISBN 0-521-78019-5.

[15] F. Daeyaert, H. Moereels, and P. Lewi. Classification and identification of proteins by means of common and specific amino acid n-tuples in unaligned sequences. *Computer Methods and Programs in Biomedicine*, 56(3):221–233, June 1998. doi: doi:10.1016/S0169-2607(98)00031-5.

[16] A. R. Davidson. Multiple sequence alignment as a guideline for protein engineering strategies, 2006. URL `http://biomed.humanapress.com/index.php?option=com_opbookdetails\&task=chapterdetails\&chapter_code=1-59745-116-9:171\&category=biomedprotocols`.

[17] W. DeLano. The PyMOL molecular graphics system, 2002. URL `http://www.pymol.org`.

[18] A. Deltour. Tertius extension to Weka. Technical Report CSTR-01-001, Department of Computer Science, University of Bristol, September 2001. URL `http://www.cs.bris.ac.uk/Publications/Papers/1000568.pdf`.

[19] G. Drutel, J. Arrang, J. Diaz, C. Wisnewsky, K. Schwartz, and J. Schwartz. Cloning of OL1, a putative olfactory receptor and its expression in the developing rat heart. *Receptors Channels*, 3(1):33–40, 1995.

[20] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999. ISBN 0-521-62971-3.

[21] E. Feldmesser, T. Olender, M. Khen, I. Yanai, R. Ophir, and D. Lancet. Widespread ectopic expression of olfactory receptor genes. *BMC Genomics*, 7:121, 2006. URL `http://www.citebase.org/abstract?id=oai:biomedcentral.com:1471-2164-7-121`.

[22] P. A. Flach and N. Lachiche. Confirmation-guided discovery of first-order rules with Tertius. *Mach. Learn.*, 42(1/2):61–95, 2001. ISSN 0885-6125. doi: 10.1023/A:1007656703224.

[23] A. A. Freitas. A critical review of multi-objective optimization in data mining: a position paper. *SIGKDD Explorations*, 6(2):77–86, December 2004. URL `http://www.cs.kent.ac.uk/pubs/2004/2042`.

[24] A. A. Freitas. Understanding the crucial differences between classification and discovery of association rules - a position paper. *SIGKDD Explorations*, 2(1):65–69, 2000. URL `http://citeseer.ist.psu.edu/freitas00understanding.html`.

[25] I. Gaillard, S. Rouquier, and D. Giorgi. Olfactory receptors. *Cellular and Molecular Life Sciences (CMLS)*, 61(4):456–469, February 2004. doi: 10.1007/s00018-003-3273-7.

[26] M. R. Garey and D. S. Johnson. *Computer and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, 1979. ISBN 0-7167-1044-7.

[27] J. Gorodkin, L. J. Heyer, S. Brunak, and G. D. Stormo. Displaying the information contents of structural RNA alignments: the structure logos. *Computer Applications in Biosciences*, 13(6):583–586, 1997. URL `http://www.cbs.dtu.dk/~gorodkin/appl/slogo.html`.

[28] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[29] M. Hegland. Algorithms for association rules. pages 226–234, 2003.

[30] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. E. Cohen, and G. Vriend. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Research*, 31(1):294–297, 2003. URL `http://www.gpcr.org/7tm/`.

[31] H. Hu and J. Li. Using association rules to make rule-based classifiers robust. In *CRPIT '39: Proceedings of the sixteenth Australasian conference on Database technologies*, pages 47–54, Darlinghurst, Australia, Australia, 2005. Australian Computer Society, Inc. ISBN 1-920-68221-X.

[32] A. Kandel. *Fuzzy mathematical techniques with applications*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1986. ISBN 0-201-11752-5.

[33] R. Karchin, K. Karplus, and D. Haussler. Classifying G-protein coupled receptors with support vector machines . *Bioinformatics*, 18(1):147–159, 2002. doi: 10.1093/bioinformatics/18.1.147. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/1/147`.

[34] K. A. Kaufman and R. S. Michalski. Learning from inconsistent and noisy data: The aq18 approach. In *International Syposium on Methodologies for Intelligent Systems*, pages 411–419, 1999. URL `http://citeseer.ist.psu.edu/article/kaufman99learning.html`.

[35] Y. S. Kim, W. N. Street, and F. Menczer. Feature Selection in Unsupervised Learning via Evolutionary Search. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000. URL `http://citeseer.ist.psu.edu/kim00feature.html`.

[36] P. S. Klosterman, M. Tamura, S. R. Holbrook, and S. E. Brenner. Scor: a structural classification of rna database. *Nucleic Acids Research*, 30(1):392–394, 2002. URL `http://scor.lbl.gov/`.

[37] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Heidelberg, 1995. ISBN 3-540-67921-9.

[38] E. Krissinel and K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D*, 60(12 Part 1):2256–2268, December 2004. doi: 10.1107/S0907444904026460. URL http://dx.doi.org/10.1107/S0907444904026460.

[39] M. Lapinsh, A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, and J. E. Wikberg. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Science*, 11(4): 795–805, 2002. doi: 10.1110/ps.2500102. URL http://www.proteinscience.org/cgi/content/abstract/11/4/795.

[40] J. Li, R. Topor, and H. Shen. Construct robust rule sets for classification. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 564–569, New York, NY, USA, 2002. ACM Press. ISBN 1-58113-567-X. doi: 10.1145/775047.775130.

[41] J. Li, P. C. Edwards, M. Burghammer, C. Villa, and G. F. Schertler. Structure of bovine rhodopsin in a trigonal crystal form. *Journal of Molecular Biology*, 343(5): 1409–1438, November 2004. ISSN 0022-2836. doi: 10.1016/j.jmb.2004.08.090. URL http://dx.doi.org/10.1016/j.jmb.2004.08.090.

[42] J. C. Madden and M. T. Cronin. Structure-based methods for the prediction of drug metabolism. *Expert Opinion on Drug Metabolism & Toxicology*, 2(4):545–557, 2006. doi: 10.1517/17425255.2.4.545. URL http://www.expertopin.com/doi/abs/10.1517/17425255.2.4.545.

[43] S. Martello and P. Toth. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., New York, NY, USA, 1990. ISBN 0-471-92420-2.

[44] S. McGinnis and T. L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucl. Acids Res.*, 32(suppl 2):W20–25, 2004. doi: 10.1093/nar/gkh435. URL http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_2/W20.

[45] P. Mombaerts. Seven-transmembrane proteins as odorant and chemosensory receptors. *Science*, 286:707–711, 1999. doi: 10.1126/science.286.5440.707.

[46] S. Moro, F. Deflorian, M. Bacilieri, and G. Spalluto. Ligand-based homology modeling as attractive tool to inspect gpcr structural plasticity. *Current Pharmaceutical Design*, 12(17):2175–2185, June 2006. doi: 10.2174/138161206777585265.

[47] P. Nambi and N. Aiyar. G protein-coupled receptors in drug discovery. *ASSAY and Drug Development Technologies*, 1(2):305–310, 2003. doi: 10.1089/15406580360545116. URL http://www.liebertonline.com/doi/abs/10.1089/15406580360545116.

[48] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*. Worth Publishers, third edition, 2000. ISBN 1572599316.

[49] T. Okada, M. Sugihara, A.-N. Bondar, M. Elstner, P. Entel, and V. Buss. The retinal conformation and its environment in rhodopsin in light of a new 2.2 å crystal structure. *Journal of Molecular Biology*, 342(2):571–583, September 2004.

[50] L. Oliveira, A. C. M. Paiva, and G. Vriend. Correlated mutation analyses on very large sequence families. *ChemBioChem*, 3(10):1010–1017, 2002. doi: 10.1002/1439-7633(20021004)3:10⟨1010::AID-CBIC1010⟩3.0.CO;2-T.

[51] L. Oliveira, P. B. Paiva, A. C. M. Paiva, and G. Vriend. Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein. *Proteins*, 52 (4):553–560, September 2003. doi: 10.1002/prot.10489.

[52] C. Ordonez. Association rule discovery with the train and test approach for heart disease prediction. *Information Technology in Biomedicine, IEEE Transactions on information technology in biomedicine [1089-7771]*, 10(2):334–343, April 2006. doi: doi:10.1109/TITB.2006.864475.

[53] J. M. Otaki, A. Mori, Y. Itoh, T. Nakayama, and H. Yamamoto. Alignment-free classification of G-protein-coupled receptors using self-organizing maps. *Journal of chemical information and modeling*, 46(3):1479–1490, 2006. doi: 10.1021/ci050382y.

[54] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, 289(5480): 739–745, 2000. doi: 10.1126/science.289.5480.739.

[55] G. L. Pappa, A. A. Freitas, and C. A. A. Kaestner. A multiobjective genetic algorithm for attribute selection. In A. Lofti, J. Garibaldi, and R. John, editors, *Proc. 4th Int. Conf. on Recent Advances in Soft Computing (RASC-2002)*, pages 116–121. Nottingham Trent University, December 2002. ISBN 1842330764. URL `http://www.cs.kent.ac.uk/pubs/2002/1789`.

[56] A. Patny, P. V. Desai, and M. A. Avery. Homology modeling of G-protein-coupled receptors and implications in drug design. *Current Medicinal Chemistry*, 13(14): 1667–1691, June 2006. doi: 10.2174/092986706777442002.

[57] J. Pei and N. V. Grishin. MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. *Nucl. Acids Res.*, page gkl514, 2006. doi: 10.1093/nar/gkl514. URL `http://nar.oxfordjournals.org/cgi/content/abstract/gkl514v1`.

[58] T. H. Pham, J. C. Clemente, K. Satou, and T. B. Ho. Computational discovery of transcriptional regulatory rules. *Bioinformatics*, 21(suppl 2):ii101–107, 2005. doi: 10.1093/bioinformatics/bti1117. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/suppl_2/ii101`.

[59] V. Podgorelec, P. Kokol, M. M. Stiglic, M. Hericko, and I. Rozman. Knowledge discovery with classification rules in a cardiovascular dataset. *Computer Methods and Programs in Biomedicine*, 80(suppl 1):S39–S49, December 2005. doi: 10.1016/S0169-2607(05)80005-7. URL http://www.sciencedirect.com/science/article/B6T5J-4JD1XJX-5/2/79640be8a307b50a1788fd8a0e811b8f.

[60] P. H. Reggio. Computational methods in drug design: modeling G protein-coupled receptor monomers, dimers, and oligomers. *AAPS Journal*, 8(2):E322–E336, May 2006. doi: 10.1208/aapsj080237. URL http://www.aapsj.org/view.asp?art=aapsj080237.

[61] U. Roshan and D. R. Livesay. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, page btl472, 2006. doi: 10.1093/bioinformatics/btl472. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btl472v1.

[62] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, first edition, 1995. ISBN 0131038052.

[63] E. V. Samsonova, P. Krause, T. Bäck, and A. P. IJzerman. Characteristic amino acid combinations in olfactory G protein-coupled receptors. *Proteins: Structure, Function, and Bioinformatics*, January 2007. doi: 10.1002/prot.21112. URL http://dx.doi.org/10.1002/prot.21112.

[64] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990. URL http://citeseer.ist.psu.edu/schneider90sequence.html.

[65] K. Schwarzenbacher, J. Fleischer, H. Breer, and S. Conzelmann. Expression of olfactory receptors in the cribriform mesenchyme during prenatal development. *Gene Expr Patterns*, 4(5):543–552, 2004. doi: 10.1016/j.modgep.2004.02.004.

[66] M. R. Segal, M. P. Cummings, and A. E. Hubbard. Relating amino acid sequences of phenotype: Analysis of peptide binding data. *Biometrics*, 57(2):632–643, June 2001. URL http://repositories.cdlib.org/cbmb/peptide_binding/.

[67] N. G. Sgourakis, P. G. Bagos, P. K. Papasaikas, and S. J. Hamodrakas. A method for the prediction of gpcrs coupling specificity to G-proteins using refined profile hidden markov models. *BMC Bioinformatics*, 6(1), 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-104. URL http://dx.doi.org/10.1186/1471-2105-6-104.

[68] W.-H. Steeb. *The Nonlinear Workbook*, chapter 11.4: Kohonen Networks. World Scientific, 1999. ISBN 9810240260.

[69] S. Takeda, S. Kadowaki, T. Haga, H. Takaesu, and S. Mitaku. Identification of G protein-coupled receptor genes from the human genome sequence. *FEBS Letters*, 520(1-3):97–101, June 2002. doi: 10.1016/S0014-5793(02)02775-8.

[70] The Persistence Of Vision Development Team. POV-Ray - the Persistence of Vision Raytracer. URL http://www.povray.org/.

[71] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22(22):4673–4680, 1994. doi: 10.1093/nar/22.22.4673. URL http://nar.oxfordjournals.org/cgi/content/abstract/22/22/4673.

[72] H. Vafaie and K. DeJong. Robust feature selection algorithms. *Proc. 5th Intl. Conf. on Tools with Artifical Intelligence*, pages 356–363, 1993. URL http://citeseer.ist.psu.edu/vafaie93robust.html.

[73] J. T. Wang, S. Rozen, B. A. Shapiro, D. Shasha, Z. Wang, and M. Yin. New techniques for DNA sequence classification. *Journal of Computational Biology*, 6(2):209–218, 1999. URL http://citeseer.ist.psu.edu/article/wang99new.html.

[74] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, June 2005. ISBN 0-12-088407-0. URL http://www.cs.waikato.ac.nz/~ml/weka/.

[75] Y. Yabuki, T. Muramatsu, T. Hirokawa, H. Mukai, and M. Suwa. GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. *Nucl. Acids Res.*, 33(suppl 2):W148–153, 2005. doi: 10.1093/nar/gki495.

[76] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, 13:44–49, 1998. URL http://citeseer.ist.psu.edu/yang98feature.html.

[77] X. Yuan and C. Bystroff. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics*, 21(7):1010–1019, November 2005. doi: 10.1093/bioinformatics/bti128. URL http://dx.doi.org/10.1093/bioinformatics/bti128.

[78] Y. Zhang, M. E. E. Devries, and J. Skolnick. Structure modeling of all identified g protein-coupled receptors in the human genome. *PLoS Comput Biol*, 2(2), February 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020013. URL http://dx.doi.org/10.1371/journal.pcbi.0020013.

[79] S. Zozulya, F. Echeverri, and T. Nguyen. The human olfactory receptor repertoire. *Genome Biology*, 2(6):research0018.1–0018.12, 2001. doi: 10.1186/gb-2001-2-6-research0018.