UPTEC X 02 002
JAN 2002

ISSN 1401-2138

INGRID GUNNARSSON

Multivariate analysis of G protein-coupled receptors

Master's degree project

Molecular Biotechnology Programme
Uppsala University School of Engineering

UPTEC X 02 002 | **Date of issue 2002-01**

Author

Ingrid Gunnarsson

Title (English)

Multivariate analysis of G protein-coupled receptors

Title (Swedish)

Abstract

A large number of G protein-coupled receptors have been investigated with respect to groupings among the sequences based on their physicochemical properties using multivariate methods. Initially, the trans membrane (TM) regions of roughly 900 receptors were examined. In addition, the complete sequences of a smaller subset of receptors were examined in the same way and the results compared. The sequences were multivariately characterised using five zz-scales, and for handling sequences of varying length Auto Cross Covariances (ACC) were used. The methods used include Principal Components Analysis (PCA), partial least squares Projections to Latent Structures (PLS) and Soft Independent Modelling of Class Analogies (SIMCA).

Keywords

G protein-coupled receptor, PCA, PLS, SIMCA, ACC

Supervisors

Per Andersson
Melacure Therapeutics AB

Examiner

Torbjörn Lundstedt
Melacure Therapeutics AB

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **53**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Multivariate analysis of G protein-coupled receptors

Ingrid Gunnarsson

Sammanfattning

Receptorer är stora proteinmolekyler som sitter på ytan av de flesta celler, och hjälper dem att kommunicera med varandra och sin omgivning. Alla proteiner består av 20 st. olika byggstenar som kallas aminosyror. Aminosyrorna sitter ihop i en lång kedja och sekvensen, dvs. hur olika aminosyror följer på varandra i kedjan, är avgörande för proteinets egenskaper. G proteinkopplade receptorer är en familj av receptorer som är viktig för många av människokroppens centrala funktioner, och därför medicinskt intressanta.

I det här projektet har ett stort antal G-proteinkopplade receptorer analyserats med multivariat analys, ett samlingsnamn för olika metoder som används för att extrahera användbar information ur stora datatabeller. För att kunna göra en sådan analys måste varje receptor beskrivas numeriskt. Här har varje aminosyra beskrivits med deskriptorer som motsvarar aminosyrans fysiokemiska egenskaper, och en receptor beskrivs genom att byta ut namnen på varje aminosyra i sekvensen mot motsvarande deskriptorer.

Syftet med projektet var att undersöka om det finns kvantifierbara skillnader mellan olika receptortyper inom familjen, och resultaten tyder på att så är fallet.

Examensarbete 20 p i Molekylär bioteknikprogrammet
Uppsala universitet, januari 2002

Contents

1. Background

G protein-coupled receptors are a large and varied family of receptors in fungi, plants and animals, with the ability to bind many different types of ligands [1]. They are crucial for many of the central functions of our body, including sight, smell, and taste. All GPCR's share a common structure with 7 transmembrane regions (Fig 1), but other than that little is known about their 3D-structure [2]. As for all membrane proteins, determining the crystal structure of the receptors is very difficult and it has only been made for one member of the family, Bovine Rhodopsin [3]. This structure therefore serves as a model for the structure of all members of the family, an assumption that, given the diversity of function of the different receptors in the family, is not necessarily very accurate. In order to learn more about this important family of receptors, other methods must be tried, and one alternative is the approach used here, a multivariate analysis.

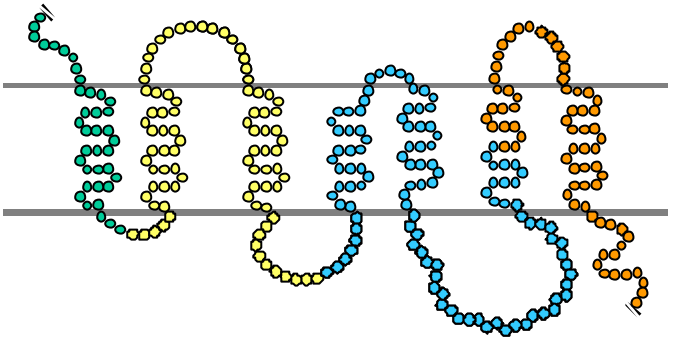

Fig 1. 7 TM structure of a GPCR.

2. Sequence data

The sequence data used initially is an in house collection of the transmembrane (TM) regions of 897 G-coupled receptors. Hence, initially, the loops were ignored and only the seven transmembrane regions of each receptor were investigated. The data set is divided by function into 12 classes, most of which are further divided into several sub classes. The twelve main classes are amine (am), peptide (pe), hormone protein (hp), rhodopsin (op), olfactory (ol), nucleotide like (nu), cannabis (cb), platelet activating factor (pa), gonadotropin releasing hormone (gr), thyrotropin releasing hormone (tr), melatonin (ml), and orphan (or).
In addition commercially available databases were used to retrieve the whole sequences of a smaller subset of G-protein coupled receptors.
With all sequence data downloaded from the Internet, it is important to bear in mind that the information might be of varying quality, and should not be regarded as 100% accurate.

3. Methods

The methods used include Principal Component Analysis (PCA), Projections to Latent Structures (PLS) and SIMCA modelling. The amino acid sequences have been quantitatively described using the five zz-scales described by Sandberg *et.al.* [4].

3.1 Models in general

A model is a description of important characteristics of a system, such as its components, interactions with the environment and sequences of events. A model is by definition incomplete, but should contain the essential structure of the system it describes. The aim is often to reveal systematic information such as structures and phenomena, and to present complex phenomena in a form that is easy to understand [5].

3.2 The zz-scales

The zz-scales describe each amino acid with numerical values, descriptors, which represent the physicochemical properties of the amino acid. In this project, the descriptors used are the five principal properties described by Sandberg *et al*. Three z-scales for the 20 coded amino acids were described by Hellberg *et al* and have subsequently been extended by Sandberg *et al* to include 87 non-coded amino acids and a total of five zz-scales. The zz-scales are derived from a multiproperty matrix, a matrix that consists of a number of physiochemical properties measured and calculated for each amino acid. A PC analysis of this matrix yields principal components or descriptors, referred to as zz-scales, which describe the intrinsic properties of the amino acids. The first zz-scale represents the hydrophilicity of the amino acid, the second represents the bulk of the side-chain, and the third represents the electronic properties. The fourth and fifth are more difficult to interpret [4].

The practical use of the zz-scales is very straightforward. The one-letter code used to describe each amino acid in a protein or peptide is simply replaced by the corresponding numerical descriptors (Fig 2). A sequence of length p will thus be represented by 5*p variables in a so-called multipositional description [6].

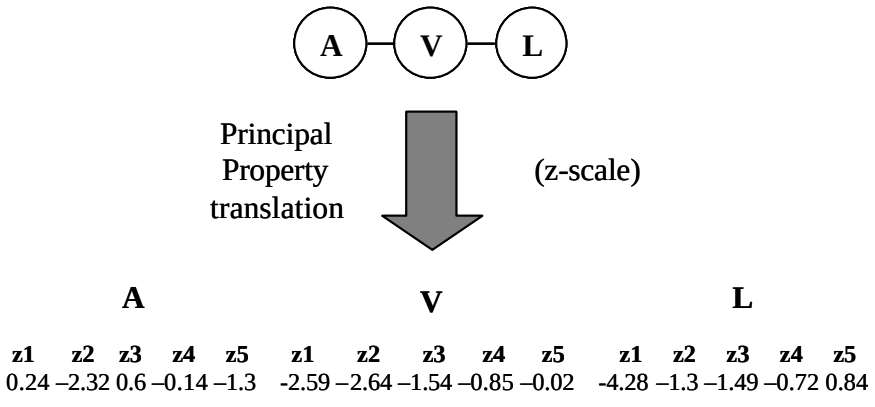

Fig 2. Translation of a tripeptide to five zz-scales.

3.3 Auto Crossed Covariances

When analysing sequences of different lengths, alignment independent methods such as Auto Crossed Covariances (ACC) are often used. The advantage of using an alignment independent method is that it can be used without pre-treatment of data such as identification of TM regions, gaps etc.

ACC calculates the average interaction between an amino acid and its neighbour some positions away in a sliding window. Two kinds of variables are calculated: Auto covariances

(Eq. 1), between the same principal property in each position, and crossed covariances (Eq. 2), between two different principal properties. The lag used can be varied, but the maximum lag is determined by the shortest sequence [7]. ACC's are calculated with lags 1 … L, and the resulting number of variables is d^2*L, where d is the number of descriptors and L the lag.

$$ACC_{j,lag} = \sum_{i}^{n-lag} \frac{z_{j,i} * z_{j,i+lag}}{n-lag} \quad \text{Eq. 1}$$

$$ACC_{j\neq k,lag} = \sum_{i}^{n-lag} \frac{z_{j,i} * z_{k,i+lag}}{n-lag} \quad \text{Eq. 2}$$

By calculating ACC the information in sequences of different length is summarized in vectors of equal length [5]. ACC takes neighbouring effects, i.e. lack of independence between subsequent positions, into account [8].

There is a variation of ACC that can be used to describe interactions in circular and branched amino acid sequences. The formulae are similar to those used in the linear case, only the denominator changes. In a circular sequence where every amino acid is joined to two others without branches, *n-lag* is replaced by *(n-lag)/2*. For more irregular protein structures, the interactions are instead divided by the number of interaction terms, M (Eq. 3-4) [9].

$$ACC_{j,lag} = \sum_{i}^{n} \frac{z_{j,i} * z_{j,i+lag}}{M} \quad \text{Eq. 3}$$

$$ACC_{j\neq k,lag} = \sum_{i}^{n} \frac{z_{j,i} * z_{k,i+lag}}{M} \quad \text{Eq. 4}$$

3.4 PCA

PCA is a projection method used to visualise data in high dimensions by reducing the dimension of the data. The starting point is a matrix of data, **X**, with N rows (observations) and K columns (variables). PCA finds the line/plane/hyper plane in the K-dimensional space that best approximates the data, by finding the directions of the largest variation in the data, referred to as principal components. The orientation of the model plane in the K-dimensional variable space is explained by the loadings, explaining how much each of the original variables contributes to the principal components. The principal components form the basis in a new coordinate system into which the data points are projected (Fig 3). The coordinates of the data points in this new coordinate system are called scores (Fig 4-5) [10]. The principal components are the eigenvectors of the covariance matrix of the data matrix **X**, and are thus orthogonal. The eigenvectors associated with the largest eigenvalues of the data correspond to the directions of the largest variation of the data [11].

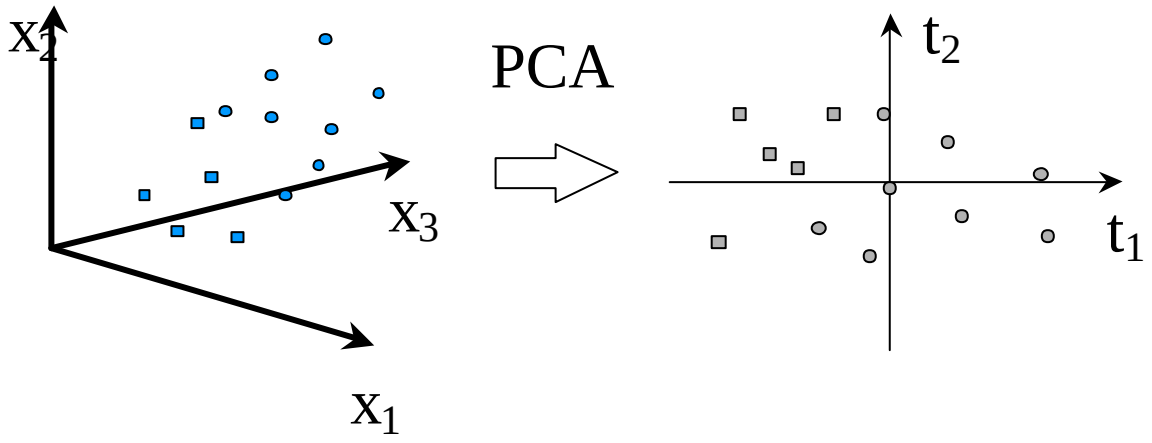

Fig 3. Illustration of PCA: A dataset in three dimensions is projected down to two.

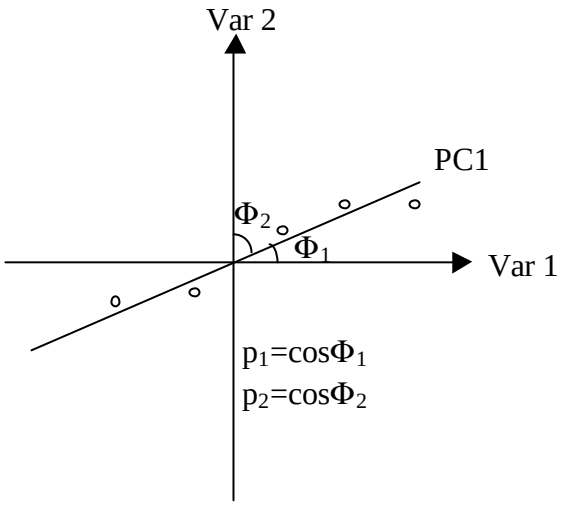

Fig 4. The first principal component is the line in K-dimensional space that best approximates the data. The components in the loading vector are the cosines of the angles F_1 and F_2.

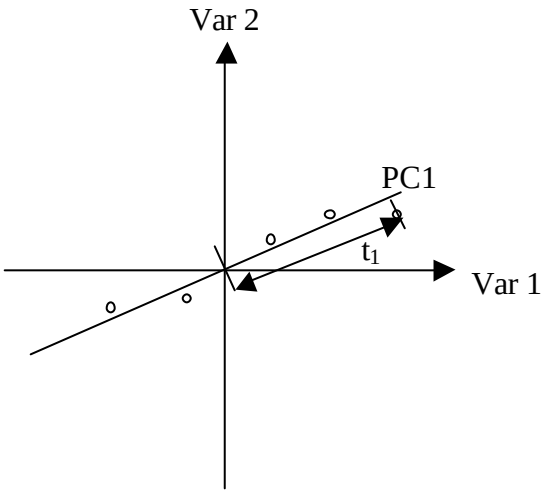

Fig 5. The scores are the projections of the data points on the principal component.

Before applying PCA, data is normally pre-treated. The most common treatments are mean-centering and scaling to unit variance. The variables of a dataset often have different numerical ranges and thus different variances. A variable with a wide range has a high variance whereas a short range will give a low variance. Unless data are normalised, variables with high variance will dominate over variables with low variance. Therefore, the standard deviation, σ_k, is calculated for each variable and each column is multiplied by $1/\sigma_k$ to give all variables unit variance. Mean centering improves the interpretability of the model. It is done by calculating the average value of each variable and subtracting it from the data [10].

Mathematically the model plane can be expressed as

$$\mathbf{X} = \mathbf{x}' + \mathbf{TP}' + \mathbf{E} \qquad \text{(Eq. 5)}$$

Here **x** is the mean of the variables, **T** the scores, **P** the loadings and **E** the residuals [10].

When interpreting a PCA model, plots of the scores and loadings are useful. A score plot shows the projection of the observations in a model plane, and are helpful in revealing any groupings of the data. A loading plot shows which original variables are important for the separation between groups. However, these plots can illustrate only three model dimensions at a time.

Observations that do not fit the PCA model are referred to as outliers. Strong outliers are identified from score plots using the Hotelling T^2 ellipse. The Hotelling T^2 ellipse drawn in score plots defines the area corresponding to (for instance) a 95% confidence interval. Observations that fall outside this ellipse are strong outliers. Moderate outliers do not show up in a score plot, but can be identified by the residuals of each observation, DModX. DModX is an acronym for Distance to the model in the X-block. It is based on the elements of the residual matrix E (Eq. 5) summarized row-by-row. DModX can be calculated for each observation in the data set, and plotted in a control chart where the tolerance limit of the class, Dcrit, is given. If the DModX of an observation is higher than Dcrit, the observation is a moderate outlier [10].

3.4.1 Cross validation and eigenvalues

To determine the appropriate number of components in the PCA model, an internal validation method called cross validation is used. In cross validation, the dataset is divided into a number of groups, and a reduced dataset is formed by excluding one of the groups. For a starting value of $S=S_0$, where S is the numbers of components, a model is estimated on the basis of the reduced dataset, predicted values are calculated for the excluded objects and the sum of squares of prediction errors is calculated from the predicted and observed values of the excluded objects. This is then repeated with another group excluded, until all groups have been excluded once and only once. Finally, a total sum of squares of prediction errors is calculated. S is then changed, and the process repeated, until a minimum total prediction error is found for $S=S_n$. S_n is then the optimum choice of components for the given data set [12].

Cross validation is often used in combination with looking at eigenvalues; for a component to be significant the corresponding eigenvalue should preferably be larger than two.

3.4.2 Hierarchical PCA

Hierarchical PCA modelling is a variant of PCA that is useful for data with many variables, where the results often are difficult to interpret. The variables are divided into conceptually meaningful blocks (in this project: TM or loop regions), and a PCA model is fitted to each block. The principal components from each of these models then become the new variables, and the PCA model fitted to this data is the hierarchical PCA model. The interpretation of a hierarchical model has to be done in two steps. First, the loading plots of the hierarchical model reveals which of the blocks that are most important for any groupings that can be seen in the hierarchical score plot. Second, the loading plots for the blocks of interest are studied to see which of the original variables this corresponds to [10, 13, 14].

3.4.3 SIMCA modelling

Soft Independent Modelling of Class Analogies (SIMCA) is a method where separate PCA models are made for each of the known classes. Tolerance intervals can be constructed around the PCA hyperplanes, such that new objects are assigned to a certain class if they fall inside

the tolerance interval. An object that falls outside the tolerance limits of all class models is called an outlier (Fig 6) [15].

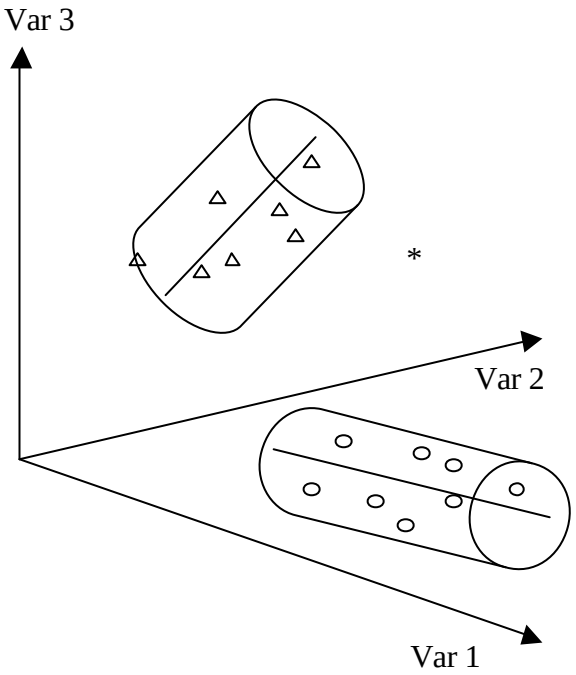

Fig 6. SIMCA modelling, two well separated classes and one outlier.

3.4.4 Cooman's plots

In a Cooman's plot, the DModX for two PC models are plotted against each other in a scatter plot. Giving Dcrit for both classes in the plot creates four areas of interest in the plot. Observations found in the lower left-hand area of the plot fit both models. Observations found in the upper left-hand or lower right-hand area fit the corresponding model, and observations found in the upper right-hand corner fit neither of the models (Fig 7) [10].

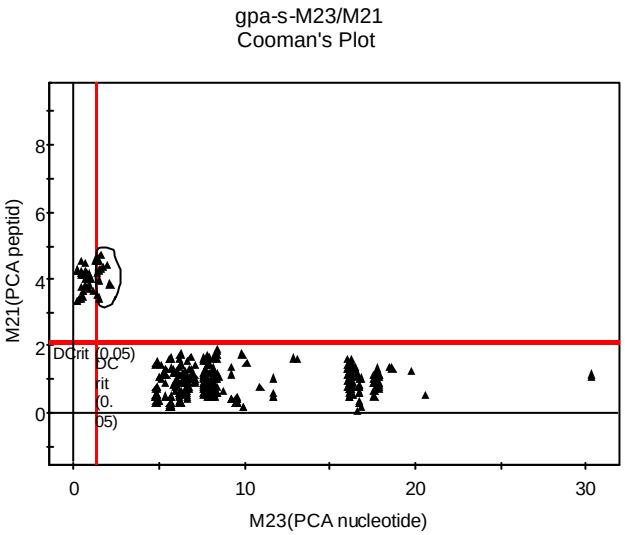

Fig 7. Example of a Cooman's plot. The peptide class fits its model well. The encircled observations, part of the nucleotide class, are moderate outliers.

3.5 Multivariate design

Multivariate design, MVD, is a method for selecting a set of representative observations among a large set of data. First, a multivariate characterisation must be made. Next, a PCA model is fitted to the data, to find the principal properties that best describe the data. Then, a representative choice of observations according to the principal properties can be made. The principal components in a PCA model are mathematically independent (orthogonal) and limited in number, properties that make them well suited for statistical experimental design schemes. There are a number of approaches for a multivariate design, all with the aim of maximizing the information content of the selected observations. If a dataset consists of several classes, it is important to make sure that all classes are represented by the design. In this case, it is therefore not enough to make one MVD, but rather local designs have to be made for each class. A weighting of the classes might be appropriate, giving large classes more representatives than small ones [10, 16, 17].

Using MVD, a smaller number of representative objects can be selected from a large dataset and used for model foundation. This is called a training set, and if the multivariate design has been successful the model based on the training set should be as good as one based on the whole dataset. A test set selected in the same way is used to validate the model, that is, to test if it is able to predict objects not included in the model building correctly. The aim is to be able to use the validated model for prediction, in this case classification, of new objects.

3.6 PLS

Partial Least Squares Projections to Latent Structures, PLS, is a method used to find relationships between two matrixes, **X** (variables) and **Y** (a response, e.g. biological activity). It is similar to PCA, in that it is also a projection method, but when calculating the principal properties of the **X** matrix, the correlation between the **X** and **Y** matrices is also taken into account. Thus, each principal component is in a direction that has both a large variance in **X** and is correlated to **Y**. This is achieved by introducing an inner relation, linking the two blocks by exchanging information on their respective scores.

The outer relations for the **X** and **Y** blocks are:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \qquad \text{Eq. 6}$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F} \qquad \text{Eq. 7}$$

Where **T** and **U** are the scores for **X** and **Y** respectively, **P** and **Q** are the loadings and **E** and **F** the residuals. To obtain orthogonal **t** values with the algorithm used, the loadings **p** are replaced by weights **w**.

The inner relation between **X** and **Y** is:

$$\mathbf{u}_h = b_h \mathbf{t}_h \qquad \text{Eq. 8}$$

Where b_h is a regression coefficient [18].

3.6.1 PLS-Discriminate Analysis

In PLS-Discriminate Analysis (PLS-DA) the **Y** matrix contains information about which class each observation belongs to. Using this method, the variables in **X** that are important for separating the classes can be identified (Fig 8) [10].

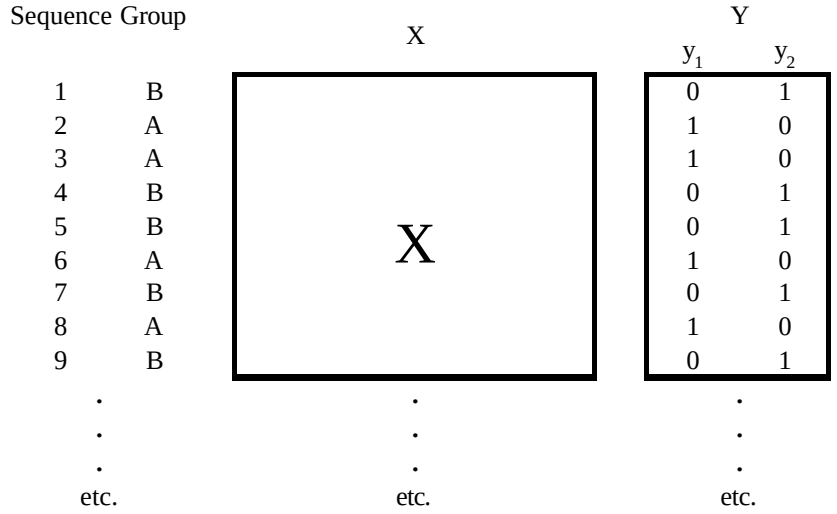

Fig 8. Illustration of PLS-DA for two classes.

3.6.2 Hierarchical PLS

Hierarchical PLS modelling is a method similar to hierarchical PCA. First, as in hierarchical PCA, individual PCA models are made for each transmembrane (or loop) region. Components from these models are used as variables, and a PLS model is fitted to the data.

3.7 Software

The software used in this project is: Simca-P 8.0 (Umetrics AB, Box 7960, SE-907 19, Umeå, Sweden, www.umetrics.com, [2000]), Seqan 1.1 (Infex, Rödhakevägen 52b, SE-906 51 Umeå, Sweden), SPOC-SEQ.EXE, and SPOC-CRO.EXE (Michael Sjöström, Research Group for Chemometrics, Umeå University, SE-901 87 Umeå, Sweden).

4. Analysis of transmembrane regions

4.1 Global model

First, a global PCA model was made using all the sequences in the data set. The data set used consists of 897 sequences, each described by 675 variables (135 amino acid positions, each represented by 5 zz-scales). The model obtained had 109 components according to cross validation [Simca 8.0], 99 of which were saved[1], explaining 85% of the variance. Explained variance and eigenvalue for each component have been plotted in graphs (Figs 9-10). Of the 99 components plotted, 77 have an eigenvalue larger than two. These components explain 79% of the variance. A closer look at the first 20 components, explaining 47% of the variance, reveals that the first 15 describe mainly one or two classes each, but after that the pattern

[1] It is not possible to save more than 99 components per model in Simca 8.0.

becomes blurred (Table 1).

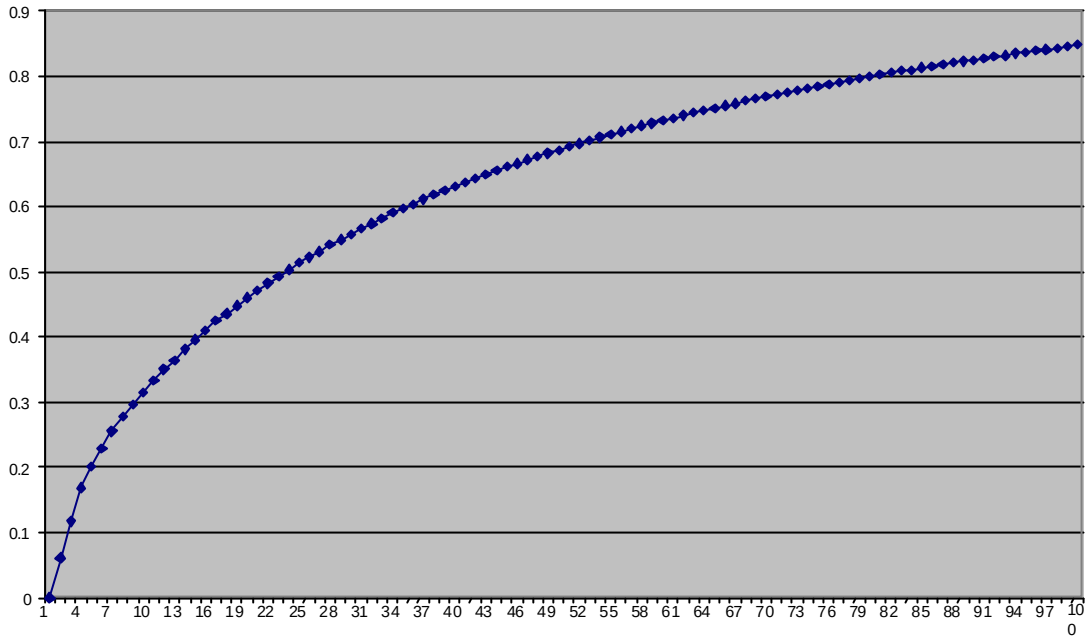

Fig 9. Explained variance for global PCA model with 99 components.

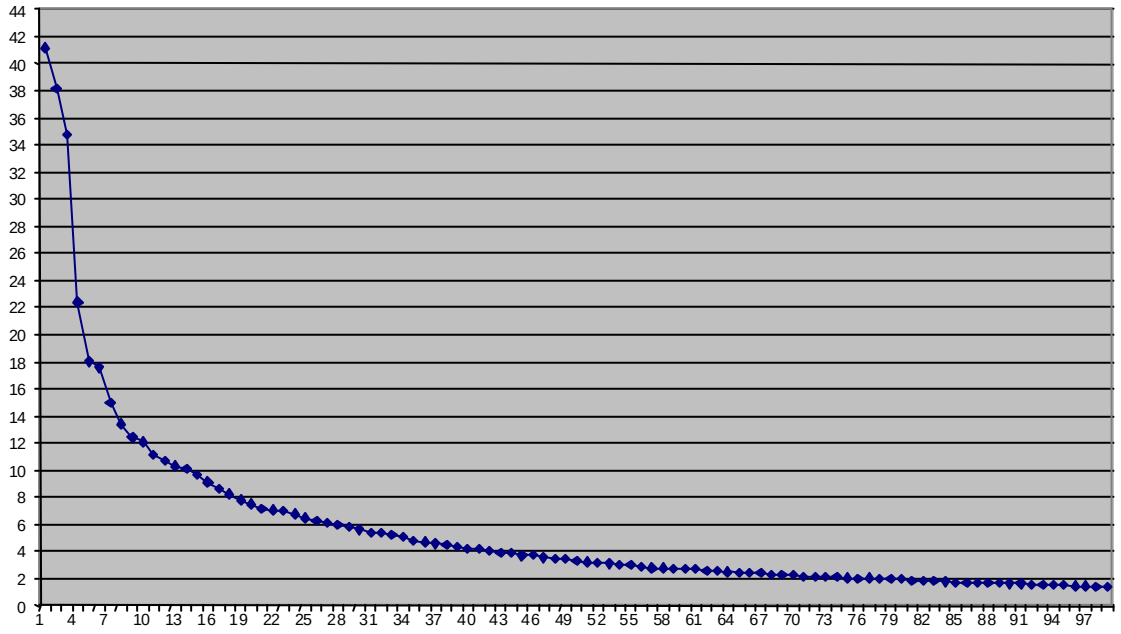

Fig 10. Eigenvalues for global PCA model with 99 components.

Component	Classes described +	-
t1	amine	olfactory
t2	olfactory	rhodopsin
t3	nucleotide/peptide (ck)	olfactory
t4	olfactory	hormone protein
t5	hormone protein	peptide (mc)
t6	peptide (mc)	peptide (et, bm)
t7	rhodopsin (opsa, opsm)	peptide (et)/rhodopsin (opsv)
t8	amine/peptide (ck)	rhodopsin (opsa)
t9	nucleotide/peptide (thr)	nucleotide
t10	peptide (vsl)/gonadotropin	peptide (et)
t11	peptide (op,ss)	melatonin/peptide (ck)
t12	rhodopsin	melatonin/peptide (tk)
t13	rhodopsin/peptide (ck)	melatonin/peptide (ag)
t14	cannabis/nucleotide/melatonin	peptide (mc, tk)
t15	peptide (op,br)	peptide (tk)/nucleotide
t16	rhodopsin/olfactory/nucleotide/peptide/thyrotropin	cannabis/peptide (ag)
t17	peptide/rhodopsin/nucleotide/gonadotropin	peptide (ny)/rhodopsin
t18	rhodopsin/cannabis/amine	rhodopsin/olfactory/orphan
t19	peptide/thyrotropin/orphan/cannabis	rhodopsin/peptide (ck)/amine
t20	rhodopsin/amine	thyrotropin/rhodopsin/peptide/ gonadotropin

Table 1. Classes described by the first 20 components of the global PCA model.

In the t1/t2 score plot for the global model, the rhodopsin, amine and olfactory classes form separate clusters (Fig 11), and in the t3/t4 score plot the olfactory and hormone protein classes (Fig 12). Remaining classes form a big cluster in the centre of the score plot. Looking at score plots t1/t3, t1/t4, t2/t3 and t2/t4 did not reveal any further groupings. It is interesting to note that the rhodopsin class is so well separated from the rest, bearing in mind that the 3D structure that all GPCR's are aligned towards belongs to this class. In an attempt to further separate the central cluster, a global PLS-Discriminate analysis was made, resulting in a model with 21 components. This led to a better separation of the clusters already seen in the PCA model, but gave no further separation of the remaining classes.

4.1.1 Reduced model

Next, all well separated clusters were removed from the work set, and a new model fitted to the remaining data, in the hope that this would help in separating remaining classes. This was repeated in several steps, and resulted in the separation of the melatonin class, as well as parts of other classes but did not, as hoped, give a good class separation for all classes (Figs 13-14). For example, the peptide class forms several clusters, each representing a subclass of the peptide group.

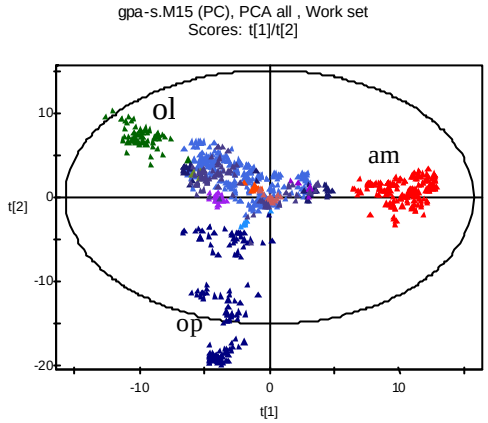

Fig 11. t1/t2 score plot for the global PCA model, showing the amine, rhodopsin and olfactory classes to be well separated. The red data points in the centre of the plot do not belong to the amine class.

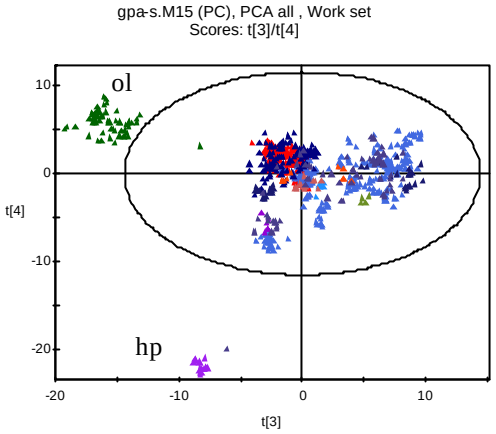

Fig 12. t3/t4 score plot for the global PCA model, showing the olfactory and hormone protein classes to be well separated. The blue data point near the hp cluster belongs to the orphan class.

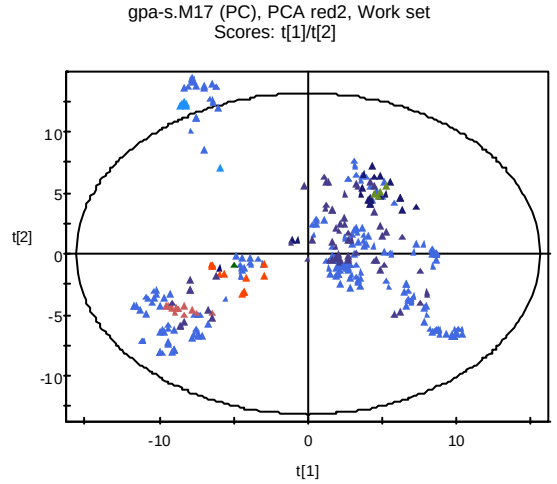

Fig 13. t1/t2 score plot for the reduced PCA model, showing an extensive overlap between the remaining classes.

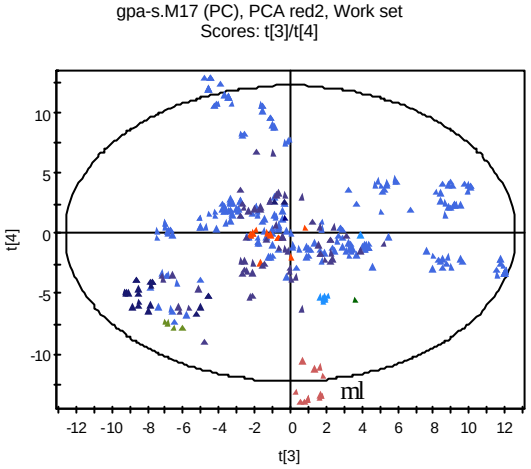

Fig 14. t3/t4 score plot for the reduced PCA model, showing the melatonin class to be clearly separated from the rest.

4.1.2 SIMCA modelling

In the next step, individual PCA models were made for those classes that the global PCA model could not separate. The peptide class consists of 302 sequences described by 675 variables, giving a model with 85 components explaining 96% of the variance. The nucleotide class, 45 sequences, gave a model with 16 components explaining 95% of the variance and the orphan class, 82 sequences, a model with 28 components explaining 85% of the variance. Score plots for the peptide and nucleotide classes show six and four clusters respectively, each representing one or more sub classes (Figs 15-16). The score plot for the orphan class, by contrast, shows a fairly even distribution of data points with few clear clusters (Fig 17). Since the peptide and nucleotide classes showed considerable overlap in the score plot for the global PCA model an attempt to class the peptide group into the nucleotide model was made. Judging by the score plot, the peptide class fits well into the nucleotide model (Fig 18), but a plot of DModX (16 components) reveals that this is not the case (Fig 19). A Cooman plot

confirms that the two classes are in fact well separated (Fig 20). Cooman plots for the peptide/orphan, peptide/thyrotropin, thyrotropin/gonadotropin, peptide/cannabis and orphan/ cannabis classes showed that these classes are also well separated, even though the global model cannot separate them.

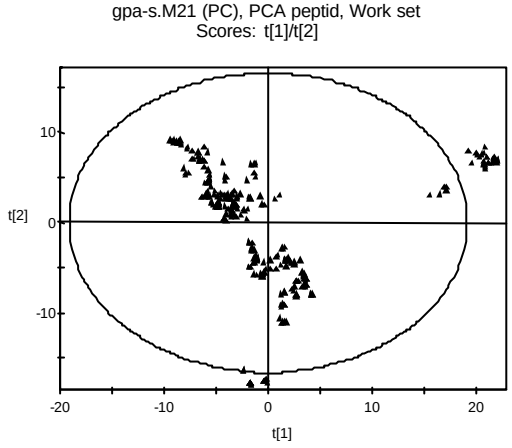

Fig 15. Score plot for PCA model of the peptide class, showing six clear clusters.

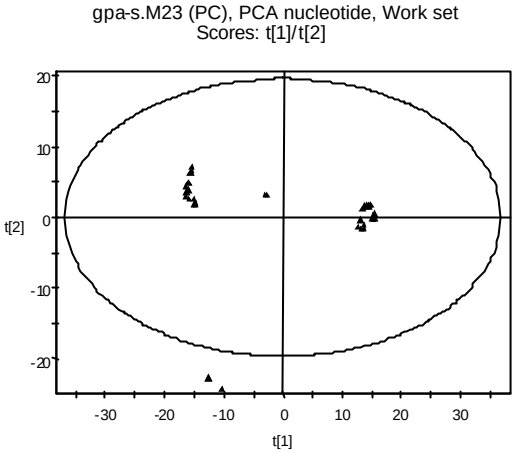

Fig 16. Score plot for PCA model of the nucleotide class, showing four clear clusters.

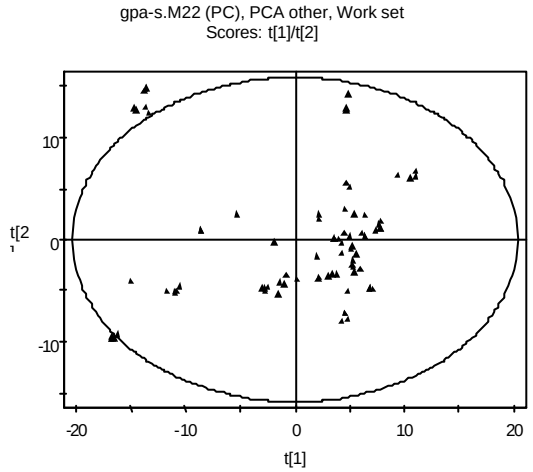

Fig 17. Score plot for PCA model of the orphan class. Compared to other classes, no strong groupings are observed.

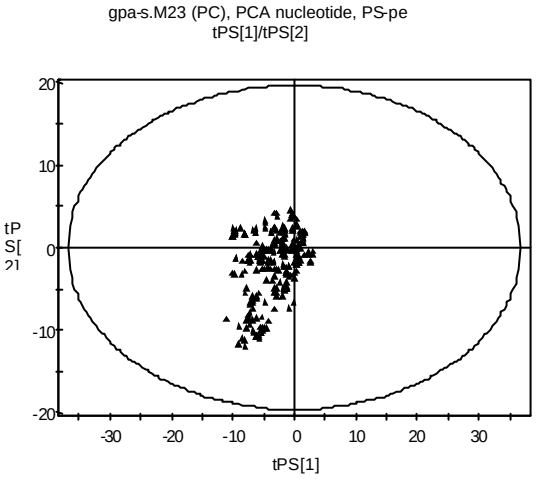

Fig 18. Predicted scores for the peptide class predicted in the nucleotide PCA model, showing a good fit in the score space.

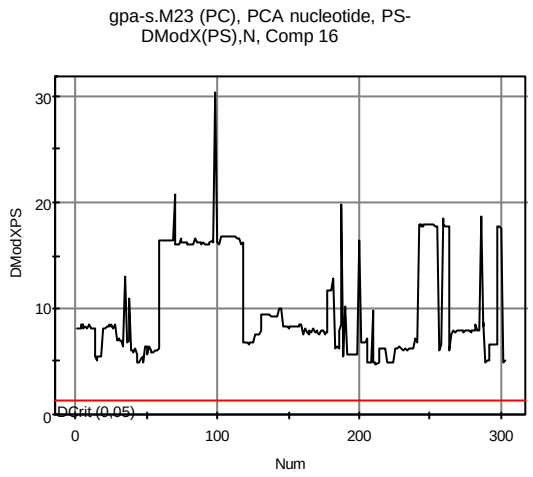

Fig 19. DModX plot for the peptide class predicted in the nucleotide model, showing a large distance for all peptide sequences to the nucleotide model.

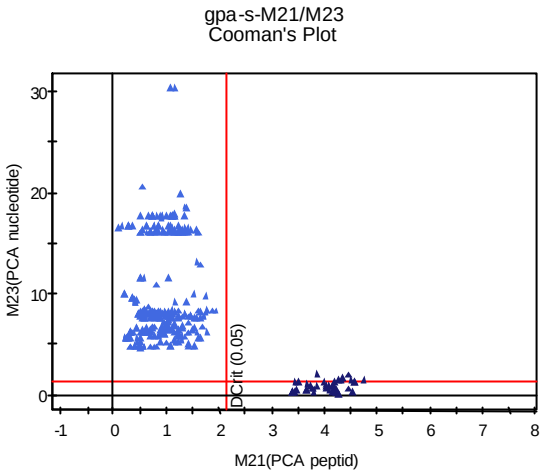

Fig 20. Cooman plot for the peptide and nucleotide classes, showing a good separation between the classes.

A separate PCA model was also made for the rhodopsin class (131 sequences), with 44 components explaining 92% of the variance. The t1/t2 score plot for the rhodopsin PCA model shows five well-separated clusters, and does not reveal any outliers (Fig 21). Each of the clusters represents one or more sub classes, and only one of the sub classes, a small sub class of only nine sequences, is split between two clusters. Higher component score plots, e.g. t3/t4, show the model to have several outliers (Fig 22), as does a plot of DModX (Fig 23).

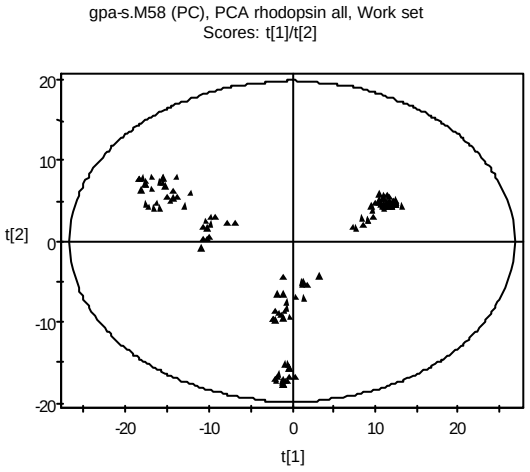

Fig 21. t1/t2 score plot for PCA model of the Rhodopsin class, showing five well-separated clusters.

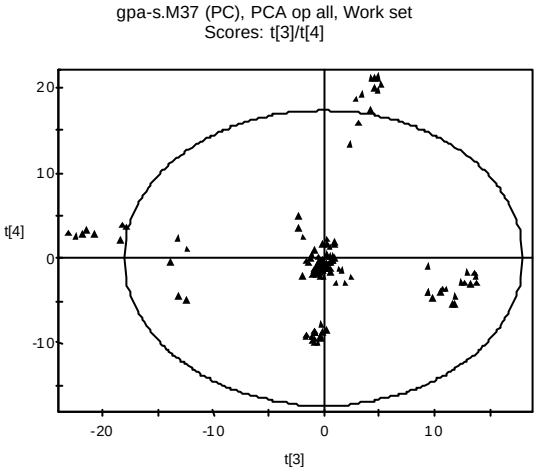

Fig 22. t3/t4 score plot for PCA model of the Rhodopsin class, showing a few outliers.

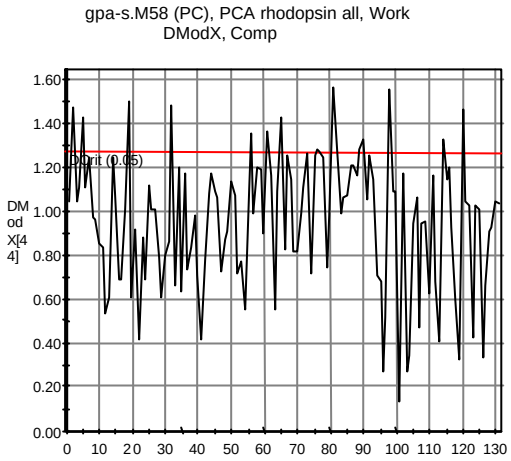

Fig 23. DModX plot for PCA model of the Rhodopsin class, showing only moderate outliers.

4.1.3 Local models

The t1/t2 score plots for the peptide, nucleotide and rhodopsin PCA models show well-separated clusters, each representing one or more sub classes. For a more detailed picture, a model can be fitted to the sequences of one of the clusters. This will give a separation between the different sub classes in the cluster. Similarly, a model can be fitted to one of the sub classes for a separation between different receptor types in the sub class, and, finally, a model based on sequences from one receptor type gives a separation between receptors of the same type but from different species. Thus, the more local a model is, the more detailed information it can give. Figs 24-27 illustrate this. Fig 24 is a t1/t2 score plot for a PCA model of the peptide class, and the encircled cluster contains sequences from the peptide sub classes bm, ny and tk. These three classes contain 48 sequences described by 590 variables[2], and a PCA model based on these data has 17 components according to cross-validation, explaining 96% of the variance. In a t1/t2 score plot for this model, the three sub classes are well separated from each other (Fig 25). The encircled sequences belong to the sub class tk, a sub class with 16 sequences described by 450 variables. A PCA model based on these data gives a model with 6 components, explaining 93% of the variance. A t1/t2 score plot for this model shows four distinct clusters, one for each receptor type (tk1, tk2, tk3 and tkl) in the sub class (Fig 26). The encircled sequences belong to the receptor type tk2, a receptor type with 7 sequences described by 125 variables. A PCA model based on these data gives a model with 3 components, explaining 69% of the variance. A t1/t2 score plot for this PCA model shows two clusters containing sequences from the species human, rabit and bovin, and rat, mesau and mouse, respectively (Fig 27), and one sequence separated from the others, cavpo. Contribution plots show that these sequences differ only in a handful of places. Within the two clusters, there are seven and nine positions respectively where the sequences differ, and between them there are 21 positions that differ.

[2] For models based on few sequences that are similar, a number of variables are usually excluded due to small or zero variance.

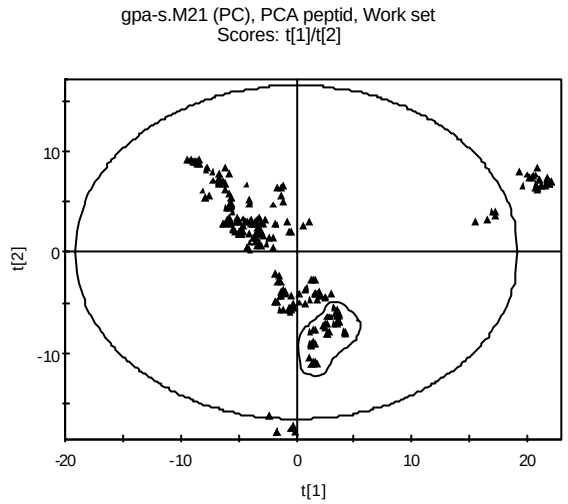

Fig 24. Score plot for PCA model of the peptide class. Encircled sequences (sub classes bm, ny and tk) are modelled separately.

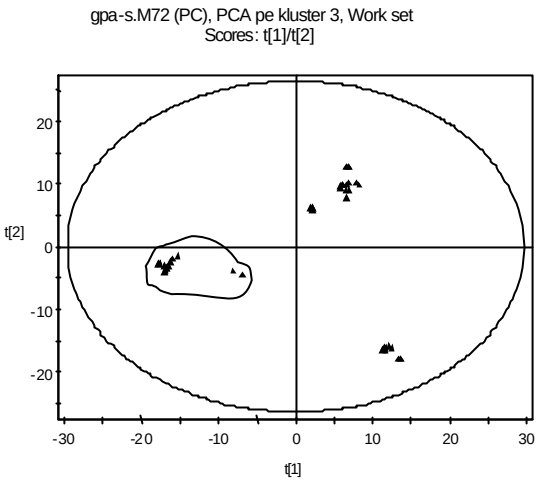

Fig 25. Score plot for PCA model of the sequences encircled in fig 24. The three sub classes form well-separated clusters. Encircled in this plot is sub class tk.

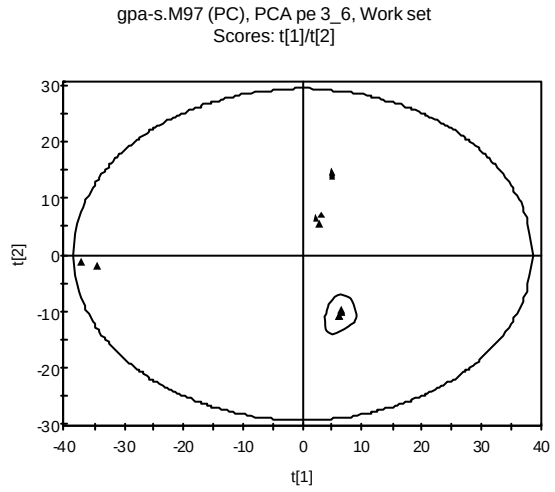

Fig 26. Score plot for PCA model of the peptide sub class tk. The four receptor types (tk1, tk2, tk3 and tkl) form well-separated clusters. Encircled is receptor type tk2.

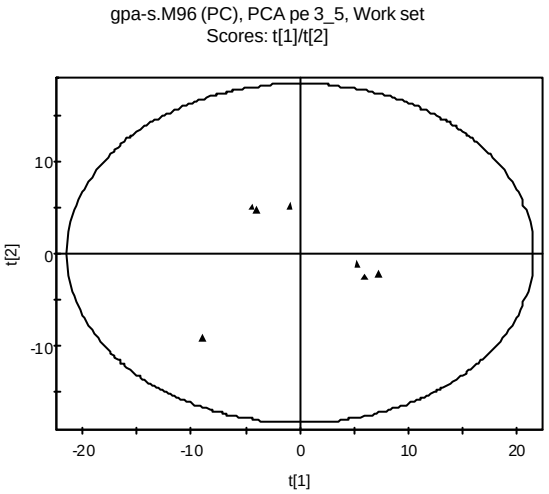

Fig 27. Score plot for PCA model of the receptor type tk2.

4.2 Hierarchical model

To investigate whether there is a particular TM region that is responsible for the separation between the classes, a hierarchical model was made. First, separate PCA models for the seven TM regions were made, which gave around 30 components for each model, with explained variances in the range of 79-89%. Explained variance and eigenvalue for each component in the models have been plotted in graphs (Figs 28-29). The components from the separate models were then joined into a new dataset for the hierarchical PCA model. The dataset consisted of 897 sequences, described by 230 variables, and the model had 77 components, explaining 83% of the variance. The score plot was similar to that for the global model (Fig 30), and the loading plot shows that the first four components in the separate PCA models, explaining 24-32% of the variance, are the most important for the separation (Fig 31). Hence,

a hierarchical model was made where only the first four components from each of the models for the seven TM regions were used. The dataset for this model consisted of 897 sequences described by 28 variables, and the model had 5 components explaining 66% of the variance. The two classes with the best separation are the amine and rhodopsin classes (Fig 32) and the following investigation will therefore be focused on them.

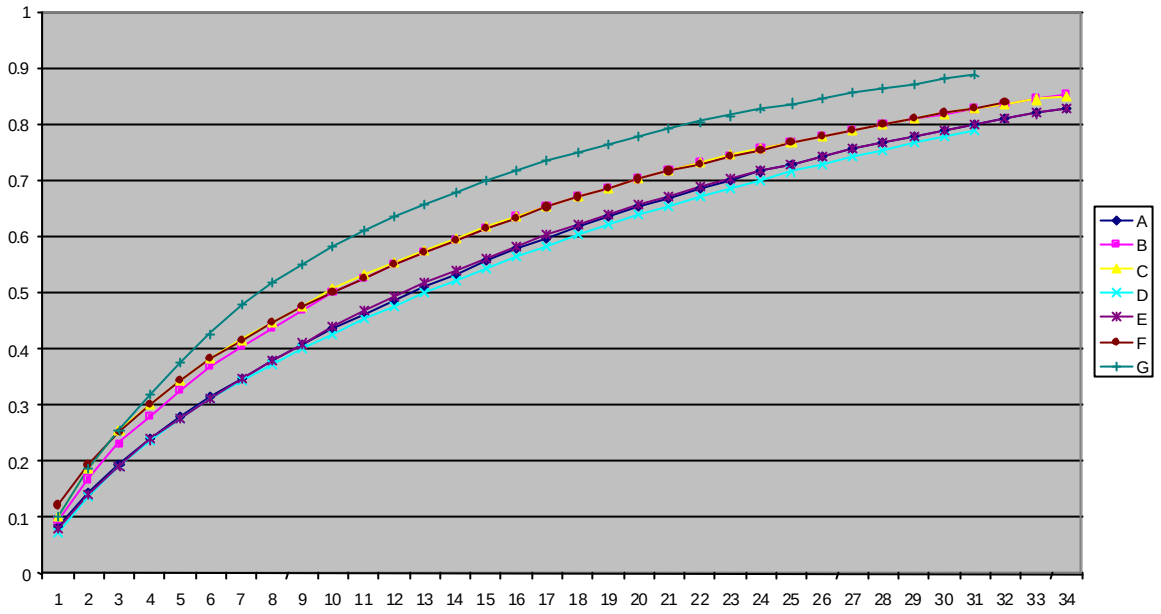

Fig 28. Explained variance for separate PCA models for transmembrane regions A-G.

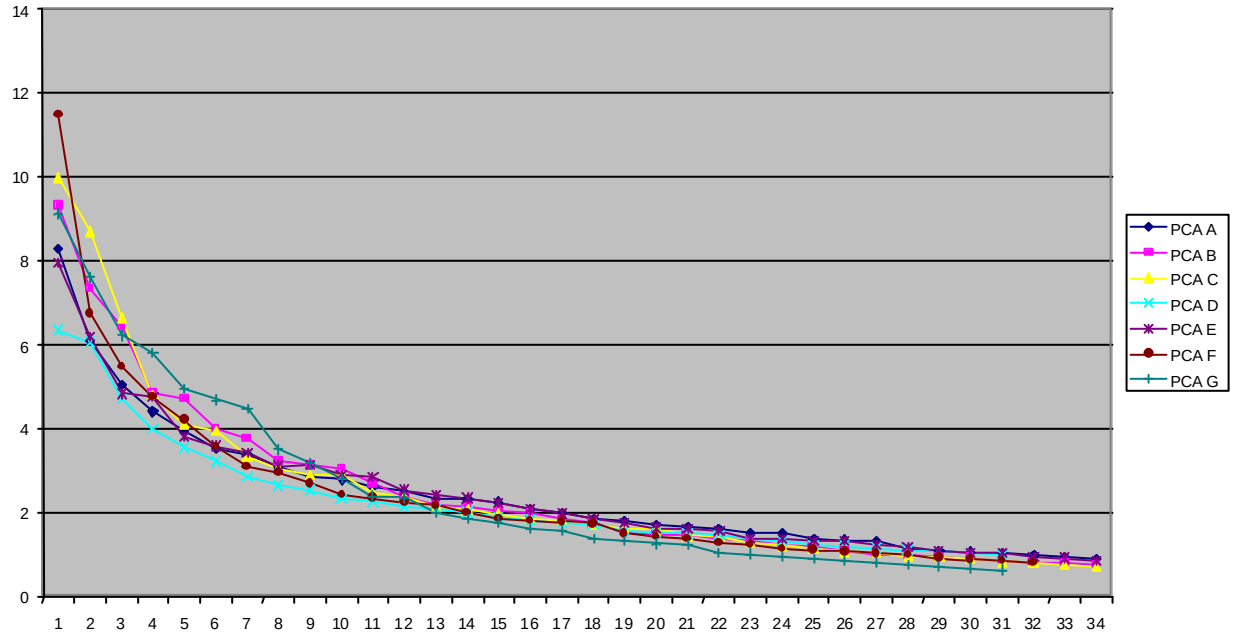

Fig 29. Eigenvalues for separate PCA models for transmembrane regions A-G.

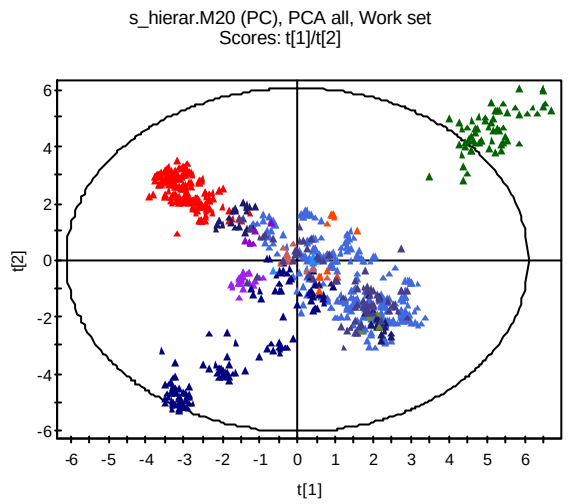

Fig 30. Score plot for the hierarchical PCA model based on all components significant by cross-validation. The score plot is similar tho that for the global model.

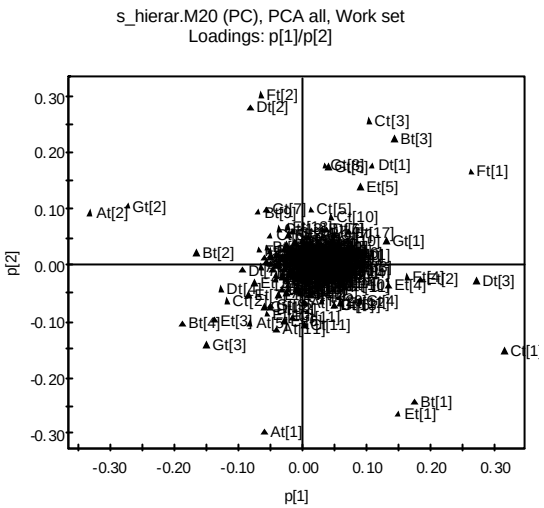

Fig 31. Loading plot for the hierarchical PCA model based on all components significant by cross-validation. The first four TM components are shown to be the most important.

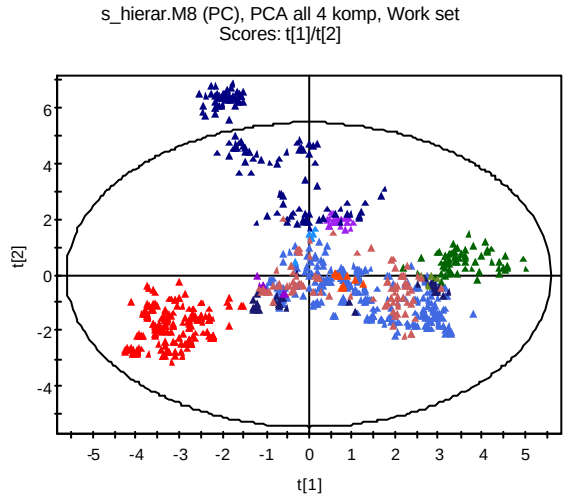

Fig 32. Score plot for the hierarchical PCA model based on four components from each TM region. The amine (red) and rhodopsin (blue) classes are well separated.

4.3 Hierarchical model for amine and rhodopsin

A new hierarchical PCA model was made for the amine and rhodopsin classes only. The dataset used consisted of 337 sequences, described by 230 variables, and this gave a model with 64 components describing 93% of the variance. The two classes are well separated (Fig 33), and the loading plots show that the first seven components in the separate PCA models are the most important for the separation (Fig 34). To make the interpretation of the loading plots easier a hierarchical model was made where only the first four components from each of the models for the seven TM regions were used, and this is the model that is used in the following analysis.

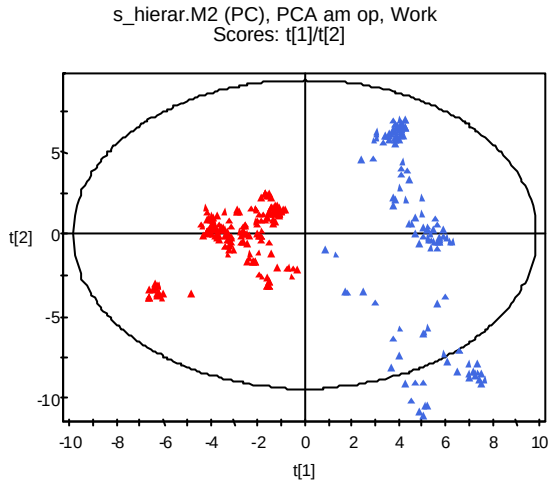

Fig 33. Score plot for the hierarchical rhodopsin (blue) + amine (red) PCA model, based on all components significant by cross validation.

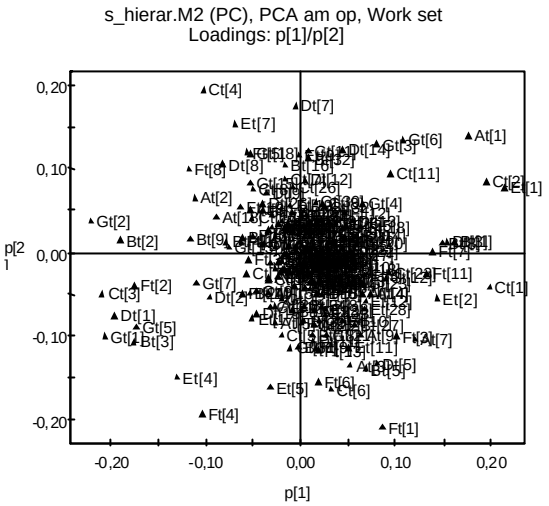

Fig 34. Loading plot for the hierarchical rhodopsin + amine PCA model, based on all components significant by cross validation.

A hierarchical PLS-Discriminate analysis was also made for the amine and rhodopsin classes only, resulting in a model with 2 components, explaining 51% of the variance in X and 97% of the variance in Y. The t1/t2 score plot shows that the two classes are well separated by this model (Fig 35). A coefficient plot for the PLS-DA model shows which of the variables in the hierarchical model that are most important for the separation between the two classes (Table 2). This confirms the information given in the loading plots (Fig 34), that the first seven components in the separate PCA models for the TM regions are the most significant.

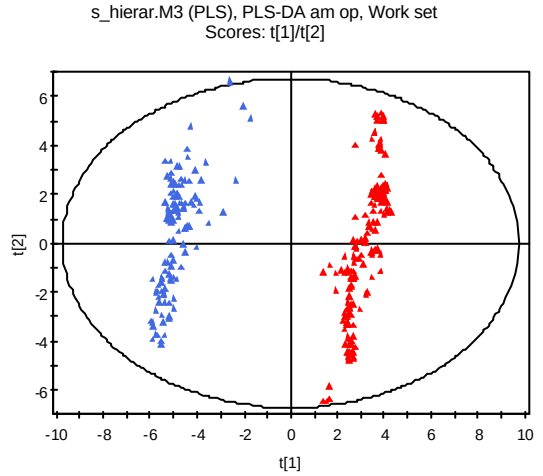

Fig 35. Score plot for hierarchical PLS-DA model for the amine (red) and rhodopsin (blue) classes, based on all components significant by cross validation.

TM region	Significant components
A	1,2,7
B	1,2,3,8,9
C	1,2,3
D	1,2,3
E	1,2,4
F	1,2,3,4,7,8
G	1,2,5,7,8

Table 2. Hierarchical variables important for the separation between amine and rhodopsin, as determined by a coefficient plot.

4.4 Specific amino acids of interest

To find out which amino acids that are conserved within the classes, separate PCA models were made for the amine and rhodopsin classes, for the three rhodopsin sub clusters as seen in a score plot for the global PCA model (Fig 11) and for the amine sub class acm. A few

observations were excluded from the amine and rhodopsin classes; those that form separate clusters away from the main cluster in the score plots for the hierarchical model (Fig 36). For the amine class, this is the acm and hh1 sub classes. In each of these models a number of variables were excluded because of small or zero variance, these correspond to amino acids conserved within the group or cluster in these positions. The excluded variables are listed in Table 3.

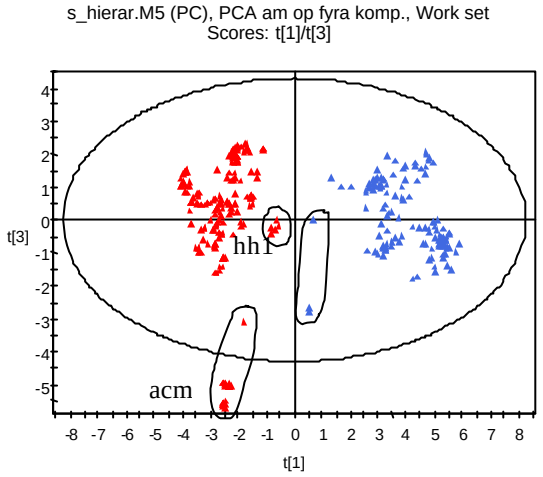

Fig 36. Score plot for the hierarchical rhodopsin (blue) + amine (red) PCA model, based on four components from each TM region. Encircled sequences were excluded in the separate PCA models.

The amino acids A15, E10, F14, F15, G6 and G16 are conserved in all three rhodopsin sub clusters (Table 3). A15, F14 and G16 are conserved also in the amine class, and looking at the sequence data reveals that both classes have the same conserved amino acid in each of these positions. Indeed, they appear to be well conserved throughout all GPCR's. Position A15 is conserved in all 897 receptor sequences in this data set, F14 is conserved in all classes except for the olfactory and G16 is conserved in all sequences except for two peptide sub classes and a handful of other sequences. Thus, it seems likely that amino acid positions A15, F14 and G16 are crucial for the function of all GPCR's, and that E10, F15 and G6 are important for the function of GPCR's belonging to the rhodopsin class. When comparing the list of excluded variables in the amine model to that of the amine/acm model, it is interesting to note that the variable G3 is excluded because of close to zero variance in the amine class as a whole, but not in the amine sub group acm. This might seem strange, but could be explained by the small size of the acm sub group. The acm sub group makes up approximately 10% of the total amine class, and a variation in the acm sub group might therefore be too small to be noticeable in the amine class as a whole.

Model	Excluded variables / conserved positions
Amine	A15 B11 C4, C11, C12 D7 E7 F8, F12, F14 G3, G5, G9, G12, G13, G16
Amine/acm	A11, A14, A15, A18 B1, B2, B4, B6, B7, B10, B11, B14, B15, B18, B19 C1, C4, C5, C8, C9, C10, C11, C12, C14, C15, C16, C18 D2, D3, D6, D7, D10, D13, D14, D16, D18 E1, E2, E3, E5, E6, E7, E8, E10, E11, E14, E17, E18 F3, F4, F5, F8, F9, F11, F12, F14, F15, F16, F19 G2, G4, G5, G6, G8, G9, G10, G12, G13, G16, G17
Rhodopsin	A15 B1 C8 E10 F14, F15 G6, G13, G16
Rhodopsin cluster 1 (furthest from centre of plot)	A3, A5, A7, A11, A13, A15, A18, A19, B1, B3, B6, B7, B18, B19, C1, C4, C5, C7, C8, C12, C15, C18, D2, D7, D9, D13, D17, D18, E1, E2, E3, E6, E7, E10, E16, E18, E19, F12, F14, F15, F16, G1, G3, G4, G6, G12, G13, G15, G16
Rhodopsin cluster 2	A4, A15, B1, B6, C8, C14, D7, D16, E6, E10, E14, E18, F4, F14, F15, G1, G6, G12, G13, G15, G16
Rhodopsin cluster 3 (nearest centre of plot)	A15, B7, B8, E10, F12, F14, F15, G6, G16

Table 3. Conserved amino acid positions identified in the separate class PCA models.

Contribution plots were made for rhodopsin in the global model. One data point in each of the three rhodopsin clusters was compared to the data point nearest the centre of the score plot. The variables with the highest contribution scores are those that contribute most to the separation of the rhodopsin class from the centre of the plot. The variables with contribution scores >0,20 and <-0,20 respectively can be seen in Table 4. In general, the variables with high contribution scores do not correspond very well with the conserved amino acid positions of rhodopsin, as might have been expected.

Data point	Variables with high contribution scores in the global model
548 O93441	+ A3t2, A18t1, B11t4, C13t4, E15t2, F4t3, F4t4, G11t2 - A3t5, A14t1, A14t3, B3t2, B3t3, D2t2, D13t5, F4t5, F16t2, G6t3, G11t5, G17t1
625 OPSB_CHICK	+A3t2, A18t1, B11t4, C13t4, F4t3, F4t4, G11t2 -A3t5, A14t1, A14t3, B3t2, B3t3, D2t2, D9t4, E6t4, F4t5, F16t2, G6t3, G11t5
645 OPSD_APIME	+A6t4, A13t4, B11t4, G11t2 -A1t2, D13t5, E6t4, G6t3, G11t5

Table 4. Variables with high scores in contribution plots for rhodopsin in global PCA model. Interpretation of variables: A3t2, for example, refers to the second principal property (zz-scale) of the third amino acid in transmembrane region A.

Contribution plots for rhodopsin were also made in the hierarchical model (Fig 37). The same three rhodopsin data points were compared to the same point near the centre as in the global model. In the hierarchical model this was not the closest to the centre, but still reasonably close. The loading plots for the hierarchical variables that had high contribution scores (>0,50 or <-0,50) were examined to find what original variables they corresponded to (Fig 38 and Table 5). This list corresponds reasonably well with both excluded variables in the local models (Table 3) and variables with high contribution scores in the global model (Table 4).

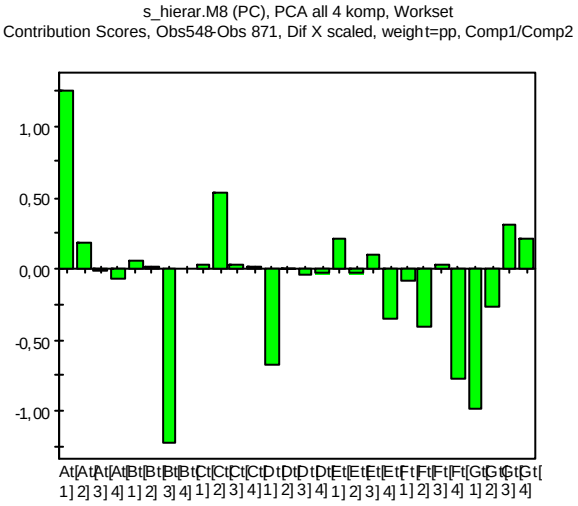

Fig 37. Contribution plot for rhodopsin in hierarchical PCA model.

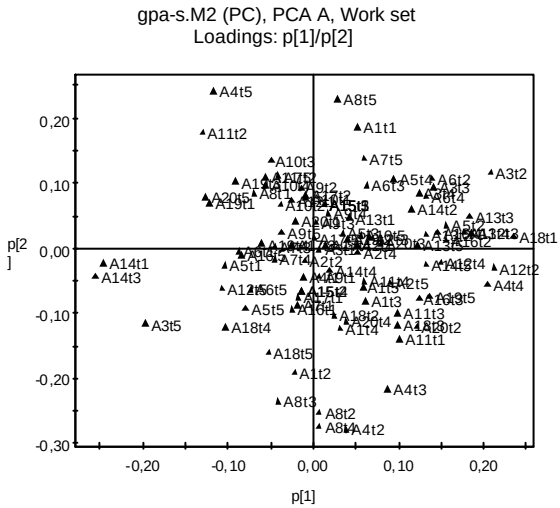

Fig 38. Loading plot for the PCA model for transmembrane region A.

Data point	Variables with high contribution scores in the hierarchical model
548 O93441	+ A3t2, A4t4, A12t2, A18t1, C13t4, B3t4, B3t2, B3t3, D9t4, D13t3, D5t2, D9t2, D13t4, D9t3, F4t4, F4t3, F19t3, G6t5 - A3t5, A14t1, A14t3, C8t2, B11t4, D9t5, D13t5, F2t2, F2t4, F16t1, F11t2, G6t2, G3t4, G3t2
625 OPSB_CHICK	+ A3t2, A4t4, A12t2, A18t1, B3t4, B3t2, B3t3, F4t4, F4t3, F19t3, G6t5 - A3t5, A14t1, A14t3, B11t4, F2t2, F2t4, F16t1, F11t2, G6t2, G3t4, G3t2
645 OPSD_APIME	+ B3t4, B3t2, B3t3, G6t5 - B11t4, G6t2, G3t4, G3t2

Table 5. Variables with high scores in contribution plots for rhodopsin in hierarchical PCA model.

For the amine/acm sub class, contribution plots were made within the group. These corresponded well, as expected, with the list of conserved amino acids.

A closer study of the loading plots for the hierarchical model for amine and rhodopsin was made, to see which components are significant for the separation between the two classes. The separation between the classes is mainly given by the first component in the hierarchical PCA model (Fig 39), and according to the loading plots is mainly due to the hierarchical variables At1, Bt2, Bt3, Ct1, Ct2, Ct3, Dt1, Et1, Et4, Ft2, Gt1 and Gt2 (Fig 40). This corresponds well with the rhodopsin contribution plots (Fig 37), though the contribution plots show fewer variables to be important than the loading plots do. This might be because the loading plots represent all sequences whereas the contribution plots look at two sequences only at a time. The sub class acm is separated from the rest of the amine class, a separation mainly given by component 3 (Fig 41). According to the loading plot the separation is mainly due to the hierarchical variables Bt4, Dt3, Dt4 and Gt4 (Fig 42). To see which of the original variables, and hence amino acid positions, the hierarchical variables correspond to, the loading plots for the separate PCA models for the seven TM regions were studied (Table 6 and Fig 38). Usually there are only a few variables, and hence amino acid positions, that have a large influence on the separation.

The amino acid positions that are represented in Table 6 by two or more principal properties (for example A3, represented by t2 and t5) can be assumed to be particularly influential and were further investigated. Looking at the amino acid score plots for the principal properties of interest reveals what kind of amino acid and thus what properties that are important for the separation between the amine and rhodopsin classes. For position A3, for example, the important properties are given by amino acids with positive values of principal property 2 and negative values of principal property 5, and the score plot (Fig 43) shows that Tryptophan (Trp) and Tyrosine (Tyr), both aromatic amino acids, fit this description. Looking at the sequence data, the amine class mainly has amino acids T, V, L and I (neutral, hydrophobic) in position A3, whereas the rhodopsin class mainly has W, Y and F (aromatic). This is consistent with the finding that aromaticity is important for the separation between the classes. Not all positions investigated gave as clear results however. A few examples are given in Table 7.

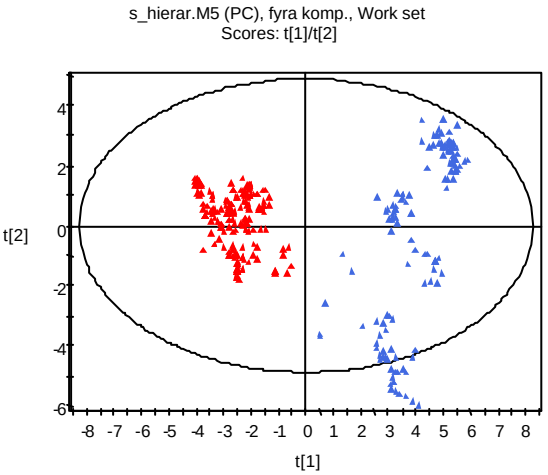

Fig 39. t1/t2 score plot for hierarchical PCA model for the rhodopsin and amine classes, based on four components from each TM region.

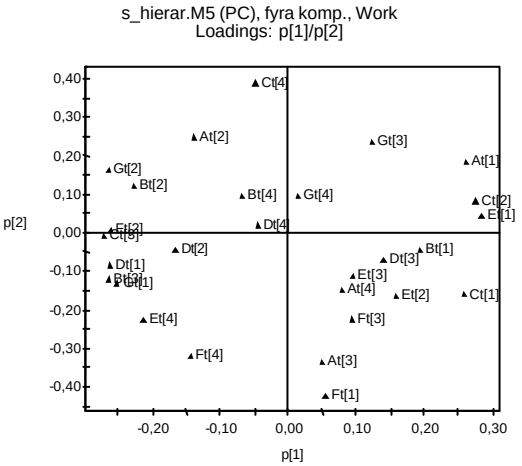

Fig 40. t1/t2 loading plot for hierarchical PCA model for the rhodopsin and amine classes, based on four components from each TM region.

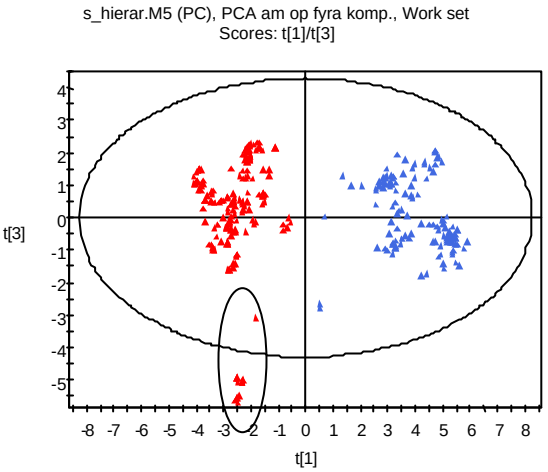

Fig 41. t1/t3 score plot for hierarchical PCA model for the rhodopsin and amine classes. The acm sub class, encircled, separates from the rest of the amine class.

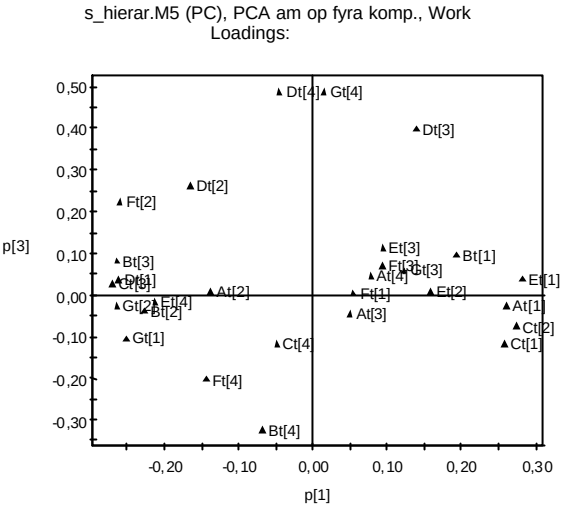

Fig 42. t1/t3 loading plot for hierarchical PCA model for the rhodopsin and amine classes.

Variable in hierarchical model	Original variables
At1	+A3t2, A4t4, A12t2, A18t1 -A3t5, A14t1, A14t3
Bt2	+B1t3, B1t5 -B13t3, B9t2
Bt3	+B3t2, B3t3, B3t4 -B11t4
Ct1	+C4t4, C7t5, C16t5 -C4t1, C8t3, C14t1, C16t3
Ct2	+C13t4 -C8t2
Ct3	+C12t1, C12t3, C18t1, C18t3, C18t4 -C1t1, C7t1, C18t5
Dt1	+D9t5, D13t5 -D5t2, D9t2, D9t3, D9t4, D13t3, D13t4
Et1	+E14t5, E17t3 -E8t2, E14t2, E14t3, E14t4, E17t5
Et4	-E2t3, E6t2, E6t4, E20t1, E20t2
Ft2	+F2t1, F6t1, F16t2, F19t2 -F4t3, F4t4, F16t1, F16t5
Gt1	+G6t3, G8t2, G10t1, G10t3 -G11t2, G4t2
Gt2	+G3t2, G3t4, G6t2 -G6t5
Bt4	+B4t3, B4t4 -B4t5
Dt3	+D2t5, D13t1 -D2t3, D2t4, D10t2, D14t2, D15t3
Dt4	+D7t2, D7t4, D14t1 -D7t5, D11t2, D11t4, D14t4
Gt4	+G13t5, G18t2 -G15t1, G15t3

Table 6. List of which original variables some of the hierarchical variables correspond to. Interpretation of variables: At1, for example, refers to the first principal component of transmembrane region A, and A3t2, refers to the second principal property (zz-scale) of the third amino acid in transmembrane region A.

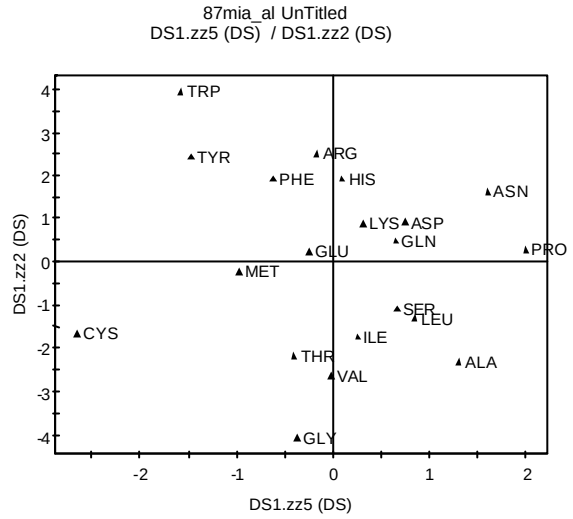

Fig 43. zz-scales 2 and 5 plotted against each other.

Aa position, pp's	**Amino acids with appropriate properties**	**Amino acids in amine class**	**Amino acids in rhodopsin class**
A3 + t2 – t5	W, Y	T, V, L, I	Y, W, F
A14 – t1, t3	V, I, L	G	V, I, T, G
B1 + t3, t5	P, N	N	N
B3 + t2, t3, t4	W, H	F, Y, L	I
D13 + t5 – t2, t3, t4	I, L	L, V, I	C, W
E6 – t2, t4	T, V, G	A, S	H, C, V, S
E20 – t1, t2	V, I, L	R, H, K, L	R, Q, F
F16 + t2 – t1, t5	W, Y	N, F	A, T, L
F4 – t3, t4	T, Q	I, V	M, T, I
G3 + t2, t4	R, W, H	W	F, Y, C
G6 + t2 – t5	W, Y	Y, W	K

Table 7. Amino acids with properties that are important for the separation of the amine and rhodopsin classes.

5. Analysis of whole sequences

The analysis of the transmembrane sequences is alignment dependent, and also depends on the division between TM-regions and loops being correct. To make the analysis alignment-independent, the complete amino acid sequences of the receptors were investigated. Three of the classes were selected for this, the amine, rhodopsin and peptide classes.

5.1 Selection of training and test data

Two sets of sequences were chosen, one training and one test set, each with approximately 16-20 observations from each of the three classes. In order to have a systematic strategy for selecting sequences, a multivariate design was made in the 7 TM models for each class (Figs 44-45). This approach assumes that a set of sequences selected by MVD in a 7 TM model will be representative also in a model where the complete sequences are considered, an

assumption that proved to be incorrect. A PCA model made on the whole sequences for the training and test sets together showed that the peptide training set has properties quite different from those of the peptide test set. In the t1/t2 score plot for this model, the training and test sets have a similar distribution, and completely overlap, as would be expected (Fig 46). In the t2/t3 score plot however, the peptide training set forms a well separated cluster (Fig 47).

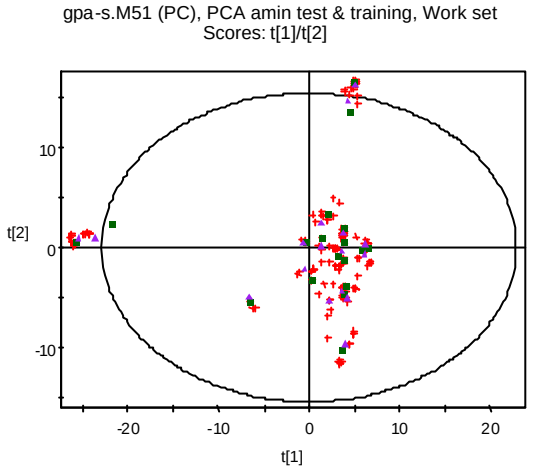

Fig 44. t1/t2 score plot for amine TM PCA model, used for multivariate design. The green squares and purple triangles represent the training and test sets, respectively.

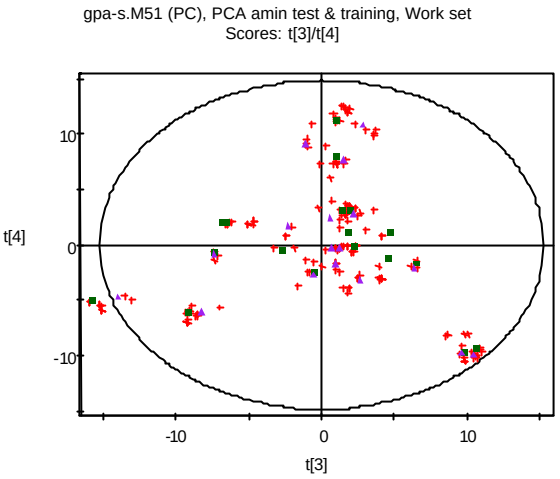

Fig 45. t3/t4 score plot for amine TM PCA model, used for multivariate design. The green squares and purple triangles represent the training and test sets, respectively.

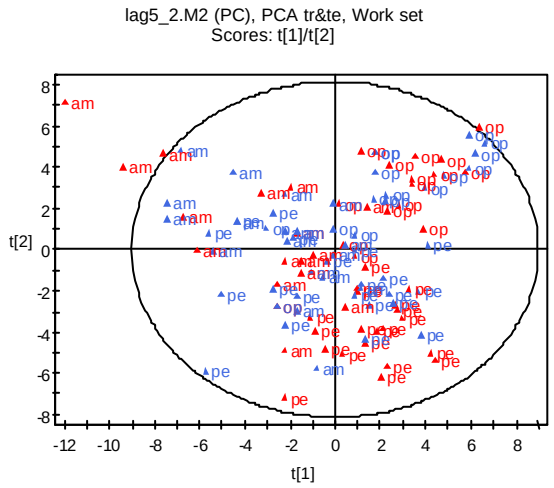

Fig 46. t1/t2 score plot for PCA model of whole sequence training (red) and test (blue) data from the amine, rhodopsin and peptide classes.

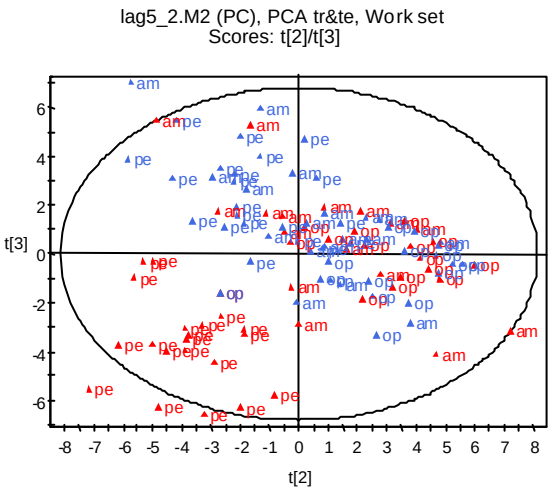

Fig 47. t3/t4 score plot for PCA model of whole sequence training (red) and test (blue) data from the amine, rhodopsin and peptide classes. The peptide training set forms a separate cluster.

5.2 Modelling

Complete sequences for the selected receptors were downloaded from the Internet [19], [20]. The amino acid sequences were then translated to the 5 zz-scales using the program SPOC-

SEQ.EXE, and Auto Cross Covariances were calculated using the SPOC-CRO.EXE program. Different lags were tried, to see how this would affect the result.

Initially lags 3, 5 and 7 were tried. For each of these, PCA and PLS-DA models were made for all training sequences, followed by SIMCA modelling. Lag 5 gave the best separation between the classes in the PCA and PLS-DA models, as determined by visual inspection of score plots. Lag 3 and lag 7 resulted in a bigger overlap between the classes. For comparison, lag 4 and lag 6 were also tried, and lag 20 was tried to see if a significantly larger lag would give a much better separation. Neither lag 4 nor lag 6 gave as good a separation as lag 5. Lag 20 gave a separation comparable to that of lag 5, but not quite as good (Figs 48-49). For all lags, Cooman plots give a clear separation between the amine and rhodopsin classes, and between rhodopsin and peptide, though one or other of the two classes often has observations in the lower left corner of the plot, indicating that they would fit either model. For the amine and peptide classes, however, the model with lag 5 is the only one that gives a good separation (Figs 50-53).

The reason that a lag of five seems to work so well might be connected to the fact that five amino acids corresponds to about one and a half turns in a protein alpha-helix, which would be a suitable distance for interactions between two amino acids. Therefore, lag 10 was also tried, to see if all multiples of five work well. This gives a result similar to lag 20, a reasonably good class separation, but not quite as good as lag 5.

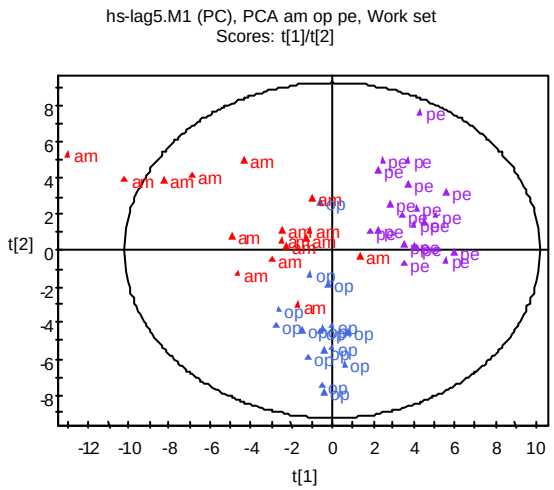

Fig 48. Score plot for PCA model of whole sequence training data, lag 5. There is a small overlap between the classes.

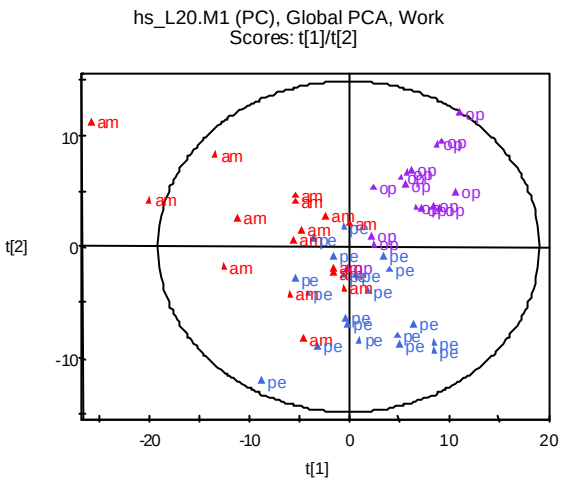

Fig 49. Score plot for PCA model of whole sequence training data, lag 20. There is a bigger overlap between the classes compared to the model with lag 5.

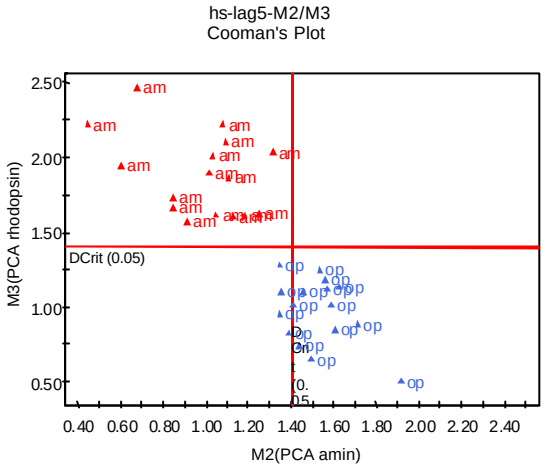

Fig 50. Cooman plot for the whole sequence amine and rhodopsin PCA models, lag 5. Each class fits its model well.

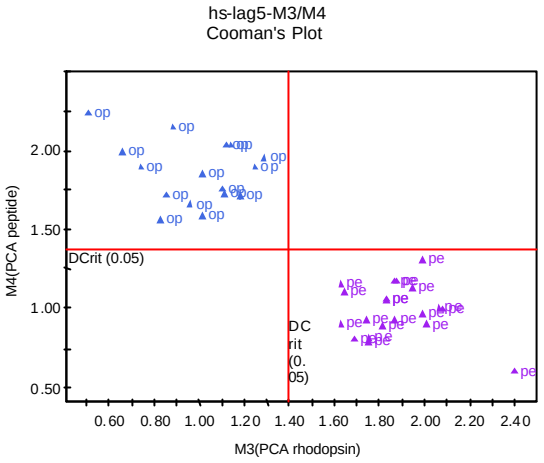

Fig 51. Cooman plot for the whole sequence peptide and rhodopsin PCA models, lag 5. Each class fits its model well.

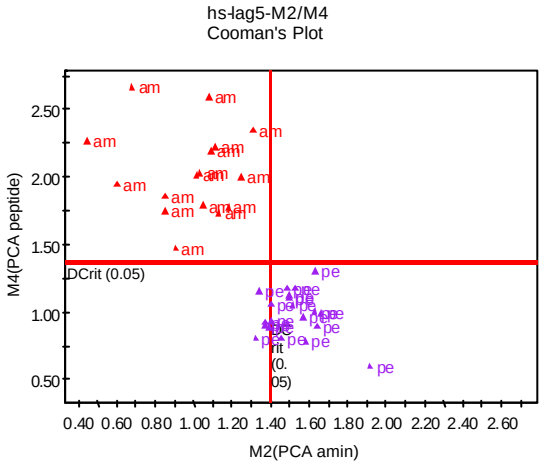

Fig 52. Cooman plot for the whole sequence amine and peptide PCA models, lag 5. Each class fits its model well.

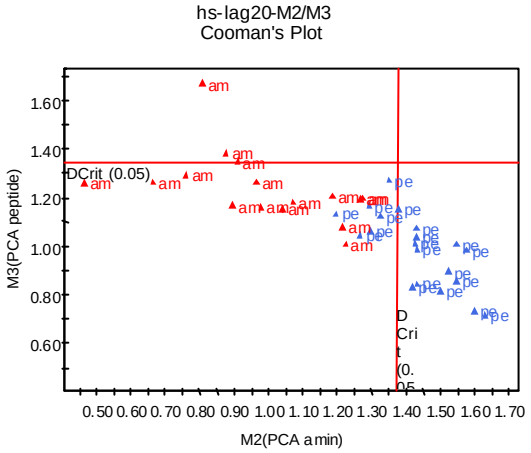

Fig 53. Cooman plot for the whole sequence amine and peptide PCA models, lag 20. Most amine and a few peptide sequences fit both models equally well.

5.3 Validation

The test set was used to test the predictive ability of the whole-sequence model based on amine, rhodopsin and peptide sequences. Generally, in plots of predicted t-values, the rhodopsin class forms a separate cluster while the amine and peptide classes overlap. The models for lags 3-7 show similar results, while lag 20 gives a slightly better separation (Figs 54-55).

The models for each class separately were also validated. Figs 56-59 show predicted t-values and DModX for the amine test set, predicted using the amine and peptide models (lag 5) respectively. The predicted t1/t2 score plots are very similar, and show no outliers for either model. DModX, however, is high for both models. The amine model gives a slightly lower DModX, but there are a high proportion of outliers. Thus, the amine test data does not fit the amine model significantly better than it does the other models, and the same applies to the

rhodopsin and peptide test data. The peptide test set actually fits the rhodopsin model better (Figs 60-63) than the peptide model. In Cooman plots made for the test data, there is no separation between the classes (Figs 64-66).

Thus, the predictive ability of these models is very poor. In an attempt to improve the predictive ability of the model, all outliers were removed from the training data set and a new model fitted (lag 5). This did not, however, make much of a difference.

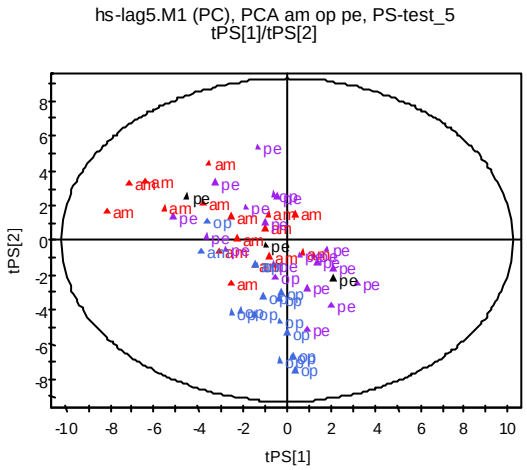

Fig 54. Predicted scores for the test set predicted in the whole sequence PCA model, lag 5. There is an extensive overlap between the classes.

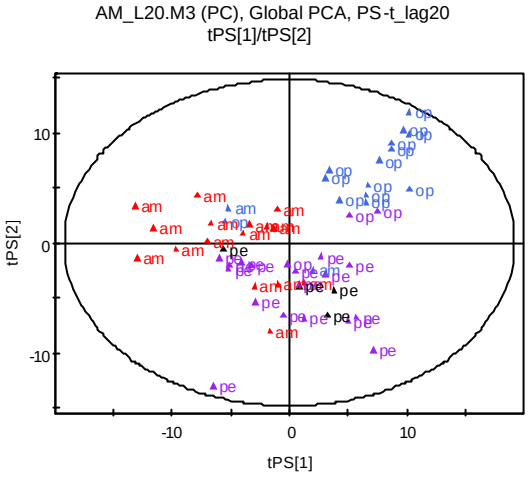

Fig 55. Predicted scores for the test set predicted in the whole sequence PCA model, lag 20. The rhodopsin class is well separated, but the amine and peptide classes overlap.

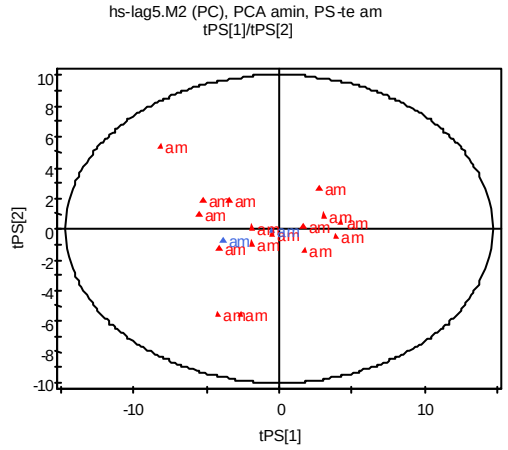

Fig 56. Predicted scores for the amine test set predicted in the whole sequence amine PCA model, lag 5, showing a good fit in the score space.

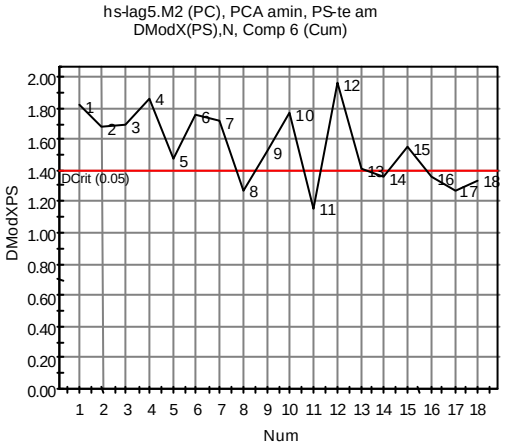

Fig 57. Predicted DModX for the amine test set predicted in the whole sequence amine PCA model, showing a large proportion of outliers.

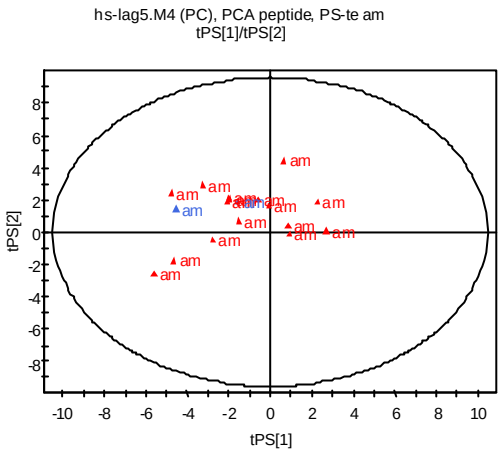

Fig 58. Predicted scores for the amine test set predicted in the whole sequence peptide PCA model, lag 5, showing a good fit in the score space.

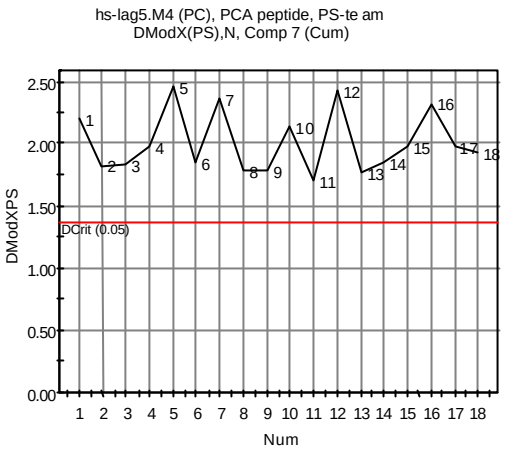

Fig 59. Predicted DModX for the amine test set predicted in the whole sequence peptide PCA model. None of the amine sequences fit this model.

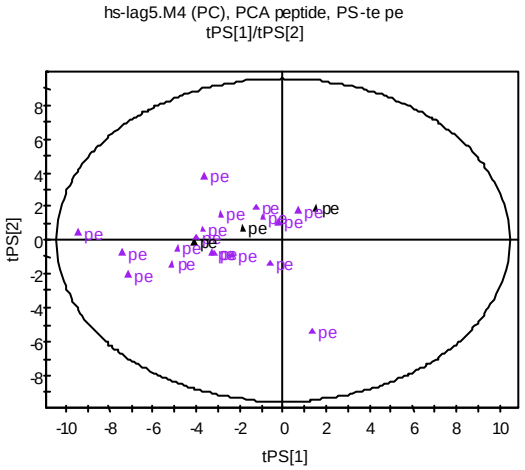

Fig 60. Predicted scores for the peptide test set predicted in the whole sequence peptide PCA model, lag 5, showing a good fit in the score space.

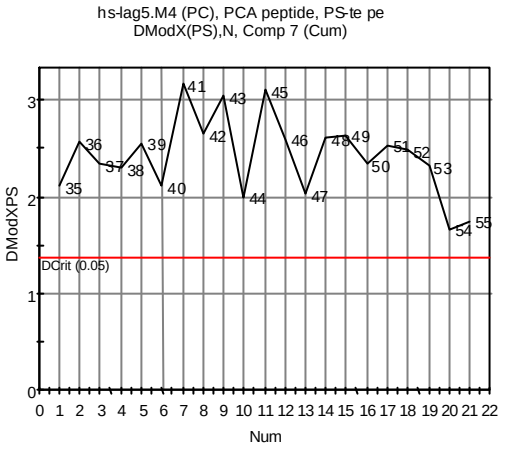

Fig 61. Predicted DModX for the peptide test set predicted in the whole sequence peptide PCA model. None of the peptide sequences fit this model.

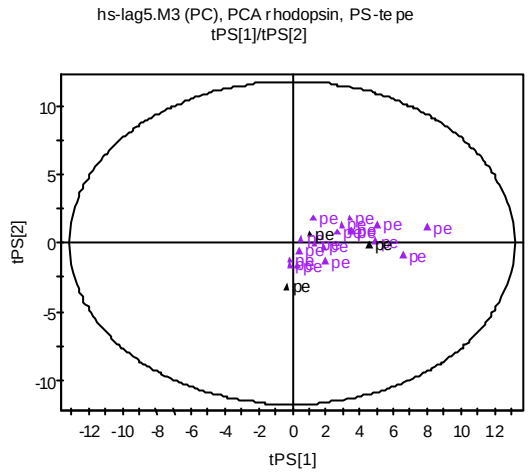

Fig 62. Predicted scores for the peptide test set predicted in the whole sequence rhodopsin PCA model, lag 5, showing a good fit in the score space.

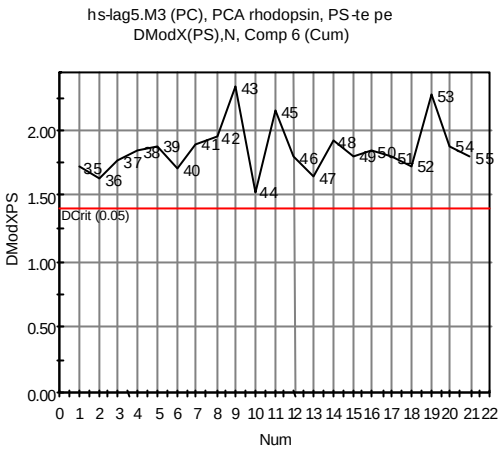

Fig 63. Predicted DModX for the peptide test set predicted in the whole sequence rhodopsin PCA model. None of the peptide sequences fit this model.

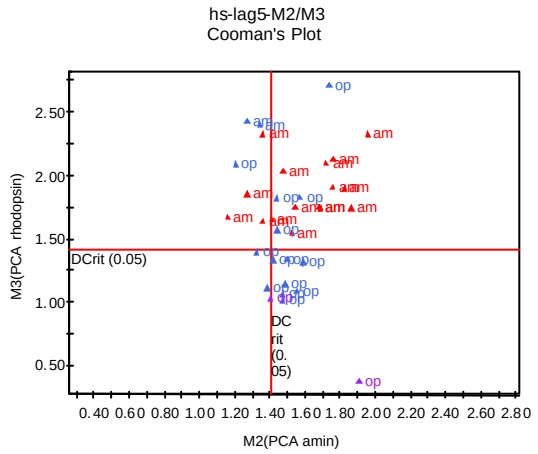

Fig 64. Cooman plot for the amine and rhodopsin test set. Most of the amine and a few of the rhodopsin sequences fit neither model.

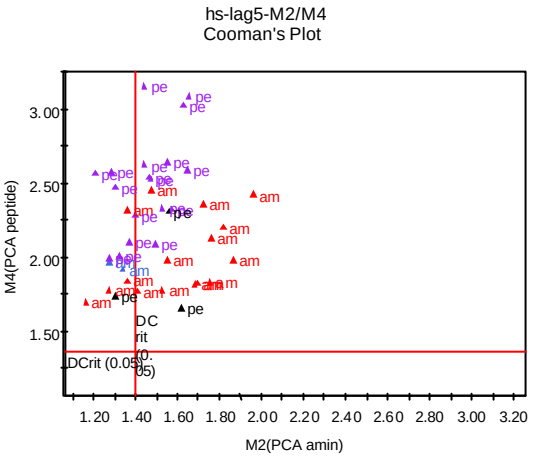

Fig 65. Cooman plot for the amine and peptide test set. Most of the sequences from both classes fit neither model.

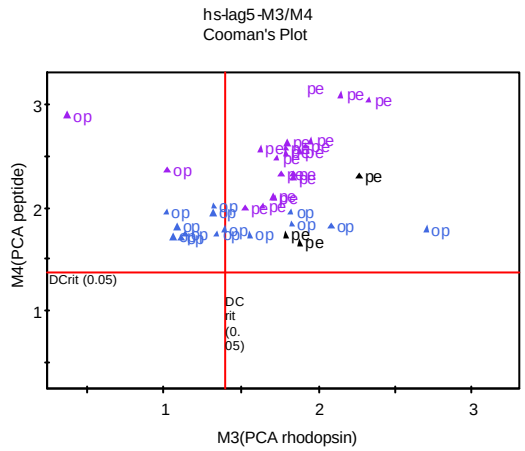

Fig 66. Cooman plot for the rhodopsin and peptide test set. All of the peptide and most of the rhodopsin sequences fit neither model.

For comparison, the same training and test set was used to make a model based only on the transmembrane regions. This gave significantly better class separation in the model score plots and Cooman plots, a better DModX for the predicted test set and a clear separation between the classes in the Cooman plots for the test set (Figs 67-69). Also, each class in the test group fitted its own class model significantly better than the other models (Figs 70-73).

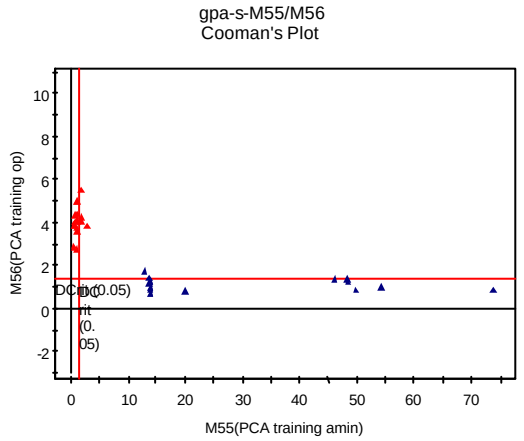

Fig 67. Cooman plot for the amine and rhodopsin test set in the TM model. Each class fits its model well but both have a few moderate outliers.

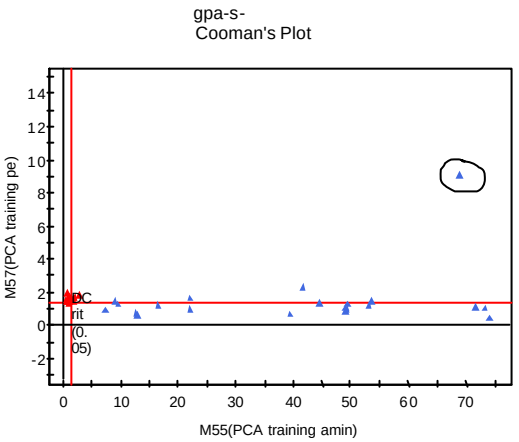

Fig 68. Cooman plot for the amine and peptide test set in the TM model. Each class fits its model well but both have a few moderate outliers. The encircled data point is the peptide sequence O54689, a strong outlier.

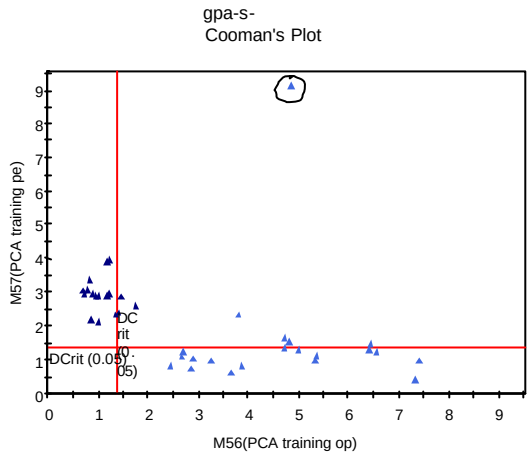

Fig 69. Cooman plot for the rhodopsin and peptide test set in the TM model. Each class fits its model well but both have a few moderate outliers. The encircled data point is the peptide sequence O54689, a strong outlier.

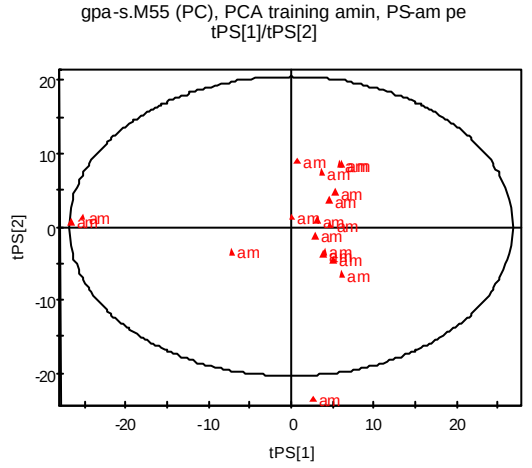

Fig 70. Predicted scores for the amine test set predicted in the amine TM model, showing a good fit in the score space.

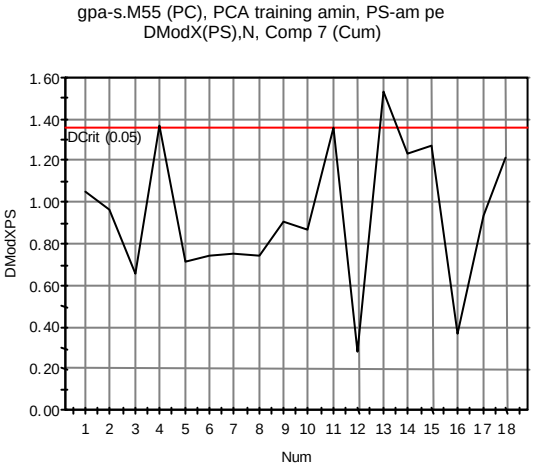

Fig 71. Predicted DModX for the amine test set predicted in the amine TM model, showing a good fit to the model, with just one moderate outlier.

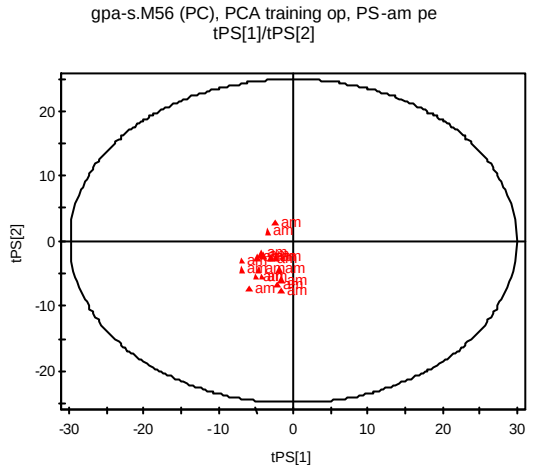

Fig 72. Predicted scores for the amine test set predicted in the rhodopsin TM model, showing a good fit in the score space.

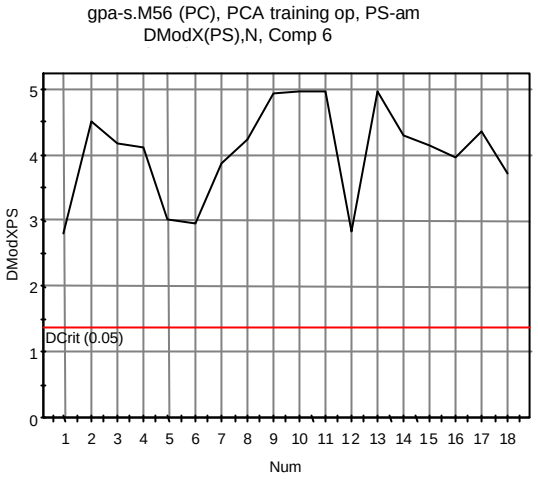

Fig 73.Predicted DModX for the amine test set predicted in the rhodopsin TM model. None of the amine sequences fit this model.

5.4 Extension of the training data

The reason for the poor predictive ability of the whole sequence model might be connected to the way in which the training and test sets were selected. A larger training set might help overcome this problem. Therefore, the complete sequences of all amine, rhodopsin and peptide sequences in the original dataset were downloaded from the Internet [21]. The sequences were translated to zz-scales and ACC calculated using Seqan 1.1. A lag of five was used here, as this seemed to work best in the previous analysis. The dataset consists of 626 sequences and 125 variables, a PCA model fitted to this data has 42 components, explaining 81% of the variance. The t1/t2 score plot of the PCA model shows a big overlap between the peptide and amine classes, and a plot of DModX shows many outliers (Figs 74-75). To validate the model, the sequences previously used as a training set were left out, and a new model fitted to the remaining data. This model, based on 555 sequences, also had 42 components, explaining 81% of the variance, and was used to predict the sequences that had been left out.

Although the model itself seems to be of poorer quality, with a big overlap between the amine and peptide classes and a high proportion of outliers, compared to the model based on the smaller training set that shows well-separated classes, its predictive ability appears to be better. The training set for each class fits its own model much better than the others (Figs 76-79), and Cooman plots for the test set shows well-separated classes, although many observations are found in the upper right-hand corner of the plot, indicating that they fit neither of the models (Figs 80-82).

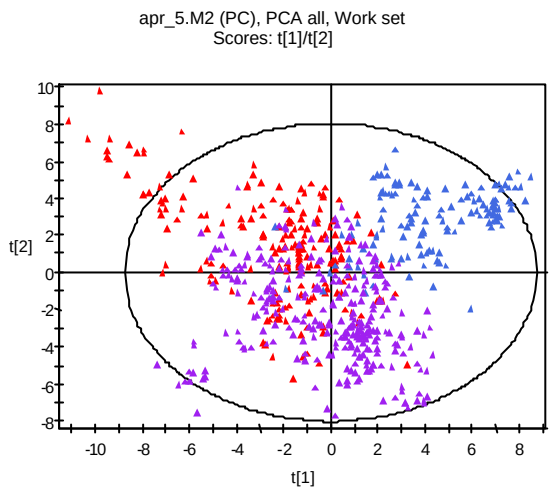

Fig 74. Score plot for the whole sequence PCA model for the amine, rhodopsin and peptide classes, lag 5, using the extended training set. The amine (red) and peptide (purple) classes overlap extensively.

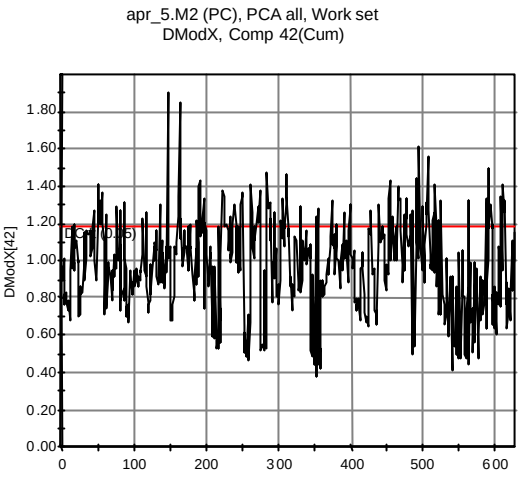

Fig 75. DModX for the whole sequence PCA model for the amine, rhodopsin and peptide classes, lag 5, using the extended training set. The plot shows several moderate and a few strong outliers.

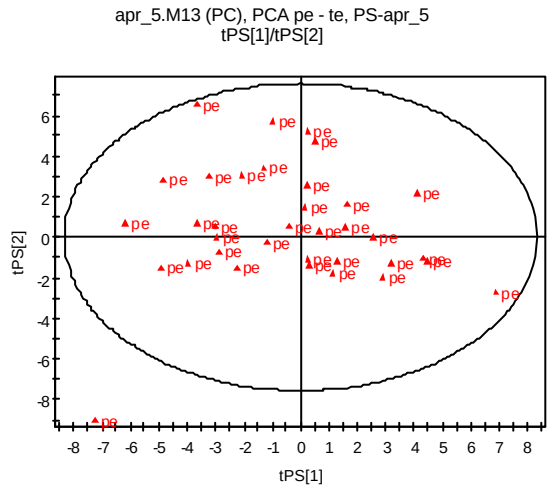

Fig 76. Predicted scores for the whole sequence peptide test set predicted in peptide PCA model, showing a good fit in the score space.

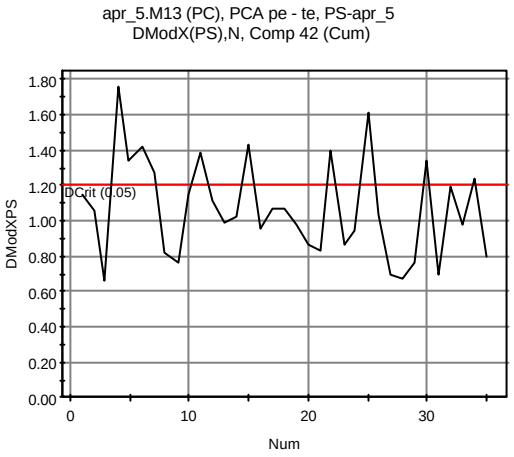

Fig 77. Predicted DModX for the whole sequence peptide test set predicted in peptide PCA model, showing a few moderate outliers.

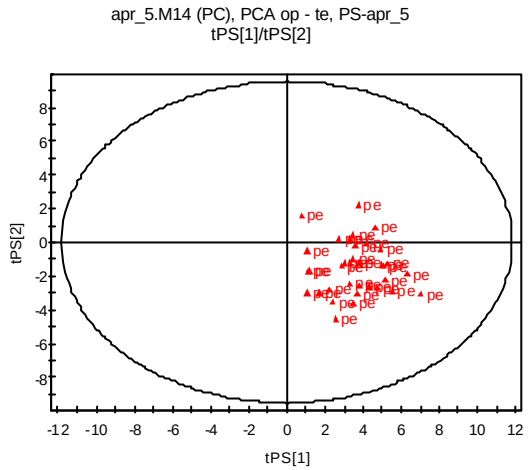

Fig 78. Predicted scores for the whole sequence peptide test set predicted in rhodopsin model, showing a good fit in the score space.

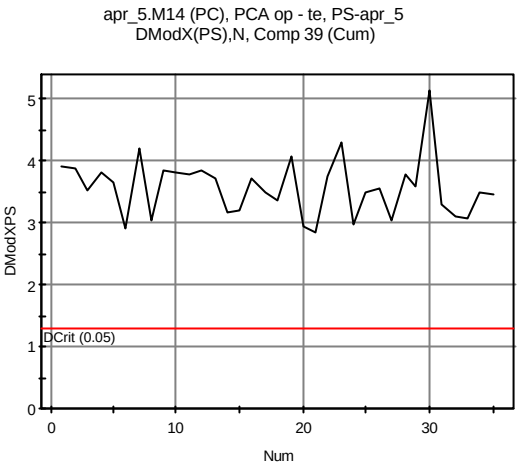

Fig 79. Predicted DModX for the whole sequence peptide test set predicted in rhodopsin model. None of the peptide sequences fit this model.

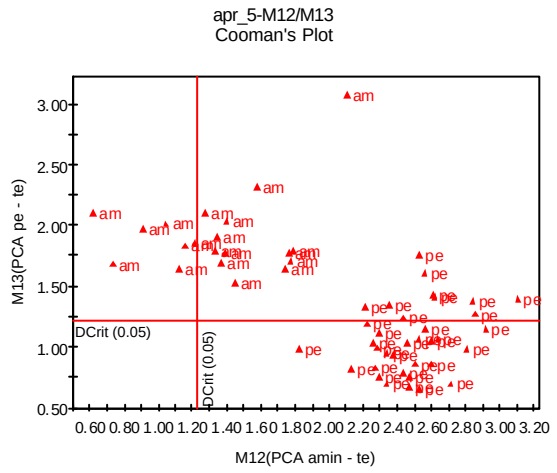

Fig 80. Cooman plot for the amine and peptide test sets. Each class fits its own model best, but both have several outliers.

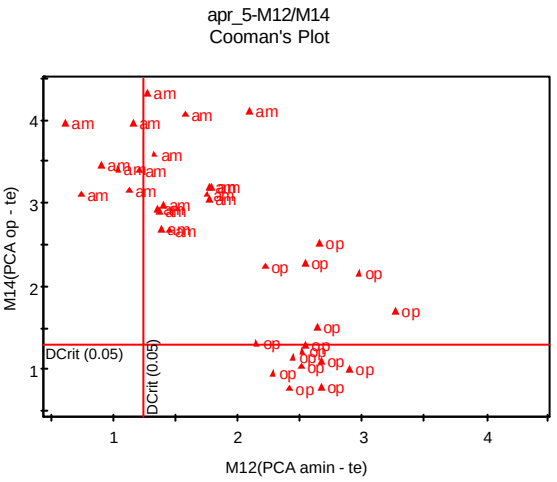

Fig 81. Cooman plot for the amine and rhodopsin test sets. Each class fits its own model best, but both have several outliers.

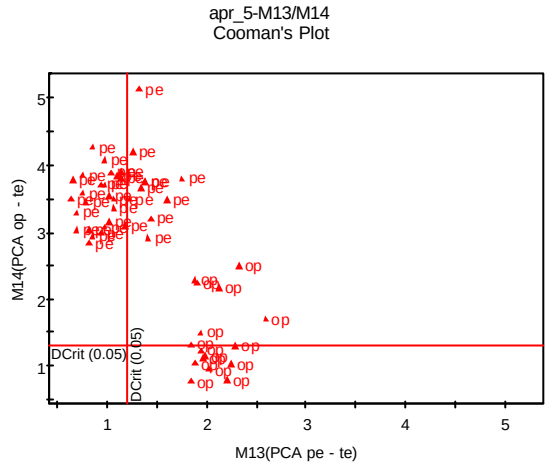

Fig 82. Cooman plot for the rhodopsin and peptide test sets. Each class fits its own model best, but both have several outliers.

6. Analysis of loop regions

6.1 Modelling

Using the same training and test set, the intracellular and extracellular loops connecting the transmembrane sequences were identified and analysed separately. Sequences, along with information about which are the transmembrane regions and which are the loops, were downloaded from the Internet[3] [22]. A few of the sequences had to be omitted due to lack of information about which regions are which. These sequences are in the training set O42317 and O88721, and in the test set O88537, O93237, O73671, O42324, O46554 and O76123. For the remaining sequences, the loop regions were identified, translated to zz-scales and ACC was calculated, separately for each of the eight loops. A lag of five was used, since this worked best in the analysis of the complete sequences. However, the lag can never be bigger than the shortest sequence minus one, and there were a few very short sequences of 2-5 amino acids. These sequences were also excluded. The resulting dataset had 52 sequences and 1000 variables, 125 for each of the eight loops.

A PCA model was fitted to the training data, resulting in a model with 18 components explaining 59% of the variance. In score plots for this model, two rhodopsin clusters are clearly separated from the rest while the amine and peptide classes, as well as part of the rhodopsin class, are inseparable (Figs 83-84). A PLS-DA gives a model with only four components, explaining 18% of the variance in X and 98% of the variance in Y. A score plot for this model shows a good separation between the classes (Fig 85). Cooman plots show a good separation between the amine and rhodopsin, and between the rhodopsin and peptide classes. Between the amine and peptide classes there is a small overlap.

[3] The separation into transmembrane regions and loops in the original data set differs from that suggested at the database used (EXPASY), which is used in this chapter.

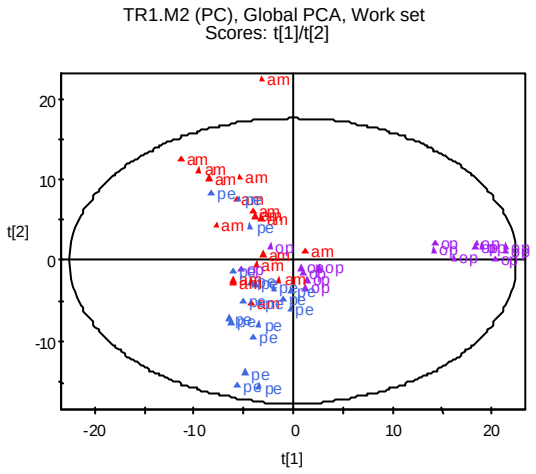

Fig 83. t1/t2 score plot for PCA model of loop training sequences. Part of the rhodopsin class forms a well separated cluster while the amine and peptide classes, as well as remaining rhodopsin sequences, overlap.

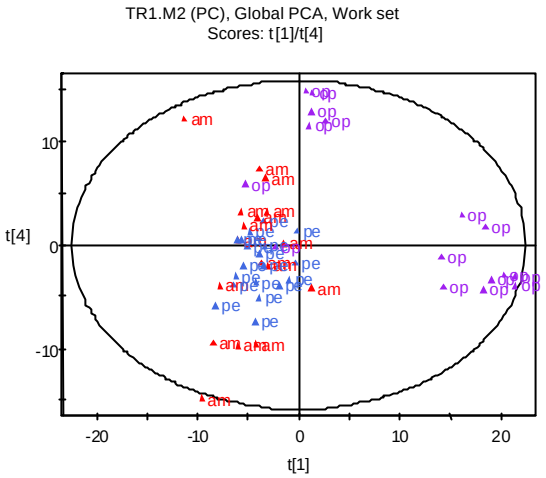

Fig 84. t1/t4 score plot for PCA model of loop training sequences. Two well separated rhodopsin clusters can be seen, the amine and peptide classes overlap.

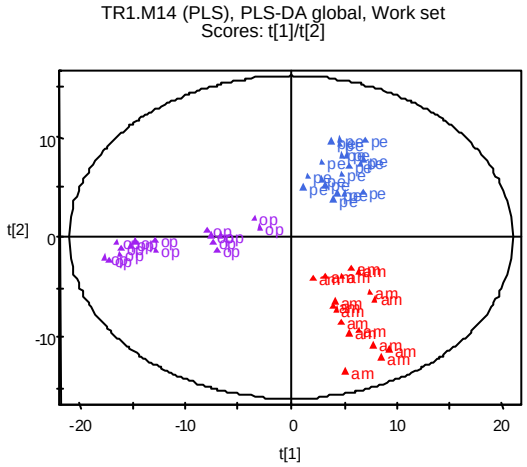

Fig 85. Score plot for PLS-DA model of loop training sequences, showing a good separation of all three classes..

6.2 Validation

The PCA and PLS-DA models were both used for predicting the test set. In a plot of t-values predicted using the PCA model, only one rhodopsin cluster is clearly separated. Using the PLS-DA-model for prediction gives a better separation of clusters, with only a slight overlap between the classes. Cooman plots for the test set show no separation between classes. Each class in the test set was predicted using the separate models for each class. Again, the test data from one class fitted its own class model only slightly better than the other class models, and the predictive ability of the model is not very good (result not shown).

6.3 Hierarchical modelling

Separate PCA models were made for the eight loop sequences, and the components from these models were used as variables in a hierarchical model with around 18 components per loop model. Again, one well-separated rhodopsin cluster can be seen in the score plot for the PCA model, and the amine and peptide classes overlap completely (Fig 86). The first component in each separate loop model is clearly the most important for the separation of the rhodopsin cluster (Fig 87). A PLS-DA model gives a good separation between the classes (Fig 88).

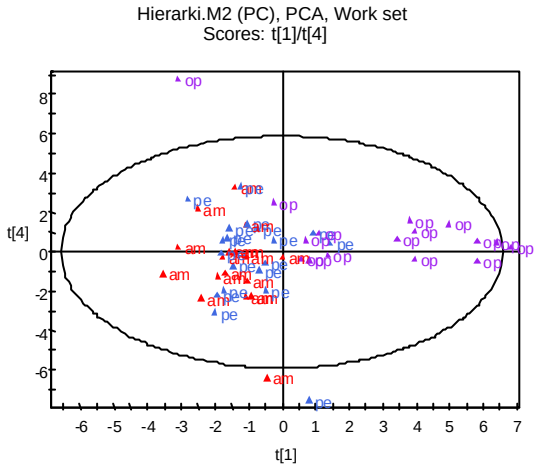

Fig 86. Score plot for hierarchical PCA model for loop training sequences based on all components significant by cross validation. Part of the rhodopsin class forms a well separated cluster while the amine and peptide classes, as well as remaining rhodopsin sequences, overlap.

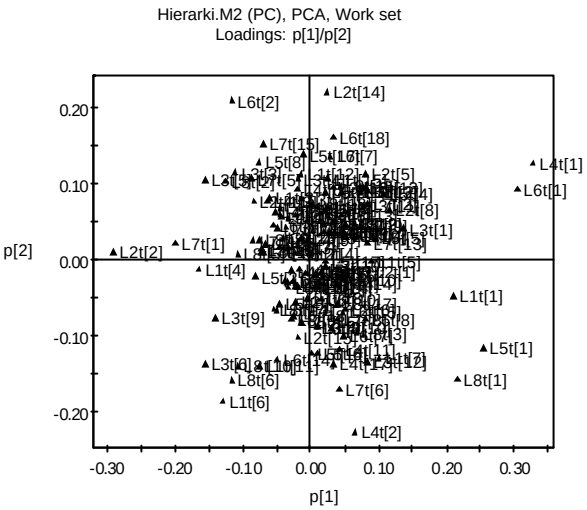

Fig 87. Loading plot for hierarchical PCA model for loop training sequences based on all components significant by cross validation.

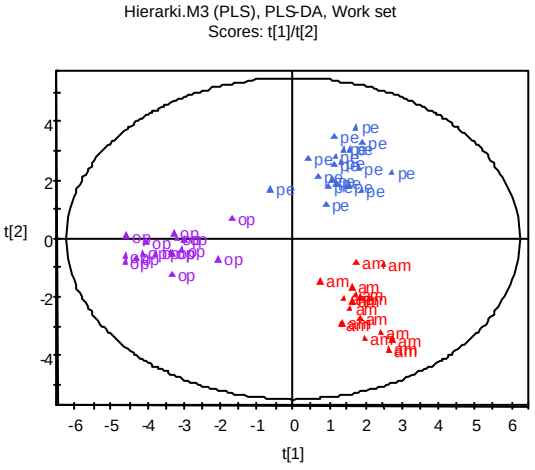

Fig 88. Score plot for hierarchical PLS-DA model for loop training sequences based on all components significant by cross validation. The three classes are well separated

7. Analysis of transmembrane and loop regions

7.1 Modelling

A hierarchical model where the TM and loop regions were combined was made, to see if combining the information in this way gives more information than looking at the complete sequences directly does. The same training data set as in the loop models was used. All components from the separate models for each of the 7 TM regions and 8 loop regions that are significant according to cross validation, 17-18 each, were put together to form a new set of variables. Thus a dataset with 52 observations characterised by 269 variables was obtained. A PCA model was fitted to this data, resulting in a model with 18 components, explaining 59% of the variance. For comparison, a dataset composed of the first four components from each loop/TM region, in total 60 variables was also investigated, resulting in a PCA model with 12 components, explaining 75% of the variance. t1/t2 score plots for these models show that the model based on four components only from each region gives a clearer separation between the classes (Figs 89-90).

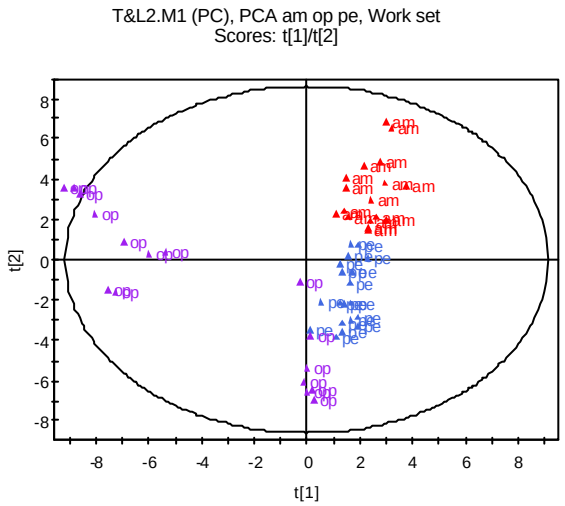

Fig 89. Score plot for hierarchical PCA model for TM and loop regions, based on all components significant by cross validation. There is a separation between the classes.

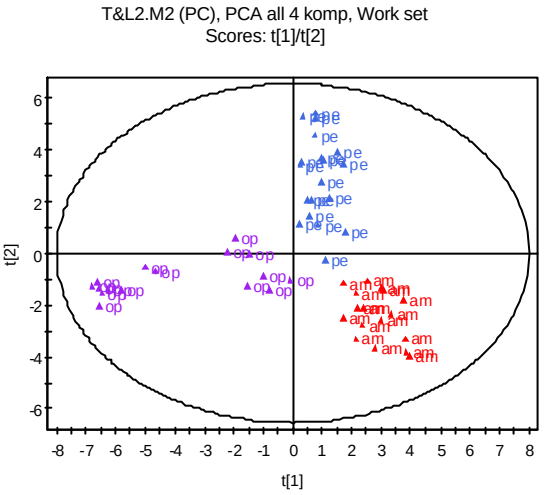

Fig 90. Score plot for hierarchical PCA model for TM and loop, based on four components from each region. The three classes are well separated.

For both model approaches, i.e. using four components or the number of components as determined by cross validation, Cooman plots were made, and in both cases the three classes were well separated. However, for models based on all components there were a few outliers, and observations were found in the lower left corner of the Cooman plots, whereas the model based on four components from each region had neither of those problems (Figs 91-92).

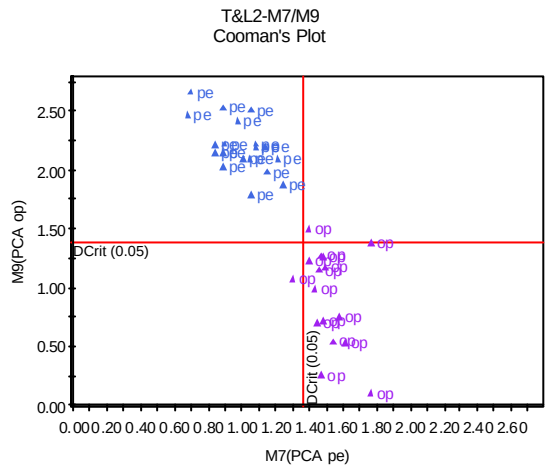

Fig 91. Cooman plot for the peptide and rhodopsin classes, hierarchical PCA model for TM and loop regions based on all significant components. Both classes fit their model well, but the rhodopsin class has a couple of moderate outliers.

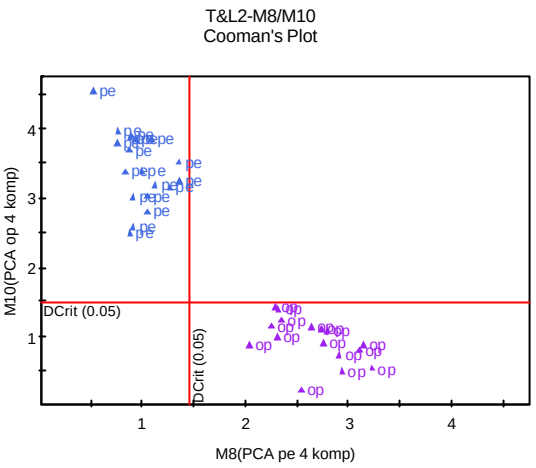

Fig 92. Cooman plot for the peptide and rhodopsin classes, hierarchical PCA model for TM and loop regions based on four components from each region. Both classes fit their model well, with no outliers.

7.2 Validation

For both model approaches, models were validated using a test set with sequences from the amine, rhodopsin and peptide classes, as well as a few sequences from each of the other classes in the original data set. The prediction of a test set in a hierarchical model has to be done in two steps. First, the sequences in the test set are divided into TM regions and loops, and each region predicted in the corresponding model. Next, the predicted t-values are used as a test set, and t-values in the hierarchical model predicted. Separate hierarchical models for each class were used to predict t-values and DModX for the test set.

The models based on all components from each TM and loop region have a very poor predictivity. The amine test set fits the amine and peptide models equally well and the rhodopsin test set also fits the peptide model as well as its own model. Cooman plots do show a separation between the three classes, though with many observations in both the upper right and the lower left part of the plot, i.e. observations that either fit both models in the plot or neither (Figs 93-95). Only the peptide test set fits its own model and not the others. The test set sequences from other classes (hormone protein (hp), olfactory (ol), nucleotide like (nu), cannabis (cb), platelet activating factor (pa), gonadotropin releasing hormone (gr), thyrotropin releasing hormone (tr), melatonin (ml), and orphan (or)) fitted well into both the amine and peptide class models.

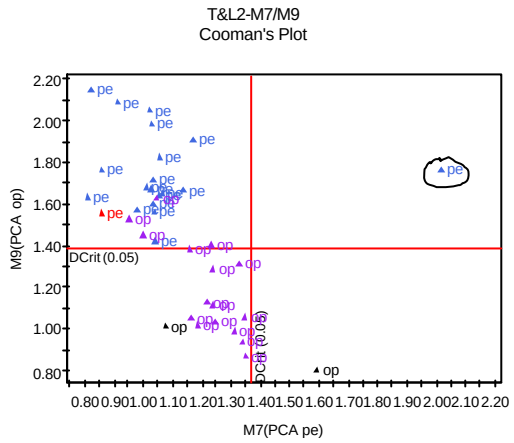

Fig 93. Cooman plot for the rhodopsin and peptide test set. Hierarchical models for TM and loop regions based on all components significant by cross validation. The peptide class fits its model well, apart from one strong outlier, O54689 (encircled). The rhodopsin class fits both models.

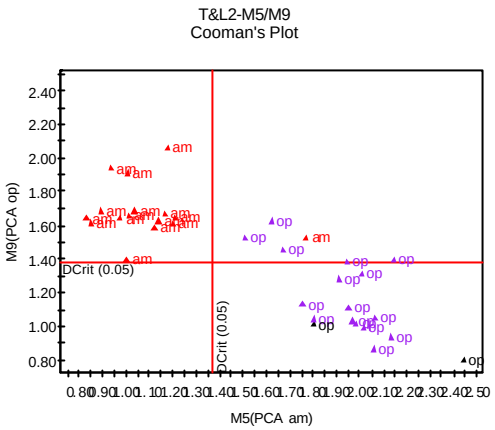

Fig 94. Cooman plot for the rhodopsin and amine test set. Hierarchical models for TM and loop regions based on all components significant by cross validation. The amine class fits its model well apart from one moderate outlier, the rhodopsin class has several moderate outliers.

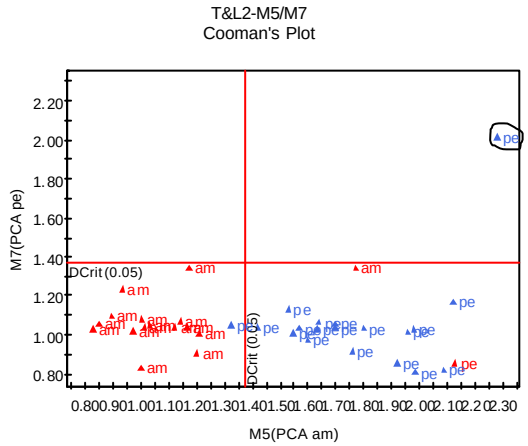

Fig 95. Cooman plot for the amine and peptide test set. Hierarchical models for TM and loop regions based on all components significant by cross validation. The peptide class fits its model well, apart from one strong outlier, O54689 (encircled). The amine class fits both models.

The models based on four components from each region have a better, though not perfect, predictive ability. The amine, rhodopsin and peptide test sets all fit their own model and not the others. Cooman plots show well separated classes, with only a few sequences found in the upper right or lower left part of the plot (Figs 96-98). The test set with sequences from other classes, however, that ideally should fit none of the models, fits perfectly into the peptide model (result not shown).

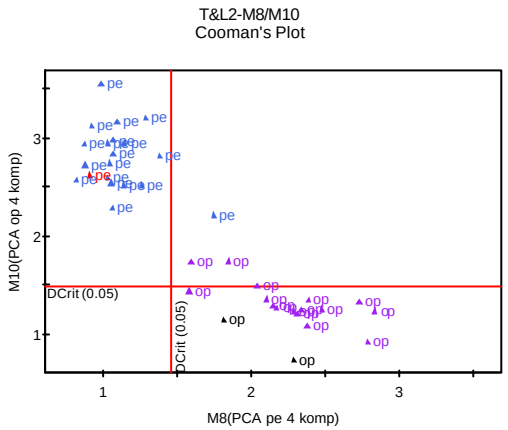

Fig 96. Cooman plot for the rhodopsin and peptide test set. Hierarchical models based on four components from each TM and loop region. Both classes fit their model well, with a few moderate outliers.

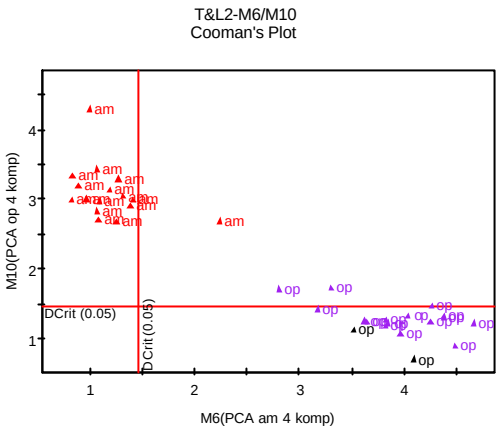

Fig 97. Cooman plot for the rhodopsin and amine test set. Hierarchical models based on four components from each TM and loop region. Both classes fit their model well, with a few moderate outliers.

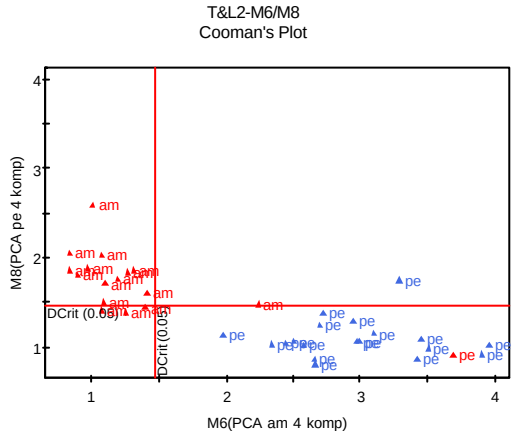

Fig 98. Cooman plot for the amine and peptide test set. Hierarchical models based on four components from each TM and loop region. Both classes fit their model well, with a few moderate outliers. A few amine sequences fit both models.

Clearly the predictive ability of the hierarchical models based on transmembrane and loop regions is not very good. They can classify correctly sequences from the three classes the models are based on, but fail to recognise as outliers sequences from other classes.

This is a first attempt to model a full sequence (TM and loop regions) hierarchically, and the results reported are preliminary. Some amino acids are missing from the sequences in the training data set due to the previously mentioned difference in definition of transmembrane regions and loops between the original data set and the database used (EXPASY).

The TM & Loop hierarchical model was then compared to the hierarchical models for the TM and loop regions separately, and to the model for the whole sequences, all for the same training and test set. The combined hierarchical model shows a better separation between the

classes in the t1/t2 score plot than either of the separate hierarchical models (Figs 99-101), and the whole sequence model (Fig 102).

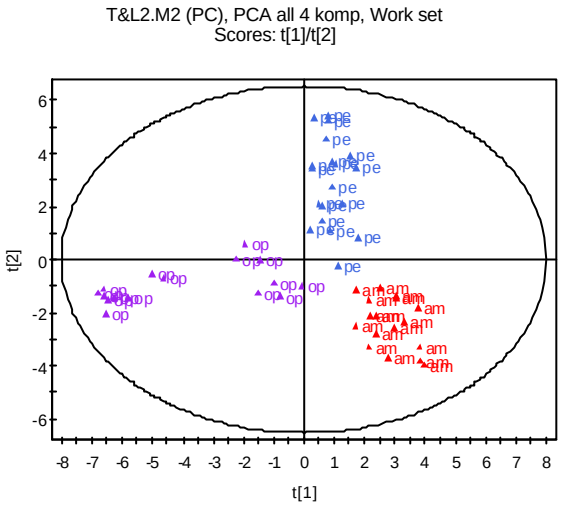

Fig 99. Score plot for hierarchical TM and loop model based on four components from each TM and loop region. The three classes are well separated.

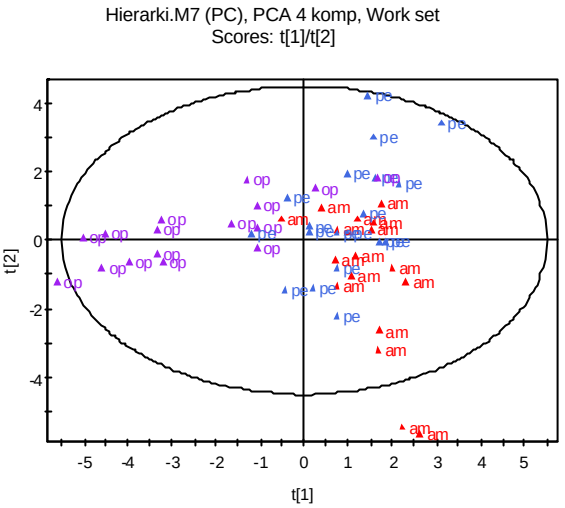

Fig 100. Score plot for hierarchical loop model based on four components from each loop region. The amine and peptide classes overlap.

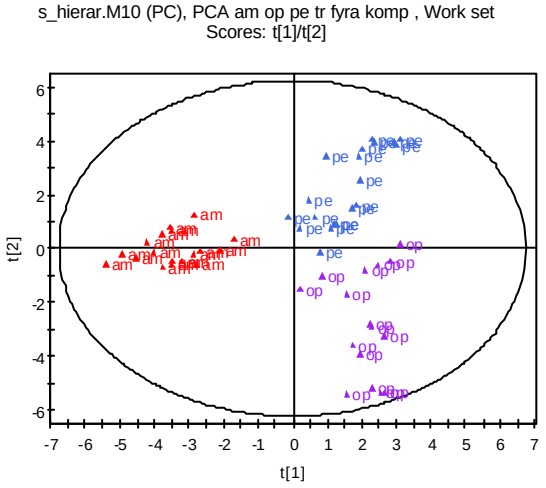

Fig 101. Score plot for hierarchical TM model based on four components from each TM region. The three classes are well separated.

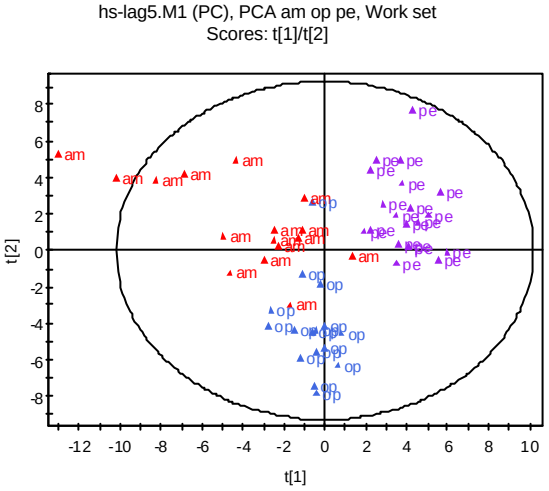

Fig 102. Score plot for whole sequence model, lag 5. There is a small overlap between classes.

The same training set was used for modelling using transmembrane regions only, both global and hierarchical models were made for the amine, rhodopsin and peptide classes. The test set containing sequences from the other classes in the original data set was then predicted and the results compared. Looking at the Cooman plots showed quite different results for the different models. In the global 7TM model, none of the sequences in the test set was predicted as belonging to any of the class models (Figs 103-105). In the hierarchical 7 TM model, a few of the sequences were classified as belonging to the peptide model (Figs 106-108), and as mentioned above, in the hierarchical TM & loop model, all the sequences in the test set were predicted as belonging to the peptide model (Figs 109-111).

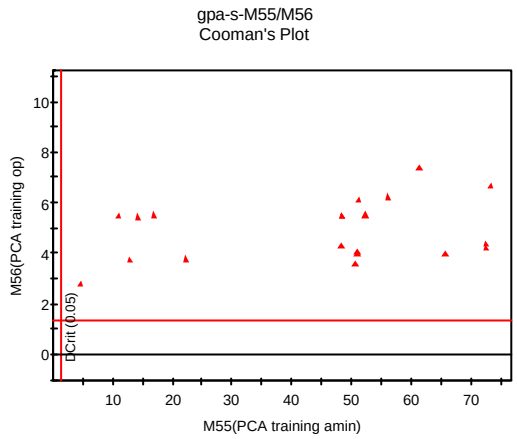

Fig 103. Cooman plot for amine and rhodopsin 7TM models, test set from other classes. None of the test sequences are predicted as belonging to either class.

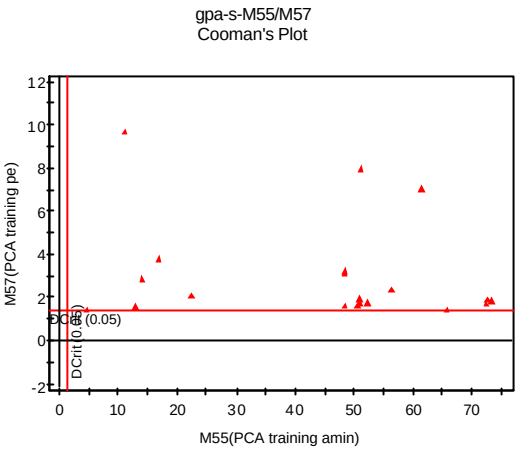

Fig 104. Cooman plot for amine and peptide7TM models, test set from other classes. None of the test sequences are predicted as belonging to either class.

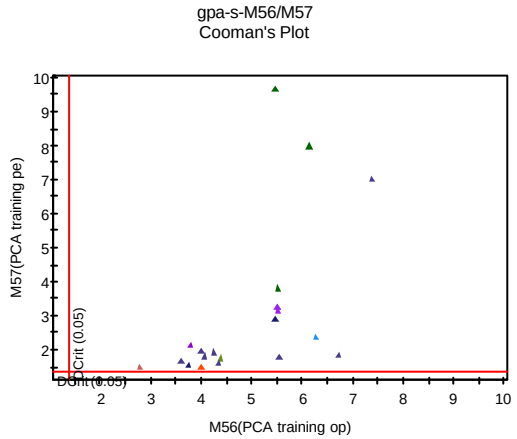

Fig 105. Cooman plot for peptide and rhodopsin 7TM models, test set from other classes. None of the test sequences are predicted as belonging to either class.

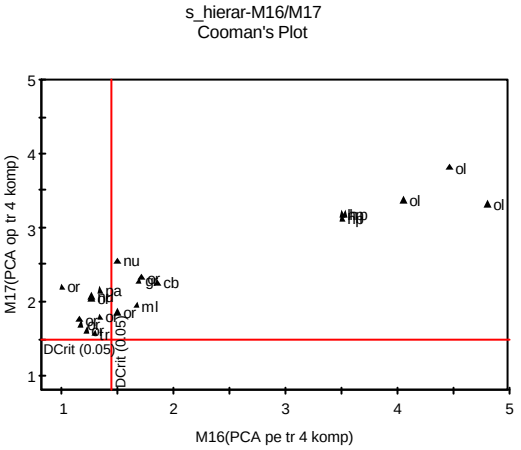

Fig 106. Cooman plot for peptide and rhodopsin hierarchical 7TM models, test set from other classes. A few sequences, belonging to the or, nu, tr and pa classes, are predicted as belonging to the peptide class.

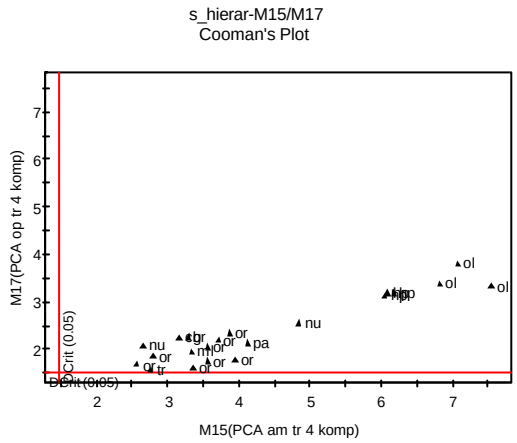

Fig 107. Cooman plot for amine and rhodopsin hierarchical 7TM models, test set from other classes. None of the test sequences are predicted as belonging to either class.

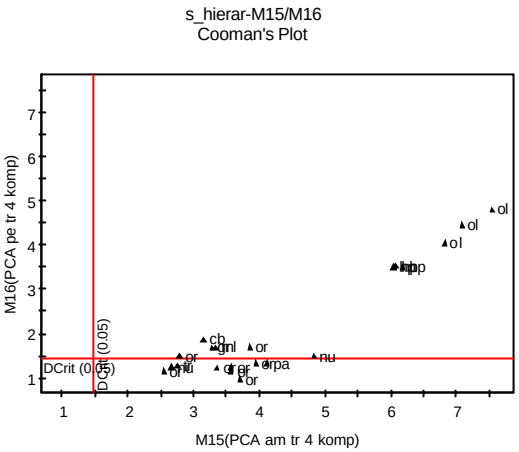

Fig 108. Cooman plot for peptide and amine hierarchical 7TM models, test set from other classes. A few sequences, belonging to the or, nu, tr and pa classes, are predicted as belonging to the peptide class.

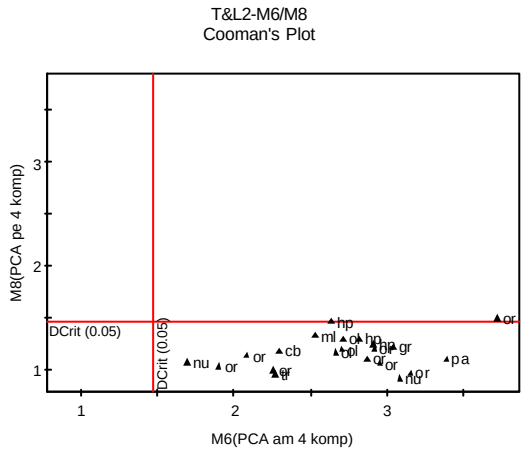

Fig 109. Cooman plot for amine and peptide hierarchical TM & loop models, test set from other classes. All test sequences are predicted as belonging to the peptide class.

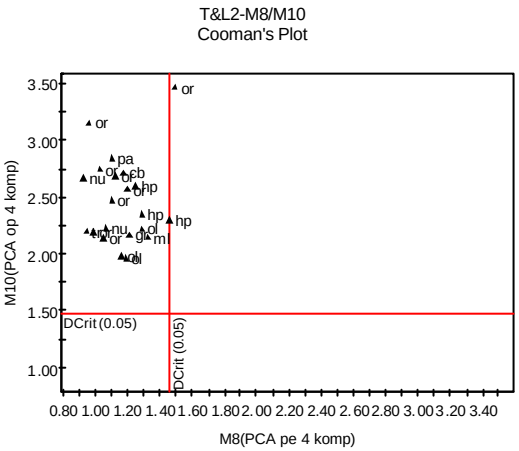

Fig 110. Cooman plot for rhodopsin and peptide hierarchical TM & loop models, test set from other classes. All test sequences are predicted as belonging to the peptide class.

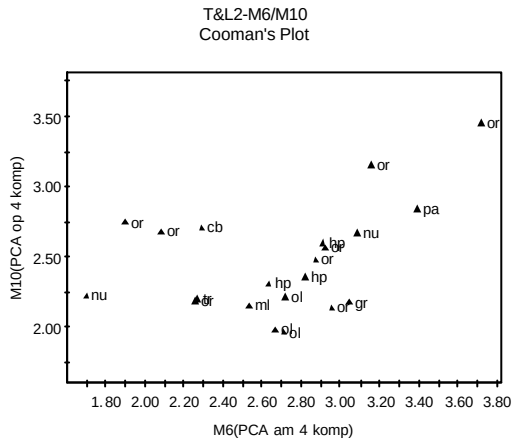

Fig 111. Cooman plot for amine and rhodopsin hierarchical TM & loop models, test set from other classes. None of the test sequences are predicted as belonging to either class.

8. Conclusions

In this project a multivariate approach has been used to compare and classify G protein-coupled receptor (GPCR) sequences based on their physicochemical properties. The aim was to find out whether the receptors separate into groups according to biological function, and the results do indeed show groupings of the receptors according to receptor type and thus biological function.

A global model of the dataset resulted in the separation of three classes from the rest: amine, olfactory and rhodopsin. The remaining classes could not be separated by the global model, but with SIMCA modelling all classes could be shown to be well separated. By making the models increasingly local, by focusing on one cluster of interest and making a new model for the sequences of that cluster only, more detailed information about sub clusters etc can be obtained.

Since bovine rhodopsin is the only GPCR with a known 3D structure, all other GPCR's are aligned towards that 3D structure. It is therefore interesting to note that the rhodopsin class is so well separated from the others in this study. This suggests that its structure is not actually such a good model for the structure of receptors belonging to other classes.

Hierarchical models were made, to see if any transmembrane region was particularly important for the differences between classes. However, it appears that all regions are equally important.

Different models have been compared – global or hierarchical models, based on TM regions, whole sequences or loop regions. The degree of separation between classes clearly depends on the type of model, with models based only on TM regions giving the best separation. Test data have been classified using the different models, but unfortunately there is no consensus between the different models. Thus, using any of these models for the classification of new sequences would be difficult.

9. Future studies

There are scales other than the five zz-scales used here that could be used for a multivariate characterization of the sequences. It would be interesting to compare different scales to see how the choice of scales affects the result of modelling and classification.

The poor predictivity of the models in this study might be due to the large size of the classes. It would therefore be interesting to further study classification using models based on smaller sub groups.

A possible correlation between the sequence of receptors and their biological function could be investigated using a response matrix containing e.g. data on receptor binding. There is data available for a number of receptors for which binding assays have been performed for many different ligands.

Acknowledgements

First of all, I would like to thank everyone at Melacure Therapeutics for making me feel welcome, and for making my time here so enjoyable. In particular, I would like to thank my supervisor, Per Andersson, for always taking the time to answer my questions and help me solve any problems, and for valuable input into the writing of the report. Thanks also to my examiner, Prof. Torbjörn Lundstedt, for giving me the opportunity to work with this project, for guiding me through my work, and for many interesting discussions.

List of abbreviations

ACC	Auto Crossed Covariances
DModX	Distance to the model in the X block
GPCR	G protein-coupled receptor
MVD	Multivariate Design
PCA	Principal Component Analysis
PLS	Partial Least Squares Projections to Latent Structures
PLS-DA	PLS Discriminate Analysis
SIMCA	Soft Independent Modelling of Class Analogies
TM	Transmembrane

References

1. R. C. Graul and W. Sadée, Evolutionary relationships among G protein-coupled receptors using a clustered database approach. *AAPS PharmSci.* **3**(2), article 12 (2001)
2. http://www.nobel.se/medicine/laureates/1994/illpres/index.html (010906)
3. K. Palczewski et al, Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor, *Science* **289,** 739-745 (2000).
4. M. Sandberg et al, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **41**(14), 2481-2491 (1998)
5. M. Sandberg, Deciphering Sequence Data A Multivariate Approach. Ph.D. thesis, *Research Group for Chemometrics, Department of Organic Chemistry,* Umeå University, Umeå, Sweden (1997)
6. M. Sjöström et al, Polypeptide sequence property relationships in Escherichia coli based on auto cross covariances, *Chemometrics and intelligent laboratory systems* **29,** 295-305 (1995)
7. M. Edman, Detection of sequence patterns in membrane proteins. Ph.D. thesis, *Biochemistry and Organic Chemistry/Research Group for Chemometrics, Department of Chemistry,* Umeå University, Umeå, Sweden (2001)
8. S. Wold et al, DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures, *Analytica Chimica Acta* **277,** 239-253 (1993)
9. Å. Nyström, P. M. Andersson and T. Lundstedt, Multivariate Data Analysis of Topographically Modified α-Melanotropin Analogues using Auto and Cross Auto Covariances (ACC). *Quant. Struct-Act. Relat.* **19,** 264-269 (2000)
10. L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, *Multi-and megavariate data analysis,* Umetrics AB (2001)
11. C. M. Bishop, *Neural Networks for pattern recognition*, Oxford University Press (1995)
12. S. Wold, Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models, *Technometrics* **20**(4), (1978)
13. S. Wold, S. Hellberg, T. Lundstedt, M. Sjöström and H. Wold, PLS modelling with latent variables in two or more dimensions. *The PLS-meeting Frankfurt* (1986)
14. K. Janne, J. Pettersen, N.-O. Lindberg and T. Lundstedt, Hierarchical principal component analysis (PCA) and projection to latent structure (PLS) technique on spectroscopic data as a data pretreatment for calibration. *J. chemometrics* **15**, 203-213 (2001)
15. B. Thelin et al, Classification of Estradurin® batches: correlation between ^{31}P NMR and a biological duration test for batch approval. *Chemometrics and Intelligent laboratory Systems* **00** (1994)
16. S. Wold, M. Sjöström, R. Carlsson, T. Lundstedt, S. Hellberg, B. Skageberg and C. Wikström, Multivariate design. *Analytica Chimica Acta* **191,** 17 (1986)
17. R. Carlsson and T. Lundstedt, Scope of organic synthetic reactions. Multivariate methods for exploring the reaction space. An example of the Willergodt-Kindler reaction. *Acta Chem. Scand.* **B41**, 164 (1987)
18. P. Geladi and B. R. Kowalski, Partial Least-Squares Regression: A tutorial, *Analytica Chimica Acta* **185,** 1-17(1986)
19. ExPASy: "Expert Protein Analysis System proteomics server of the Swiss Institute of Bioinformatics", http://www.expasy.org/srs5bin/cgi-bin/wgetz 010920

20. GPCRDB: "Information system for G protein-coupled receptors", www.gpcr.org 010920
21. ExPASy: “Expert Protein Analysis System proteomics server of the Swiss Institute of Bioinformatics”, http://www.expasy.org/srs5bin/cgi-bin/wgetz 011114
22. ExPASy: “Expert Protein Analysis System proteomics server of the Swiss Institute of Bioinformatics”, http://www.expasy.org/srs5bin/cgi-bin/wgetz 011008