JOURNAL OF CHEMOMETRICS
J. Chemometrics 2003; **17**: 82–92
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.777

Multivariate analysis of G protein-coupled receptors†

Ingrid Gunnarsson[1], Per Andersson[1], Jarl Wikberg[2] and Torbjörn Lundstedt[1,3]*

[1]Melacure Therapeutics AB, Ulleråkersvägen 38, SE-756 43 Uppsala, Sweden
[2]Department of Pharmaceutical Pharmacology, BMC, Uppsala University, PO Box 591, SE-751 24 Uppsala, Sweden
[3]Department of Organic Pharmaceutic Chemistry, BMC, Uppsala University, PO Box 574, SE-751 23 Uppsala, Sweden

Received 17 August 2002; Accepted 18 November 2002

With the rapidly increasing amount of data of a bioinformatic nature, it is of great interest to find different methods to help overview, classify, identify outliers and in other ways extract information from the large amounts of data. In this paper, chemometric methods often applied in statistical process control are utilized to analyse a large number of G protein-coupled receptors. They have been investigated with respect to groupings among the sequences based on their physicochemical properties using multivariate methods. The transmembrane (TM) regions of 897 receptors were examined. The sequences were multivariately characterized using principal properties for amino acids. The methods used include principal component analysis (PCA), hierarchical principal component analysis (HPCA) and partial least squares projections to latent structures (PLS). The results show that groups of receptors belonging to the same functional class can be identified using this approach, and that the rhodopsin class of G protein-coupled receptors is very different from other classes present in the study with regard to their physicochemical properties. Copyright © 2003 John Wiley & Sons, Ltd.

KEYWORDS: G protein-coupled receptor; PCA; PLS; 7TM; multivariate sequence analysis

1. INTRODUCTION

G protein-coupled receptors (GPCRs) are a large and varied family of receptors in fungi, plants and animals, with the ability to bind many different types of ligand [1]. They are crucial for many of the central functions of our body, including sight, smell and taste. All GPCRs share a common structure with seven transmembrane (TM) regions (Figure 1), but other than that little is known about their 3D structure [2]. In the present study the seven TM regions are abbreviated A, B, C, D, E, F, G. As for all membrane proteins, determining the crystal structure of the receptors is very difficult, and it has only been done 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.

2. MATERIALS AND METHODS

The methods used include principal component analysis (PCA), hierarchical PCA (H-PCA), projections to latent structures (PLS), PLS discriminant analysis (PLS-DA) and hierarchical PLS (H-PLS). All variables were mean centred and scaled to unit variance. The number of significant components was determined using cross-validation or an eigenvalue larger than two, unless otherwise stated. The amino acid sequences have been quantitatively described using the five zz-scales described by Sandberg *et al.* [4].

2.1. 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

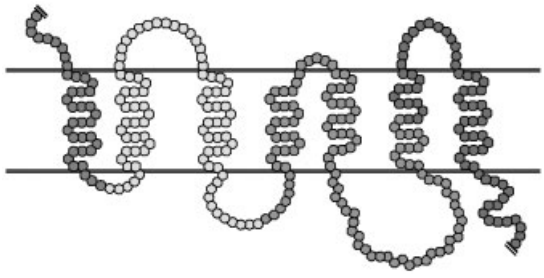

Figure 1. 7TM structure of a GPCR.

*Correspondence to: T. Lundstedt, Melacure Therapeutics AB, Ulleråkersvägen 38, SE-756 43 Uppsala, Sweden.
E-mail: torbjorn.lundstedt@melacure.com
†Dedicated to Professor John F. MacGregor: a pioneer of multivariate statistical process control and recipient of the fourth Herman Wold medal.

phenomena and to present complex phenomena in a form that is easy to understand [5].

2.2. Sequence data

The sequence data used initially are an in-house collection of the transmembrane (TM) regions of 897 G protein-coupled receptors from different species. 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 subclasses. The 12 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.

2.3. 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.* [4]. Three z-scales for the 20 coded amino acids were described by Hellberg *et al.* [6] and have subsequently been extended by Sandberg *et al.* [4] 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 physicochemical 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 from a physicochemical point of view. In our study they are, however, useful and will be used in our interpretation and identification of the importance of different amino acids in different positions [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. A sequence of length p amino acids will thus be represented by $5p$ variables in a so-called multipositional description [7].

2.4. PCA

PCA is a projection method used to visualize data in high dimensions by reducing the dimension of the data. The starting point is a matrix of data, $\mathbf{X}$, with N rows (observations) and K columns (variables). PCA finds the line/plane/hyperplane 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 of a new co-ordinate system into which the data points are projected. The co-ordinates of the data points in this new co-ordinate system are called scores [8]. The principal components are the eigenvectors of the covariance matrix of the data matrix $\mathbf{X}$ and are thus orthogonal. The eigenvectors associated with the largest eigenvalues of the data correspond to the directions of the largest variation in the data [9].

Before applying PCA, data are normally pretreated. The most common treatments are mean centring and scaling to unit variance. The variables of a data set 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 normalized, 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 centring improves the interpretability of the model and is done by calculating the average value of each variable and subtracting it from the data [8].

Mathematically, the model plane can be expressed as

$$\mathbf{X} = \mathbf{x}^{\mathrm{T}} + \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

Here $\mathbf{x}$ is the mean of the variables, $\mathbf{T}$ the scores, $\mathbf{P}$ the loadings and $\mathbf{E}$ the residuals [8]. 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 is 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 $\mathbf{E}$ (Equation (1)) 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 DModX of an observation is higher than Dcrit, the observation is a moderate outlier [8].

2.4.1. *Cross-validation and eigenvalues*

To determine the appropriate number of components in the PCA model, an internal validation method called cross-validation (CV) is used. In cross-validation the data set is divided into a number of groups, and a reduced data set is formed by excluding one of the groups. For a starting value of $S = S_0$, where S is the number of components, a model is estimated on the basis of the reduced data set, predicted values are calculated for the excluded objects, and the sum of

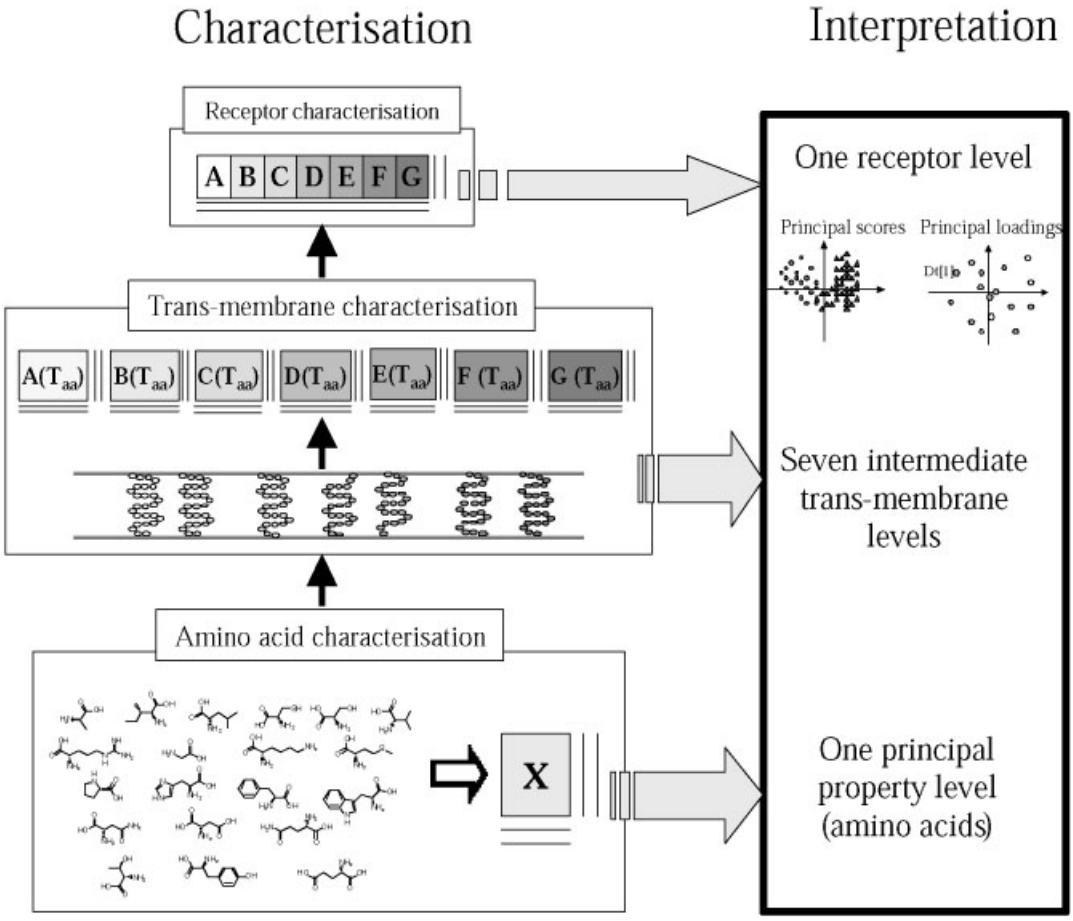

Figure 2. An overview of hierarchical PCA modelling for 7TM receptors.

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 is 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 [10].

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.

2.4.2. *Hierarchical PCA*

Hierarchical PCA modelling is a variant of PCA that is useful for data with many variables, where the results are often 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 these data is the hierarchical PCA model. Hierarchical models have to be interpreted in three stages. First, the loading plots of the hierarchical model reveal which of the blocks 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 they correspond to. Third, important amino acids can be identified [8,11,12]. An overview is given in Figure 2.

2.5. PLS

Partial least squares projections to latent structures (PLS) is a method used to find relationships between two matrices **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 with **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

$$\mathbf{X} = \mathbf{x}^{\mathrm{T}} + \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{2}$$

$$\mathbf{Y} = \mathbf{y}^{\mathrm{T}} + \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{3}$$

where **x** and **y** are the means of the variables, **T** and **U** the scores for **X** and **Y** respectively, **P** and **Q** the loadings and **E** and **F** the residuals.

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

$$\mathbf{u}_h = b_h \mathbf{t}_h \tag{4}$$

where b_h is a regression coefficient [13].

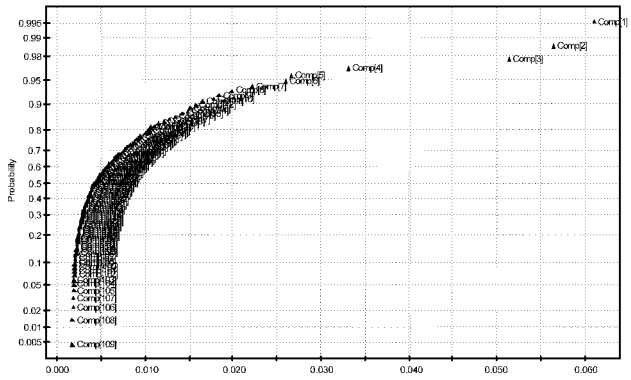

Figure 3. Normal probability plot of R2X for global PCA model with 109 components.

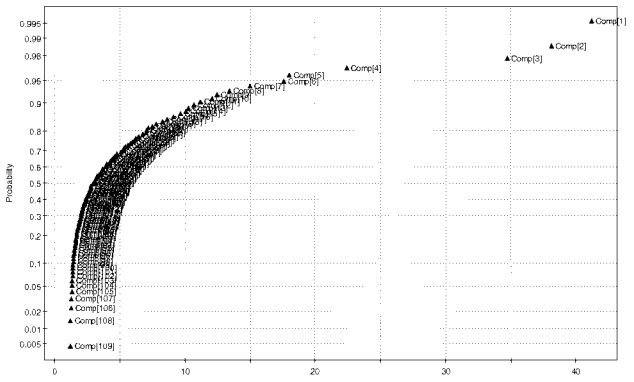

Figure 4. Normal probability plot of eigenvalues for global PCA model with 109 components.

2.5.1. PLS discriminant analysis

In PLS discriminant 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 [8].

2.5.2. 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 region. Components from these models are used as variables, and a PLS model is fitted to the data.

2.6. Software

The software used is Simca-P 8.0 and Simca-P+ (Umetrics AB, Umeå, Sweden; www.umetrics.com).

3. RESULTS AND DISCUSSION

3.1. 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 five zz-scales). The model obtained had 109 components according to cross-validation, explaining 85% of the variance. Seventy-seven components have an eigenvalue larger than two, explaining 79% of the variance. Normal probability plots for the explained variance and eigenvalues show that the first three components are clearly the most significant (Figures 3 and 4). 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 becomes blurred (Table I).

In the t1/t2 score plot for the global model the rhodopsin, amine and olfactory classes form separate clusters (Figure 5), and in the t3/t4 score plot the olfactory and hormone protein

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

Component	Classes described +	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

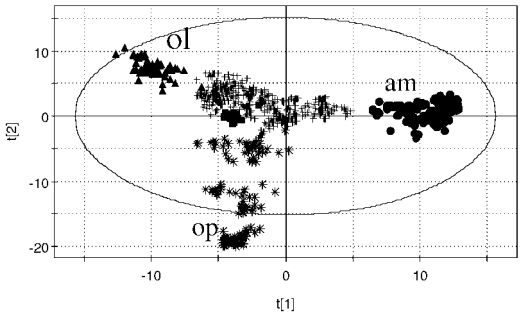

Figure 5. t1/t2 score plot for the global PCA model, showing the amine (am, circles), rhodopsin (op, asterisks) and olfactory (ol, triangles) classes to be well separated from the remaining classes (crosses, rectangles).

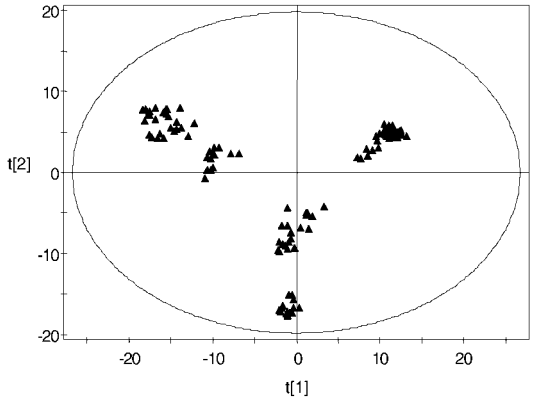

Figure 7. t1/t2 score plot for PCA model of the rhodopsin class, showing five well-separated clusters.

classes form separate clusters (Figure 6). The 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 all GPCRs are aligned towards one receptor in this class. In an attempt to further separate the central cluster, a global PLS discriminant 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 (result not shown).

3.1.1. *Reduced model*

Next, all well-separated clusters were removed from the work set, and a new model was fitted to the remaining data in the hope that this would help in separating the 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. For example, the peptide class forms several clusters, each representing a subclass of the peptide group.

3.1.2. *Local models*

A separate PCA model was also made for the rhodopsin class (131 sequences, four main classes), 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 (Figure 7). Each of the clusters represents one or more subclasses, and only one of the subclasses, a small subclass of only nine sequences, is split between two clusters. Higher-component score plots show the model to have several outliers, as does a plot of DModX (Figure 8).

The t1/t2 score plots for the peptide, nucleotide and rhodopsin PCA models show well-separated clusters, each representing one or more subclasses. 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 subclasses in the cluster. Similarly, a model can be fitted to one of the subclasses for a separation between different receptor types in the subclass, 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 can be extracted (Figures 9–12). Figure 9 is a t1/t2 score plot for a PCA model of the peptide class, and the encircled cluster contains sequences from the peptide subclasses bm, ny and tk. These three classes contain 48 sequences described by 590

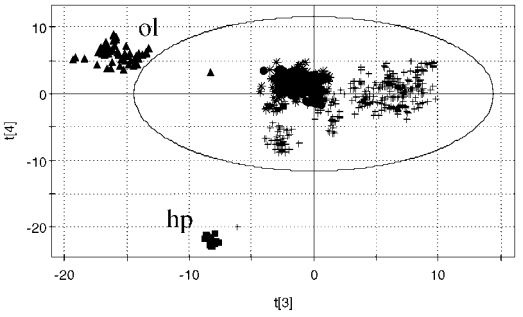

Figure 6. t3/t4 score plot for the global PCA model, showing the olfactory (ol, triangles) and hormone protein (hp, rectangles) classes to be well separated. The data point near the hp cluster belongs to the orphan class.

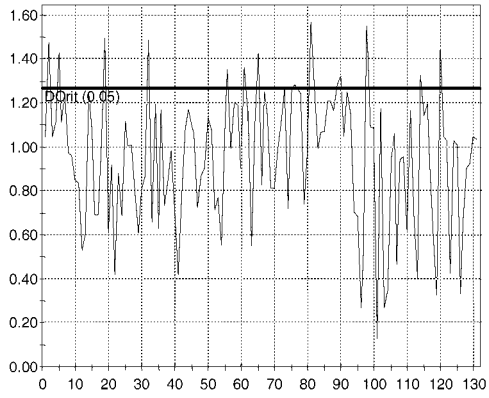

Figure 8. DModX plot for PCA model of the rhodopsin class, showing only moderate outliers.

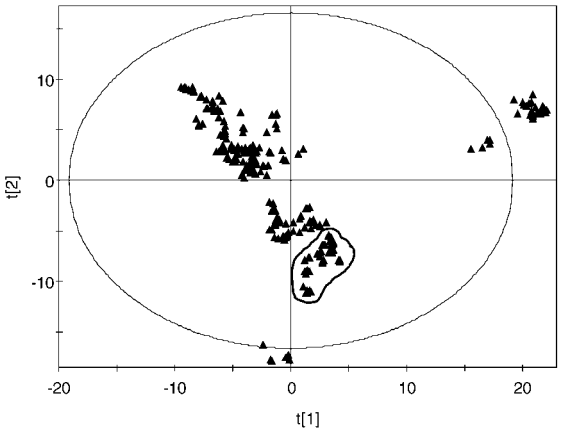

Figure 9. Score plot for PCA model of the peptide class. Encircled sequences (subclasses bm, ny and tk) are modelled separately.

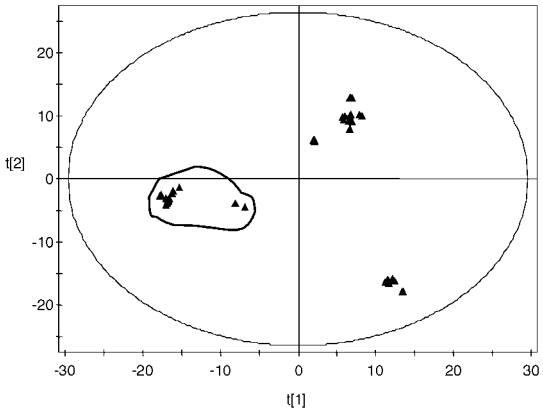

Figure 10. Score plot for PCA model of the sequences encircled in Figure 9. The three subclasses form well-separated clusters. Encircled in this plot is subclass tk.

variables*, and a PCA model based on these data has 17 components explaining 96% of the variance (CV). In a t1/t2 score plot for this model the three subclasses are well separated from each other (Figure 10). The encircled sequences belong to the subclass tk, a subclass with 16 sequences described by 450 variables. A PCA model based on these data gives a model with six 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 tk1) in the subclass (Figure 11). The encircled sequences belong to the receptor type tk2, a receptor type with seven sequences described by 125 variables. A PCA model based on these data gives a model with three components explaining 69% of the variance (CV). A t1/t2 score plot for this PCA model shows two clusters containing sequences from the species human, rabbit and bovine, and rat, golden hamster and mouse, respectively (Figure 12), and one sequence separated from the others, guinea pig. Contribution plots show that these sequences differ only in a handful

* For models based on few sequences that are similar, a number of variables are usually excluded owing to small or zero variance.

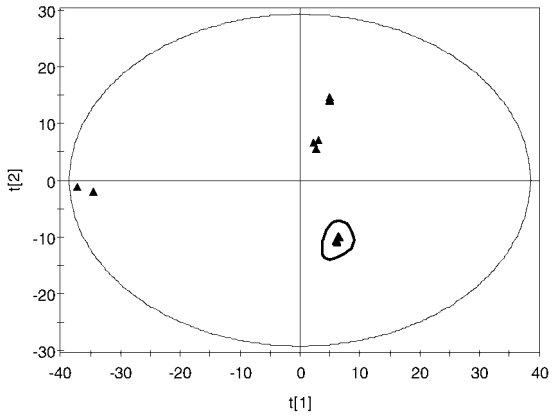

Figure 11. Score plot for PCA model of the peptide subclass tk. The four receptor types (tk1, tk2, tk3 and tkl) form well-separated clusters. Encircled is receptor type tk2.

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.

3.2. 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% (CV). The components from the separate models were then joined into a new data set for the hierarchical PCA model. The data set consisted of 897 sequences described by 230 variables, and the model had 77 components explaining 83% of the variance (CV). The score plot was similar to that for the global model, 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. 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 data set for this model consisted of 897 sequences described by 28 variables, and the model had five components

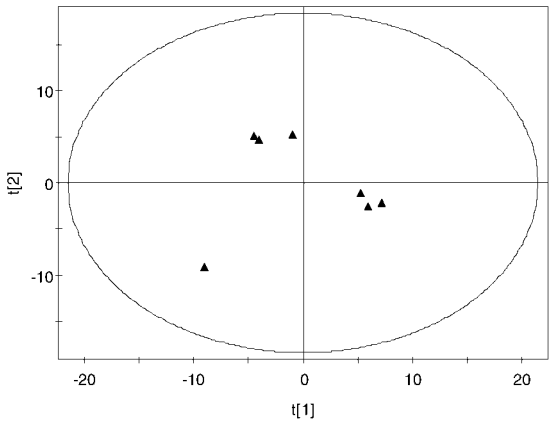

Figure 12. Score plot for PCA model of the receptor type tk2.

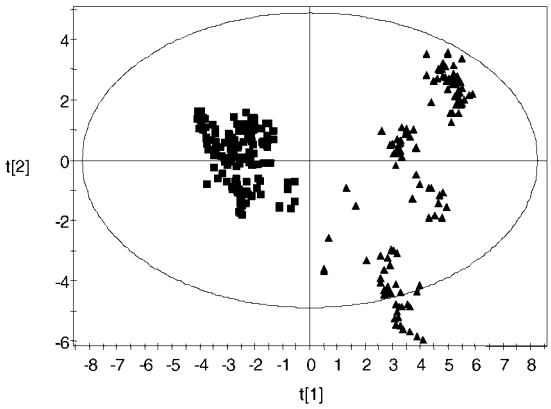

Figure 13. t1/t2 score plot for hierarchical PCA model for the rhodopsin (triangles) and amine (rectangles) classes, based on four components from each TM region.

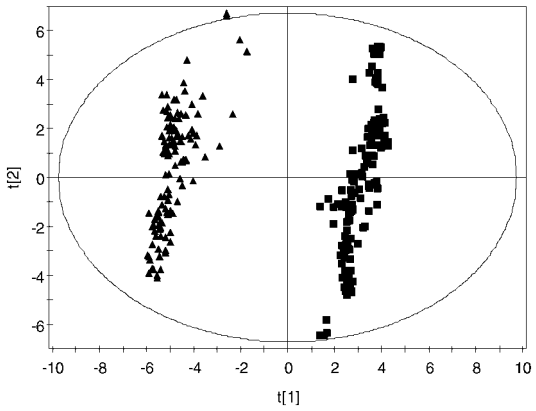

Figure 15. Score plot for hierarchical PLS-DA model for the amine (rectangles) and rhodopsin (triangles) classes, based on all components significant by cross-validation.

explaining 66% of the variance (CV). Again the score plot was very similar to that for the global model. The two classes with the best separation are the amine and rhodopsin classes, and the following investigation will therefore be focused on them.

3.3. Hierarchical model for amine and rhodopsin

A new hierarchical PCA model was made for the amine and rhodopsin classes only. The data set used consisted of 337 sequences described by 230 variables, and this gave a model with 64 components describing 93% of the variance (CV). The two classes are well separated and the loading plots show that the first seven components in the separate PCA models are the most important for the separation. 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. The reduced data set used consisted of 337 sequences described by 28 variables, and this gave a model with 10 components describing 90% of the variance (CV). According to a t1/t2 score plot, the two classes are well separated and three rhodopsin subclusters can be identified (Figure 13). The loading plot shows that all seven TM regions are equally important for the separation between the classes (Figure 14).

A hierarchical PLS discriminant analysis was also made for the amine and rhodopsin classes only (337 sequences described by 230 variables), resulting in a model with two components (CV) 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 (Figure 15). A coefficient plot for the PLS-DA model shows which of the variables in the hierarchical model are most important for the separation between the two classes (Table II). This confirms the information given in the loading plot for the PCA model, that the first seven components in the separate PCA models for the TM regions are the most significant.

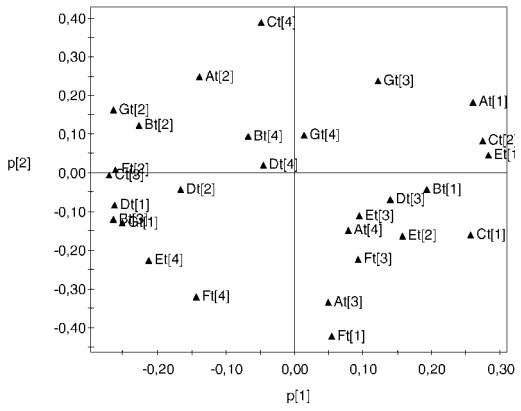

Figure 14. p1/p2 loading plot for hierarchical PCA model for the rhodopsin and amine classes, based on four components from each TM region.

3.4. Specific amino acids of interest

To find out which amino acids are conserved within the classes, separate PCA models were made for the amine and rhodopsin classes, for the three rhodopsin subclusters as seen in a score plot for the global PCA model (Figure 5), and for the amine subclass 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. For the amine class this is the acm and hh1 subclasses. In each of these models a

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

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

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

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 III.

The amino acids A15, E10, F14, F15, G6 and G16 are conserved in all three rhodopsin subclusters (Table III). 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 GPCRs in this compilation. 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 subclasses 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 GPCRs, and that E10, F15 and G6 are important for the function of GPCRs belonging to the rhodopsin class. When comparing the list of excluded variables in the amine model with that in 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 subgroup acm. This might seem strange, but could be explained by the small size of the acm subgroup. The acm subgroup makes up approximately 10% of the total amine class, and a variation in the acm subgroup might therefore be too small to be noticeable in the amine class as a whole.

Contribution plots were made for rhodopsin in the global model. One data point in each of the three rhodopsin clusters was compared with 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. Variables with contribution scores >0.20 and <-0.20 respectively can be seen in Table IV. 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.

Contribution plots for rhodopsin were also made in the hierarchical model. The same three rhodopsin data points were compared with 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 (Figure 16 and Table V). This list corresponds reasonably well with both excluded variables in the local models (Table III) and variables with high contribution scores in the global model (Table IV).

For the amine/acm subclass, 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 (Figure 13), and, according to the loading plots, is mainly

Table IV. 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

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

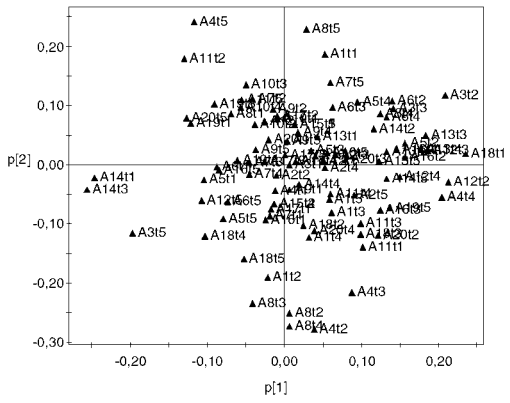

Figure 16. Loading plot for the PCA model for transmembrane region A.

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

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

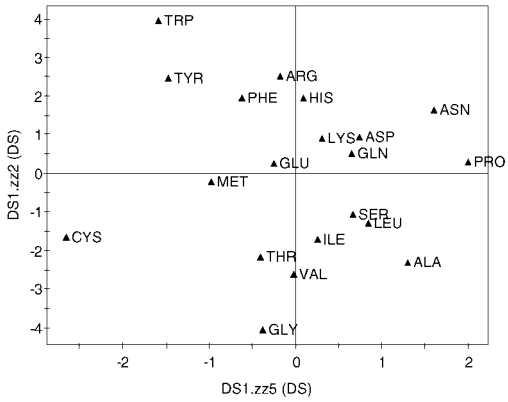

Figure 17. zz-Scales 2 and 5 plotted against each other.

Table VI. 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

Variable in hierarchical model	Original variables
At1	+ A3t2, A4t4, A12t2, A18t1 − A3t5, A14t1, A14t3
Bt2	+ B1t3, B1t5 − B13t3, B9t2
Bt3	+ B3t2, B3t3, B3t4 − B11t4
Dt1	+ D9t5, D13t5 − D5t2, D9t2, D9t3, D9t4, D13t3, D13t4
Gt1	+ G6t3, G8t2, G10t1, G10t3 − G11t2, G4t2
Gt2	+ G3t2, G3t4, G6t2 − G6t5

due to the hierarchical variables At1, Bt2, Bt3, Ct1, Ct2, Ct3, Dt1, Et1, Et4, Ft2, Gt1 and Gt2 (Figure 14). This corresponds well with the rhodopsin contribution plots, 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 subclass acm is separated from the rest of the amine class, a separation mainly given by component 3. According to the loading plot, the separation is mainly due to the hierarchical variables Bt4, Dt3, Dt4 and Gt4. 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 VI). 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 VI by two or more principal properties (e.g. 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 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 amino acid score plot (Figure 17) 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 Thr, Val, Leu and Ile (neutral, hydrophobic) in position A3, whereas the rhodopsin class mainly has Trp, Tyr and Phe (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 VII. The reason for the discrepancy in the prediction of important amino acids between the zz-scales and the sequences is most likely due to the presence of subgroups in both the amine and, particularly, the rhodopsin class.

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

Amino acid position, principal properties	Amino acids with appropriate properties for rhodopsin class based on zz-plots	Amino acids in rhodopsin class based on sequence	Amino acids with appropriate properties for amine class based on zz-plots	Amino acids in amine class based on sequence
A3 + t2 − t5	W, Y, F	Y, W, F	S, L, I, A	T, V, L, I
A14 − t1, t3	N, D, S	V, I, T, G	V, I, L	G
B1 + t3, t5	T, E, R	N	P, N	N, H
B3 + t2, t3, t4	L, V, I, T	I	W, H	F, Y, L
D13 + t5 − t2, t3, t4	W, Y, F	C, W	I, L	L, V, I
G3 + t2, t4	I, V, L, S, T, G	F, Y, C	R, W, H	W
G6 + t2 − t5	S, I, L, A	K	W, Y	Y, W

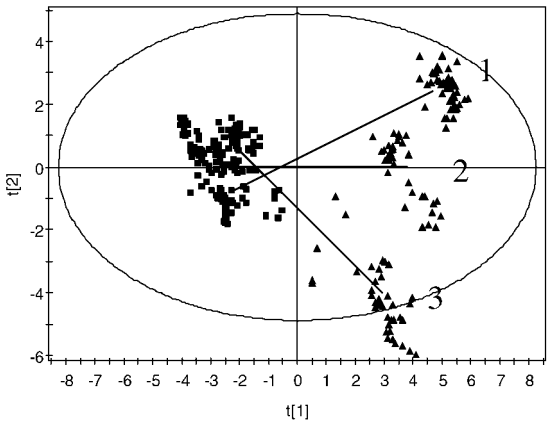

Figure 18. t1/t2 score plot for hierarchical PCA model for the rhodopsin (triangles) and amine (rectangles) classes, based on four components from each TM region. The rhodopsin clusters are numbered 1–3, and lines show the direction where the main separation between the amine class and each rhodopsin subclass lies.

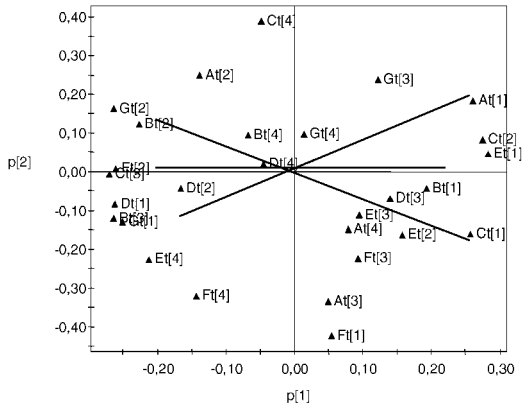

Figure 19. p1/p2 loading plot for hierarchical PCA model for the rhodopsin and amine classes, based on four components from each TM region. Lines correspond to those in the score plot and show which variables are important for the separation between the amine class and each rhodopsin subclass.

It can be seen that different variables are important for the separation of the amine class from the three rhodopsin subclusters: for cluster 1, At1, Bt3, Dt1, Et4 and Gt1; for cluster 2, Ct2, Ct3, Et1 and Ft2; and for cluster 3, Ct1, Bt2 and Gt2 (Figures 18 and 19).

4. CONCLUSIONS

In this study a multivariate approach has been used to compare and classify G protein-coupled receptor (GPCR) sequences based on the physicochemical properties of the TM regions. 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 data set 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. 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 subclusters can be obtained.

Since bovine rhodopsin is the only GPCR with a known 3D structure, all other GPCRs 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 probably not such a good model for the structure of receptors belonging to other classes.

The use of hierarchical models has given us the possibility to handle large amounts of sequence data and has helped identify the differences between classes or even subclasses of GPCRs. Using hierarchical models, it is also possible to identify which specific TM regions differ and what types of amino acids are involved.

In the future this will be used to investigate and identify the binding site in any type of 7TM receptor and might thus be of substantial help in the design of new ligands.

APPENDIX I. LIST OF ABBREVIATIONS

ACC	auto-crossed covariances
CV	cross-validation
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 discriminant analysis
SIMCA	soft independent modelling of class analogies
TM	transmembrane

APPENDIX II. THE AMINO ACIDS

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

REFERENCES

1. Graul RC and Sadée W. Evolutionary relationships among G protein-coupled receptors using a clustered database approach. *AAPS PharmSci* 2001; **3**(2): 612.
2. Available: http://www.nobel.se/medicine/laureates/1994/illpres/index.html. Access date 2001-09-06.
3. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M and Miyano M. Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* 2000; **289**: 739–745.
4. Sandberg M, Eriksson L, Jonsson J, Sjöström M and Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* 1998; **41**: 2481–2491.
5. Sandberg M. *Deciphering sequence data, a multivariate approach. PhD Thesis*, Umeå University, 1997.
6. Hellberg S, Sjöström M, Skagerberg B and Wold S. Peptide quantitative structure–activity relationships, a multivariate approach. *J. Med. Chem.* 1987; **30**: 1126–1135.
7. Sjöström M, Rännar S and Wieslander Å. Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances. *Chemometrics Intell. Lab. Syst.* 1995; **29**: 295–305.
8. Eriksson L, Johansson E, Kettaneh-Wold N and Wold S. *Multi- and Megavariate Data Analysis*. Umetrics AB, Umeå, Sweden, 2001.
9. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford University Press: Oxford, 1995.
10. Wold S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 1978; **20**(4) 397–405.
11. Wold S, Hellberg S, Lundstedt T, Sjöström M and Wold H. *PLS modelling with latent variables in two or more dimensions*. *The PLS Meeting*, Frankfurt, 1986.
12. Janne K, Pettersen J, Lindberg N-O and Lundstedt T. Hierarchical principal component analysis (PCA) and projection to latent structure (PLS) technique on spectroscopic data as a data pretreatment for calibration. *J. Chemometrics* 2001; **15**: 203–213.
13. Geladi P and Kowalski BR. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.