Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 73 (2004) 95–104
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Multivariate analysis of five GPCR receptor classes

Ingrid Long*, Per Andersson, Elisabeth Seifert, Torbjörn Lundstedt

Melacure Therapeutics AB, Ulleråkersvägen 38, SE-756 43, Uppsala, Sweden

Received 17 June 2003
Available online 10 March 2004

Abstract

The family of 7-transmembrane (TM) G-protein coupled receptors (GPCRs) represent the largest family of proteins in the human genome and are the target for many of the best selling drugs used today. Here a classification analysis of five GPCR receptor classes is made based on their physico-chemical properties defined by the principal properties of the amino acids in their sequences. The studied classes all belong to the Rhodopsin family, they are the muscarinic, adrenergic, dopamine and serotonin from the amine class and melanocortin from the peptide class. The whole sequences are studied, as well as the 7TM regions separately, and the results compared. The pattern of grouping for the different classes is studied. A few examples are shown and the results are discussed.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Multivariate analysis; 7TM regions; GPCR

1. Introduction

The 7-transmembrane 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 ligands. All GPCRs share a common structure with 7-transmembrane (TM) regions. In the present study, the seven TM regions are abbreviated A, B, C, D, E, F and G. The GPCRs are divided into five families; Rhodopsin-like, Secretin-like, Metabotropic glutamate, Fungal Pheromeone and cAMP receptors [1]. This study focuses on the Rhodopsin family. The Rhodopsin family of GPCRs is divided by function, i.e. type of ligand, into 16 classes, most of which are further divided into several sub-classes. The main classes are amine, peptide, hormone protein, Rhodopsin, olfactory, prostanoid, nucleotide like, cannabinoid, platelet activating factor, gonadotropin releasing hormone, thyrotropin releasing hormone, melatonin, Viral, Lysosphingolipid and LPA, Leukotriene B4 receptor, and orphan [1].

The GPCRs are an interesting group of receptors to study, since they are important as targets for drug discovery. They have been very successful as targets in the past; among the 100 top-selling drugs today, 25% are targeted at GPCRs, as are 50% of all recently launched drugs. There are several hundred GPCRs with known function, and only 30 are the targets of currently marketed drugs. In addition, there are hundreds of orphan receptors, whose ligand and function are as yet unknown [2].

In the present study, two main classes from the Rhodopsin family are represented, amine and peptide. The sub-classes of receptors included are muscarinic, adrenergic, dopamine and serotonin from the amine class and melanocortin from the peptide class. These are further divided into three levels of sub-class/type. A total of 23 muscarinic, 56 adrenergic, 42 dopamine, 67 serotonin and 32 melanocortin receptor sequences have been investigated. A complete list of the different classes included and the number of receptor sequences in each class is found in Appendix A, and a list of all receptor sequences included in Appendix B.

Using multivariate methods, a classification analysis of the receptor sequences is made. The classification is based either on the TM region only, or on the whole receptor sequence. The aims of the study are to investigate whether the classes and sub-classes of GPCRs, based on function, can be identified in models based on physico-chemical properties, and if there are differences in the results depending on whether the whole sequences or the 7TM regions are studied.

* Corresponding author. Tel.: +46-18-567218; fax: +46-81-567201.
E-mail address: ingrid.long@melacure.com (I. Long).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2003.12.008

2. Experimental

2.1. Sequence data

Both the 7TM regions and the complete sequences of the receptors have been studied. The analysis of the transmembrane sequences is alignment dependent, and depends also 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.

The amino acid sequences were quantitatively described using the five zz-scales described by Sandberg et al. [3].

The data consisting of the TM regions only are part of an in-house collection of GPCR sequences. The seven TM regions have a total of 135 amino acid positions, with five zz-scales, this results in 675 variables to describe each receptor. The TM regions are assigned according to Refs. [4,5].

Complete sequences for the selected receptors were downloaded from the Internet [1,6]. The number of amino acids in the sequences varies, therefore, Auto Cross Covariances were calculated to give the same number of variables for each receptor. A lag of five was used when calculating the ACC, on the basis that five amino acids correspond to about one and a half turns in a protein alpha-helix, which would be a suitable distance for interactions between two amino acids. Thus, 125 variables ($d^2*L = 5^2*5 = 125$) are used when considering the whole sequences.

2.2. zz-scales

The zz-scales describe each amino acid with numerical values, descriptors, which represent the physico-chemical properties of the amino acid. In this project, the descriptors used are the five principal properties described by Sandberg et al. [3]. Three z-scales for the 20 coded amino acids were described by Hellberg et al. [7] and have subsequently been extended by Jonsson et al. [8] and Sandberg et al. [3] 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 physio-chemical properties measured and calculated for each amino acid. A PCA 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 physico-chemical point of view [3]. In our study, they are, however, useful.

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 5*p variables in a so-called multipositional description [9].

2.3. 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 indices j and k are used for the zz-scales ($j = 1\ldots, 5$, $k = 1\ldots, 5$), index i is for the amino acid position ($i = 1\ldots, n$) and n is the number of amino acids in the sequence. The lag used can be varied, but the maximum lag is determined by the shortest sequence [10]. ACCs are calculated with lags $1\ldots, L$, and the resulting number of variables is d^2*L, where d is the number of descriptors and L the lag.

$$\mathrm{ACC}_{j,\mathrm{lag}} = \sum_{i}^{n-\mathrm{lag}} \frac{Z^*_{j,i} Z_{j,i+\mathrm{lag}}}{n - \mathrm{lag}} \tag{1}$$

$$\mathrm{ACC}_{j\neq k,\mathrm{lag}} = \sum_{i}^{n-\mathrm{lag}} \frac{Z^*_{j,i} Z_{k,i+\mathrm{lag}}}{n - \mathrm{lag}} \tag{2}$$

By calculating ACC, the information in sequences of different length is summarized in vectors of equal length [11]. ACC takes neighbouring effects, i.e. lack of independence between subsequent positions, into account [12].

2.4. Methods

The methods used are Principal Component Analysis (PCA) [13,14] and Partial Least Squares Projections to Latent Structures Discriminant Analysis (PLS-DA) [15,16], with the aim of identifying groupings among the receptor sequences based on physico-chemical properties. All variables were mean centred and scaled to unit variance. The number of significant components was determined using eigenvalues (that is, components should have an eigenvalue (ev)>2), unless otherwise stated.

2.5. Software

The software used is: Simca-P + 10.0 (Umetrics AB, Box 7960, SE-907 19, Umeå, Sweden, http://www.umetrics.com, [2000]), SPOC-SEQ.EXE, and SPOC-CRO.EXE (Michael Sjöström, Research Group for Chemometrics, Umeå University, SE-901 87 Umeå, Sweden).

3. Results and discussion

3.1. The melanocortin receptor sequences

There are 32 receptor sequences in this group. They are further divided into three subgroups, mcsh (MC_1R, 14 receptor sequences), mca (MC_2R, 5 receptor sequences) and mch (MC_3R, MC_4R and MC_5R, 13 receptor sequences). A PCA model based on the TM regions of these receptor sequences has four significant components based on ev. For this model, $R^2X = 0.65$ and $Q^2 = 0.40$. A $t1/t2$ score plot for this model shows a clear separation between the three sub-groups (Fig. 1). In the $p1/p2$ loading plot, the 675 variables evenly spread out, with no apparent groupings (result not shown). It would be very difficult to try and interpret this plot. A separate model made for the mch sub-group has two significant components according to ev. For this model, $R^2X = 0.53$ and $Q^2 = 0.36$. A $t1/t2$ score plot for this model shows a clear separation between the three types of receptor sequences in this sub-group (MC_3R, MC_4R and MC_5R) (Fig. 2).

Models based on the whole sequences of the melanocortin receptors show a similar pattern to those based on the 7TM regions only. A model based on all sequences has four significant components according to cross validation. For this model, $R^2X = 0.66$ and $Q^2 = 0.47$. A $t1/t2$ score plot for this model shows a clear separation between the three sub-groups (Fig. 3). However, one member of the mcsh group is found closer to the mca group. Again, in the $p1/p2$ loading plot, the variables are evenly spread out with no apparent groupings (result not shown). An additional difficulty with these variables is that they were generated using ACC, and hence in order to get an interpretation of which amino acids that are important, the individual ACC terms must be investigated separately [17]. A separate model made for the mch sub-group has two significant variables according to cv. For this model, $R^2X = 0.50$ and $Q^2 = 0.22$. A $t1/t2$ score plot for this model shows a clear separation between the three types of receptor sequences in this sub-group (MC_3R, MC_4R and MC_5R) (Fig. 4).

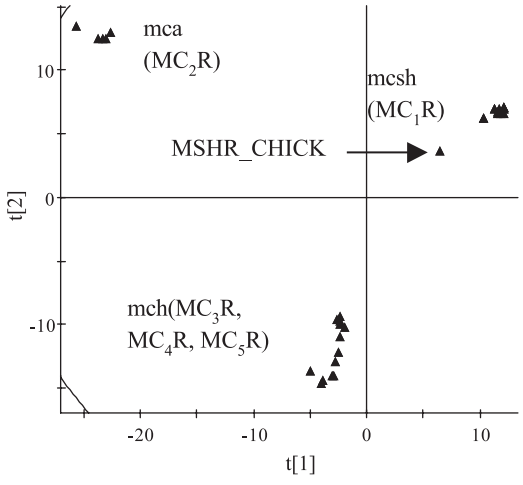

Fig. 1. $t1/t2$ score plot for 7TM model of the melanocortin receptor sequences. The three sub-groups form well-separated clusters.

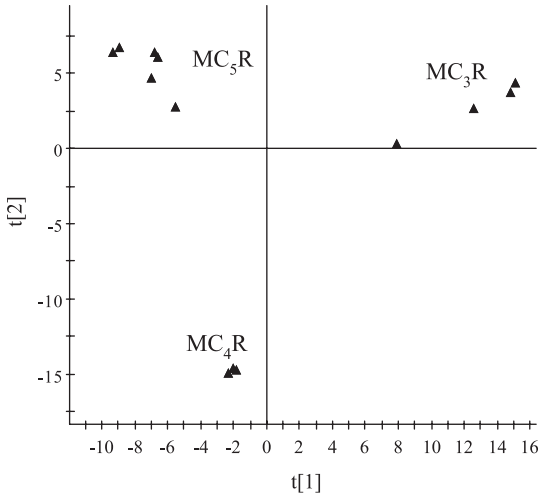

Fig. 2. $t1/t2$ score plot for 7TM model of the melanocortin sub-group mch. The three receptor types form well-separated clusters.

These results suggest that for the melanocortin receptor sequences, both the 7TM region and the whole sequence is well-conserved, since both the 7TM and the whole sequence models give similar and distinct groupings of the sub-groups in two levels.

PLS-DA models have also been calculated, for both 7TM and the whole sequence. The same pattern of groupings could be seen in the score plots as for the PCA model.

3.2. Muscarinic receptor sequences

There are 23 receptor sequences in this group. They are further divided into six sub-groups, acm1, acm2, acm3,

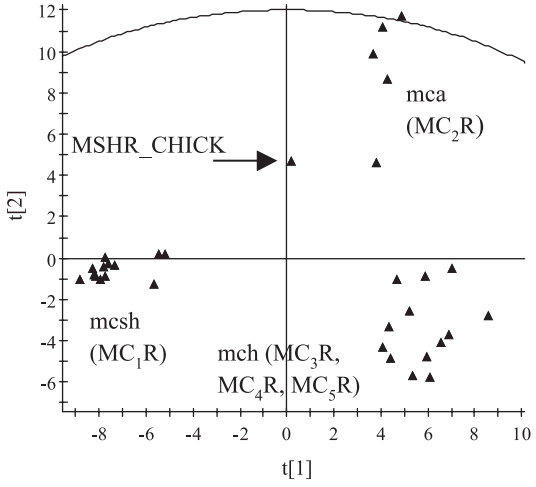

Fig. 3. $t1/t2$ score plot for whole sequence model of the melanocortin receptor sequences. The three sub-groups form well-separated clusters. One of the receptors in the mcsh group seems to be more similar to the mca sub-group.

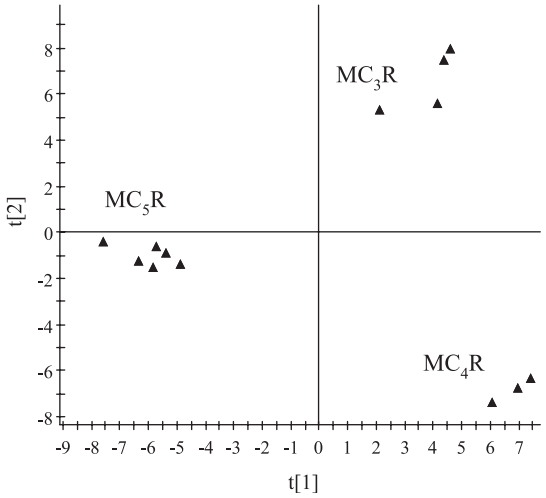

Fig. 4. *t*1/*t*2 score plot for whole sequence model of the melanocortin sub-group mch. The three receptor types form well-separated clusters.

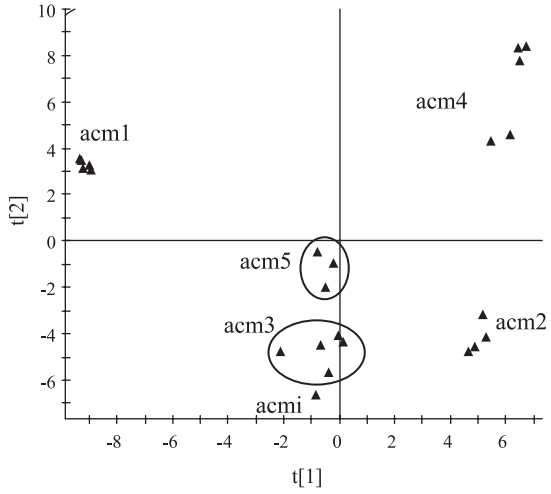

Fig. 6. *t*1/*t*2 score plot for whole sequence model of the muscarinic receptor sequences. The six sub-groups form well-separated clusters.

acm4, acm5 and acmi. A PCA model based on the TM regions of these receptors has four significant components based on ev. For this model, $R^2X = 0.81$ and $Q^2 = 0.61$. A *t*1/*t*2 score plot shows three clear groups, with acm1, acm3 and acm5 in one, acm2 and acm4 in one, and acmi on its own. This sub-class, consisting of only one observation, is an outlier to the model in the score space (Fig. 5). Higher component score plots show a good separation for the sub-groups acm1, acm2, acm3 and acm5, while sub-groups acm4 and acmi overlap (result not shown).

Models based on the whole sequences show a slightly different pattern. A model based on all sequences has four significant components according to ev. For this model, $R^2X = 0.75$ and $Q^2 = 0.51$. A *t*1/*t*2 score plot for this model shows that the sub-groups acm1, acm2 and acm4 are well separated. The sub-groups amc3, acm5 and acmi are grouped closer together, but do not overlap (Fig. 6).

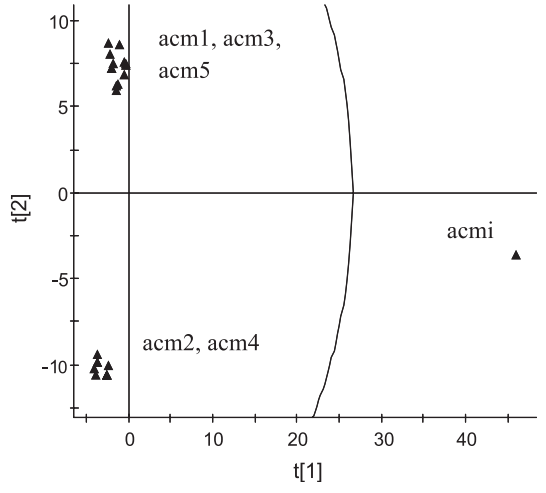

Fig. 5. *t*1/*t*2 score plot for 7TM model of the muscarinic receptor sequences. The six sub-groups form three well-separated clusters.

Thus, for the muscarinic receptors, the pattern of groupings is not the same when only the 7TM regions are studied as when the whole sequences are considered. In a PCA model based on the TM regions, sub-classes acm1, acm3 and acm5, for example, have very similar scores, while the one sequence in sub-class acmi is an outlier (Fig. 5). The explanation for this might be that this is the only non-vertebrae receptor included, it is of species *Drosophila melanogaster* (fruit fly). When the whole sequence is considered, this receptor is more similar to the others. In a PCA model based on the whole sequences, acm3, acm5 and acmi have similar scores, and acm1 form a distinct cluster on its own (Fig. 6).

For the muscarinic receptor sequences, there are only three groups in the score plot for the model based on the 7TM regions. This suggests that the 7TM regions are very well conserved in this group of receptors. The whole sequences are less well conserved, but are well conserved within the sub-groups, since they form separate groups in the score plot.

PLS-DA models have also been calculated, for both 7TM and the whole sequence. The same pattern of groupings could be seen in the score plots as for the PCA model.

3.3. Serotonin receptor sequences

There are 67 receptor sequences in this group, divided into eight sub-groups, sh1, sh2, sh4, sh5, sh6, sh7, shi, sho. The sub-group shi is further divided into two groups, shi1 and shi2. A PCA model based on the TM regions of these receptor sequences has nine significant components by ev, for this model $R^2X = 0.78$ and $Q^2 = 0.59$. A *t*1/*t*2 score plot shows several distinct groupings. Sub-groups sh6 and sh4 have similar scores, but there is a separation between the groups. Sh2 is well separated from the other groups. Sh1, sh5, sh7, shi and sho all have similar scores,

but sh1 is separated from the other sub-groups. Two observations belonging to the sho group form a group of their own in the score plot (Fig. 7). Higher dimension score plots show a good separation for the sh4, sh5, sh6, and sh7 sub-groups, and the shi sub-group together with two observations from the sho sub-group (result not shown). A separate model is made for the overlapping sub-groups sh5, sh7, shi and sho, to see if it would be possible to separate them. The dataset has 21 observations, a model fitted to this data has three significant components according to ev, giving a model with $R^2X = 0.66$ and $Q^2 = 0.43$. A $t1/t2$ score plot shows that sh5 and sh7 form well-separated groups in the score space, while shi and sho still overlap. Also, the same two observations belonging to the sho group that formed a separate group in the score space in the previous model do so here as well (Fig. 8).

Models based on the whole sequence show a very different pattern. The model is based on 64 sequences, and has 11 significant components according to ev, giving a model with $R^2X = 0.75$ and $Q^2 = 0.32$. A $t1/t2$ score plot for this model shows a lot less structure in the data compared to the 7TM model. Two rather large and spread out clusters can be identified (Fig. 9). The smaller group contains data from sub-groups sh2, sh4, sh7 and sho, the larger group contains data from sub-groups sh1, sh5, sh6, shi and sho.

Thus, for the serotonin receptor sequences, the structure of data as seen in the score plot is not the same when looking at the whole sequence as when only the 7TM regions are considered. There are fewer distinct groupings, and different sub-groups appear close to each other in the score plots. For example, sh4 and sh6 have very similar scores in a model based on TM regions only, but are found in different groups in the score plot for a model based on the whole sequences.

This suggests that the 7TM regions are very well conserved in this group of receptors, and within the sub-groups. The whole sequences are less well conserved, as they are very spread out and with no clear groupings.

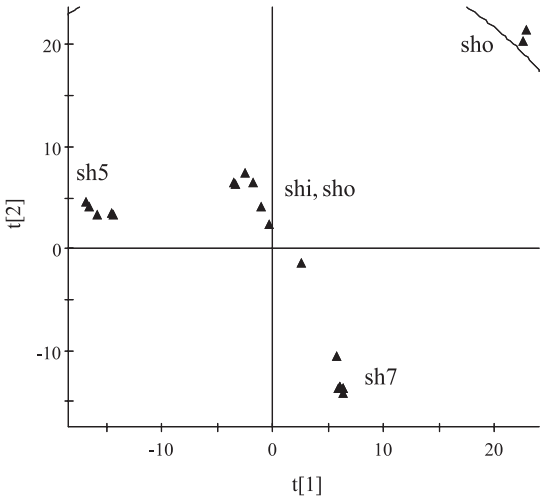

Fig. 8. $t1/t2$ score plot for 7TM model of the serotonin subgroups sh5, sh7, shi and sho. The sub-groups shi and sho still overlap.

PLS-DA models have also been calculated, for both 7TM and the whole sequence. For 7TM sequences the same pattern of groupings could be seen in the score plots as for the PCA model, but for the whole sequence it looks slightly different. There is a better, though not perfect, separation between the sub-classes in the PLS-DA score plots. However, the same level of separation can be achieved with PCA if local models are made for the two groups that can be identified in Fig. 9.

It is interesting to note that the 5HT2c sub-group of serotonin receptors, part of the sh2 sub-group, previously called 5HT1c, part of sh1, which was shown by Julius et al. [18] to belong to the 5HT2 (sh2) sub-group does indeed group together with the other 5HT2 receptors in the sh2 sub-

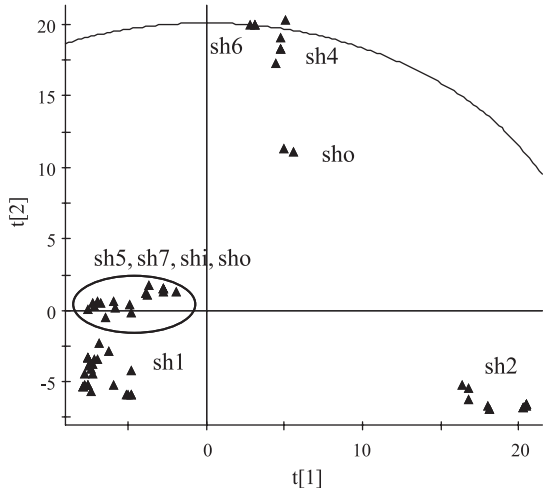

Fig. 7. $t1/t2$ score plot for 7TM model of the serotonin receptor sequences. The sub-groups sh5, sh7, shi and sho overlap.

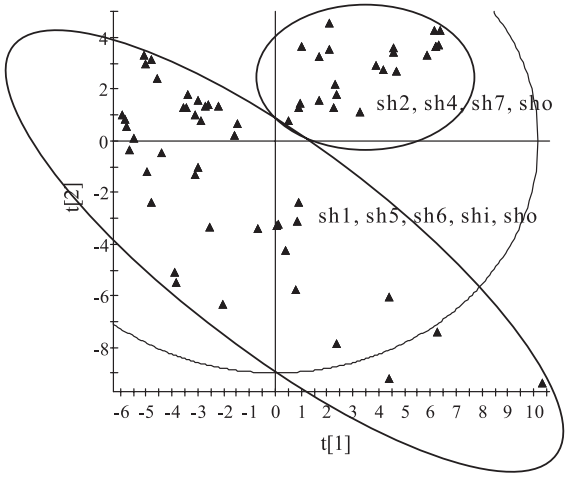

Fig. 9. $t1/t2$ score plot for whole sequence model of the serotonin receptor sequences. Two groups can be identified. In one sub-groups sh1, sh5, sh6, shi and sho is found, and in the other sub-groups sh2, sh4, sh7 and sho.

group, rather than with the sh1 sub-group. This could be seen for both 7TM and whole sequence analysis.

3.4. Adrenergic receptor sequences

There are 56 receptor sequences sequences in this group, divided into two sub-groups, adra (alpha) and adrb (beta). These are further divided into subgroups; adra1 and adra2, and adrb1, adrb2, adrb3 and adrb4, respectively. A model based on the 7TM data for these receptor sequences has seven significant components according to ev, for this model $R^2X=0.82$ and $Q^2=0.75$. A $t1/t2$ score plot shows three distinct groups of data. The sub-groups adra1 and adra2 forms two separate groups in the score plot, whereas the four sub-groups of adrb all form one group in the score plot (Fig. 10). A model based on the adrb group only (23 sequences) has three significant components according to ev, giving a model with $R^2X=0.71$ and $Q^2=0.45$. A $t1/t2$ score plot for this model shows four groups of data, the four sub-groups each form a separate group in the score plot (Fig. 11).

A model based on the whole sequences (54 observations) has eight significant components according to ev, giving a model with $R^2X=0.77$ and $Q^2=0.58$. A $t1/t2$ score plot for this model shows less structure in the data as compared to the model based on the 7TM regions (Fig. 12). The data points are more spread out, but there is a fairly good separation between the alpha and beta receptors. The sub-groups adra1 and adra2 overlap extensively. The beta sub-group adrb2 forms a distinct group, adrb1 and adrb3 form well-separated but spread out groups. A model based on the adra group only (33 sequences) has five significant components according to ev, giving a model with $R^2X=0.75$ and $Q^2=0.59$. A $t1/t2$ score plot for this model shows a clear separation between the two adra sub-groups, adra1 and adra2 (Fig. 13).

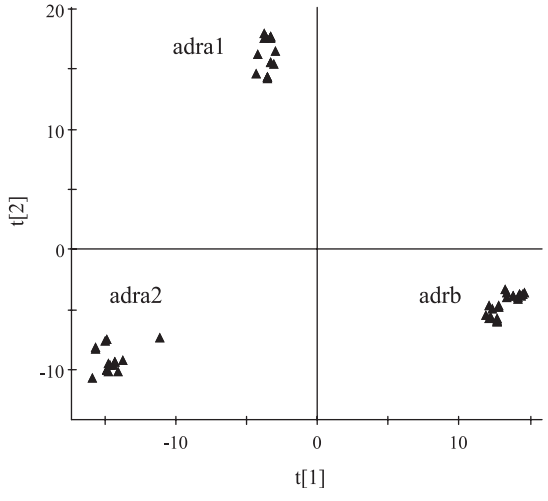

Fig. 10. $t1/t2$ score plot for 7TM model of the adrenergic receptor sequences. The sub-groups adra1, adra2 and adrb form well-separated clusters.

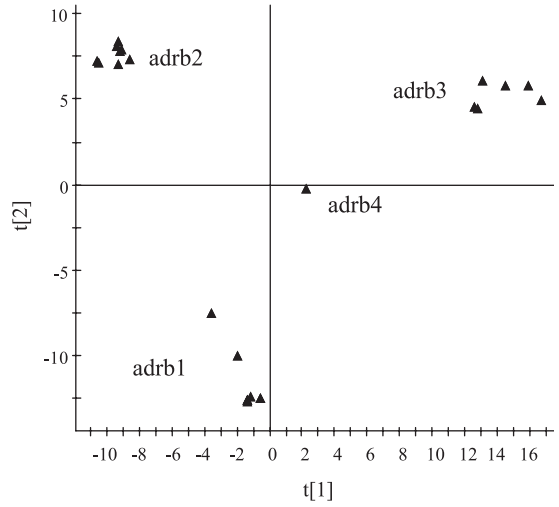

Fig. 11. $t1/t2$ score plot for 7TM model of the adrenergic sub-group adrb. The sub-groups adrb1, adrb2, adrb3 and adrb4 form well-separated clusters.

These results suggests that the 7TM regions are very well conserved for the adrenergic receptors, and within the sub-groups. The whole sequences are less well conserved within the sub-groups.

PLS-DA models have also been calculated, for both 7TM and the whole sequence. For 7TM sequences, the same pattern of groupings could be seen in the score plots as for the PCA model, but for the whole sequence, it looks slightly different. The PLS-DA score plots, as opposed to the PCA score plots, show a clear separation of the sub-group adra1 from adra2, however, the same level of separation can be

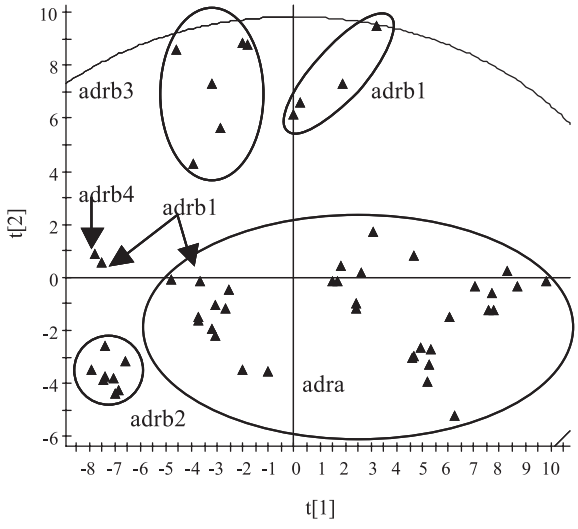

Fig. 12. $t1/t2$ score plot for whole sequence model of the adrenergic receptor sequences. Sub-groups adra and adrb are, with one exception, indicated by an arrow, well separated. The sub-groups adrb2 and adrb3 form well-separated groups. A few of the adrb1 receptors also form a group. Sub-classes adra1 and adra2 overlap completely.

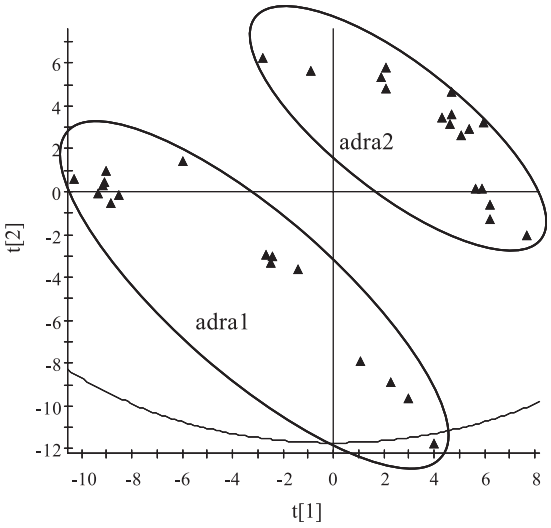

Fig. 13. $t1/t2$ score plot for whole sequence model of the adrenergic sub-group adra. The two sub-groups adra1 and adra2 are well separated.

achieved with PCA if a local models is made for the adra sub-group (Fig. 13).

3.5. Dopamine receptor sequences

There are 42 receptor sequences in this group, further divided into six subgroups, dop1, dop2, dop3, dop4, dopi and dopo. A PCA model based on the 7TM regions of these receptor sequences has six significant components according to ev, for this model $R^2X = 0.81$ and $Q^2 = 0.52$. A $t1/t2$ score plot shows sub-groups 1–4 to be well separated, whereas sub-groups dopi and dopo are close, but do not overlap (Fig. 14).

A model based on the whole sequences (40 sequences) has five significant components according to ev, giving a

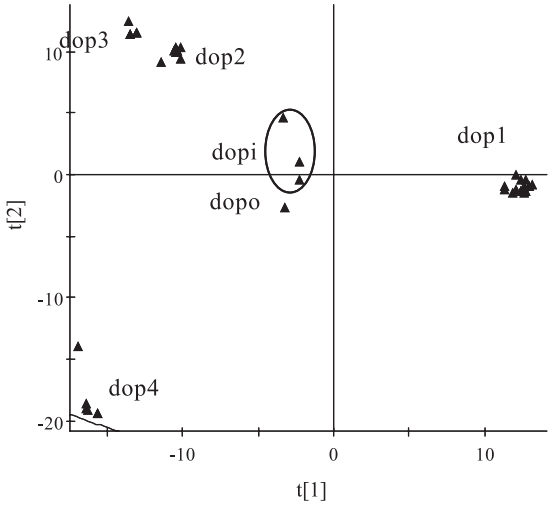

Fig. 14. $t1/t2$ score plot for 7TM model of the dopamine receptor sequences. The sub-groups dop1, dop2, dop3 and dop4 form well-separated clusters. Sub-groups dopi and dopo are close but do not overlap.

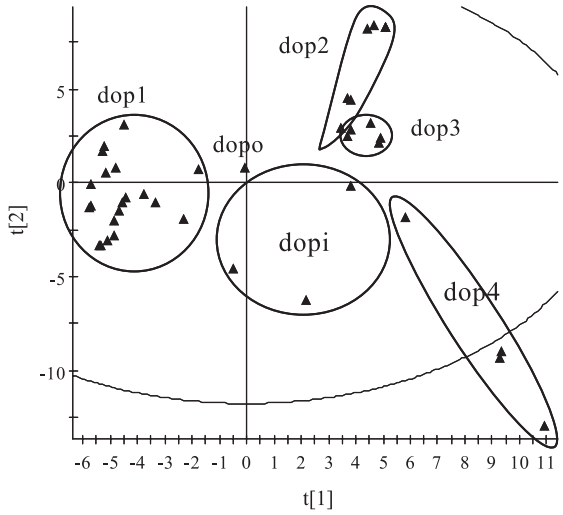

Fig. 15. $t1/t2$ score plot for whole sequence model of the dopamine receptor sequences. The sub-groups dop1, dop2, dop3 and dop4 form clusters that are more spread out than in the 7TM model.

model with $R^2X = 0.62$ and $Q^2 = 0.33$. A $t1/t2$ score plot for this model shows similar structure in the data as compared to the model based on the 7TM regions (Fig. 15). The data points are more spread out, but there is a separation between the different sub-groups. As in the score plot for the model based on 7TM data, subgroups dop2 and dop3 have similar scores, as do sub-groups dopi and dopo.

This suggests that the dopamine sequences are well conserved, within the group and the sub-groups, both in the 7TM regions and the whole sequences.

PLS-DA models have also been calculated, for both 7TM and the whole sequence. The same pattern of groupings could be seen in the score plots as for the PCA model.

4. Conclusions

The loading plots are not very informative, and have not been investigated in this study. In the models based on 7TM regions, there is a large number of variables, making the score plots very crowded and difficult to interpret. One way to facilitate the interpretation of the score plots would be to make hierarchical models [19]. In the models based on the whole sequences, there are fewer variables, but the interpretation of each variable is more difficult due to the ACC origin.

The ACC variables are useful in this study, since they make it possible to compare sequences of varying lengths. The disadvantage when using this type of variable is the poor interpretability. In the present case, interpretation would be especially difficult, as each sequence is several hundred amino acids long and each ACC variable thus composed of a very large number of cross terms.

Using multivariate methods, and working with small groups of receptor sequences, it has been possible to separate sub-groups. Several levels of sub-groups can be identified by making the models increasingly local. Both 7TM regions and whole sequences have been studied, and some differences in the classification can be noted between the two approaches. For the melanocortin and dopamine receptor sequences, the groupings in the score plot are similar for the 7TM and whole sequence models. This suggests that for those classes of receptors, the whole sequences are well conserved within the class and within the sub-class. For the serotonin, adrenergic and muscarinic receptor sequences, the groupings in the score plots are not the same for the 7TM and whole sequence models, which suggests that the 7TM regions are better conserved than the whole sequences. In future work, it would be interesting to study the influence of the intracellular and extracellular loops on the groupings.

Based on these results, it would appear as sho in the serotonin group could be divided into two groups, or alternatively that two of the sho receptors actually belong to the shi sub-group (Fig. 8).

Appendix A. Number of receptors in each group

Main class	Sub-class 1	Number of sequences	Type	Number of sequences	Sub-class 2	Number of sequences
Peptide-pe	Melanocortin-mc	32	Mcsh	14		
			Mca	5		
			Mch	13		
Amine-am	Muscarinic-acm	23	acm1	5		
			acm2	4		
			acm3	5		
			acm4	5		
			acm5	3		
			Acmi	1		
	Adrenergic-adr	56	Adra	33	adra1	16
					adra2	17
			Adrb	23	adrb1	8
					adrb2	8
					adrb3	6
					adrb4	1
	Dopamine-dop	42	dop1	20		
			dop2	7		
			dop3	5		
			dop4	6		
			Dopi	3		
			Dopo	1		
	Serotonin-sh	67	sh1	25		
			sh2	13		
			sh4	5		
			sh5	6		
			sh6	3		
			sh7	6		
			Shi	5	shi1	3
					shi2	2
			Sho	4		

Appendix B. List of receptors included

Names in bold are not included in the analysis of the whole sequences.

Main class	Sub-class 1	Type	Sub-class 2	Receptor name
am	acm	Acm1	–	ACM1_HUMAN
am	acm	Acm1	–	ACM1_RAT
am	acm	Acm1	–	ACM1_MACMU
am	acm	Acm1	–	ACM1_PIG
am	acm	Acm1	–	ACM1_MOUSE
am	acm	Acm2	–	ACM2_HUMAN
am	acm	Acm2	–	ACM2_PIG
am	acm	Acm2	–	ACM2_RAT
am	acm	Acm2	–	ACM2_CHICK
am	acm	Acm3	–	ACM3_PIG
am	acm	Acm3	–	ACM3_HUMAN
am	acm	Acm3	–	ACM3_BOVIN
am	acm	Acm3	–	ACM3_RAT
am	acm	Acm3	–	ACM3_CHICK
am	acm	Acm4	–	ACM4_RAT
am	acm	Acm4	–	ACM4_MOUSE
am	acm	Acm4	–	ACM4_HUMAN
am	acm	Acm5	–	ACM5_HUMAN
am	acm	Acm5	–	ACM5_RAT
am	acm	Acm5	–	ACM5_MACMU
am	acm	Acmi	–	ACM1_DROME
am	acm	Acm4	–	ACM4_CHICK
am	acm	Acm4	–	ACM4_XENLA
am	adr	Adra	adra1	A1AD_HUMAN
am	adr	Adra	adra1	A1AB_MESAU
am	adr	Adra	adra1	A1AB_RAT
am	adr	Adra	adra1	A1AD_RAT
am	adr	Adra	adra1	A1AB_HUMAN
am	adr	Adra	adra1	A1AB_MOUSE
am	adr	Adra	adra1	A1AD_MOUSE
am	adr	Adra	adra1	A1AA_HUMAN
am	adr	Adra	adra1	A1AA_RAT
am	adr	Adra	adra1	A1AA_RABIT
am	adr	Adra	adra1	A1AA_BOVIN
am	adr	Adra	adra1	O54913
am	adr	Adra	adra1	A1AD_RABIT
am	adr	Adra	adra1	Q13675
am	adr	Adra	adra1	O60451
am	adr	Adra	adra1	A1AA_ORYLA
am	adr	Adra	adra2	A2AC_RAT
am	adr	Adra	adra2	A2AD_HUMAN
am	adr	Adra	adra2	A2AC_MOUSE
am	adr	Adra	adra2	A2AC_HUMAN
am	adr	Adra	adra2	A2AC_CAVPO
am	adr	Adra	adra2	A2AA_CAVPO
am	adr	Adra	adra2	A2AA_MOUSE
am	adr	Adra	adra2	A2AA_RAT
am	adr	Adra	adra2	A2AA_PIG
am	adr	Adra	adra2	A2AA_HUMAN
am	adr	Adra	adra2	A2AC_DIDMA
am	adr	Adra	adra2	A2AB_CAVPO
am	adr	Adra	adra2	A2AB_MOUSE
am	adr	Adra	adra2	A2AB_HUMAN
am	adr	Adra	adra2	A2AB_RAT
am	adr	Adra	adra2	A2AR_LABOS
am	adr	Adra	adra2	A2AR_CARAU
am	**adr**	**Adrb**	**adrb1**	**B1AR_MACMU**

Appendix B (*continued*)

Main class	Sub-class 1	Type	Sub-class 2	Receptor name
am	**adr**	**Adrb**	**adrb1**	**B1AR_HUMAN**
am	adr	Adrb	adrb1	B1AR_MOUSE
am	adr	Adrb	adrb1	B1AR_RAT
am	adr	Adrb	adrb1	B1AR_PIG
am	adr	Adrb	adrb1	B1AR_CANFA
am	adr	Adrb	adrb1	B1AR_XENLA
am	adr	Adrb	adrb1	B1AR_MELGA
am	adr	Adrb	adrb2	B2AR_RAT
am	adr	Adrb	adrb2	B2AR_CANFA
am	adr	Adrb	adrb2	B2AR_MOUSE
am	adr	Adrb	adrb2	B2AR_MESAU
am	adr	Adrb	adrb2	B2AR_MACMU
am	adr	Adrb	adrb2	B2AR_BOVIN
am	adr	Adrb	adrb2	B2AR_HUMAN
am	adr	Adrb	adrb2	B2AR_PIG
am	adr	Adrb	adrb3	B3AR_MACMU
am	adr	Adrb	adrb3	B3AR_BOVIN
am	adr	Adrb	adrb3	B3AR_HUMAN
am	adr	Adrb	adrb3	B3AR_CANFA
am	adr	Adrb	adrb3	B3AR_MOUSE
am	adr	Adrb	adrb3	B3AR_RAT
am	adr	Adrb	adrb4	B4AR_MELGA
am	dop	Dop1	–	DADR_PIG
am	dop	Dop1	–	O77680
am	dop	Dop1	–	DADR_HUMAN
am	dop	Dop1	–	DADR_RAT
am	dop	Dop1	–	DADR_XENLA
am	dop	Dop1	–	DADR_DIDMA
am	dop	Dop1	–	Q98841
am	dop	Dop1	–	D1DR_FUGRU
am	dop	Dop1	–	O42315
am	dop	Dop1	–	DBDR_RAT
am	dop	Dop1	–	Q98842
am	dop	Dop1	–	DBDR_HUMAN
am	dop	Dop1	–	DBDR_XENLA
am	dop	Dop1	–	DCDR_XENLA
am	dop	Dop1	–	D5DR_FUGRU
am	dop	Dop1	–	D1DR_CARAU
am	dop	Dop1	–	Q98844
am	dop	Dop1	–	Q98843
am	dop	Dop1	–	D1DR_OREMO
am	dop	Dop1	–	O42317
am	dop	Dop2	–	D2DR_BOVIN
am	dop	Dop2	–	D2DR_MOUSE
am	dop	Dop2	–	D2DR_HUMAN
am	dop	Dop2	–	D2DR_CERAE
am	dop	Dop2	–	D2D1_XENLA
am	dop	Dop2	–	D2DR_FUGRU
am	dop	Dop2	–	O73810
am	dop	Dop3	–	D3DR_CERAE
am	dop	Dop3	–	D3DR_HUMAN
am	dop	Dop3	–	D3DR_MOUSE
am	dop	Dop3	–	D3DR_RAT
am	dop	Dop3	–	Q13167
am	dop	Dop4	–	D4DR_MOUSE
am	**dop**	**Dop4**	**–**	**O35838**
am	**dop**	**Dop4**	**–**	**Q62610**
am	dop	Dop4	–	D4DR_RAT
am	dop	Dop4	–	D4DR_HUMAN
am	dop	Dop4	–	O42322
am	dop	Dopi	–	DOP1_DROME
am	dop	Dopi	–	DOP2_DROME
am	dop	Dopi	–	O44198

Appendix B (*continued*)

Main class	Sub-class 1	Type	Sub-class 2	Receptor name
am	dop	Dopo	–	O02146
am	sh	sh1	–	5H1B_MOUSE
am	sh	sh1	–	5H1B_SPAEH
am	sh	sh1	–	5H1B_RAT
am	sh	sh1	–	5H1B_HUMAN
am	sh	sh1	–	5H1B_CRIGR
am	sh	sh1	–	5H1B_RABIT
am	sh	sh1	–	5H1B_CAVPO
am	sh	sh1	–	5H1B_DIDMA
am	sh	sh1	–	5H1D_HUMAN
am	sh	sh1	–	5H1D_RABIT
am	sh	sh1	–	5H1D_MOUSE
am	sh	sh1	–	5H1D_RAT
am	sh	sh1	–	5H1D_CANFA
am	sh	sh1	–	5H1D_CAVPO
am	sh	sh1	–	5H1A_MOUSE
am	sh	sh1	–	5H1A_RAT
am	sh	sh1	–	5H1D_FUGRU
am	sh	sh1	–	5H1A_HUMAN
am	sh	sh1	–	5H1F_MOUSE
am	sh	sh1	–	5H1F_RAT
am	sh	sh1	–	5H1F_HUMAN
am	sh	sh1	–	5H1F_CAVPO
am	**sh**	**sh1**	**–**	**O42384**
am	sh	sh1	–	Q98998
am	sh	sh1	–	5H1E_HUMAN
am	sh	sh2	–	5H2A_HUMAN
am	sh	sh2	–	5H2A_RAT
am	sh	sh2	–	5H2A_MACMU
am	sh	sh2	–	5H2A_MOUSE
am	sh	sh2	–	5H2A_CRIGR
am	sh	sh2	–	5H2A_PIG
am	**sh**	**sh2**	**–**	**O42385**
am	sh	sh2	–	5H2C_RAT
am	sh	sh2	–	5H2C_MOUSE
am	sh	sh2	–	5H2C_HUMAN
am	sh	sh2	–	5H2B_HUMAN
am	sh	sh2	–	5H2B_RAT
am	sh	sh2	–	5H2B_MOUSE
am	sh	sh4	–	O89003
am	sh	sh4	–	5H4_MOUSE
am	sh	sh4	–	O70528
am	sh	sh4	–	O89034
am	sh	sh4	–	5H4_RAT
am	sh	sh5	–	5H5A_RAT
am	sh	sh5	–	5H7_XENLA
am	sh	sh5	–	5H5A_HUMAN
am	sh	sh5	–	5H5A_MOUSE
am	sh	sh5	–	5H5B_MOUSE
am	sh	sh5	–	5H5B_RAT
am	sh	sh6	–	5H6_HUMAN
am	sh	sh6	–	5H6_RAT
am	sh	sh6	–	Q63004
am	sh	sh7	–	5H7_HUMAN
am	sh	sh7	–	P78336
am	sh	sh7	–	5H7_RAT
am	sh	sh7	–	P97842
am	sh	sh7	–	5H7_MOUSE
am	sh	sh7	–	5H7_CAVPO
am	sh	Shi	shi1	5HT_HELVI
am	sh	Shi	shi1	5HT_BOMMO
am	sh	Shi	shi1	5HT1_DROME

(continued on next page)

Appendix B (*continued*)

Main class	Sub-class 1	Type	Sub-class 2	Receptor name
am	sh	Shi	shi2	5HTA_DROME
am	sh	Shi	shi2	5HTB_DROME
am	sh	Sho	–	5HT_LYMST
am	sh	Sho	–	O76267
am	sh	Sho	–	5HT1_APLCA
am	sh	Sho	–	5HT2_APLCA
pe	mc	Mcsh	–	MSHR_CAPCA
pe	mc	Mcsh	–	MSHR_DAMDA
pe	mc	Mcsh	–	MSHR_CEREL
pe	mc	Mcsh	–	MSHR_SHEEP
pe	mc	Mcsh	–	MSHR_RANTA
pe	mc	Mcsh	–	MSHR_CAPHI
pe	mc	Mcsh	–	MSHR_ALCAA
pe	mc	Mcsh	–	MSHR_OVIMO
pe	mc	Mcsh	–	MSHR_BOVIN
pe	mc	Mcsh	–	MSHR_VULVU
pe	mc	Mcsh	–	MSHR_HUMAN
pe	mc	Mcsh	–	O77616
pe	mc	Mcsh	–	MSHR_MOUSE
pe	mc	Mcsh	–	MSHR_CHICK
pe	mc	Mca	–	ACTR_HUMAN
pe	mc	Mca	–	ACTR_MOUSE
pe	mc	Mca	–	ACTR_MESAU
pe	mc	Mca	–	ACTR_BOVIN
pe	mc	Mca	–	O57317
pe	mc	Mch	–	MC5R_HUMAN
pe	mc	Mch	–	MC5R_SHEEP
pe	mc	Mch	–	MC5R_RAT
pe	mc	Mch	–	MC5R_MOUSE
pe	mc	Mch	–	MC5R_BOVIN
pe	mc	Mch	–	O73671
pe	mc	Mch	–	MC4R_RAT
pe	mc	Mch	–	MC4R_HUMAN
pe	mc	Mch	–	MC3R_HUMAN
pe	mc	Mch	–	MC3R_MOUSE
pe	mc	Mch	–	O73667
pe	mc	Mch	–	MC3R_RAT
pe	mc	Mch	–	O93259

References

[1] GPCRDB: Information system for G protein-coupled receptors, http://www.gpcr.org, 2001901.

[2] T. Klabunde, G. Hesler, Drug design strategies for targeting G-protein-coupled receptors, ChemBioChem 3 (2002) 928–944.

[3] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, J. Med. Chem. 41 (14) (1998) 2481–2491.

[4] M. Lapinsh, A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, J. Wikberg, Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences, Protein Sci. 11 (2002) 795–805.

[5] J.M. Baldwin, G.F.X. Schertler, V.M. Unger, An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors, J. Mol. Biol. 272 (1997) 144–164.

[6] ExPASy: "Expert Protein Analysis System proteomics server of the Swiss Institute of Bioinformatics", http://www.expasy.org/srs5bin/cgi-bin/wgetz, 2001901.

[7] S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, Peptide quantitative structure–activity relationships, a multivariate approach, J. Med. Chem. 30 (1987) 1126–1135.

[8] J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström, S. Wold, Multivariate parametrization of 55 coded and non-coded amino acids, Quant. Sruct.-Act. Relat. Pharmacol. Chem. Biol. 8 (1989) 204–209.

[9] M. Sjöström, S. Rännar, Å. Wieslander, Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 29 (1995) 295–305.

[10] M. Edman, Detection of sequence patterns in membrane proteins, Department of Chemistry, Research Group of Chemometrics, Umeå University, Umeå, 2001, Thesis.

[11] M. Sandberg, Deciphering sequence data a multivariate approach, Department of Organic Chemistry, Umeå University, Umeå, 1997, Thesis.

[12] S. Wold, J. Jonsson, M. Sjöström, M. Sandberg, S. Rännar, DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures, Anal. Chim. Acta 277 (1993) 239–253.

[13] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst., Lab. Inf. Manag. 2 (1987) 37–52.

[14] J.E. Jackson, A Users Guide to Principal Components, Wiley, New York, 1991.

[15] M. Sjöström, S. Wold, B. Söderström, PLS discriminant plots, PARC in Practice, Elsevier, Amsterdam, 1986, pp. 461–470.

[16] L. Ståhle, S. Wold, Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo Study, J. Chemom. 1 (1987) 185–196.

[17] P.M. Andersson, M. Sjöström, T. Lundstedt, Preprocessing peptide sequences for multivariate sequence-property analysis, Chemometr. Intell. Lab. Syst. 42 (1998) 41–50.

[18] D. Julius, K.N. Huang, T.J. Livelli, R. Axel, T.M. Jessell, The 5HT2 receptor defines a family of structurally distinct but functionally conserved serotonin receptors, Proc. Natl. Acad. Sci. 87 (1990) 928–932.

[19] I. Gunnarsson, P. Andersson, J. Wikberg, T. Lundstedt, Multivariate analysis of G protein-coupled receptors, J. Chemom. 17 (2003) 82–92.