Multiway calibration in 3D QSAR
Nilsson, Jonas

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
1998

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 31-12-2018
Summary

There are many opinions about how medicinal chemistry should be practiced. The procedure described in this thesis for the design of new drugs comprises several steps including the selection of a lead compound, experimental design, syntheses of new compounds, pharmacological in vitro testing, molecular modeling and multivariate statistical evaluation. Each step has achieved more or less attention but the main emphasis has been put on the multivariate statistical evaluation.

In the pursuit of potent and selective dopamine D₃ antagonists the trans-N-(n-propyl)-7-[(trifluoromethyl)sulfonyl]oxy]-OHBfQ (1 in Table 3.3) was initially considered as a lead compound for this investigation. The compound displayed presynaptic DA receptor antagonistic properties in rats (see Table 1.4), although the 7-hydroxy analogue was a potent agonist. The racemic 1 showed a 10-fold selectivity for the DA D₃ over the D₂ receptor. The influence of the 7-triflate group on these effects was of special interest.

Several OHBfQs were designed and described with theoretical physicochemical descriptors using compound 1 as the template. Only a fraction of the most different compounds was selected by means of an experimental design in the descriptor space (see Chapter 3). Accordingly, the 16 compounds were synthesized and tested for in vitro affinities for the DA D₂, D₃ and D₄ receptor subtypes. None of the compounds were real selective for any of the receptors and, in general, compounds with a hydroxy group at the seven position displayed significant high affinities for all three dopamine receptors, while the compounds with a sulfon ester group were less potent. In addition, the sulfon ester group suppressed the affinity for the DA D₄ receptor. The nitrogen substituent may be as large as a phenylethyl group without detrimentally affecting the affinity for the DA receptors. Finally, a compound with a 7-OH group and an N-propargyl group lacks affinity for the DA D₄ receptor. The somewhat rigid N-propargyl group and the low pKₐ value (6.1) may be contributing factors to the low D₄ affinity.

At this point, a 3D QSAR model may provide information about how to proceed further. Which compound should be synthesized next? However, due to lack of time further investigation of this data set was not pursued. Instead, a data set was retrieved from Dr. Shelly Glase at Parke-Davis in the USA with which the multivariate statistical analyses were investigated. This theoretical part of the thesis is focused mainly on the optimization of multivariate and multiway regression analysis methods in 3D QSAR.

In Chapter 4, conformational analyses and alignments of mutual and potential interaction points with the DA D₃ receptor of Dr. Glase’s flexible compounds were carried out. The absolute configuration of the compound (1 in Table 4.1) used as the template to fit the remaining 29 compounds on, was determined with help from X-ray crystallographic structures. It is important to stress that the emphasis and aim of a 3D QSAR study is to measure the differences (e.g. steric and electrostatic fields) between the compounds under investigation and, consequently, the aligned compounds (Figure 5.1) do not necessarily fit into the active site of the receptor. This is in particular...
the case when the 3D structure of the target receptor is not known, which is the case for all dopamine receptors.

The Grid program was used to generate molecular fields from ten different probe atoms for all the aligned compounds. Accordingly, the three probes that generated the most different molecular fields were selected by means of a Principal Component Analysis (see Figure 4.2). The OH2, C3 and CA+2 probes were selected and reflect the hydrogen bonding, steric and electrostatic interactions, respectively, between the target receptor protein and the ligands. Each molecular field, in form of a 3D grid (see Figure 1.9), is unfolded to form a row vector with as many elements as there are grid points in the grid. The complete data set, X in Figure 4.4, comprises 30 molecules described by 25110 molecular descriptors and one dependent variable, y, i.e. the affinity for the DA D3 receptor subtype. The very essence of 3D QSAR modeling is to establish a regression model between X and y.

In Chapter 4, the GOLPE program was used for the variable selection, data pretreatment, and subsequently the PLS regression analysis. The necessity of eliminating grid points with more or less insignificant variation, e.g. grid points at large distances from the ligands placed in the periphery of the grid, was investigated. It turned out that the crossvalidated \( Q^2 \) increased from 0.45 to 0.65 when the number of variables was reduced from 19180 (after pretreatment) to 784. The variable reduction was carried out in two steps: first by means of D-optimal preselection in the loading space from an initial PLS model followed by a fractional factorial design selection procedure. The GOLPE analysis is based on the two-way PLS method and, as a consequence, the descriptor grids must be unfolded into a two-way matrix prior to the analysis. Actually, the raw data set looks like in Figure 5.2 where five different directions or modes are defined; one mode represents the 30 molecules, the x, y, and z modes correspond to the axes of the three dimensional grids and the fifth mode represents the three probes. For the analysis of a five-way data set the Multilinear PLS method (N-PLS) intuitively is a better choice than the two-way PLS method, since the unfolding procedure is unnecessary. Instead, the data set is directly decomposed into a score vector and four loading vectors (see Figure 5.3) corresponding to the five modes as defined above.

In Chapter 5, it was shown that the N-PLS models are easier to interpret due to the loading vectors obtained in each mode, and that they possess higher predictability than the PLS models. However, the fit was worse with N-PLS, as compared with PLS, possible due to the smaller number of parameters that needed to be estimated. PLS probably overfits and therefore N-PLS reflects better the relationship between X and y. At this point it was important to verify the excellent performance of the N-PLS method by analyzing yet another data set. In Chapter 6, a very well known data set, consisting of 58 benzamides (raclopride analogues) with affinity for the DA D2 receptor subtype, was re-analyzed using GRID descriptors and N-PLS as the regression method. The result was convincing and the final model had a predicted \( Q^2 \) of 0.62. It was concluded that the N-PLS method certainly is an alternative to PLS for the analysis of 3D QSAR data sets.

Both PLS and N-PLS are so component wise regression methods since each component is calculated from the residuals of X, after removing the variation accounted for by the previous
component. In Chapter 7, two methods that calculate all components simultaneously in an alternating least squares algorithm have been evaluated. The two methods, PCovR/PARAFAC and PCovR/Tucker, are combinations of the Principal Covariate Regression (PCovR) method with PARAFAC and Tucker decompositions, respectively. These algorithms balance between reconstructing $X$ and explaining $y$ by adjusting the $\alpha$ value (see Equation 7.4). Therefore, in order to find the models with optimal predictability the $\alpha$ value and the number of components were optimized simultaneously. When the sum of squares of $X$ and $y$ were normalized, the most predictive models had $\alpha$ values close to unity. Interestingly, the size of $\alpha$ that produced the most predictive models could be tuned with the ratio of the sum of squares of $X$ and $y$ ($s_r$). The best PCovR/Tucker5 model had a predictive $Q^2 = 0.31$ and was found when $\alpha = 0.4$, $s_r = 100$ using (6,7,3,4,2) number of components in the five modes. This model was considered stable, reproducible and it had a high predictability. The corresponding best PCovR/PARAFAC model was unstable, possibly since the data set has a Tucker structure. It was shown that the convergence of this latter algorithm strongly depended on the starting parameters used and frequently converged into local minima. This observation reduced significantly the reliability of the method, at least for the analysis of the present 3D QSAR data set. The question is how reliable is the N-PLS method since it too, is a PARAFAC derived method? It is clear, however, that the predictability of the N-PLS and the PCovR/Tucker5 models are comparable but N-PLS models are easier to interpret and provides simpler solutions.

Finally, two of the three multiway regression methods, evaluated in this thesis, definitely are alternatives to consider along with the traditional PLS method when 3D QSAR data sets need to be analyzed in the future.

In Chapter 8, three additional fields of research, where possible multiway data are generated, are proposed. Suggestions for the analysis of these types of data and a future prospective for multiway methods in medicinal chemistry are given.
Summary