Journal of Computer-Aided Molecular Design, 4 (1990) 381–389
ESCOM

J-CAMD 096

Experimental design based 3-D QSAR analysis of steroid–protein interactions: Application to human CBG complexes

Ulf Norinder

Astra Research Centre Södertälje, S-151 85 Södertälje, Sweden

Received 2 April 1990
Accepted 1 June 1990

Key words: PCA; PLS; Factorial design

SUMMARY

An experimental design based 3-D QSAR analysis using a combination of principal component and PLS analysis is presented and applied to human corticosteroid-binding globulin complexes. The predictive capability of the created model is good. The technique can also be used as guidance when selecting new compounds to be investigated.

INTRODUCTION

The use of quantitative structure-activity relationships (QSARs) has proven to be a powerful approach in drug design [1, 2]. Recently, two new and promising 3-D QSAR techniques have been presented by Ghose et al. [3] and by Cramer et al. [4]. An important factor in deriving a successful model (relationship) is the choice of initial compounds (the training set).

Experimental design based methods provide efficient means of varying the structures in an organized way, i.e. obtaining much information using a small number of observations (experiments) [5].

This paper describes the combined approach of a principal component analysis (PCA) based selection of the training set compounds followed by a PLS-supported 3-D QSAR analysis applied to some human corticosteroid-binding globulin (CBG) complexes (for another 3-D QSAR study of CBG complexes, see Ref. 4).

METHODS

Affinity constants

The affinity constants of human CBG complexes for compounds **1-37** (Table 1) were obtained from Mickelson et al. [6].

TABLE 1
NAMES, EXPERIMENTAL AND CALCULATED AFFINITY CONSTANTS OF COMPOUNDS 1–37

No.	Name	Affinity constants[a]	
		Exp.	Calc.[b]
1	11β,17,21-Trihydroxy-4-pregnene-3,20-dione	8.85	*8.16* 8.22
2	14α,17,21-Trihydroxy-4-pregnene-3,20-dione	6.85	*6.87* 6.74
3	11β,17,21-Trihydroxy-1,4-pregnadiene-3,20-dione	8.57	7.90 7.99
4	11β,17,21-Trihydroxy-2α-methyl-4-pregnene-3,20-dione	8.78	8.37 8.35
5	11β,17,21-Trihydroxy-2α-methyl-9α-fluoro-4-pregnene-3,20-dione	6.23	*6.59* 6.91
6	17,21-Dihydroxy-4-pregnene-3,11,20-trione	7.70	*8.11* 8.16
7	11α,21-Dihydroxy-4-pregnene-3,20-dione	8.15	8.33 8.86
8	11β,21-Dihydroxy-4-pregnene-3,20-dione	8.98	8.70 8.98
9	16α,17-Dihydroxy-4-pregnene-3,20-dione	6.85	7.35 6.55
10	17,21-Dihydroxy-4-pregnene-3,20-dione	8.81	8.54 8.30
11	11β,21-Dihydroxy-5β-pregnane-3,20-dione	7.70	*7.76* 7.86
12	2α-Hydroxy-4-pregnene-3,20-dione	8.43	8.62 8.12
13	6α-Hydroxy-4-pregnene-3,20-dione	7.15	7.37 7.38
14	6β-Hydroxy-4-pregnene-3,20-dione	6.49	7.21 7.22
15	11α-Hydroxy-4-pregnene-3,20-dione	8.00	7.66 8.15
16	16α-Hydroxy-4-pregnene-3,20-dione	7.00	7.50 7.19
17	17-Hydroxy-4-pregnene-3,20-dione	8.80	8.17 8.24
18	12α-Hydroxy-5β-pregnane-3,20-dione	6.00	*5.79* 6.18
19	21-Hydroxy-4-pregnene-3,20-dione	8.83	8.77 9.12
20	17-Hydroxy-6α-methyl-4-pregnene-3,20-dione	7.41	7.73 7.64
21	17-Hydroxy-16α-methyl-4-pregnene-3,20-dione	7.69	7.47 7.27
22	4-Pregnene-3,11,20-trione	7.57	7.94 8.17
23	4-Pregnene-3,20-dione	8.77	8.03 8.34

TABLE 1 (*cont.*)
NAMES, EXPERIMENTAL AND CALCULATED AFFINITY CONSTANTS OF COMPOUNDS 1–37

No.	Name	Affinity constants[a]	
		Exp.	Calc.[b]
24	5-Pregnene-3,20-dione	8.11	6.58 6.85
25	5β-Pregnane-3,20-dione	7.62	7.03 7.25
26	3β-Hydroxy-5-pregnen-20-one	5.70	*5.93* 6.32
27	3α-Hydroxy-5β-pregnan-20-one	6.36	*6.32* 6.42
28	2α-Methyl-4-pregnene-3,20-dione	8.53	*8.48* 8.62
29	6α-Methyl-4-pregnene-3,20-dione	7.85	*8.14* 8.33
30	16α-Methyl-4-pregnene-3,20-dione	8.04	*7.81* 7.85
31	19-Nor-4-pregnene-3,20-dione	7.70	7.42 7.50
32	9α-Fluoro-16α-methyl-11β,17,21-trihydroxy-1,4-pregnadiene-3,20-dione	5.59	*5.58* 6.19
33	17,21-Dimethyl-19-norpregna-4,9-diene-3,20-dione	6.70	*6.61* 7.45
34	17β,19-Dihydroxy-4-androsten-3-one	6.70	*6.59* 6.40
35	17β-Hydroxy-4-androsten-3-one	7.70	*7.87* 7.60
36	17β-Hydroxy-4-estren-3-one	6.70	6.96 6.88
37	3,17β-Dihydroxy-1,3,5(10)-estratriene	4.90	*4.88* 4.76

[a] The original affinity constants [6] were measured in molar units and they are here shown as common (Briggsian) logarithmic values.

[b] The upper and lower values refer to the experimental design based model and to the model with all 37 steroids included, respectively. Values in italics indicate the training set compounds.

Computational methods

A conformational analysis of the side chains (60-deg increments) with complete minimization of the entire structures was carried out using the molecular mechanics program MM2PI2 [7].

Carbon atoms 3, 5, 6, 13, 14 and 17 of compounds **2-37** were fitted in a least squares manner to the corresponding atoms of compound **1**.

A 3-D grid space with 1.5 Å between the points was spanned around the steroids (2475 points) and a methyl probe, similar to the one used by Cramer et al. [4], was utilized for calculating the non-bonded interactions with the equation of Del Re et al. [8] and an upper cut-off value of 30 kcal/mol. The charges used for the charge-charge interactions were calculated at the CNDO-level

[9]. The charge-charge terms were assigned a 'missing value' if the distance between the probe and the closest atom was less than the sum of the van der Waals radii. A molecular lipophilicity potential (MLP) calculated by the equation of Furet et al. [10] with the atomic lipophilicity factors of Viswanadhan et al. [11] was also computed for each point in the grid space. The three fields (nonbonded interactions, charge-charge interactions and MLP) were then transformed to a common scale by choosing the field with the numerically largest scale (the difference between the extreme values) as norm and rescaling the remaining ones accordingly. This was done to ensure that each field would contribute equally and not be biased by the original numerical values. The resulting descriptor matrix was finally mean value-centered before further statistical operations were performed.

Experimental design

A principal component analysis (PCA) was performed on the descriptor matrix with each molecule represented by the three fields (7425 points). Cross-validation was used to determine the number of significant principal components [12].

3-D QSAR analysis

The relationship between binding affinity and field descriptors was analyzed by PLS [13]. The number of significant components was determined with cross-validation by leaving one compound out of the analysis each round [12, 14] (the 'leave-one-out' method was suggested by S. Wold due to the large number of variables used in the analysis). Following the PLS analysis a retransformation of the PLS loadings into ordinary regression coefficients was performed (see Appendix) [15].

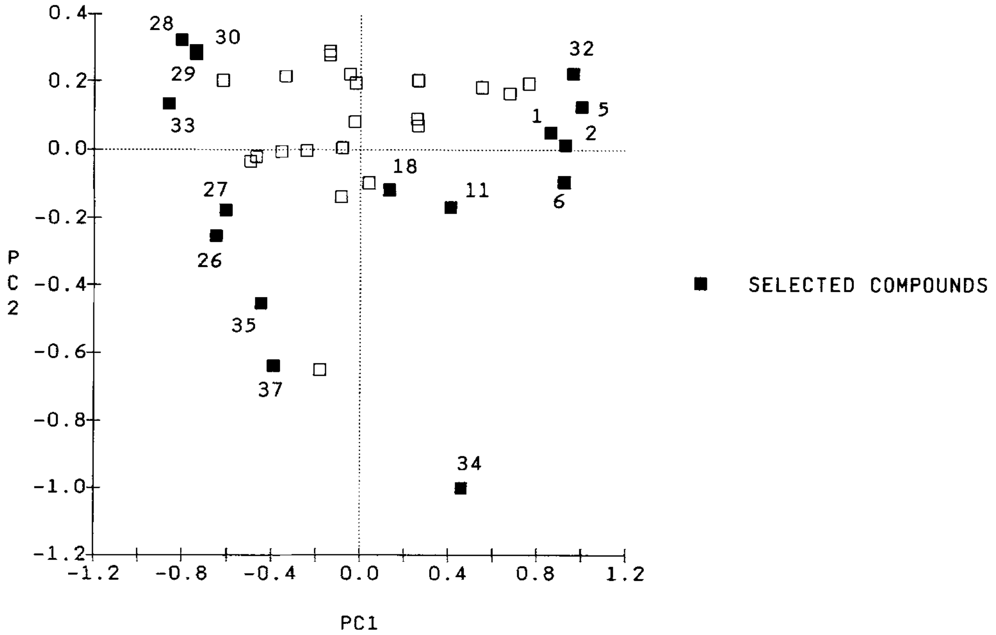

Fig. 1. The first two principal components of the PCA of compounds **1–37**.

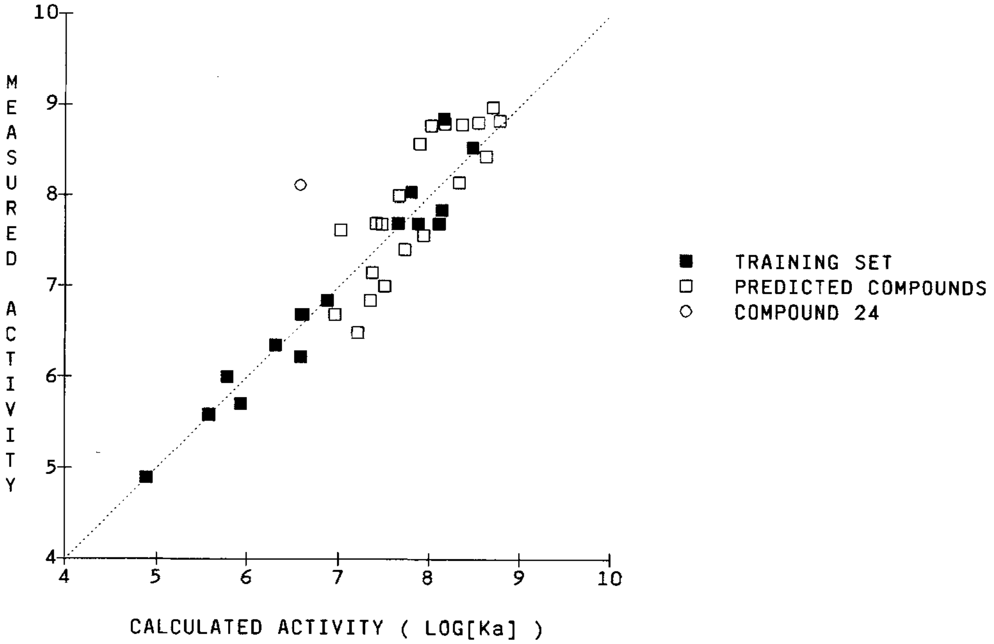

Fig. 2. Calculated vs. experimental affinity constants for the experimental design based model.

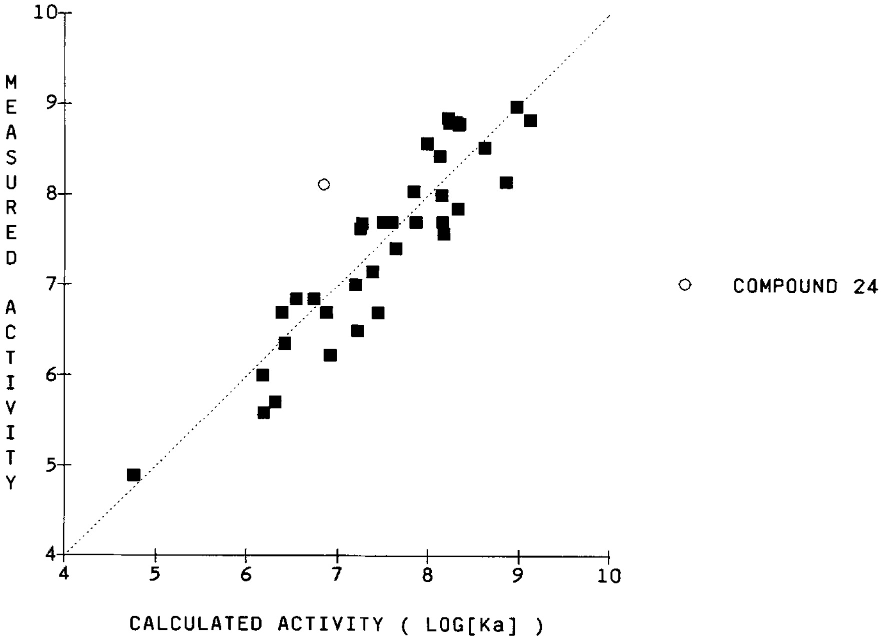

Fig. 3. Calculated vs. experimental affinity constants for the model with all the 37 compounds included.

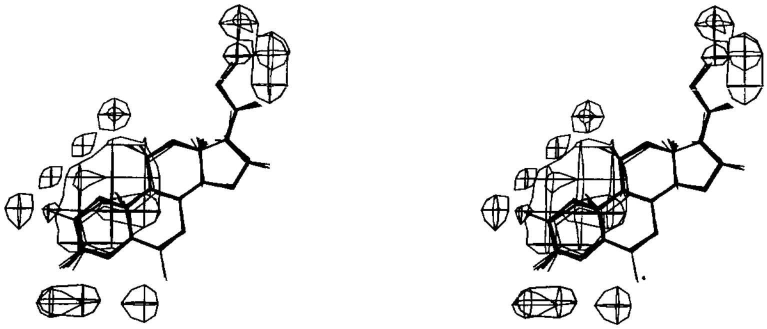

Fig. 4. Stereoscopic views of the major positive non-bonded interactions of human CBG complexes (37 steroids).

The results are depicted as 3-D contour maps of the grid points corresponding to the coefficients of each field using CHEM-X [16]. For display purposes the regression coefficients were transformed in such a way that the largest absolute value (i.e. ignoring the sign) was set to +100 or −100 depending upon the sign and the rest of the values were then scaled accordingly. Figures 4–7 show coefficients with an absolute value larger than 30.

RESULTS AND DISCUSSION

Experimental design

The first four principal components (PCs), all significant, were calculated. They described 89.6% of the variance in the data set with individual contributions of 79.1%, 5.6%, 2.8% and 2.1%, respectively. The four PCs were used to construct a 2^4 factorial design. The 16 compounds of the training set were selected to span the 4-D hyperspace as efficiently as possible. Since the first two PCs explain a major part of the variance (84.7%), care was taken to effectively cover the PC1–PC2 plane (Fig. 1).

The analysis resulted in choosing compounds **1**, **2**, **5**, **6**, **11**, **18**. **26–30**, **32–35** and **37** as the training set.

3-D QSAR analysis

The PLS analysis of the training set gave four significant PCs. The 'predictive r^2' value was 0.94. The created model was then used to estimate the affinity constants of the remaining 21 steroids

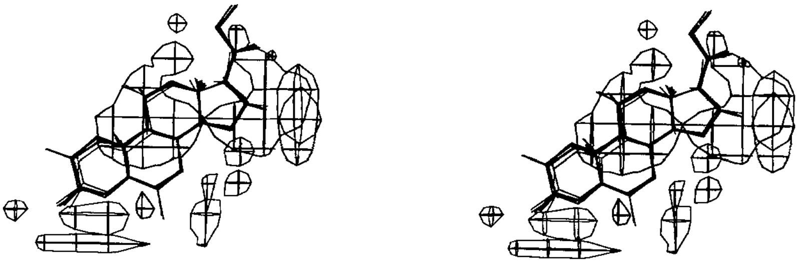

Fig. 5. Stereoscopic views of the major negative non-bonded interactions of human CBG complexes (37 steroids).

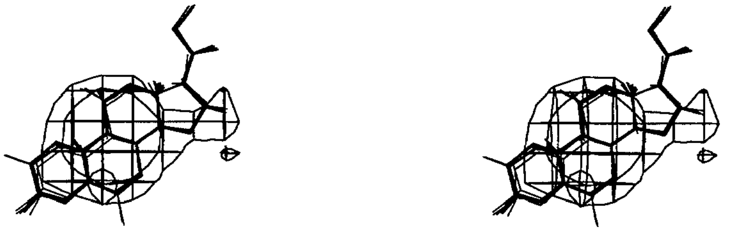

Fig. 6. Stereoscopic views of the major positive electronic interactions of human CBG complexes (37 steroids).

which resulted in a 'predictive r^2' value of 0.51 (0.70 when compound **24** was excluded). A plot of calculated vs. experimental affinity constants is depicted in Fig. 2.

The predictive power of the factorial design based relationship is satisfactory except for compound **24** which is the only steroid examined in this study with a double bond in the 5–6 position (5-pregnene-3,20-dione). It is interesting to note that the compounds of the training set cover approximately 3.5 orders of magnitude of activity while the remaining steroids only span 2 orders of magnitude.

One way of using the PCA/PLS based 3-D QSAR model when choosing new compounds to be examined (synthesized, tested, etc.) could be as follows:

(a) Examine the compounds which are calculated to have the desired activities within the predictive range (the residual standard deviation of a compound being less than two times that of the training set [17, 18]) of the model.

(b) For compounds outside the predictive range, choose the ones that contribute most information to the model, i.e. structures that probe areas in the PCA space not previously or poorly covered.

Inclusion of all 37 steroids into the model gave four significant PCs with a 'predictive r^2' value of 0.80 (Fig. 3). Compound **24** is still somewhat underestimated even when included in the training set.

The most important interactions of the CBG-steroid complexes are depicted in Figs. 4–7. From these contour plots, it can be seen that non-bonded terms dominate the interactions and that charge-charge terms are also of importance while those originating from molecular lipophilic po-

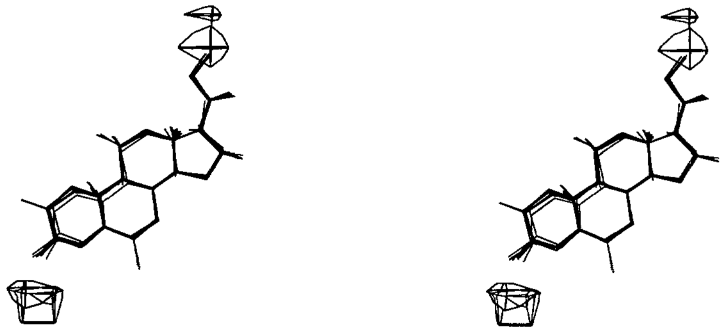

Fig. 7. Stereoscopic views of the major negative electronic interactions of human CBG complexes (37 steroids).

tentials contribute to a lesser extent. Figures of MLP are not included since the absolute values of the regression coefficients are < 30. However, other examples have demonstrated the importance of the MLPs (Norinder, U., unpublished results). Thus, the contour plots are useful for locating the areas of special importance when attempting to design new molecules with some desired activity.

CONCLUSIONS

This combined approach seems to represent a rational technique which enables one to construct a relatively small and widely spanned training set using PCA, to create a 3-D QSAR model with PLS, and subsequently predict the biological activities of new compounds with reasonable accuracy as well as expand the model in a somewhat systematic manner.

APPENDIX

The PLS methodology relates a descriptor matrix ($\sum x_{ik}$) to a dependent y-variable using the equations:

$$x_{ik} = \bar{x}_i + \sum_{a=1}^{A} (t_{ak} p_{ai}) + e_{ik} \tag{1}$$

and

$$y_k = \bar{y} + \sum_{a=1}^{A} (t_{ak} b_a) + f_k \tag{2}$$

where $\bar{x}_i$ and $\bar{y}$ denote the average of variable x_i and y, respectively, t_{ak} the scores, p_{ai} the loadings, b_a the coefficient of the inner PLS relationship between y and t_{ak}, and A is the number of extracted principal components. The residuals are denoted by e_{ik} and f_k. Parameters t_{ak} can be substituted with the equation:

$$t_{ak} = \sum_{i=1}^{M} (w_{ai} x_{ik}) \tag{3}$$

according to the NIPALS algorithm [19] where w_{ai} is the PLS weights and M the number of x-variables.

Inserting Eq. (3) into (2), remembering that x_{ik} is updated after the extraction of each principal component according to:

$$x_{ik} = x_{ik} - t_{ak} p_{ai} \tag{4}$$

gives an expression which has the form of a normal multiple regression equation where it is possible to identify the coefficients.

REFERENCES

1 Martin, Y.C., Quantitative Drug Design, Marcel Dekker, New York, 1978.

2 Gupta, S.P., Chem. Rev., 89 (1989) 1765.

3 Ghose, A.K., Crippen, G.M., Revankar, G.R., McKernan, P.A., Smee, D.F. and Robins, R.K., J. Med. Chem., 32 (1989) 746.

4 Cramer, III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.

5 Box, G.E.D., Hunter, W.G. and Hunter, J.S., Statistics for Experimenters, Wiley and Sons, New York, 1978.

6 Mickelson, K.E., Forsthoefel, J. and Westphal, U., Biochemistry, 20 (1981) 6211.

7 Norinder, U., In-house version of MM2(87). The original program is available from QCPE, University of Indiana, Bloomington, IN 47405 (academic users) and from Molecular Design Limited, 2132 Farallon Dr., San Leandro, CA 94577 (commercial users).

8 Del Re, G., Gavuzzo, E., Giglo, E., Lelj, F., Mazza, F. and Zappia, V., Acta Crystallogr., B33 (1977) 3289.

9 Phillips, A.C., QCPE Program No. 274, QCPE, University of Indiana, Bloomington, IN 47405.

10 Furet, P., Sele, A. and Cohen, N.C., J. Mol. Graph., 6 (1988) 182.

11 Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K., J. Chem. Inf. Comput. Sci., 29 (1989) 163.

12 Wold, S., Technometrics, 20 (1978) 397.

13 Wold, S., Albano, C., Dunn, III, W.J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W. and Sjöström, M., In Kowalski, B.R. (Ed.) Chemometrics – Mathematics and Statistics in Chemistry, Reidel Publ. Co., Dordrecht, 1984, pp. 17–95.

14 Johansson, E. and Wold, S., private communication.

15 Norinder, U., In-house software. Skagerberg, B., private communication.

16 Chem-X, developed and distributed by Chemical Design Ltd., Oxford, U.K.

17 Wold, S., SIMCA-3B Manual, 1983.

18 Sjöström, M., Wold, S., Lindberg, W., Persson, J.-Å. and Martens, H., Anal. Chim. Acta, 150 (1983) 61.

19 Wold, S., Albano, C., Dunn, III, W.J., Esbensen, K., Hellberg, S., Johansson, E. and Sjöström, M., In Martens, H. and Russwurm, Jr., H. (Eds.) Food Research and Data Analysis (Proceedings of the IFOST Symposium on Data Analysis in Food Research), Applied Science Publishers Ltd., London, 1983, pp. 147-188.