CROATICA CHEMICA ACTA CCACAA **62** (3) 449—458 (1989)

YU ISSN 0011-1643
CCA-1873 UDC 541.6
Original Scientific Paper

Classification of Steroid Partition Coefficients by a Pattern Recognition Technique

A. J. Harget and D. S. Ellis*

Dept. of Computer Science and Applied Mathematics, Aston University, Aston Triangle, Birmingham B4 7ET, UK

Received August 9, 1988

A pattern recognition technique, the linear learning machine method, has been used to determine structure-activity relationships for certain steroids. The steroids used in this study were classified into two categories according to their observed partition coefficient and a correlation made with certain substructural descriptors. The linear learning machine method was employed to calculate a suitable decision surface that would classify each steroid into its correct category. The resulting structure-activity relationship and the relative contributions of the various structural variables are discussed, and a comparison made with results obtained from a study using a different approach.

INTRODUCTION

In the design and development of bioactive compounds considerable attention must be given to the particular physicochemical properties, the solubility and partition coefficient. The reason for this is that the partition coefficient together with the solubility of a drug, influences the absorption and transport processes in a biological system and so determines the biological activity of the drug. Drugs taken orally in tablet or capsule form must first disintegrate after swallowing to have any medicinal effect. Slow dissolution of drug particles then occurs and the drug is transported to the gut, where active drug molecules can diffuse into the bloodstream. The fraction of drug that reaches the desired receptor is largely dependent on the dissolution and diffusion processes which are governed by solubility and membrane-water partion coefficient of the drug.

Several methods exist for the calculation of partition coefficients such as the group contribution approach proposed by Hansch[1,2] and by a method of correlation with other molecular properties as demonstrated by Yalkowsky and Valvani[3]. These methods result in good estimates of partition coefficients for most groups of compounds. However, the group contribution approach cannot account for any interactions of non-bonded atoms or groups or for any intramolecular interaction due to branching. Hence, for certain types of compounds and certain bulky or flexible molecules, estimated values can

* To whom correspondence should be addressed.

differ considerably from the experimentally determined partition coefficients. Yalkowsky and Morozowich[4] found that estimation of partition coefficients by a group contribution method worked poorly for steroids.

In this study we have applied a pattern recognition approach, the linear learning machine method[5,6] in an attempt to predict the partition coefficients of the steroids. Pattenr recognition methods attempt to categorize data samples into the membership of their observed classes. In this study the class is defined by a particular range of partition coefficient values. Despite the criticism leveled[7] at certain pattern-recognition studies for their choice of data and representation, the pattern-recognition method when applied properly, does offer the chemist a useful means of determining the relationships existing in a large amount of high-dimensional data.

PATTERN RECOGNITION

Pattern recognition techniques have been widely and successfully applied in a number of different domains[8,9]. Such techniques are particularly useful for dealing with data of high dimensionality where the deduction of relationships in the data is difficult. Although the technique is empirical it can afford a useful insight into any relationships which may exist in the experimental data. The sole assumption made by the technique is that relationships exist within the experimental data, although even this assumption will be investigated

In this study a pattern recognition technique, the linear learning machine method, has been used in an attempt to develop classification rules capable of categorising steroids according to their experimental partition coefficients. If each steroid is represented as a point in n-dimensional space, then it might be expected that steroids with similar partition coefficients would cluster in one region of the space separated from steroids with vastly different partition coefficients. The linear learning machine method was applied in an attempt to create a linear decision surface that would separate the two clusters. The computer programs were written and developed by the authors in Algol 68 for use on the University of Aston's ICL 1904S.

The linear learning machine method has been widely discussed in the literature[5,6] and so only a brief outline of the method will be given here. The experimental data to be classified is represented as a vector.

$$X = X_1, X_2, X_3 \ldots X_N$$

where each element X_I of the pattern vector represents an experimental observation and the value of N indicates the number of observations required to describe the pattern.

Linearly separable data can be separated by a linear discriminant of the form

$$S = \sum_{I=1}^{N+1} W_I \cdot X_I$$

where X_I is an element of the pattern vector, and W_I is the element of the associated weight vector. An $N + 1$ component is added where $X_{N+1} = 1$, so that the category is determined by the dot product S, namely $S > 0$ implies category 1 and $S < 0$ implies category 2.

The procedure adopted is as follows, a training set composed of steroids was classified into one of two categories according to experimentally observed partition coefficients. The training set was used to develop an effective decision surface, that is, to determine the set of weights (W_1, W_2, $W_3 \ldots W_N$) such that each member of the training set is assigned to the correct category according to its partition coefficient. The members of the training set are presented to the learning machine program one at a time and when an incorrect classification is made the weigth vector is altered. Numerous error--correction feedback algorithms exist for modifying the weight vector. The one used in this study modifies the weight vector such that the linear decision surface is reflected about the misclassified point, in other words, the dot product S has the same magnitude but the opposite, and hence correct, sign. This algorithm was chosen because of its simplicity and because it gurantees convergence when the data is linearly separable, although the rate of convergence cannot be guaranteed. In this study the sole criterion for convergence was the correct classification of each member of the training set. In the event of convergence not being obtained the program was instructed to terminate after a predetermined number of iterations.

A requirement of the linear learning machine method is that the number of compounds in the training set should exceed by at least a factor of 3 the number of descriptors if chance separation is to be avoid[10]. A further requirement states that the population of the least populated category should be greater than the number of descriptors[11]. Both of these requirements were met in the present study and consequently the derived weight vectors must be considered meaningful.

Correct classification of the training set would allow relationships to be deduced from the resulting weight vectors. Furthermore, the weight vectors obtained from the training process may allow the partition coefficients of unknown or unsynthesized steroids to be predicted with a certain measure of confidence. The predictive power of the method is expected to increase as the training set becomes larger and more representative.

The application of the linear learning machine method to steroids of known partition coefficients and the results obtained are discussed in a subsequent section.

PARTITION COEFFICIENTS

The definition and theory of partition coefficinets is well known[12] and so only a brief description will be given here.

In general terms, if an excess of liquid or solid is added to a mixture of two immiscible liquids, it will distribute itself between the two phases so that each phase becomes saturated. In the case where the amount of material added to the immiscible solvents is insufficient to reach saturation, it will become distributed between the two layers in a definite concentration ratio.

If y_1, y_2 and c_1, c_2 are the activity coefficients and the concentrations of the two solvents respectively, then the equilibrium expression becomes

$$K = \frac{y_1 c_1}{y_2 c_2}$$

where the equilibrium constant K is known as the distribution ratio or puartition coefficient.

In dilute solutions, the activity coefficients can be approximated to unity and then the above equation reduces to

$$K = \frac{c_1}{c_2}$$

Many experimental methods have been devised for determining the partion coefficient. The most common method is simply to shake a solute with two immiscible solvents and then analyse the solute concentration in one or both phases after equilibrium has been reached. In the present case, c_2 refers to the concentration in water.

An extensive literature search was undertaken to determine the partition coefficient of as many steroids as possible. Numerous sources were found but a lack of standardisation in experimental technique precluded many of them from inclusion. Two principal sources of data were found, namely, the steroid partition coefficients measured in an ether-water system by Flynn[13] and the coefficients measured in an octanol-water system by Leo, Hansch and Elkins[2]. The latter coefficients were translated into ether-water values using the method of Leo, Hansch and Elkins[2] so that the experimental data would be sufficient to meet the criteria required of a pattern recognition study.

The data set comprised 43 steroids after eliminating those steroids containing unique substituents. This elimination was necessary if meaningful structure-activity relationships were to be obtained. The remaining steroids were classified into two categories of approximately equal membership by defining a threshold value of 1.45 for the partition coefficient. Table 1 lists the steroids together with their partition coefficients.

RESULTS AND DISCUSSION

Steroids because of their similar molecular structure facilitate computational description. Since steroids have the same basic nucleus and only differ from one another by the groups that are attached to that nucleus, then each steroid can be unambiguosly described by specifying the position and type of each of its substituents; substructural descriptors are used to indicate the presence or absence of the associated substructure.

In order that meaningful structure-activity relationships could be derived, non-essential descriptors were eliminated by the weight-sign change feature selection technique[14]. In this technique a descriptor is regarded as essential if the sign of its weight vector component obtained after training is found to be invariant to the initial weight value taken for the learning machine. Three descriptors were eliminated by this process. The descriptors removed were those representing the 17-acetyl group, saturation at the 1, 2 bond and the 16-methyl group. These findings seem reasonable in the current context because the partition coefficient of a steroid is governed by the steroid-solvent interaction with hydrogen bonding being an important factor. The presence or absence of a carbon-carbon double bond within the ring structure of the steroid is unlikely to affect such an interaction. Similarly, the 17-acetyl group will not significantly alter the interaction since it is shielded by the carbonyl group at position 20. The 16-alpha-methyl would be expected to have a

TABLE I

Steroids and Their Experimental Partition Coefficients, [Me-methyl, F-fluoro]

No.	Steroid	Partition coefficient log PC
1	Prednisolone	0.053
2	Cortisone	0.146
3	Hydrocortisone	0.204
4	6α-F-Prednisolone	0.286
5	9α-F-Hydrocortisone	0.365
6	6α-Me-Prednisolone	0.537
7	Dexamethasone	0.588
8	Prednisone (Dehydrocortisone)	0.600
9	9α-F,6α-Me-Prednisolone	0.620
10	Corticosterone	0.656
11	Prednacinolone	0.820
12	6α-F-Dexamethasone	0.875
13	Flurandrenalone acetonide	1.09
14	Triamcinolone acetonide	1.16
15	Prednisolone-21-acetate	1.32
16	Cortisone-21-acetate	1.40
17	6α-F-Triamcinolone acetonide	1.41
18	Hydrocortisone-21-acetate	1.41
19	6α-M-9α-F-21-Deoxy prednisolone	1.51
20	6α-Me-Triamcinolone acetonide	1.54
21	6α-F-Prednisolone-21-acetate	1.56
22	11α-Hydroxy progesterone	1.63
23	6α,16α-Difluoro prednisolone-21-acetate	1.66
24	9α-F-Hydrocortisone-21-acetate	1.67
25	6α-Me-9α-F-21-Deoxy hydrocortisone	1.71
26	6α-Me-9α-F-16α-Hydroxy hydrocortisone acetonide	1.73
27	6α-Me-Prednisolone-21-acetate	1.83
28	6α,9α-Difluoro-21-Deoxy hydrocortisone-17-acetate	1.91
29	6α-Me-9α-F-Prednisolone-21-acetate	1.92
30	6α,9α,16α-Trifluoro-Prednisolone-21-acetate	1.93
31	6α-Me,9α-F-21-Deoxy prednisolone-17-acetate	1.97
32	Hydrocortisone-21-propionate	1.98
33	6α-F-Dexamethasone-21-acetate	2.16
34	Dexamethasone acetate	2.25
35	Hydrocortisone-21-isobutyrate	2.35
36	Hydrocortisone-21-butyrate	2.38
37	9α-F-11β-Hydroxy-6α-Me-4-pregnen-3,20-dione	2.43
38	6α-Me-9α-F-16α-Hydroxy prednisolone-16,17-acetonide--21-acetate	2.93
39	6α-Me-9α-F-16α-Hydroxy hydrocortisone-16,17-acetonide--21-acetate	2.98
40	6α-Me-9α-F-16α-Hydroxy hydrocortisone-16,17-acetonide--21-propionate	3.14
41	6α-F-Dexamethasone-21-butyrate	3.18
42	6α-Me-Triamcinolone acetonide-21-propionate	3.23
43	6α-F-Dexamethasone-21-isobutyrate	3.24

similar effect to the 6-alpha-methyl group, but this is not shown by our study and can only be due to the environment of the 16-alpha-methyl group compared to that of its 6-alpha-methyl counterpart. The steroids considered together with their partition coefficients are shown in Table I, and the substructural descriptors in Table II.

TABLE II

Substructural Descriptors and Their Associated Weight Values and Scaled Group Contribution Factors

Descriptor	Weight vector	Group Contribution
11 α/β-hydroxy	—1	—1
6A-fluoro	4	6
6A-methyl	7	11
9A-fluoro	3	4
17-hydroxy	—13	—14
16A-fluoro	1	4
16,17-acetonide	—7	—9
21-deoxy	9	21
21-acetate	12	29
21-propionate	15	
21-butyrate	15	
21-isobutyrate	15	

The linear learning machine method was applied to the steroids to determine those weight vectors which would correctly categorise each steroid according to the partition coefficient given in Table I. Complete convergence was obtained in the training procedure. The resulting weight vectors are shown in Table II. The magnitude of the weight vector is indicative of the contribution the feature makes to the classification while the sign of the weight vector indicates whether the contribution increases the partition coefficient(given by the positive values) or decreases the partition coefficient (given by the negative values). Thus if we consider the weight vectors obtained it can be seen that there is general agreement between the results and experiment. An increase in the hydrogen bonding of the system, by the introduction of more electronegative groups to the steroid molecule will cause a decrease in the partition coefficient, whilst the removal of hydrogen bonding groups from the steroid molecule will cause an increase in the partition coefficient. Other factor affecting the hydrogen bonding ability of the steroid include the steric effect (shielding of electronegative groups by inert methyl groups producing an increase in the partition coefficient) and effects due to conformations. The cyclic acetal group (16, 17-acetonide) decreases markedly the partition coefficients of the parent compounds. More particularly it can be seen from the weight vector values that the presence of a 17-hydroxy group is predicted to reduce the partition coefficient quite strongly, as it increases the hydrogen bonding. Removal of the 21-hydroxy group from the steroid, the 21-deoxy feature, or even more strongly its esterification, on the other hand reduces the hydrogen bonding and results in an increase in the partition coefficient. The relatively low contribution of the 11-hydroxy (-1), as opposed to the value of -13 for the 17-hydroxy is explained because the 11-keto group is equally active in hydrogen bonding, so its replacement by a hydroxyl group has little effect on the partition coefficient. Apart from hydrogen bonding factors, the branching of an aliphatic chain in a molecule usually affects the partition coefficient[2], with a straight chain normally having a higher partition coefficient than a branched chain. However, the weight vector values of the 21-butyrate and 21-isobutyrate groups are identical. This we believe is due to the fact that here the main

factor is the replacement of OH by OCOR, where R are nonpolar groups of approximately equal size; indeed the 21-propionate also has the same weight vector value. It should be remembered that the weight vector values shown in Table II are relative values, that is, the inclusion of a 21-ester group will increase the partition coefficient but that the final value for the partition coefficient is given by the other substituents present in the steriod molecule.

The linear pattern classifier obtained after training can be tested according to its ability to correctly classify the partition coefficient of a steroid not belonging to the training set. The predictive ability of the classifier was determined by the so-called 'leave-one-out' procedure[15]. In this procedure, a steroid was removed from the training set, and the remaining steroids were subjected to training in the usual manner. The steroid was then classified and returned to the training set and a second steroid was removed, and the training and classification procedure repeated. This process was repeated for all members of the training set, whereupon the predictive ability of the linear pattern classifier could be determined. Although this procedure is computationally expensive, it does give a good indication of the performance of the linear pattern classifier. The predictive ability was found to be 81.4% which is gratifying since the probability of guessing the correct classification is 50% for a binary pattern classifier. The following steroids were misclassified: 9, 16, 17, 21, 24, 32, 35 and 36 (the numbering refers to compounds in Table I). The misclassification of some of the compounds can be explained by examination of the training set; if a particular descriptor does not occur in many patterns, then the weight vector component associated with that descriptor may be inaccurate. For example, hydrocortisone-21-butyrate (steroid 36) when removed from the training set during the 'leave-one-out' procedure left only one other steroid containing the 21-butyrate group (steroid 41). Hence, when the weight vector component corresponding to the 21-butyrate group was calculated, three was insufficient data available on that descriptor to give an accurate result. When the calculated weigth vector was used to classify the steroid, the error caused misclassification. A similar situation exists with steroid 35, since removal of this leaves only one other steroid (steroid 43) containing the 21-isobutyrate group. Thus it would be expected that in the 'leave-one-out' procedure, as the size of the training set would increase the predictive ability would improve, particularly if the expanded training set includes steroids containing some of the poorly represented groups.

As stated earlier, the size of the weight vector component may give some indication of the relative contribution that the associated descriptor makes to the value of the partition coefficient, and the sign indicates whether the contribution is positive or negative. The idea of the relative contribution made by a particular group to the effect being studied is similar to that used in 'group contribution' approaches.

The majority of the data used in this study was taken from a report by Flynn[13] which was concerned with the estimation of steroid partition coefficients by molecular constitution. This is a group contribution method which is particularly suitable for comparison as the data used for both studies is identical.

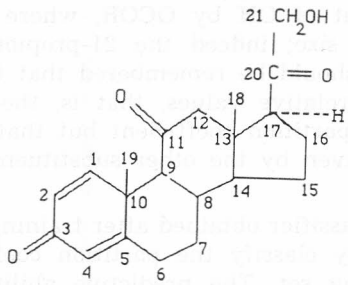
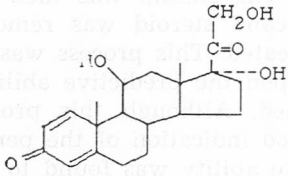
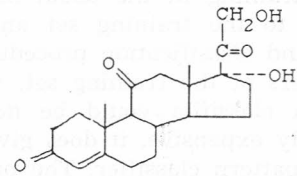


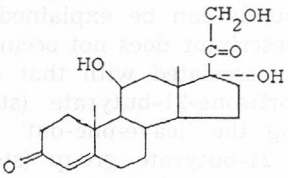
Figure 1. 17-deoxy prednisone.



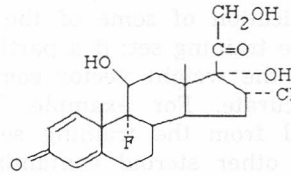
Steroid 1 - Prednisolone



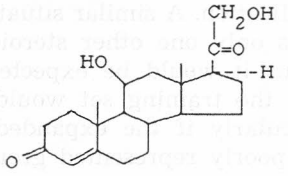
Steroid 2 - Cortisone



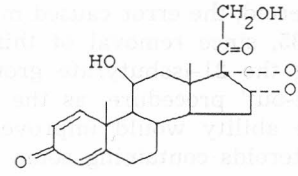
Steroid 3 - Hydrocortisone



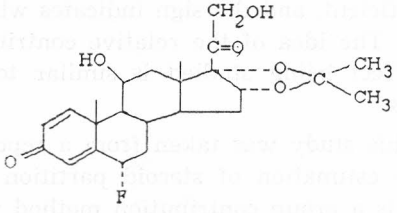
Steroid 7 - Dexamethasone



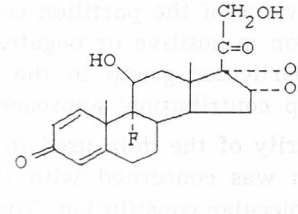
Steroid 10 - Corticosterone



Steroid 11 - Prednacinolone



Steroid 13 - Flurandrenalone acetonide



Steroid 14 - Triamcinolone acetonide

Figure 2. Structures of selected steroids.

The contributions to the partition coefficient shown by the weight vectors are relative to that of the steroid 17-deoxy prednisone (Figure 1), whilst Flynn's group contribution values are relative to that of prednisolone (steroid 1 in Figure 2.).

The only differences between the two 'standard' steroids are the presence in prednisolone of a hydroxy group at positions 11 and 17. Hence, only two group contribution factors require alteration to make the two sets compatible.

Table 2 shows some of the original group contribution factors calculated by Flinn which have been scaled-up for comparison with the weight vectors derived in this experiment; only the group contribution factors pertinent to this study are given. It can be seen that the correlation is good, since both methods agree on the features which affect the partition coefficient and the orders of magnitude of the factors are similar. The group contribution factors relate the effect of a group on the partition coefficient as a multiple of the reference steroid's partition coefficient. For a group which decreases the partition coefficient, the factor is less than 1, while for a group which causes an increase, the factor is greater than 1.

The results of the present study can clearly only be applied to those steroids containing the particular substituents considered in this paper. Steroids containing new substituents, or identical substitutents in new positions would have to be subjected to the pattern-recognition process.

The above evidence suggests that the derived pattern classifier could be used to predict the pattern coefficient of unknown, or untested, steroids. Hence the likely biological solubility of a steroid and hence its mobility to the activity site, can be very quickly determined. The medicinal chemist thus has another tool at his disposal to aid in a more rational approach to drug design.

CONCLUSION

A pattern-recognition technique has been applied successfully to the problem of predicting the partition coefficients of the steroids. The results obtained were shown to be in good agreement with experiment and with the results derived from another, quite different, study. The elements of the weight vector produced for classification were comparable with the equivalent group contribution factors calculated from a structural approach. The predictive ability of the method was found to be good, although improvement is expected as more data becomes available. The results allow the medicinal chemist to adopt a more rational approach to drug desing.

REFERENCES

1. T. Fujita, J. Isawa, and C. Hansch, *J. Amer. Chem. Soc.* **86** (1964) 5175.
2. A. Leo, C. Hansch, and D. Elkins, *Chem. Rev.* **71** (1971) 525.
3. S. H. Yalkowsky and S. C. Valvani, *J. Med. Chem.* **19** (1976) 727.
4. S. H. Yalkowsky and W. Morozowich in *Drug Design,* E. J. Ariens (Ed.), Academic Press, New York, vol. 9 1980, p121.
5. N. J. Nilsson, *Learning Machines,* New York, McGraw Hill, 1965.
6. T. L. Isenhour and P. C. Jurs, *Anal. Chem.* 43(10) (1971) 20A.
7. C. L. Perrin, *Science,* **183** (1974) 551; J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia* **27** (1973) 639.
8. P. C. Jurs and T. L. Isenhour, *Chemical Applications of Pattern Recognition,* New York, Wiley, 1975.
9. N. Bodor, A. J. Harget, and E. W. Phillips, *J. Med. Chem.* **26** (1983) 318.

10. C. F. Bender, H. D. Shepherd, and B. R. Kowalski, *Anal. Chem.* **45** (1973) 617 ;D. H. Foley, *IEEE Trans. Inf. Theory* **IT-18** (1972) 618.
11. E. K. Whalen-Pedersen-and P. C. Jurs, *J. Chem. Inf. Comput. Sci* **19** (1979) 264.
12. S. H. Yalkowsky, *Physical chemical properties of drugs,* New York, Marcel Dekker, Inc., 1980.
13. G. L. Flynn, *J. Pharm. Sci.* **60** (1971) 345.
14. T. L. Isenhour, B. R. Kowalski, and P. C. Jurs, *CRC Crit. Rev., Anal. Chem.* **4** (1974) 1.
15. B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.* **94** (1972) 5632.

SAŽETAK

Klasifikacija koeficijenata razdjeljenja steroida s pomoću tehnike prepoznavanja obrazaca

A. J. Harget i D. S. Ellis

Jedna od tehnika za prepoznavanje obrazaca — metoda linearnog učećeg stroja (LLM) — primijenjena je za određivanje relacija struktura-aktivnost (SAR) za neke steroide. Ti su steroidi razvrstani u 2 kategorije na osnovi njihovih koeficijenata razdjeljenja i nekih (sub)strukturnih deskriptora. Metodom (LLM) izračunana je razlučna ploha, s pomoću koje se svaki od proučavanih steroida razvrstava u ispravnu kategoriju. Tako dobivena SAR kao i relativni doprinosi raznih strukturnih varijabli uspoređeni su s rezultatima polučenima različitim pristupom.