ELSEVIER

Journal of Chromatography B, 687 (1996) 151–156

JOURNAL OF
CHROMATOGRAPHY B:
BIOMEDICAL APPLICATIONS

Quantitative structure–retention relationships in doping control

C.G. Georgakopoulos*, J.C. Kiburis

Doping Control Laboratory of Athens, Olympic Athletic Centre of Athens, Kifissias 37, 15123 Maroussi, Greece

Abstract

Regression equation modelling was used for the correlation of gas chromatographic relative retention times t_{RR} of anabolic steroids, stimulants and narcotics with their molecular characteristics in order to create a model for the prediction of t_{RR} values of unanalysed molecules. Predicting chromatographic retention parameters is one of the main goals of the quantitative structure–retention relationships (QSRR) methodology. To be performed, QSRR studies require two tools; a methodology for the extraction of the structural characteristics and a statistical program for the correlation of these characteristics with the chromatographic data.

Keywords: Quantitative structure–retention relationships; Anabolic steroids; Steroids

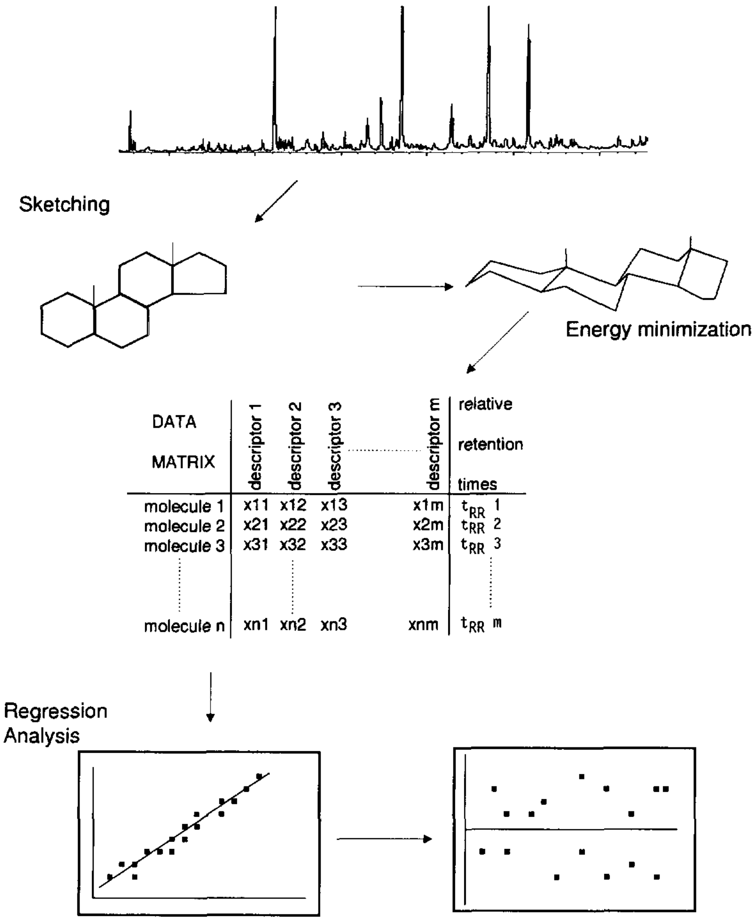
1. Introduction

Quantitative structure–retention relationships (QSRR) [1] is one specialised branch of the quantitative structure–activity relationships (QSAR) [2], involved in chromatographic retention. The goals of QSRR studies are the prediction of chromatographic retention parameters for new solutes, the calculation of the structural characteristics or properties (descriptors) of the solutes, the elucidation of the molecular mechanisms of a particular chromatographic system and the evaluation of properties, other than chromatographic, and activities of the solutes. The application of QSRR studies in the field of doping control is useful, because the International Olympic Committee (IOC) accredited doping control laboratories [3] rely on chromatographic systems in order to analyse processed urine collected from competing athletes.

The creation of a predictive QSRR model requires that several steps are performed. Firstly, the availability of a chromatographic system with a sufficiently large set of solutes and reliable data is required in order for statistical analysis to be carried out. Fortunately, chromatography can provide a great amount of precise and reproducible data. If all chromatographic conditions are kept constant (which is true), then the solute molecule is the independent variable and the retention parameter is the dependent variable. Secondly, the molecules of the solutes should be sketched in the computer using specialised software in order to calculate solute structural descriptors. Thirdly, the chromatographic data (dependent variable) and the descriptors (independent variables) are correlated and statistical models are created. This

*Corresponding author.

PII S0378-4347(96)00167-3



DATA MATRIX	descriptor 1	descriptor 2	descriptor 3	....	descriptor m	relative retention times
molecule 1	x11	x12	x13		x1m	t_{RR} 1
molecule 2	x21	x22	x23		x2m	t_{RR} 2
molecule 3	x31	x32	x33		x3m	t_{RR} 3
⋮		⋮				⋮
molecule n	xn1	xn2	xn3		xnm	t_{RR} m

Fig. 1. Flow-chart of the procedure for a QSRR model generation.

last task is repeated until the creation of a model with satisfactory statistics. In Fig. 1 the flow-chart of the entire task of the model-building is presented.

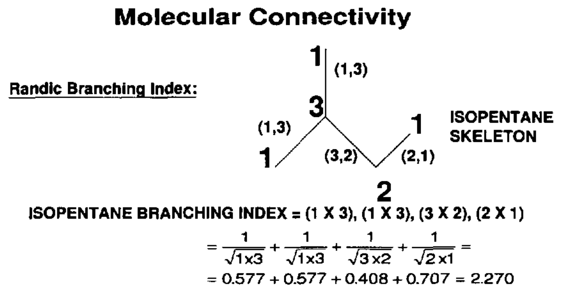


Fig. 2. Sample of the computation of a molecular connectivity descriptor.

2. Methodology

Statistical models were created by our laboratory using the QSRR methodology, and three of these have been published [4,5]. These models refer to the GC–MS analysis of steroids and gas chromatography–nitrogen–phosphorus detection (GC–NPD) analysis of stimulants and narcotics. The routine sets of chromatographic data (relative retention time, t_{RR}) of the doping substances were used as the first kind of input data. The second kind of input data were the structural descriptors of these substances. The ADAPT software, created by Prof. P. Jurs and coworkers at the Pennsylvania State University, PA, USA, was used for the studies. The followed structure input procedure comprise the sketching of the molecules in the computer, the correction of the molecules' conformation and the calculation of their descriptors. Sketching of molecules is performed in a graphical two-dimensional environment and the structures are saved as connection and distances tables (ADAPT's internal code).

After sketching, the next level of the structure input information is the conformational and geometrical properties of the molecule. This task is performed by molecular mechanics algorithms [6], which minimise the strain energy of the molecule by changing the atomic positions in a three-dimensional environment. ADAPT software includes an MM2 algorithm for strain energy minimisation [7]. This algorithm calculates the strain energy of a molecule, taking into consideration the co-ordinates of each atom. Then, successively, each atom's coordinates are changed following a predefined movement step with calculation of strain energy changes, until an acceptable minimum is reached. The better the result achieved by the minimisation energy algorithm, the better the quality of the information produced by the algorithms which calculate the descriptors' values. Another significant molecular mechanics algorithm is MOPAC [8].

Table 1
Summary of the QSRR models concerned with doping analysis

Variable[a]	Regression coefficient	Standard error of regression coefficient	Partial *F*
Regression model I for 57 stimulants and narcotics			
DPSA 3	0.00714	0.00058	149.252
NBND	0.04369	0.00298	215.466
MOLC 8	−0.09395	0.01897	24.529
V6C	4.58991	0.82314	31.093
S6C	−2.06121	0.23663	75.873
ETOT	0.01676	0.00207	65.553
Intercept	−0.12898	0.02771	21.660
R=0. 991, *n*=57 *s*=0.046 *F*(6, 50)=444.1			
Regression model II for 20 stimulants			
SRMX1	0.01396	0.00318	19.327
GEOM 5	0.00099	0.00013	55.527
S3P	0.09319	0.00900	107.227
WPSA 3	0.06142	0.00625	96.590
Intercept	−0.08667	0.05505	2.477
R=0.982 *n*=20 *s*=0.027 *F*(4,15)=99.58			
Regression model for 45 anabolic steroids			
GEOM 1	0.02182	0.00183	142.861
MOMI 4	1.01677	0.17769	74.300
V4C	−0.54379	0.06776	64.399
V5CH	2.40045	0.35070	46.851
S4P	0.12199	0.01428	72.991
S4PC	−0.06117	0.01617	14.305
S7CH	1.74499	0.19919	76.742
S6CH	−3.21262	0.39638	65.690
WTPT 3	0.02622	0.00331	62.714
Intercept	−1.65865	0.13070	161.036
R=0.991 *n*=45 *s*=0.027 *F*(9,35)=213.7			

[a] DPSA3 and WPSA3 are electronic descriptors, which encode information about polar intermolecular interactions. NBND is the number of bonds. MOLC8, V6C, S6C, S3P, V4C, V5CH, S4P, S4PC, S7CH, S6CH and WTPT3 are molecular connectivity descriptors. ETOT is an electronic descriptor, whose value is calculated from the sum of energies of all the highest and lowest occupied molecular orbitals. SRMX1 is an electronic descriptor encoding orbital concentration. GEOM5 and GEOM1 are geometric descriptors. MOMI4 is a geometric descriptor, which encodes information about moments of inertia [4,5].

The next step in the model building process, after finishing the structure input, is the calculation of the descriptors by the ADAPT software. Descriptors fall into four classes; topological, geometrical, electronic and physicochemical. Topological descriptors are derived by the data included in the connection table of the structure and encode such information as number of atoms, number of bonds, number of rings, molecular mass, substructure counts, molecular connectivity, substructure environment and path descriptors. The molecular connectivity [9] of a molecule is a measure of the size and the degree of branching of the molecule, based on the graph theory. A sample of the computation of a connectivity descriptor is presented in Fig. 2. The advantages of the topological descriptors is that they are easily computed and have strong correlations with physicochemical properties of the molecules. Geometrical descriptors, which are used mainly to differentiate between molecules with the same topological descriptors, are derived from the three-dimensional molecular models of the structures and include such information as the principal moment of inertia, molecular volume and surface area. Some three-dimensional aspects of the structures can be captured by their three orthogonal projections. Electronic descriptors, which encode the electronic structure of the molecule, are values that characterise the structure with partial atomic charges, dipole moments, bond strengths, etc. Physicochemical descriptors are values such as the logarithm of the partition coefficient of a compound between water and 1-octanol, molar refraction, molecular polarizability and others. For each QSRR study more than one hundred descriptors were calculated, from the most simple, such as the molecular mass, the number of carbon atoms, etc., to the more complicated, such as the electronic descriptors which compute the σ electron density, interatomic distance, etc.

The next task, after the computation of the descriptors, is the correlation of their values with the chromatographic data and the construction of the model. This task is performed by multiple linear regression analysis. The statistical analysis comprises several steps; reduction of the number of descriptors, model generation, model evaluation and prediction of new t_{RR} values. The reduction of the number of descriptors is performed in order to eliminate descriptors with overlapping, or minimal, information and avoid their inclusion in the final model. Descriptors having insufficient variation (e.g. descriptor values mostly identical), descriptors containing 90–95% zero values and descriptors exhibiting high pairwise correlations (e.g. from two descriptors with 95% correlated information, one is eliminated, taking into consideration criteria such as chemical significance, normal distribution, variance, etc.). Other criteria for elimination of descriptors are the multicollinearities, i.e. high intercorrelations of information between a descriptor and a set (linear combination) of descriptors. A multicollinearities test is performed by multiple linear regression analysis. Within a set of descriptors, multiple linear regression equations are built, having one descriptor from the set as the dependent variable and the remainder fo the descriptors as independent variables. This procedure is repeated for all the descriptors in the data set. The goal of this task is to predict which descriptor information is considered as the dependent variable based on the information provided by the rest. After discarding those descriptors with high correlation coefficients (e.g. 98%), the procedure is repeated until the elimination of an adequate number of descriptors has taken place. Another multicollinearities test is conducted by principal components analysis. In a set of descriptors presented in the multidimensional space, eigenanalysis is performed in order to reduce the dimensionality of the space. In a case where the difference between the initial and the final dimensionality is significant, at least 95% of the variance is sought in a reduced number of descriptors, taking into consideration the dimensionality of the space after the eigenanalysis. A third multicollinearities test is performed by the application of the Gram-Schmidt orthogonalization procedure [10]. In this procedure, descriptors are also treated as multidimensional vectors, where orthogonal vectors contain zero overlapping information, because the projection of one descriptor onto the other will be zero. The procedure of choosing the set of descriptors with the least overlapping information starts with the selection of the descriptor with the highest correlation to the dependent variable, as the initial basis vector. The next selected descriptor is that with the largest projection angle to the initial basis vector. The third descriptor selected is the most orthogonal to the plane defined by the first two descriptors, etc. This procedure continues until a user-specified maximum descriptor number is reached.

Once a refined descriptor pool has been identified, regression analysis follows. Equations/models relating the structural properties to the t_{RR} values are developed with the form:

$$t_{RR} = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots \qquad (1)$$

where a_o is the intercept and a_i represent the coefficients of various descriptors, x_i. The model is defined in such a way that the residuals between the observed t_{RR} values and the t_{RR} values calculated by the regression model are minimised. The procedure is performed by selecting, from the pool, the appropriate descriptors in order to generate a robust model with the best possible statistics. The following forward selection and stepwise multiple linear regression analysis procedure is applied [11,12]; the most highly correlated descriptor comprises the first independent variable, and the regression equation is developed. For selecting the next descriptor to enter into the model, a list with the partial correlation coefficients of descriptors, as yet unentered in the model, is examined. The descriptor with the largest correlation coefficient is considered as the next to be entered. The square of the partial T-statistic (F-value) of the new descriptor is compared with a predefined F-value (F-to-enter) and if it exceeds this limit, then the particular descriptor is entered in the equation. This procedure is repeated until no partial F-values from descriptors outside the equation exceed the F-to-enter value. Partial F-values of descriptors already participating in the equation are compared with a predefined F-to-delete, in order to decide if any descriptor must be removed from the equation (e.g., in a case where partial F-value$<F$-to-delete). This procedure is applied to all the descriptors of the final pool. After finishing the procedure of addition or deletion of variables, a new model is generated, which is validated using the following criteria; the correlation coefficient R, the standard error, s, the overall F-value for analysis of variance, the number of descriptors included in the model, the jackknifed estimates (removing, successively, one substance from the data set and calculating the *tRR* from the new generated model), the multicollinearities between these descriptors, using the multiple linear regression procedure described previously, the variance inflation factor (VIF) [13], etc.

The residuals of the generated model can be analysed to detect outliers, i.e. points with a negative influence on the coefficients of the model. Plotting the residuals can be used to judge the quality of the fit or any distribution of them (from where information could be extracted) and seek for outliers in order to delete them from the data set. For the detection of outliers, several other tests can be used: DFFITS, Cook's distance, leverage values, studentized residuals, standardised residuals, etc [11–13].

3. Results and discussion

Details about the predictive ability and the statistics of the models generated by the QSRR studies on anabolic steroids, stimulants and narcotics can be found elsewhere [4,5]. These models and their statistics are summarised briefly in Table 1.

ADAPT software was used for the entire workload of the computations of these studies. ADAPT software is an integrated tool for developing QSRR studies, because it contains both routines for descriptor generation and model construction. Several studies have been performed in the past using this software on a variety of prediction subjects; retention of polychlorinated dibenzofurans [14], boiling points [15], retention of polychlorinated biphenyls [16], odor intensity relationships [17], retention of warfare agents [18], nuclear magnetic resonance chemical shift, using neural networks [19], etc. Neural networks' methodology is an alternative of the regression analysis in the QSRR studies, but it has had less success in the predictive ability of the systems [20], because neural networks are more "sensitive" in untrained cases of prediction for new structures, compared to regression analysis.

MOLCONN-X software, by Kier and Hall, can be used instead of ADAPT software for the computation of topological descriptors. PCMODEL software, running under the MS Windows environment, can be used for determination of molecules' strain energy minimisation. Statistical computation can be performed using many other statistical software packages, including MINITAB, SPSS, SAS, etc.

Topological descriptors can be used for the creation of prediction models, concerning molecules that belong to homologous series [21]. In this case, the computation of more complicated descriptors, e.g. electronic or geometrical, in many cases seems to be unnecessary. This is very helpful when the computation of other classes of descriptors is not possible. Topological descriptors are easily computed, since their computation does not require three-dimensional sketching or strain energy minimisation. QSRR studies based only on topological descriptors are feasible for non-polar GC and reversed-phase HPLC systems.

A question always put after a predictive QSRR model has been generated, is if one can determine chromatographic mechanisms from the kind of descriptors included in the model. An answer to that question is difficult to give, mainly for two reasons: first, it is possible that several equations with similar statistics can be generated in a QSRR study, including different sets of variables (descriptors) and second, in seeking a predictive model with the best statistics, the transformation of descriptor values (e.g. raised in powers, logarithms, etc.) [22] is allowed. From such a transformed variable, a physical meaning is even more difficult to extract. However, if the objective of a QSRR study is the elucidation of the chromatographic mechanisms, then a different approach is required [1].

The number and the homogeneity of the solutes participating in a QSRR study is critical. A big number of homogeneous solutes is a good starting point for every QSRR study. As referred to previously, QSRR studies in an homologous series is much easier to perform than a study carried out in non-homologous solutes. In the case of non-homologous solutes, related solutes or solutes with common substructures would facilitate the task. There is a relationship, which should be maintained, between the number of descriptors (variables) participating in the model and the number of solutes (observations), in order to avoid correlations. This is a rule of thumb, stating that for every variable (descriptor) a minimum of five observations should be included in the data set [23]. This means that the bigger the number of observations, the bigger the number of descriptors that can be used and the bigger the amount of variation of the data set explained by the regression equation via the correlation coefficient *R*, etc. Model statistics' goals are summarised in the following set: maximum *R*, maximum overall *F*-value, minimum number of participating descriptors, minimum standard error, minimum standard deviation of the statistical coefficients and minimum correlations and multicollinearities. To follow these guidelines, the selection of *F*-to-enter and *F*-to-delete values is crucial. Low *F*-to-enter values will allow the entry of irrelevant descriptors into the equation (a value of 3–4 is mostly recommended). Similarly, a zero value of the *F*-to-delete means that descriptors can never be removed from the equation. The *F*-to-enter value should always be higher than the *F*-to-delete value, because otherwise the same descriptors would be added and deleted.

Concerning the prediction of t_{RR} values using the QSRR models, it is obvious that the structural similarity between solutes of the data set and new solutes is a limitation in the application of this methodology.

4. Conclusions

QSRR studies have practical applications, which would facilitate the standardisation of the IOC accredited doping control laboratories around the world, which was one of the life-works of Professor Donike. Modelling of the (similar) chromatographic systems of the doping control laboratories, would provide the potential to predict t_{RR} values, provide better control of experimental results, and provide a better understanding of the chromatographic systems, etc., for new metabolites. An increase in the knowledge of the molecular structure and mechanics will improve the models' performance and will result in a wider application in the future.

References

[1] R. Kaliszan, Quantitative Structure Retention Relationships, Wiley, New York, 1987.

[2] A.J. Stuper, W.E. Brugger and P.C. Jurs, Computer Assisted Studies of Chemical Structure and Biological Function, Wiley, New York, 1979.

[3] M.Donike, World Symposium on Doping in Sports — Official Proceedings, Florence, December 5–10, 1987, International Amateur Athletic Foundation, New York, 1987, pp 53–80.

[4] C.G. Georgakopoulos, J. Kiburis and P.C. Jurs, Anal. Chem., 63 (1991) 2021.

[5] C.G. Georgakopoulos, O. Tsika, J. Kiburis and P.C. Jurs, Anal. Chem., 63 (1991) 2025.

[6] P.C. Jurs, Computer Software Applications in Chemistry, Wiley, New York, 1986.

[7] U. Burker and N.L. Allinger, Molecular Mechanics, ACS Monograph, Vol. 177, American Chemical Society, Washington, DC, 1982.

[8] MOPAC, Quantum Chemistry Program Exchange, QCPE Program No. 445.

[9] L.B. Kier and L.H. Hall, Molecular Connectivity in Structure–Activity Analysis, Research Study Press, Letchworth, 1986.

[10] P.G. Ciarlet, Introduction to Numerical Linear Algebra and Optimisation, Cambridge University Press, Cambridge, 1989.

[11] N.R. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, 2nd ed., 1981.

[12] J. Neter, W. Wasserman and M. Kutner, Applied Linear Statistical Models, Irwin, Homewood, IL, 3rd ed., 1990.

[13] D.A. Belsey, E. Kuh and R.E. Welsch, Regression Diagnostics, Wiley, New York, 1980.

[14] M.D. Needham, K.C. Adams and P.C. Jurs, Anal. Chim. Acta, 258 (1992) 199.

[15] F.C. Smeeks and P.C. Jurs, Anal. Chim. Acta, 233 (1990) 111.

[16] M.N. Hasan and P.C. Jurs, Anal. Chem., 60 (1988) 978.

[17] L.M. Egolf and P.C. Jurs, Anal. Chem., 65 (1993) 3119.

[18] T.F. Woloszyn and P.C. Jurs, Anal. Chem., 64 (1992) 3059.

[19] L.S. Anker and P.C. Jurs, Anal. Chem., 64 (1992) 1157.

[20] A. Bruchmann, P. Zinn and C.M. Haffer, Anal. Chim. Acta, 283 (1993) 869.

[21] A. Robbat, Jr. and C. Kalogeropoulos, Anal. Chem., 62 (1990) 2684.

[22] M.N. Hasan and P.C. Jurs, Anal. Chem., 62 (1990) 2318.

[23] J.G. Topliss and R.P. Edwards, J. Med. Chem., 22 (1979) 1238.