Molecular Similarity Concepts. 5. Analysis of Steroid-Protein Binding Constants†

Gabrielle Rum and William C. Herndon*

Contribution from the Department of Chemistry, University of Texas at El Paso, El Paso, Texas 79968. *Received March 25, 1991*

Abstract: Molecular similarity concepts based on chemical graph theory algorithms are used to define metrics of molecular similarity ranging from unity (identical) to zero (completely dissimilar). These concepts are tested by analysis of the molecular structures of 47 steroids that bind to human corticosteroid binding globulin. The methodology has very modest computational demands, and it makes use of no physical-chemical data for independent variables other than the information contained in the customary drawings that represent molecular structure. The procedure begins with a transformation of a molecular structure into a canonically ordered symbolic matrix that is unique for each molecule up to and including stereochemical elements. Molecular similarities are then obtained by comparisons of the ordered elements of the symbolic matrices, and the molecular similarity indices are used as independent variables in multilinear regression analyses to correlate the binding data. Good correlations are obtained with reasonable numbers of similarity parameters. The results are compared with those from a more conventional analysis that uses the presence or absence of substituents and structural modifications as independent variables, and the predictive capabilities of both procedures are evaluated.

Introduction

Two distinct but complementary approaches have been developed over the past several years to deal with problems involving the relationships of molecular structure to physical, chemical, or biological properties. The widely used QSAR (quantitative structure-activity relationships) procedures normally express a molecular structure by a large set of experimental and/or theoretical numerical parameters and seek correlations of, for example, biological properties using factor and cluster analysis or multilinear regression techniques.[1-8] The second approach attempts to model the same properties by using numerical or symbolic descriptors that can be derived solely from the molecular structure as represented by the molecular drawing or graph.[9-16] In this latter case, several methodologies have been developed to obtain quantitative measures of molecular similarity. In principle, these similarity terms or indices, perhaps expressed relative to the most active compound in a data set, can be used as independent variables in a QSAR analysis. This possibility, some applications, and historical perspectives are presented in a recent edited volume of review articles on the molecular similarity concept.[17]

In previous work, we implemented and tested several simple procedures to specify molecular structure and to quantify molecular similarity.[18-23] In the initial work, several definitions were devised and compared for a small group of aliphatic alcohols.[20,21] This was followed by an application to correlate the carcinogenic potencies of a small set (16 compounds) of polycyclic benzenoid aromatic hydrocarbons.[23] In the present work, we consider a larger and structurally more diverse group of 47 steroid molecules, where the structure-dependent property of interest is the binding affinity to human corticosteroid binding globulin (CBG). The binding affinity data, obtained and previously analyzed by Mickelson et al.,[24] are given in Table I. One notes that the binding constants cover a range of 4.5 powers of 10 or almost 6 kcal in the $\Delta G°$ for binding.

Procedures

Molecular Graphs and Molecular Symbolic Matrices. The essential connectivities and three-dimensional aspects of molecules are normally represented by the conventional drawings called "constitutional formulas" or "chemical structures".[25] The labeled molecular graph is an abstract but more explicit realization of the structural drawing in which the labeled vertices of the graph denote atoms or groups of atoms and the labeled graph edges symbolize chemical bonds. We use standard atomic symbols for the atoms, and the lower case letters s, d, t, a, and h are used to designate single, double, triple, aromatic, and hydrogen bonds, respectively.[18,19] Finally, relevant, stereochemical aspects of the molecular structure are identified by adding a slash (/) to the atom or bond labels of the graph followed by conventional stereochemical notations.[21]

The molecular graph vertex labels and the bond symbols also define, respectively, the diagonal and off-diagonal elements of a symmetric square matrix. This symbolic matrix provides an alternate and completely equivalent representation of the molecular structure. The computer programs used in this work to manipulate molecular structures require the symbolic matrix for each compound under consideration. Of course, the order of the rows and columns of a molecular matrix depends on the order of the numbering of the vertices of the graph, and the numerical evaluations of similarity to be discussed also depend upon this

† Presented at the 199th National Meeting of the American Chemical Society, Division of Medicinal Chemistry, Boston, MA, Spring 1990.

(1) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(2) Olson, E. C., Cristofferson, R. E., Eds. *Computer Assisted Drug Design*; ACS Symposium Series 112; American Chemical Society: Washington, DC, 1979.
(3) Golander, V. E.; Rozenblit, A. B. *Logical and Combinatorial Algorithms for Drug Design*; Research Studies Press, Ltd.: Letchworth, England, 1983.
(4) Seydel, J. K., Ed. *QSAR and Strategies in the Design of Bioactive Compounds*; VCH: Weinheim, 1985.
(5) Valkenburg, W. V., Ed. *Biological Correlations - The Hansch Approach*; American Chemical Society: Washington, DC, 1972.
(6) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.
(7) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
(8) Jurs, P. C.; Stouch, T. S.; Czerwinski, M.; Narvaez, J. N. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 296.
(9) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press, Ltd.: Letchworth, England, 1986.
(10) Balaban, A. T., Ed. *Chemical Applications of Graph Theory*; Academic: London, 1976.
(11) Trinajstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vols. I and II.
(12) Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334.
(13) Randic, M.; Krais, G. A.; Deonauv-Jerman-Blazic, B. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; pp 192-205.
(14) Jerman-Blazic, B.; Fabic, I.; Randic, M. *J. Comput Chem.* **1986**, *7*, 176.
(15) Klopman, G. *J. Am. Chem. Soc.* **1984**, *106*, 7315.
(16) Klopman, G.; Namboodiri, K.; Kalos, A. N. In *Molecular Basis of Cancer, Part A: Macromolecular Structure, Carcinogens, and Oncogenes*; Rein, R., Ed., Alan R. Liss: New York, 1985; pp 287-298.
(17) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(18) Herndon, W. C.; Leonard, J. E. *Inorg. Chem.* **1983**, *22*, 554.
(19) Herndon, W. C. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed., Elsevier: Amsterdam, 1983; pp 231-242.
(20) Bertz, S. H.; Herndon, W. C. In *Artificial Intelligence Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1985; Chapter 15.
(21) Herndon, W. C.; Bertz, S. H. *J. Comput. Chem.* **1987**, *8*, 367.
(22) Herndon, W. C. *Comput. Math. Applic.* **1988**, *15*, 303.
(23) Herndon, W. C.; Bruce, A. J. *J. Math. Chem.* **1988**, *2*, 155.
(24) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. *Biochemistry* **1981**, *20*, 6211.
(25) Hoffmann, R.; Lazlo, P. *Diogenes* **1989**, *N147*, 23.

0002-7863/91/1513-9055$02.50/0

Table I. Binding Affinity Constants of CBG with Steroids

no.	steroid name	$10^{-7}K_a$ (M^{-1})	ln ($10^{-7}K_a$)
1	11β,17,21-trihydroxy-4-pregnene-3,20-dione	71	4.263
2	14α,17,21-trihydroxy-4-pregnene-3,20-dione	0.7	−0.357
3	11β,17,21-trihydroxy-1,4-pregnadiene-3,20-dione	37	3.611
4	11β,17,21-trihydroxy-2α-methyl-4-pregnene-3,20-dione	60	4.094
5	11β,17,21-trihydroxy-2α-methyl-9α-fluoro-4-pregnene-3,20-dione	0.17	−1.772
6	21-acetoxy-11β,17-dihydroxy-4-pregnene-3,20-dione	42	3.738
7	17,21-dihydroxy-4-pregnene-3,11,20-trione	5	1.609
8	11β,17,20α,21-tetrahydroxy-4-pregnen-3-one	1.7	0.531
9	11β,17,20β,21-tetrahydroxy-4-pregnen-3-one	0.64	0.446
10	11α,21-dihydroxy-4-pregnene-3,20-dione	14	2.639
11	11β,21-dihydroxy-4-pregnene-3,20-dione	96	4.564
12	16α,17-dihydroxy-4-pregnene-3,20-dione	0.7	−0.357
13	17,21-dihydroxy-4-pregnene-3,20-dione	64	4.159
14	11β,21-dihydroxy-5β-pregnane-3,20-dione	5	1.609
15	2α-hydroxy-4-pregnene-3,20-dione	27	3.296
16	6α-hydroxy-4-pregnene-3,20-dione	1.4	0.336
17	6β-hydroxy-4-pregnene-3,20-dione	0.31	−1.171
18	11α-hydroxy-4-pregnene-3,20-dione	10	2.303
19	16α-hydroxy-4-pregnene-3,20-dione	1.0	0.000
20	17-hydroxy-4-pregnene-3,20-dione	63	4.143
21	12α-hydroxy-5β-pregnane-3,20-dione	0.10	−2.303
22	17-acetoxy-4-pregnene-3,20-dione	0.08	−2.526
23	17-caproxy-4-pregnene-3,20-dione	0.0043	−5.449
24	21-hydroxy-4-pregnene-3,20-dione	68	4.219
25	17-hydroxy-6α-methyl-4-pregnene-3,20-dione	2.6	0.956
26	17-hydroxy-16α-methyl-4-pregnene-3,20-dione	4.9	1.589
27	4-pregnene-3,11,20-trione	3.7	1.308
28	4-pregnene-3,20-dione	59	4.078
29	5-pregnene-3,20-dione	13	2.565
30	5α-pregnane-3,20-dione	2.3	0.833
31	5β-pregnane-3,20-dione	4.2	1.435
32	3β-hydroxy-5-pregnen-20-one	0.05	−2.996
33	3α-hydroxy-5β-pregnan-20-one	0.23	−1.470
34	2α-methyl-4-pregnene-3,20-dione	34	3.526
35	6α-methyl-4-pregnene-3,20-dione	7.1	1.960
36	16α-methyl-4-pregnene-3,20-dione	11	2.398
37	19-nor-4-pregnene-3,20-dione	5	1.609
38	17-hydroxy-4-pregnen-3-one	0.6	−0.511
39	5α-pregnan-3-one	0.0025	−5.991
40	18,11-hemiacetal of 11β,21-dihydroxy-3,20-dioxo-4-pregnen-18-al	0.8	−0.223
41	9α-fluoro-16α-methyl-11β,17,21-trihydroxy-1,4-pregnadiene-3,20-dione	0.039	−3.244
42	17,21-dimethyl-19-norpregna-4,9-diene-3,20-dione	0.5	−0.693
43	17β,19-dihydroxy-4-androsten-3-one	0.5	−0.693
44	17β-hydroxy-4-androsten-3-one	5	1.609
45	17β-acetoxy-4-androsten-3-one	1.5	0.405
46	17β-hydroxy-4-estren-3-one	0.5	−0.693
47	3,17β-dihydroxy-1,3,5(10)-estratriene	0.008	−4.828

factor. A standard numbering of the molecular structure is thus required as described in the next subsection.

Canonical Numbering and Unique Linear Molecular Notations. Explicit rules and algorithms have been given previously for several types of canonical numbering systems, each of which gives rise to a unique numbering of the vertices of a molecular graph, and hence a unique arrangement of the rows and columns of the corresponding symbolic molecular matrix.[18,19,23] The basic algorithmic numbering tool used in the present work is called extended connectivity.[19] The extended connectivity number of a graph vertex can generally be assigned in an iterative process by starting with the vertex degrees and then summing the numbers already assigned to the self-same vertex with those of adjacent vertices in each iteration. It is, therefore, possible to obtain this hierarchal ordering by hand[18,19] or by making use of the appropriate computer programs. One finds that high connectivity and centrality in the molecular graph are the main factors giving priority in this numbering system. The extended connectivity numbering system along with the conventional numbering of steroidal systems are illustrated in Figure 1 for progesterone (**28** in Table I).

The matrix that results from the hierarchal numbering can be recast into a linear notation format (termed LN1).[18] The caption to Figure 1 gives the progesterone LN1 notation, and one notes that the notation fully represents the structure since either the matrix or the molecular graph can be recovered starting from the LN1 string of symbols. The inverse of the LN1 numbering results in a linear notation (INVLN1) which gives

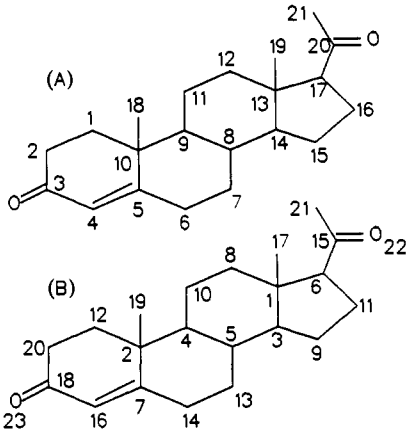

Figure 1. Conventional (A) and LN1 notation (B) numbering for progesterone. LN1 notation: (C)-03s-06s-08s-17s-(C)-04s-07s-12s-19s-(CH/β)-05s-09s-(CH/β)-05s-10s-(CH/α)-13s-(CH/β)-11s-15s-(C)-14s-16d-(CH2)-10s-(CH2)-11s-(CH2)-(CH2)-(CH2)-20s-(CH2)-14s-(CH2)-(C)-21s-22d-(CH)-18s-(CH3/α)-(C)-20s-23d-(CH3/α)-(CH2)-(CH3)-(O)-(O). TN1 notation: (Cssss)-(Cssss)-(CH/βsss)-(CH/βsss)-(CH/αsss)-(CH/βsss)-(Cssd)-(CH2ss)-(CH2ss)-(CH2ss)-(CH2ss)-(CH2ss)-(CH2ss)-(CH2ss)-(CH2ss)-(Cssd)-(CH2ss)-(CH3/αS)-(Cssd)-(CH3/αs)-(CH2ss)-(CH3s)-(Od)-(Od).

priority to atoms or groups on the periphery of a molecular structure (not illustrated). Finally, shorter notational forms (TN1 in Figure 1 or INVTN1) can be obtained by removing atom locants and coalescing atom symbols with the symbols of attached bonds. These shorter notational forms are simply lists of augmented atom descriptors ordered according to the extended connectivity numbering system.

Both the LN and TN notations are unique for all of the compounds investigated in the present paper. However, some topological structural information is suppressed in obtaining the TN-type notations, and it is possible to draw pairs of small isomeric molecules that would possess identical TN1 notations.

Similarity Indices. Our methodology for obtaining a quantifiable correlation of a physical, chemical, or biochemical property with molecular structure requires two main steps: the first is to obtain metrics of molecular similarity (similarity indices), and the second to employ statistical techniques to find valid correlations of the indices with the measured values of the property. We make the basic assumption that the overall intrinsic similarity between the linear notations for two molecular structures is a gauge for the actual molecular similarity, and we calculate the similarity of the linear notations by standard text comparison computer procedures.[26-29] The assumptions bear obvious relationships to those made by previous investigators who used the presence of particular sequences contained within linear molecular codes to correlate with properties.[30] The present work differs in that we assess the pairwise homologies of two entire molecular codes in order to define the pairwise molecular similarity.

In a previous application[23] we designated an optimal alignment of a pair of molecular notations by drawing the maximum number of noncrossing lines between corresponding elements. After counting the number of insertions and deletions (indels) required to convert one linear notation into the other while preserving the alignment, the similarity (S) could be computed as unity minus the number of indels divided by the total number of terms in the two notations.

$$S_{ij} = 1 - (\text{indels})/(N_i + N_j) \quad (1)$$

This definition of a similarity index, which gives values that range from unity (identical molecules) to zero (completely different compounds), is simple and rational. Many other options to calculate similarity are possible; for example, our computer programs can require that

(26) Sankoff, E. *Proc. Nat. Acad. Sci.* **1972**, *69*, 4.
(27) Wong, A. K. C.; Reinchert, T. A.; Cohen, D. N.; Aygun, B. O. *Comput. Biol. Med.* **1974**, *4*, 43.
(28) Sankoff, D., Kruskal, J. B., Eds. *Time Warps, String Edits, and Macromolecules; The Theory and Practice of Sequence Comparison*; Addison-Wesley: Reading, MA, 1983.
(29) Waterman, M. S. *Bull. Math. Biol.* **1984**, *46*, 473.
(30) An early example using Wiswesser line notation is the work of Adamson, W. A.; Bawden, D. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 215. Several studies by Klopman and co-workers comprise additional recent examples. For references see: Klopman, G.; Raychaudhury, C. *J. Comput. Chem.* **1988**, *9*, 232.

Table II. Linear Correlations of Similarity Indices with CBG Binding Data

similarity type	correln coeff	R^2	no. of param	std error
LN1	0.864	0.747	7	1.456
LN1	0.928	0.861	16	1.231
INVLN1	0.903	0.815	14	1.373
INVLN1	0.947	0.897	25	1.265
TN1	0.893	0.797	8	1.322
TN1	0.953	0.908	15	0.999
INVTN1	0.892	0.795	9	1.448
INVTN1	0.952	0.908	20	1.086

long homologous sequence of terms receive higher weights in calculating similarity than the same number of terms not in sequence. Thus, a corresponding sequence of terms of length l can be given a weight equal to $l\,(l+1)/2$, and the calculated similarity index for a pair of notations (molecules) is adjusted accordingly. However, only results using the simpler definition will be reported here.

Statistical Procedures. The definition of similarity given in eq 1 was applied to the structures of all the compounds listed in Table I. Recursive versions of the computer programs allowed the construction of four 47 × 47 similarity matrices, one matrix for each type of linear notation (LN1, INVLN1, TN1, and INVTN1). A column in any of these matrices represents the calculated similarities of a single compound to all of the other compounds in the data set. The calculated pairwise similarities range from 0.314 to 0.987 in the case of the LN1-type notations and from 0.115 to 0.958 for the TN1-type notations. The actual similarity matrices and code for the various computer programs are available from us upon request.

The relationship between molecular similarities (S) and the CBG binding constants is assumed to have the following multilinear form

$$\ln (10^{-7}K_a) = a_0 + a_1S_1 + a_2S_2 + \ldots + a_{47}S_{47} \qquad (2)$$

where several of the a_j coefficients of the independent S_j variables are expected to have statistically insignificant values. All similarity terms were entered into a potential model correlation equation (eq 2), and the individual terms were screened for inclusion in a final regression equation using a standard stepwise multivariate linear regression analysis. The forward entry method was used, coupled with the backward elimination procedure. The values of the statistical options controlling the criteria for inclusion of independent variables were those recommended on the basis of Monte Carlo studies of regression models.[31] Also, during the course of this work, a decision was made to critically evaluate model equations only if the number of compounds exceeded the number of included independent variables by at least a factor of three.

Tests of Prediction Capabilities. If one accepts the implicit assumption that molecular structure and activity are related, the above procedures must be expected to provide some degree of reasonable correlations of the CBG binding data. However, additional tests are required in order to determine the actual predictive value of the above methods, and perhaps to assess the possibility that the correlations are due to chance. The main test procedure used in this work was cross-validation, which involved the following steps: (a) the omission of each compound and its CBG binding constant, in turn, from the database; (b) the evaluation of new regression equations for each of the 47 new data sets; and (c) the calculation of the CBG activity for each steroid using the correlation equation obtained from the data set in which its activity was omitted. Finally, a linear regression of this leave-one-out set of calculated CBG binding values against the experimental values gave regression parameters that allowed a judgement of true predictive capabilities.

Results and Discussion

The characteristic features of the multilinear relationships between similarity parameters, defined as outlined in the previous section, and the steroid–CBG binding constants (Table I) are summarized in Table II. We list the results for two correlations for each type of similarity definition if allowed by the statistical criteria, one at the level where ca. 80% of the variance in the binding data is correlated and the other at the 90% level (R^2 = 0.8 and 0.9, respectively).

In general, all of the results listed in Table II are comparable and could be characterized as reasonable rectifications of the binding data. For example, an acceptable small number of the independent similarity parameters are required for the 80% level correlations in each case. However, even though the overall results are somewhat similar, the TN1 correlations incorporate the smallest number of parameters necessary to reach the chosen levels of correlation, and also correspond to our chosen levels for critical evaluation. Therefore, rather than discuss each one of the correlations summarized in Table II, only the details of the TN1 case will be presented.

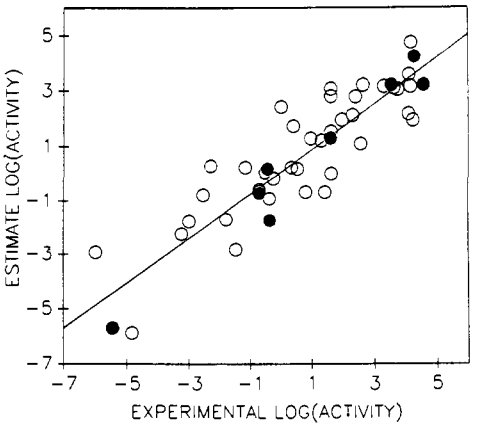

Figure 2. TN1 similarity regression model (R^2 = 0.80). The closed circles indicate compounds whose similarity terms define the model.

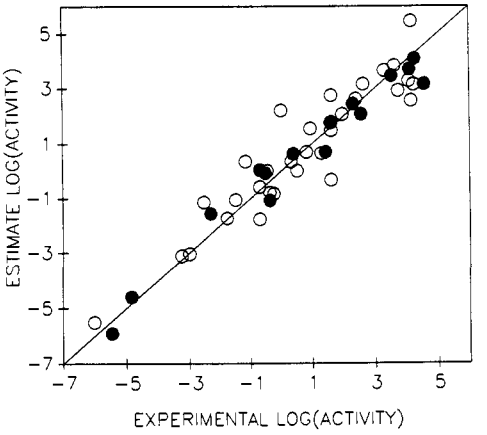

Figure 3. TN1 similarity regression model (R^2 = 0.90). The closed circles indicate compounds whose similarity terms define the model.

We will also examine a previously suggested approach to modeling the same data which is based on estimating the contributions of specific polar and nonpolar groups, along with other structural changes, to the free energy of CBG binding.[23] This procedure, which we will elaborate and term the "functional group model", uses the presence or absence of functional groups (represented by the variables unity or zero, respectively) as the independent parameters. Finally, we will compare the results of the more abstract similarity analysis with the functional group approach and attempt to assess the advantages and relative suitabilities of procedures of both types.

TN1 Similarity. The TN1 notation system uses extended connectivity to establish the order of the terms in the linear descriptor of a molecule. There is one term for each atom or atom group, and each term is composed of the symbol for the atom (atoms) followed by symbols for the bonds attached to that atom (see Figure 1). The interior atoms are grouped at the front of the notation, and terminal atoms are last. The numbering of the atoms in the molecular structure, which determines the order within the list of atom symbols, is a consequential component of the similarity definition, since indels (eq 1) are enumerated with preservation of the maximum correspondence between notation elements. However, it is probable that any consistent numbering system based on the molecular structure could serve as the starting point for the similarity analysis.

The TN1 results are illustrated in Figures 2 and 3, in which the compounds whose similarity parameters appear in the list of

(31) Bendel, R. B.; Afifi, A. A. *J. Am. Stat. Assoc.* **1977**, *72*, 46.

Table III. TN1 Similarity: CBG Binding Constant Model (R^2 = 0.90)

compd	name	reg coeff	std error
	constant term	+26.833	4.218
1[a]	11β,17,21-trihydroxy-4-pregnene-3,20-dione	−26.104	3.218
34[a]	2α-methyl-4-pregnene-3,20-dione	+24.168	3.525
29	5-pregnene-3,20-dione	−21.031	6.329
31	5β-pregnane-3,20-dione	+21.179	4.878
4	11β,17,21-trihydroxy-2α-methyl-4-pregnene-3,20-dione	+15.264	3.973
11[a]	11β,21-dihydroxy-4-pregnene-3,20-dione	+14.093	4.209
26[a]	17-hydroxy-16α-methyl-4-pregnene-3,20-dione	+7.547	4.386
47	3,17β-dihydroxy-1,3,5(10)-estratriene	+6.808	2.865
23[a]	17-caproxy-4-pregnene-3,20-dione	−5.745	3.236
42[a]	17,21-dimethyl-19-norpregna-4,9-diene-3,20-dione	−11.864	3.786
12[a]	16α,17-dihydroxy-4-pregnene-3,20-dione	−13.958	3.692
18	11α-hydroxy-4-pregnene-3,20-dione	−17.616	5.399
45	17β-acetoxy-4-androsten-3-one	−17.930	6.169
38	17-hydroxy-4-pregnen-3-one	−18.989	4.330
21	12α-hydroxy-5-pregnane-3,20-dione	−20.713	4.149

[a] Compound also included in the R^2 = 0.80 model.

regression coefficients are denoted as filled circles. The details of the TN1 model equation (R^2 = 0.90) are given in Table III where names of compounds whose similarity indices contribute to the model are listed in order of the values of the regression coefficients of the multilinear model in order to facilitate analysis. We expected that the majority of compounds included in the correlation by the stepwise regression procedure would have either very large or very small activities and that the size and sign of regression coefficients would reflect activities. An examination of Tables I and III shows that this generalization is not correct in all cases. Exceptions are compounds **18**, **45**, and **47**, whose regression coefficients indicate contributions to activity not in order of experimental magnitudes of ln ($10^{-7}K_a$). One also notes that several compounds which have a unique structural feature or functional group, i.e., **23**, **42**, and **47**, are necessary in the 80 or 90% models, while it is not necessary to involve others, notably **40** (hemiacetal) and **41** (9α-F).

It is actually very difficult to discern single specific structural features that are responsible for the differential activities based on an examination of the regression equations generated by using the similarity matrix. Perhaps this points to a weakness of this quite abstract protocol for quantification of a structure–activity relationship. However, the correlations are quite acceptable with a reasonable number of parameters. The more conventional functional group model, which will be examined below, is found to require approximately the same number of parameters to yield a comparable degree of acceptability, but several of the parameters have to be evaluated using binding constant data from single compounds.

Functional Group Model. Mickelson et al.[24] assigned numerical contributions of 24 substituent groups and other structural features for the CBG binding constants of the compounds in Table I. The analysis was carried out by examining pairs of compounds that differed in only one molecular structural change, and it was necessary to use binding data[32,33] for other than human CBG for two of the structure descriptors. The dependent variable was taken to be the $\Delta G°$ of complexation (4 °C), which is, of course, linearly related to the ln ($10^{-7}K_a$) values. Table IV contains a list of these substituents, their numerical values derived in the previous work, and the statistical results for the correlation of the Table I data using these parameters.

The substituent constants in Table IV give a very poor correlation of the Table I data compared to the TN1 similarity approach, even though the analysis employs nine additional substituent parameters. The reason for this result is that the substituent constants were not adjusted by any statistical procedure. That is, the values in the table are those based on distinctive pairs of compounds. In fact, the variables for groups that are common to several compounds are actually excluded from the model.

Table IV. Free Energy Contributions of Steroid Substituents for Binding to Human CBG[a]

subst[b]	$\Delta G°$	subst[b]	$\Delta G°$	subst[b]	$\Delta G°$
2α-OH	+0.4	11-(C=O)	+1.5	6α-CH_3	+1.5
3α-OH	+1.6	14α-OH	+2.4	10β-CH_3	−1.4
3β-OH	+3.1	17α-OH	+0.3	16α-CH_3	+1.2
3-(C=O)	−2.7[c]	20α-OH	+2.1	17α-OAc	+3.6
6α-OH	+2.1	20β-OH	+2.6	9α-F	+3.2
6β-OH	+2.9	20-(C=O)	−3.2	1(2) C=C	+0.4
11α-OH	+0.9	21-OH	−0.1	4(5) C=C	−1.6
11β-OH	−0.1	2α-CH_3	+0.2	5(6) C=C	+0.1[c]

[a] Mickelson et al.[24] [b] See Figure 1A for numbering. [c] Estimated from other types of binding data.[31,32]

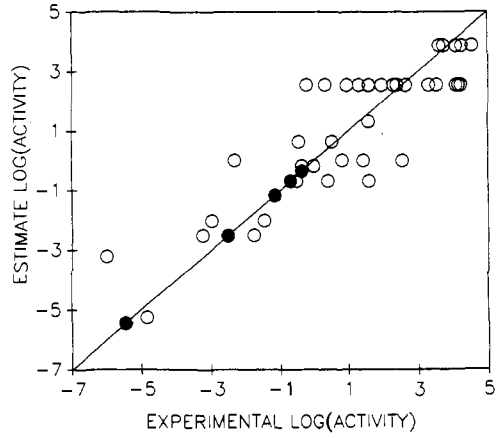

Figure 4. Functional group regression model (R^2 = 0.80). Closed circles indicate parameters unique for a single compound.

In order to extend the functional group correlation model, and to compare and evaluate its performance fairly, we carried out a more complete analysis as follows:

(a) The fundamental steroid substructure without substituents was considered to be 17β-ethyl-10β,13β-dimethylcyclopentaperhydrophenanthrene (see Figure 1A) with trans junctions at the B–C and C–D ring fusions (8β, 9α, and 14α hydrogen atoms).

(b) In order to avoid bias as to what constitutes a functional group, any deviation from this basic skeleton was considered to define a group parameter. Thirty-seven functional descriptors were identified, and a matrix of these descriptors was constructed by assigning a value of 1.0 as an "indicator variable" if the functional group was present; otherwise, the value 0.0 was assigned. Nineteen of these descriptors are unique; that is, they are each present in only a single steroid in the data set. One compound (**42**, Table I) has three such unique structural features, and two of these descriptors were eliminated from the data matrix.

(c) Stepwise linear regression with ln ($10^{-7}K_a$) as the dependent variable was then used to test this augmented functional group model.

Standard statistical criteria allow one to obtain several acceptable multilinear relationships between the steroid CBG binding constants (Table I) and the structural variables. Results at the 80% and 90% levels of correlation are depicted in Figures 4 and 5, and the regression coefficients of these multilinear models are listed in Table V. .

The 90% model with eight adjustable parameters and eight coefficients for unique functional groups is a good correlation of the binding data by any of the statistical criteria. An interesting aspect of the coefficient values is that they are all negative except for the parameters representing the 4(5) CC double bond, the 11β hydroxy group, and the carbonyl groups at C-3 and C-20. Note that the carbonyl group parameters (3.6 and 3.4 kcal) have values that would be expected for enthalpies associated with formation

(32) Blanford, A. T.; Wittman, W.; Strupe, S. D.; Westphal, U. *J. Steroid Biochem.* **1978**, *9*, 187.
(33) Mickelson, K. E.; Westphal, U. *Biochemistry* **1979**, *18*, 2685.

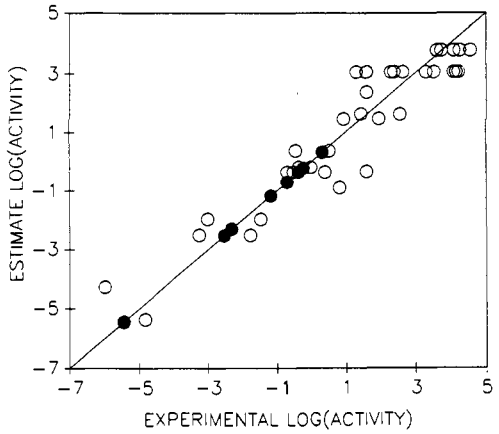

Figure 5. Functional group regression model (R^2 = 0.90). Closed circles indicate parameters unique for a single compound.

Table V. Analysis of Functional Group Modes for Human CBG Binding

structural descriptor	regression coeff (std error)	
	80% model	90% model
constant term	-5.249 (0.860)	-5.357 (0.663)
4(5) C=C	2.532 (0.613)	1.412 (0.648)
5α-H		-2.511 (0.985)
3-(C=O)	2.030 (0.971)	3.595 (0.874)
6α-CH3		-1.580 (0.784)
6α-OH[a]		-2.072 (1.075)
6β-OH[a]	-3.711 (1.397)	-4.209 (1.075)
11β-OH	1.329 (0.550)	0.731 (0.436)
12α-OH[a]		-3.929 (1.211)
14α-OH[a]	-2.897 (1.397)	-3.395 (1.075)
16α-OH	-2.719 (1.012)	-3.216 (0.784)
9α-F	-6.377 (1.088)	-6.277 (0.836)
18(11)-hemiacetal[a]		-3.261 (1.075)
17-caproxy[a]	-7.989 (1.397)	-8.487 (1.075)
17α-acetoxy[a]	-5.066 (1.397)	-5.564 (1.075)
20-(C=O)	3.227 (0.522)	3.388 (0.418)
17-CH_3[a]	-3.233 (1.397)	-3.731 (1.075)
no. of terms	11	16
correln coeff	0.895	0.949
R^2	0.802	0.900
std error	1.362	1.042
F ratio	12.854	16.968

[a] Unique substituents.

of hydrogen bonds. Therefore, the results of this model may be taken as partial support for a qualitative picture of the human CBG steroid binding site that was derived by Mickelson et al.[24] which involved hydrogen bonding at both extreme ends of an active steroid molecule. However, a requirement for hydrophobic binding at all other locations on the steroid nucleus, postulated in the former work, is not supported by the present calculated models. The small positive 11β-OH and 4(5) CC terms and the remaining uniformly negative larger values for both protic and nonprotic substituents may simply indicate that steric constraints are quite rigorous for the CBG binding process.

A deficiency of the functional group regression model is the fact that 5 of 11 (80% model) or 8 or 16 (90% model) regression coefficients in Table V are for substituents which are each present in only a single compound. Therefore, the "predicted" values of ln ($10^{-7}K_a$) for the compounds with these substituents are corrected by the regression coefficient of the respective functional group parameter to their exact experimental values. A reviewer has suggested that compounds with unique substituents should be left out of the statistical comparisons. However, our method for defining a substituent group is more rigorous than the usual subjective choice, and complete exclusion of compounds with unique groups would, unacceptably, reduce the size of the data set from 47 to 30 compounds.

Nevertheless, an evaluation along these lines can be obtained using the statistical parameters already given in Table V. In this case, the statistical parameters for the 80% model (5 unique parameters out of 11 for 47 compounds) are identical to regression statistics for an 80% correlation model for 41 compounds using only six parameters. Similarly, a 90% model with 8 parameters for 39 compounds is obtained after eliminating the eight compounds in the table with unique groups. These results sustain the conclusion that the correlative capabilities of the functional group model are quite acceptable. However, the actual predictive power of the procedure is still difficult to ascertain, based on this analysis.

Table VI. "Predictive" Correlations of CBG Binding Data

model (R^2 = 0.90)	correln coeff	R^2	std error
TN1[a]	0.673	0.453	1.994
TN1[b]	0.613	0.376	2.129
functional gp	0.536	0.287	2.276
functional gp[c]	0.765	0.585	1.649

[a] Includes the similarity indices (as an independent variable) for the compound with excluded dependent variable. See text. [b] Excludes similarity indices for excluded dependent variable. [c] Excluding seven outliers. See Figure 7 and text.

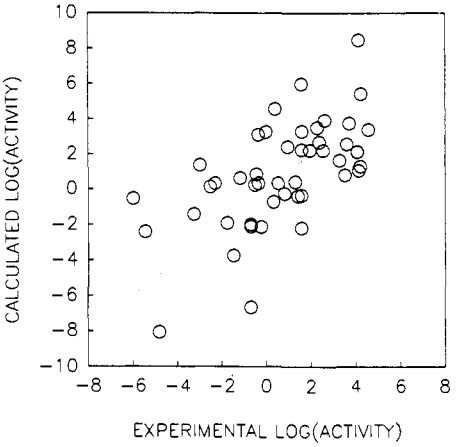

Figure 6. Similarity model cross-validated predictions of CBG binding affinities (with inclusion of similarity indices for excluded compounds).

Predictions of CBG Binding Affinities. The leave-out cross-validation protocol outlined in Procedures can be applied to the problem of evaluating the predictive capabilities of both types of modeling procedures. Every ln ($10^{-7}K_a$) value is calculated from a regression equation derived from a set of data that excludes the dependent variable of the compound under consideration. A 90% model equation was obtained for each leave-out-one set, and this equation was then used to predict the ln ($10^{-7}K_a$) value for CBG binding to the excluded compound.

The similarity analysis allows the leave-out procedure to be implemented in two different ways, i.e., with or without the similarity indices for the excluded structure included as one of the possible independent variables in the optimum regression equation. This is possible because the structure of the left-out compound is known and available even though its activity is postulated to be unknown. However, in the case of the functional group approach, the value of a regressor indicator variable for a particular compound cannot be included in the regression analysis unless the dependent variable for that compound is also used.

The statistical results for the three possible cross-validation analyses (90% models) are summarized in Table VI. Plots of experimental versus predicted ln ($10^{-7}K_a$) values given in Figures 6 and 7 illustrate the quality of the two types of predictive calculations.

The similarity analysis leave-out procedure correlates approximately twice the fraction of the variance in the binding data as does the functional group model. An examination of Figure 7 shows that a key reason for the relatively poor performance for the latter model is the group of seven compounds (filled circles in Figure 7) with experimental values of ln ($10^{-7}K_a$) that cover a range from -6 to 0, all predicted to have values of ln ($10^{-7}K_a$)

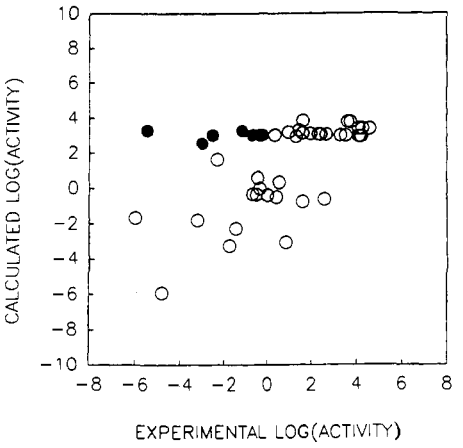

Figure 7. Functional group model cross-validated predictions of CBG binding affinities. Filled circles indicate poorly predicted values discussed in the text.

between +2 and +4. Six of these compounds have a single unique substituent, and the regression coefficient for each of these substituents is included in the original 90% functional group correlation model, accounting in substantial part for the good performance of that model equation. The removal of the seven outliers from the multilinear regression equation improves R^2 from 0.287 to 0.585. However, this action is tantamount to presuming that a unique substituent obviates the use of the functional group model for prediction. We surmise that the performance of any model equation with a large number of unique regressors should be tested by cross-validation before acceptance for predictive purposes.

Concluding Remarks

We have presented some developments of novel procedures that can be used to consider the general problem of defining molecular similarity and have used these procedures to develop an unusual type of structure–activity relationship involving the similarity indices as independent variables. The methodology has been applied to correlate data for the binding constants of steroids to human CBG. The linear regression models obtained in this work based on the similarity analysis concept give very good correlations of the experimental binding data. This is a promising result since most previous extensive studies of structure–activity relationships in steroids have been limited to qualitative classification as active or inactive (low or high potency), in some cases due to the qualitative nature of the available data.[34]

As one expects, the actual predictive power of the similarity concept method is not as satisfactory as the correlative performance. However, a comparison with cross-validated predicted results from the conventional functional group analysis of the binding data shows that the predictive mode of the similarity model approach is considerably improved and more useful. The presence of a large number of unique substituents in the regression equations of the functional group model seems to be a main factor responsible for the degradation in the quality of true predicted values of binding data. Additional evidence is required, but these results indicate that a straightforward functional group model may not be as useful as the more abstract similarity analysis in quantifying structure–activity relationships.

A study of molecular structure relationships and steroid binding data using a technique termed comparative molecular field analysis (CoMFA) has recently been carried out for a set of 31 compounds.[35] Part of this CoMFA study utilizes a subset of the CBG binding data given in Table I, and both the correlative and predictive attributes of the TN1 similarity and the CoMFA approaches seem to be comparable for this data. The main difference between the two methods is that the CoMFA method deals with the molecular structure as represented by the intersections of a preselected orientation of a three-dimensional lattice with calculated steric and electrostatic fields, whereas the present similarity analysis is based on the more simplistic molecular graph, little more than a drawing of the structure of the molecule. Of course, the added complexity of the CoMFA method could prove to have advantages, particularly in identifying new unrelated structural types that are effective for a specific application. It is difficult to see how the molecular graph similarity analysis could allow extrapolation outside of a group of congenerically related compounds.

The main purposes of the present work were to test concepts and definitions of similarity based on the molecular graph notation system and to formalize procedures to utilize the quantitative definitions of similarity. The results are encouraging, and we conclude that the overall similarity analysis procedure shows promise for development as a general QSAR tool when applied to groups of structurally related compounds. We note that this restriction to congeneric sets applies to the majority of applications of QSAR. Tentatively, we propose that definitions of molecular similarity based on structural notation similarity could be used in a laboratory environment to select promising alterations in molecular structure for an application under consideration. This would, of course, only be possible after a sufficient number of cases had been previously studied to provide the starting data for analysis. We are in the process of additional tests to establish the generality and limitations of this approach.

Acknowledgment. The financial support of the Welch Foundation (Houston, TX) and of the University of Texas at El Paso Materials Research Center of Excellence (a component of the National Science Foundation Minority Research Centers of Excellence) is gratefully acknowledged.

Registry No. 1, 50-23-7; **2**, 595-18-6; **3**, 50-24-8; **4**, 3836-17-7; **5**, 432-34-8; **6**, 50-03-3; **7**, 53-06-5; **8**, 1719-79-5; **9**, 116-58-5; **10**, 600-67-9; **11**, 50-22-6; **12**, 595-77-7; **13**, 152-58-9; **14**, 566-01-8; **15**, 604-28-4; **16**, 604-20-6; **17**, 604-19-3; **18**, 80-75-1; **19**, 438-07-3; **20**, 68-96-2; **21**, 67069-27-6; **22**, 302-23-8; **23**, 630-56-8; **24**, 64-85-7; **25**, 520-85-4; **26**, 2868-02-2; **27**, 516-15-4; **28**, 57-83-0; **29**, 1236-09-5; **30**, 566-65-4; **31**, 128-23-4; **32**, 145-13-1; **33**, 128-20-1; **34**, 2636-91-1; **35**, 903-71-9; **36**, 1239-79-8; **37**, 472-54-8; **38**, 3090-78-6; **39**, 14778-11-1; **40**, 52-39-1; **41**, 50-02-2; **42**, 34184-77-5; **43**, 2126-37-6; **44**, 58-22-0; **45**, 1045-69-8; **46**, 434-22-0; **47**, 50-28-2.

(34) For an example and leading references see: Stouch, T. R.; Jurs, P. C. *J. Med. Chem.* **1986**, *29*, 2125.

(35) Cramer, R. D.; Patterson, D.; Bunce, J. *J. Am. Chem. Soc.* **1988**, *110*, 5959.