

Computer-Aided Estimation of Synthetic Compounds Similarity with Endogenous Bioregulations

Yulia Borodina, Dmitrii Filimonov and Vladimir Poroikov*

Institute of Biomedical Chemistry Russian Academy of Medical Science, Pogodinskaya Street, 10, 119832, Moscow, Russia;
E-mail: borodina@ibmh.msk.su

Abstract

A new method for substance-to-substance similarity analysis based on topoelectric indices is described. Prediction accuracy of the method is tested using LOO cross-validation procedure on 910 biologically active compounds. The results demonstrate high discriminative ability of proposed structure description and measure of similarity.

The method is applied to compare the set of synthetic substances with well-known endogenous bioregulators. Average accuracy of active compounds recognition is 76%. Therefore, the proposed topoelectric indices and similarity measure can be used for estimating the synthetic molecules resemblance to small endogenous bioregulators.

1 Introduction

Humoral regulation of physiological processes is provided by high specific endogenous substances such as hormones, mediators, coenzymes, etc. Structural similarity of a synthetic compound with an endogenous bioregulator causes two probable consequences: 1) a compound may become a new lead; 2) a compound may have specific side/toxic effects. Thus, it is important to estimate the structural resemblance of synthetic compounds to endogenous bioregulators that allow predicting the most probable biological effects of the compounds. The number of known endogenous ligands is more than 200 [1]. Commercial and in-house databases contain hundreds of thousands of synthetic drug-like substances. Therefore, the actual task is development of method for fast and efficient computerized comparison of compounds from available databases with known endogenous bioregulators.

Methods, widely used for similarity-searching in 2-D databases, can be conditionally divided into two groups by different structure description: 1) fragment methods [2–5]; 2) graph-based methods [3, 6–10]. The main limitation of fragment methods is loss of structure integrity. Graph-based algorithms preserve integrity of structure but most of them concentrate on topology of molecule and do not take into account the particular properties of atoms [11].

Here we present a new method of substance-to-substance similarity analysis based on topoelectric description of molecule. Proposed method is tested by comparison of 32 small endogenous bioregulators' similarity with 910 synthetic analogs. The purpose of this work is to evaluate the possibilities and limitations of the method to recognize the endogenous-like substances within small organic compounds.

2 Material and Methods

2.1 Topoelectric indices

We use traditional 2-D representation of molecule that includes the list of atoms and list of bonds. Usually hydrogen atoms are omitted. In contrast, we include all hydrogen atoms and neglect atoms' charges and bonds' types. So a molecule is represented by adjacency matrix C and table of atomic properties P .

The adjacency matrix C has dimension $N \cdot N$, where N is the number of atoms in a molecule, and contains digits 0 and 1:

* To receive all correspondence

Key words: Drug-like organic compounds, endogenous bioregulators, molecular similarity estimates, topoelectric indices

Abbreviations: LOO Leave-One-Out procedure; TEM topoelectric matrix; TEI topoelectric index; IAP Independent Accuracy of Prediction; IAR Independent Accuracy of Recognition; 5-HT 5-hydroxytryptamine; GABA γ -aminobutyric acid; PAF platelet aggregation factor; n_1 the number of active compounds in the evaluation set; n_0 the number of inactive compounds in the evaluation set; p_i electronegativity; q_i equilibrium atom charge; $s(k1, k2)$ similarity between compounds $k1$ and $k2$.

$C_{ik} = C_{ki} = 1$ when i th and k th atoms are connected by bond.

In general, the properties matrix P may contain any values which characterize the atoms' specificity. Fundamental matrix of properties is presented by $A_{iz} = 1$ when atom i has atomic number z and $A_{iz} = 0$ otherwise.

In this work we characterize each atom i by its electronegativity p_i and equilibrium charge q_i which are determined on the following basis. The energy of the isolated atom can be approximated by quadratic function of its charge Q as $E(Q) = E_0 + pQ + bQ^2$. In this case the first ionization potential I_1 , the electron affinity E_a and parameters p and b are equal to:

$$I_1 = E(+1) - E(0) = p + b; \quad p = \frac{I_1 + E_a}{2};$$

$$E_a = E(0) - E(-1) = p - b; \quad b = \frac{I_1 - E_a}{2};$$

It is obvious that p coincides with the Mulliken electronegativity. The equilibrium charge gives the minimum of energy and equal to:

$$q = \text{ArgMin}E(Q) = -\frac{p}{2b}$$

So atom i is characterized by values p_i and q_i :

$$p_i = \frac{I_{1i} + E_{ai}}{2}; \quad q_i = -\frac{p_i}{I_{1i} - E_{ai}}; \quad (1)$$

where:

I_i is the first ionization potential of i^{th} atom;
 E_{ai} is the electron affinity of the i^{th} atom.

Examples of p and q values calculated for some atoms are given in Table 1.

Table 1. Examples of calculated values of p and q .

Atom	H	C	N	O
p, eV	7.1802	6.5278	7.1152	7.9052
q	-0.5599	-0.6195	-0.5489	-0.6186
Atom	F	Cl	Br	I
p, eV	10.3501	8.1482	7.4616	6.6214
q	-0.7467	-0.8661	-0.9092	-0.9204

Our experience shows that for the molecules' similarity the following normalized values are more useful:

$$\hat{p}_i = \frac{p_i - \bar{p}}{\sigma_p}; \quad \bar{p} = \frac{1}{N} \sum_i^N p_i; \quad \sigma_p^2 = \frac{1}{N} \sum_i^N (p_i - \bar{p})^2$$

$$\hat{q}_i = \frac{q_i - \bar{q}}{\sigma_q}; \quad \bar{q} = \frac{1}{N} \sum_i^N q_i; \quad \sigma_q^2 = \frac{1}{N} \sum_i^N (q_i - \bar{q})^2 \quad (2)$$

General form of the independent from atoms' numeration function of C and P is $f(P^T \cdot G(C) \cdot P)$, where f is the scalar function of matrix argument, G is the matrix function. So, different parametrization of atoms' properties, different functions f and G may give many topological indices of a molecule with general formula $f(P^T \cdot G(C) \cdot P)$.

Any analytical function can be approximated by the power series. For this reason we propose the following topoelectric indices.

Each molecule is characterized by the set of topoelectric matrices $\{TEM_1, TEM_2, \dots, TEM_{n_{\max}}\}$, calculated as:

$$TEM_n = \frac{\hat{P}^T \cdot C^n \cdot \hat{P}}{N}, \quad n = 1, 2, \dots, n_{\max} \quad (3)$$

where N is the total number of atoms in a molecule;
 C^n is the n -times product of matrix G ;
 n_{\max} is the maximal power of matrix C , taken into account;
 \hat{P} is the normalized matrix of properties, i th row of which is the vector $(1, \hat{p}_i, \hat{q}_i)$ (see Eq. 2).

The matrix \hat{P} has dimension $N \cdot 3$. Elements of matrix C^n correspond to the number of all ways of length n , including arches passed repeatedly.

In further calculations all elements of matrices $TEM_1, \dots, TEM_{n_{\max}}$, except the first elements, are used as a united set of topoelectric indices $\{TEI_1, TEI_2, \dots, TEI_m\}$, $m = 5 \cdot n_{\max}$.

The examples of such representation for three structures are given in Figure 1.

2.2 Molecule Similarity

The coefficient of similarity between compounds $k1$ and $k2$ is given by:

$$s(k1, k2) = \frac{1}{1 + \frac{1}{m} \sum_i^m [TEI_i(k1) - TEI_i(k2)]^2}, \quad (4)$$

where:

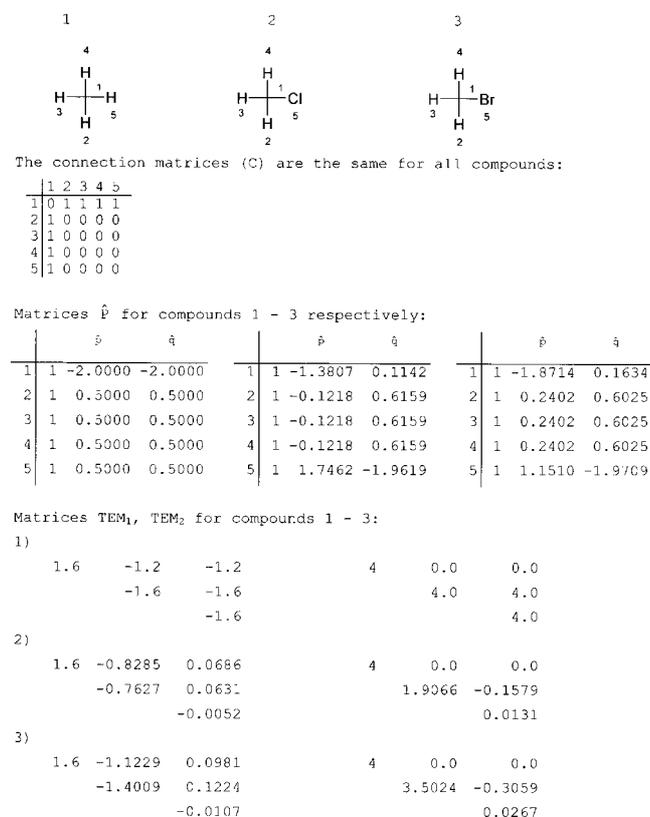


Figure 1. The examples of 3 different structure representation by the C and \hat{P} matrices.

$TEI_i(k1)$ and $TEI_i(k2)$ is the topoelectric index i for compounds $k1$ and $k2$ respectively;

m is the total number of indices and equals to $5 \cdot n_{\max}$.

Similarity coefficients of compounds 1–3 from Figure 1 are given in Table 2.

Although compounds 1, 2 and 3 differ each from the others by only one atom, $s(2, 3)$ is significantly more than $s(1, 2)$ or $s(1, 3)$ because atoms Br and Cl are more similar than H and Cl or H and Br. Thus, the example illustrates high sensitivity of the proposed similarity measure.

Table 2. The similarity coefficients of compounds 1–3.

	1	2	3
1	1	0.18	0.19
2		1	0.77
3			1

2.3 Comparison with the other methods for similarity assessment

We compare the discrimination power of the proposed method with four other methods considered by Basak and Grunwald [10]. They predicted mutagenicity of 15 nitrosamines based on the similarity assessment between the pairs of compounds by using graph invariants. We calculate our coefficient of similarity (4) for all pairs of compounds from the same data set. Following the approach of [10], the mutagenicity of each compound was estimated on the basis of this characteristic for the nearest neighbour substance. Correlation coefficients (r) and standard errors (s.e.) between the estimated and real values are used to assess the relative quality of the method for mutagenicity prediction. These results are compared with data for four methods from [10] in Table 3. It is shown that our method predicts mutagenicity better than 3 methods considered by Basak and Grunwald but is slightly worse comparing to the atom pairs (AP) approach [10]. The r and s.e. values equal to (0.944, 1.26) and (0.931, 1.43) for AP and our method respectively. Thus, our method can be applied to estimation of the biological activity of compounds.

2.4 Similarity Patterns and Evaluation data set

Evaluation set is selected from MDDR 96.1 database (MDL Information Systems, Inc.), which contains the structures and biological activities for 73707 compounds. 92.8% of them are under biological testing, 6.5% are drug candidates and 0.6% are registered drugs. The compound has been included into the evaluation set if: (1) it is agonist of the endogenous bioregulator used as the similarity pattern; (2) it should be described in details in the field "Action" of MDDR database. Thus, only compounds with activity confirmed in experiments

Table 3. Comparison of five similarity methods to select analogs for prediction of mutagenicity for 15 nitrosamines.

Similarity method	r	s.e.	p
AP ^a	0.944	1.26	< 0.0001
$TEI_{n_{\max}} = 2$	0.931	1.43	< 0.0001
TI_u^a	0.923	1.47	< 0.0001
PC_s^a	0.830	2.33	< 0.0001
TI_s^a	0.740	2.67	< 0.0016

^a the result obtained by S. Basak and G. Grunwald [10].

have been extracted from the MDDR. Evaluation set obtained in this way includes 910 agonists for 16 receptors of various endogenous bioregulators. The structures of 32 known endogenous bioregulators that interact with these 16 receptors are used as the patterns for comparison with synthetic compounds from the evaluation set. The composition of the evaluation set and names of appropriate endogenous bioregulators are given in Table 4.

Both evaluation set and the set of endogenous bioregulators are presented in the form of SD files, including structure and list of activity classes for each compound. Each structure is then coding as set of topoelectric indices TEI_1, \dots, TEI_m ; $m = 5 \cdot n_{\max}$; $n_{\max} = 7$ (all designations are explained above). The mean CPU time required to calculate the indices for one structure in PC Pentium 100 MHz is 0.05 s. Thus, the method can be used for estimating similarity of compounds in large databases.

3 Results and Discussion

3.1 Cross-validation of the method

The proposed method is tested by LOO cross-validation on the evaluation set. The similarity of each one compound

Table 4. The evaluation set and appropriate endogenous bioregulators.

NN	Activity	Number of comps. in the evaluation set	Pattern compounds
1	Adenosine Agonist	29	Adenosine
2	Adrenergic Agonist	72	Adrenaline; Noradrenaline
3	Dopamine Agonist	103	Dopamine
4	Benzodiazepine Agonist	43	β -Carboline-3-carboxylic acid ethyl ester
5	GABA B Agonist	4	GABA
6	5-HT Agonist	262	Serotonin
7	Melatonin Agonist	7	Melatonin
8	Muscarinic Agonist	137	Acetylcholine
9	Androgen	3	5 α -dihydrotestosterone; Testosterone
10	Estrogen	21	Estradiol-17 α ; Estriol; Estrone
11	Progestin	10	17-Hydroxyprogesterone; Pregnenalone; Progesterone
12	Corticosteroid	8	Aldosterone; Deoxycorticosterone; Cortizol;
13	Prostaglandin	91	PGE ₁ ; PGE ₂ ; PGA ₁ ; PGA ₂ ; PGF _{1α} ; PGF _{2α}
14	Vitamin D Analog	65	Ergocalciferol; Ergosterol
15	Retinoid	49	Retinal; Retinat; Retinol
16	PAF Analog	1	PAF

with every other compounds in the dataset is calculated. The selected compound (k) is considered as compound with “unknown” activity, and its probability to be active is estimated as:

$$\Pr(k) = \frac{\frac{1}{n_i} \sum_{i \neq k}^{n_1} s_i^1}{\frac{1}{n_1} \sum_{i \neq k}^{n_1} s_i^1 + \frac{1}{n_0} \sum_{j \neq k}^{n_0} s_j^0}, \quad (5)$$

where: s_i^1 and s_j^0 are the estimated similarities of compound k with i th active and j th inactive compound respectively, according to the Eq. 4. n_{\max} is equal to 7; n_1 and n_0 are the numbers of active and inactive compounds in the evaluation set respectively.

This procedure is repeated for each compound from evaluation set, and the resulting values Pr then form the basis for estimating the measure of activity prediction’s quality.

We estimate the prediction quality by criterion of Independent Accuracy of Prediction:

$$IAP = \frac{N\{\Pr_1 > \Pr_0\}}{n_1 \cdot n_0}, \quad (6)$$

where $N\{\Pr_1 > \Pr_0\}$ is the number of cases when Pr for active compound is more than Pr for inactive compound, when all pairs of active and inactive compounds are compared. We call this criterion “independent” because it does not depend on any additional assumptions concerning the parent population and a risk function.

Table 5. IAP (Independent Accuracy of Prediction) estimated by cross-validation of evaluation set.

Activity	IAP, %	n_1	n_0
Vitamin D Analog	97.7	65	845
Androgen	96.1	3	907
Prostaglandin	94.4	91	819
Adenosine Agonist	94.1	29	881
Progestin	93.6	10	900
Estrogen	89.5	21	889
GABA Agonist	88.5	4	906
Retinoid	88.2	49	861
Melatonin Agonist	84.2	7	903
Muscarinic Agonist	83.0	137	773
Adrenergic Agonist	82.0	72	838
Benzodiazepine Agonist	81.8	43	867
Corticosteroid	80.3	8	902
5-HT Agonist	73.5	262	648
Dopamine Agonist	72.9	103	807
PAF Analog	—	1	909
In average:	86.7		

The IAP values for 16 kinds of activity are given in Table 5. Average value of IAP equals 87% that is satisfactory to apply the method in similarity assessment.

3.2 Analysis of synthetic compounds similarity with endogenous bioregulators

We calculate the similarity of each compound from the evaluation set with every of 32 pattern ligands.

Recognition quality by criterion of Independent Accuracy of Recognition is estimated as:

$$IAR = \frac{N\{s_1 > s_0\}}{n_1 \cdot n_0}, \quad (7)$$

where: $N\{s_1 > s_0\}$ is the number of cases when the active compound is more similar to the endogenous molecule than inactive one;

n_1 and n_0 are the numbers of active and inactive compounds in the evaluation set.

The criterion is analogous to the IAP described above. If there are several endogenous patterns for one kind of activity, the overall recognition accuracy is determined by averaging the partial accuracy values. Accuracies of active compound's recognition for the evaluation set are given in Table 6.

It is clear from Table 6 that the average accuracy of recognition equal to 76%. The best results (98.8 and 98.5%) are shown for PAF Analogs and Androgens, the worst (56.2 and 54.6%) for 5-HT agonists and Dopamine Agonists. The last result is probably caused by any of two reasons: (1) significant structural similarity of dopamine D_2 and 5- HT_{1A}

Table 6. IAR (Independent Accuracy of Recognition) estimated on the basis of comparison with endogenous regulators.

Activity	IAR, %
PAF Analog	98.8
Androgen	98.5
Vitamin D Analog	84.1
Prostaglandin	83.7
Progesterin	83.3
Melatonin Agonist	80.0
Retinoid	79.5
Corticosteroid	78.4
Adenosine Agonist	75.6
Estrogen	75.0
Adrenergic Agonist	74.3
Benzodiazepine Agonist	65.5
Muscarinic Agonist	65.5
GABA Agonist	63.7
5-HT Agonist	56.2
Dopamine Agonist	54.6
In average:	76.0

agonists makes difficult the discrimination between them [12–14]; (2) similarity of dopamine and serotonin with some synthetic analogs cannot be explained on the basis of 2-D structure representation [15–19]. The examples of 5-HT agonists, which are the most similar and the least similar with serotonin molecule are given in Figure 2.

4 Conclusions

A new method for direct analysis of substance-to-substance similarity is developed. LOO cross-validation on the evaluation set of 910 compounds from MDDR database shows that the average accuracy of prediction is 86.7%.

The method is applied to estimate the similarity of 910 MDDR compounds with 32 endogenous bioregulators. Average accuracy of active compounds recognition is 76%.

Therefore, the proposed topoelectric indices and similarity measure can be used for efficient comparison of synthetic molecules with majority of small endogenous bioregulators.

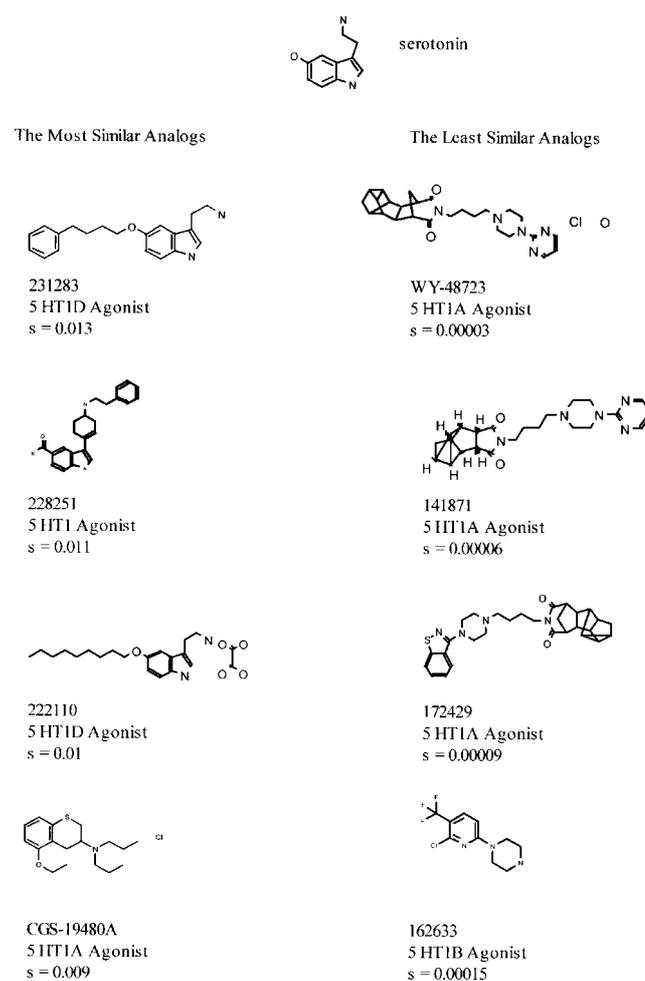


Figure 2. 5-HT Agonists, which are the most similar and the least similar with serotonin molecule. S is the similarity coefficient.

This attempt to compare synthetic substances with endogenous ligands is obviously a first step in application the proposed method, that nevertheless demonstrates the satisfactory recognition of active compounds and the possibility to estimate their probable therapeutic and toxic effects on this basis.

Acknowledgments

We gratefully acknowledge MDL Information Systems, Inc. for providing the ISIS/Base, ISIS/Host software and MDDR database. These were used both in this research and in education of graduate and post-graduate students from Russian State Medical University.

This work is supported in part by the Russian Ministry of Science and Technical Politics (Scientific Program "R & D of New Drugs by the Methods of Chemical and Biological Synthesis", Branch 04 "Computer Aided Drug Design", Project No. 04.01.13/96).

5 References

- [1] Tkachuk, V.A., Ligands and Receptors: Intracellular Signaling from Classic to Avant-garde, Abstracts of the II congress of Biochemical Society of Russian Academy of Science, Moscow, 1997, p. 14.
- [2] Willet, P., Similarity-searching and clustering algorithms for processing databases of two-dimensional and three-dimensional chemical structures. In Dean, P.M. (Ed.), *Molecular Similarity in Drug Design*, Blackie Academic & Professional, 1995, pp. 111–137.
- [3] Cheng, C., Maggiora, G., Lajiness, M. and Jonson, M., Four Association Coefficients for Relating Molecular Similarity Measures. *J. Chem. Inf. Comput. Sci.* 36, 909–915 (1996).
- [4] Dittmar, P.G., Farmer, N.A., Fisanick, W., Haines, R.C. and Mockus, J., The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* 23, 93–102 (1983).
- [5] Fisanick, W., Lipkus, A.H. and Rusinko III, A., Similarity Searching on CAS Registry Substances. 2. 2D Structural Similarity. *J. Chem. Inf. Comput. Sci.* 34, 130–140 (1994).
- [6] Randic, M., Similarity based on Extended Basis Descriptors. *J. Chem. Inf. Comput. Sci.* 32, 686–692 (1992).
- [7] Takahashi, Y., Sukekawa, M. and Sasaki, S., Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structure. *J. Chem. Inf. Comput. Sci.* 32, 639–643 (1992).
- [8] Basak, S.C., Bertelsen, S. and Grunwald, G.D., Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci.* 34, 270–276 (1994).
- [9] Cummins, D.J., Andrews, C.W., Bentley, J.A. and Cory, M., Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* 36, 750–763 (1996).
- [10] Basak, S.C. and Grunwald, G.D., Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR and QSAR in Environmental Research*. 2, 289–307 (1994).
- [11] Milne, G.W.A., Mathematics as a Basis for Chemistry. *J. Chem. Inf. Comput. Sci.* 37, 639–644 (1997).
- [12] Wikstrom, H., Andersson, B., Elebring, T. and Lagerkvist, S., 6-Hydroxy-3-n-propyl-2,3,4,5-tetrahydro-1H-3-benzazepine and Analogs: New Carefully Acting 5-HT_{1A} Receptor Agonists. *J. Med. Chem.* 35, 3984–3990 (1992).
- [13] Lin, C.-H., Haadsma-Svensson, S.R., Lahti, R.A., McCall, R.B., Piercey, M.F., Schreur, P.J., Von Voigtlander, P.F., Smith, M.W. and Chidester, C.G., Centrally Acting Serotonergic and Dopaminergic Agents. 1. Synthesis and Structure-Activity Relationships of 2,3,3a,4,5,9b-Hexahydro-1H-benz[e]indole Derivatives. *J. Med. Chem.* 36, 1053–1068 (1993).
- [14] Chidester, C.G., Lin, C.-H., Lin, C.-H., Lahti, R.A., Haadsma-Svensson, S.R. and Smith, M.W., Comparison of 5-HT_{1A} and Dopamine D₂ Pharmacophores. X-ray Structure and Affinities of Conformationally Constrained Ligands. *J. Med. Chem.* 36, 1301–1318 (1993).
- [15] Van de Waterbeemd, H., Carrupt, P.-A. and Testa, B., The Electronic Structure of Dopamine. An ab initio Electrostatic Potential Study of the Catechol Moiety. *Helv. Chim. Acta* 68, 715–723 (1985).
- [16] Bottcher, H., Barnickel, G., Hausberg, H.-H., Haase, A.F., Seyfried, C.A. and Eiermann, V., Synthesis and Dopaminergic Activity of Some 3-(1,2,3,6-Tetrahydro-1-pyridylalkyl) indoles. A Novel Conformational Model To Explain Structure-Activity Relationships. *J. Med. Chem.* 35, 4020–4026 (1992).
- [17] Buchheit, K.-H., Gamse, R., Giger, R., Hoyer, D., Klein, F., Kloppner, E., Pfannkuche, H.-J. and Mattes, H., The Serotonin 5-HT₄ Receptor. 1. Design of a New Class of Agonists and Receptor Map of the Agonist Recognition Site. *J. Med. Chem.* 38, 2326–2330 (1995).
- [18] Nelson, D.L., Structure-Activity Relationships at 5-HY_{1A} Receptors: Binding Profiles and Intrinsic Activity. *Pharmacol. Biochem. Behav.* 40, 1041–1051 (1991).
- [19] Evans, S.M., Galdes, A. and Gall, M., Molecular Modeling of 5-HT₃ Receptor Ligands. *Pharmacol. Biochem. Behav.* 40, 1033–1040 (1991).

Received on December 8, 1997; accepted on May 5, 1998