Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics

VELLARKAD N. VISWANADHAN, ARUP K. GHOSE,* GANAPATHI R. REVANKAR, and ROLAND K. ROBINS

Department of Molecular Modeling and Computers, Nucleic Acid Research Institute, 3300 Hyland Avenue, Costa Mesa, California 92626

Received June 29, 1988

We have shown previously that atomic values for physicochemical properties are an important guide for correlating the observed biological activity of the ligands to their chemical structure (Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 80-90, and references cited therein). The objective of the present work is to (i) report the hydrophobicity and the molar refractivity for phosphorus and selenium atoms at different structural environments that are ubiquitous in biologically active systems, (ii) refine the atomic values of the various elements reported earlier to satisfy the largely extended data set, and (iii) suggest a method for selecting the best superposition of different molecules on a reference structure using these atomic physicochemical properties. The octanol-water partition coefficient was used to scale the atomic hydrophobicity. The hydrophobicity values of 120 atom types were evaluated from 893 compounds. The observed and calculated octanol-water partition coefficient showed a correlation coefficient of 0.926 and a standard deviation of 0.496. The atomic refractivity values were evaluated from the molar refractivities of 538 compounds; the corresponding correlation coefficient and standard deviation were 0.999 and 0.774, respectively. The atomic values were tested by predicting the respective properties for a large number of compounds. The superposition method has been applied to certain naturally occurring nucleoside antibiotics. The algorithm presented here shows various important superpositions of two or more molecules with minimum physical assistance to avoid any personal bias.

INTRODUCTION

An important step in drug action is the interaction of the drug with a biological receptor. However, the direct study of the drug (ligand)-receptor interaction by molecular mechanics and dynamics[1,2] is not feasible in most cases because the receptor structure or the binding site is unknown. The QSAR[3] approach deals with the situation indirectly. It correlates the biological activity of the ligands with their structural or physicochemical properties and extends the correlated properties for the prediction of new active ligands. In the linear free energy relationship (LFER) approach[3] (the Hansch approach), physicochemical properties of the ligands (bioactive compounds) are used in multiparametric regression models for correlating with biological activity. Such approach was acceptable when the undertaking of the three-dimensional structure of the ligand was computationally too expensive. However, the rapid improvement of the computational facility prompted many to develop methods for three dimensional structure directed quantitative structure-activity relationships.[4-7] Application of these techniques for the three-dimensional mapping of an unknown receptor site cavity requires not only analyses of conformational flexibilities in the ligands but also a knowledge of the physicochemical characteristics of the ligands. The most important properties related to biomolecular interactions are hydrophobicity,[8] formal charge density,[9-11] and molar refractivity.[12] The basic hypotheses for receptor mapping are that (i) the interactions at different parts of the ligands with the receptor are different and (ii) the nature of interaction can be determined by correlating the local physicochemical properties with their binding free energies. To exploit the second hypothesis, we need to know the relative orientation of the ligands at the binding site (binding modes) and a method that can estimate the physicochemical properties of the ligands at any region from the occupancy of the atoms. Previously we showed[13-15] how atomic physicochemical properties can be developed and used to estimate local or overall physicochemical properties. The primary objective of the present work is to give the hydrophobicity (octanol-water partition coefficient) and the molar refractivity for phosphorus and selenium atoms in different structural environments since these atoms are found in many biologically important nucleosides and nucleotides. The atomic values of the various types of atoms that were reported earlier have been refined to satisfy the largely extended data set. The secondary objective of the work is to show a method of estimating the goodness of fit of various geometrically feasible superpositions of a molecule on a reference structure on the basis of physicochemical property matching, using these atomic parameters. Such superposition is often helpful in determining the molecular similarity and in rationalizing biological activity of compounds in diverse structure.

It is not very straightforward to decide what physicochemical properties should be considered to determine the best molecular superposition. Since understanding biological activity is the ultimate objective of our work, we should consider the properties that govern the interaction of the ligand (drug) molecules with the biological receptor.[16] In the absence of any information regarding the structure of the binding site, we considered three properties: hydrophobicity,[8] molar refractivity,[12] and formal charge density.[11] The term hydrophobicity refers to the force or corresponding energy that operates between two or more nonpolar solutes in water and arises from dispersive and electrostatic forces and the consequent entropic factor. A hydrophobic substance is soluble in nonpolar solvents but only sparingly soluble in water. Though the hydrophobic effect plays an important role in biological systems, it is not well understood theoretically. A great deal of effort is nec-

0095-2338/89/1629-0163$01.50/0

Table I. Classification of Atoms and Their Contributions to Octanol–Water Partition Coefficient Which Is a Measure of Hydrophobicity and Molar Refractivity

type	description[a]	hydrop[b]	no. of compounds	freq of use	atomic refrac	no. of compounds	freq of use
	C in						
1	:CH_3R, CH_4	−0.6771	385	589	2.9680	283	495
2	:CH_2R_2	−0.4873	245	506	2.9116	169	389
3	:CHR_3	−0.3633	46	51	2.8028	33	37
4	:CR_4	−0.1366	24	24	2.6205	11	11
5	:CH_3X	−1.0824	196	299	3.0150	76	114
6	:CH_2RX	−0.8370	302	485	2.9244	199	320
7	:CH_2X_2	−0.6015	6	6	2.6329	14	15
8	:CHR_2X	−0.5210	87	152	2.5040	42	51
9	:$CHRX_2$	−0.4042	36	36	2.3770	15	19
10	:CHX_3	0.3651	4	4	2.5559	5	5
11	:CR_3X	−0.5399	15	15	2.3030	15	16
12	:CR_2X_2	0.4011	3	3	2.3006	16	25
13	:CRX_3	0.2263	36	38	2.9627	12	15
14	:CX_4	0.8282	6	6	2.3038	4	4
15	:=CH_2	−0.1053	25	31	3.2001	38	44
16	:=CHR	−0.0681	48	70	4.2654	44	56
17	:=CR_2	−0.2287	9	10	3.9392	7	7
18	:=CHX	−0.3665	23	24	3.6005	11	11
19	:=CRX	−0.9188	13	13	4.4870	14	16
20	:=CX_2	−0.0082	5	6	3.2001	8	12
21	:≡CH	−0.1047	3	4	3.4825	12	13
22	:≡CR, R=C=R	0.1513	4	5	4.2817	18	23
23	:≡CX				3.9556	7	7
24	:R--CH--R	0.0068	584	2222	3.4491	187	896
25	:R--CR--R	0.1600	307	371	3.8821	107	144
26	:R--CX--R	−0.1033	432	737	3.7593	102	156
27	:R--CH--X	0.0598	92	142	2.5009	14	16
28	:R--CR--X	0.1290	66	73	2.5000	14	15
29	:R--CX--X	0.1652	70	85	3.0627	4	4
30	:X--CH--X	0.2975	23	23	2.5009	1	1
31	:X--CR--X	0.9421	7	7			
32	:X--CX--X	0.2074	14	14	2.6632	1	3
33	:R--CH···X	−0.1774	16	19	3.4671	19	22
34	:R--CR···X	−0.2782	39	44	3.6842	13	14
35	:R--CX···X	−0.3630	7	8	2.9372	4	5
36	:Al—CH=X	−0.0321	5	5	4.0190	15	15
37	:Ar—CH=X	0.3568	5	5	4.7770	15	15
38	:Al—C(=X)—Al	0.8255	27	31	3.9031	15	16
39	:Ar—C(=X)—R	−0.1116	23	29	3.9964	8	8
40	:R—C(=X)—X, R—C≡X, X=C=X	0.0709	289	356	3.4986	111	136
41	X—C(=X)—X	0.4571	107	115	3.4997	13	13
42	:X--CH···X	−0.1316	22	22	2.7784	6	6
43	:X--CR···X	0.0498	33	33	2.6267	4	4
44	:X--CX···X	0.1847	10	13	2.5000	2	2
45	unused	–	–	–	–	–	–
	H attached to[c]						
46	:C^0_{sp3}, having no X attached to next C	0.4418	280	1582	0.8447	185	1058
47	:C^1_{sp3}, C^0_{sp2}	0.3343	799	4252	0.8939	422	1982
48	:C^2_{sp3}, C^1_{sp2}, C^0_{sp}	0.3161	74	87	0.8005	68	93
49	:C^3_{sp3}, C^2_{sp2}, C^3_{sp2}, C^3_{sp}	−0.1488	138	209	0.8320	55	61
50	:heteroatom	−0.3260	603	1084	0.8000	107	148
51	:α-C[d]	0.2099	214	556	0.8188	123	366
52	C^0_{sp3}, having 1 X attached to next carbon	0.3695	218	790	0.9215	197	883
53	C^0_{sp3}, having 2 X attached to next carbon	0.2697	11	25	0.9769	24	83
54	C^0_{sp3}, having 3 X attached to next carbon	0.3647	3	7	0.7701	1	3
55	C^0_{sp3}, having 4 or more X attached to next carbon						
	O in						
56	:alcohol	0.1402	98	160	1.7646	20	22
57	:phenol, enol, carboxyl OH	0.4860	165	192	1.4778	35	40
58	:=O	−0.3514	464	638	1.4429	187	220
59	:Al—O—Al	0.1720	39	43	1.6191	30	44
60	:Al—O—Ar, Ar_2O R···O···R, R—O—C=X	0.2712	212	289	1.3502	146	217
61[e]	:--O	1.5810	81	178	1.9450	21	45
62–63	unused						
	Se in						
64	:Any—Se—Any	0.1473	12	12	11.1366	2	2
65	:=Se				13.1149	6	6
	N in						
66	:Al—NH_2	0.1187	24	24	2.6221	10	11
67	:Al_2NH	0.2805	23	25	2.5000	9	9
68	:Al_3N	0.3954	18	20	2.8980	8	8
69	:Ar—NH_2, X—NH_2	0.3132	84	90	3.6841	9	12

Table I (Continued)

type	description[a]	hydrop[b]	no. of compounds	freq of use	atomic refrac	no. of compounds	freq of use
70	:Ar—NH—Al	0.4238	10	10	4.2808	7	7
71	:Ar—NAl_2	0.8678	17	17	3.6189	10	11
72	:RCO—N<, >N—X=X	−0.0528	297	393	2.5000	18	19
73	:Ar_2NH, Ar_3N Ar_2N—Al, R···N···R[f]	0.4198	87	89	2.7956	7	7
74	:R≡N, R=N—	0.1461	62	90	2.7000	27	29
75	:R--N--R,[g] R--N--X	−0.1106	170	251	4.2063	24	27
76	:Ar—NO_2, R--N(--R)--O[h] RO—NO_2	−2.7640	75	87	4.0184	15	17
77	:Al—NO_2	−2.7919	6	6	3.0009	6	6
78	:Ar—N=X, X—N=X	0.5721	40	53	4.7142	10	12
79–80	unused						
	F attached to						
81	:C^1_{sp3}	0.4174	5	5	0.8725	8	8
82	:C^2_{sp3}	0.2167	8	14	1.1837	7	32
83	:C^3_{sp3}	0.2792	34	103	1.1573	7	28
84	:C^1_{sp2}	0.5839	17	26	0.8001	21	34
85	:C^{2-4}_{sp2}, C^1_{sp} C^4_{sp}, X	0.3425	1	2	1.5013	8	14
	Cl attached to						
86	:C^1_{sp3}	0.9609	20	27	5.6156	25	30
87	:C^2_{sp3}	0.5594	8	14	6.1022	16	28
88	:C^3_{sp3}	0.4656	15	37	5.9921	10	29
89	:C^1_{sp2}	0.9624	100	148	5.3885	25	28
90	:C^{2-4}_{sp2}, C^1_{sp} C^4_{sp}, X	0.6345	20	36	6.1363	32	42
	Br attached to						
91	:C^1_{sp3}	1.0242	12	13	8.5991	21	25
92	:C^2_{sp3}	0.4374	3	4	8.9188	10	21
93	:C^3_{sp3}	0.4332	2	4	8.8006	3	9
94	:C^1_{sp2}	1.2362	39	49	8.2065	14	14
95	:C^{2-4}_{sp2}, C^1_{sp} C^4_{sp}, X	0.9351	4	8	8.7352	9	9
	I attached to						
96	:C^1_{sp3}	1.4350	4	4	13.9462	7	8
97	:C^2_{sp3}				14.0792	4	7
98	:C^3_{sp3}				14.0730	3	3
99	:C^1_{sp2}	1.7018	14	14	12.9918	5	5
100	:C^{2-4}_{sp2}, C^1_{sp} C^4_{sp}, X	0.9336	1	3	13.3408	1	1
101–105	unused halogens						
	S in						
106	:R—SH	0.7268	10	10	7.8916	9	10
107	:R_2S, RS—SR	0.6145	39	42	7.7935	19	22
108	:R=S	0.3828	25	25	9.4338	7	9
109	:R—SO—R	−0.1708	2	2	7.7223	5	5
110	:R—SO_2—R	0.3717	57	61	5.7558	8	8
111–114	unused						
	P in						
115	:ylids						
116	:R_3—P=X	−1.6251	1	1	5.5306	5	5
117	:X_3—P=X (phosphate)	0.3308	49	52	5.5152	13	14
118	:PX_3 (phosphite)				6.8360	10	10
119	:PR_3 (phosphine)				10.0101	2	2
120	:C—$P(X)_2$=X (phosphonate)	0.0236	3	3	5.2806	7	7

[a] R represents any group linked through carbon; X represents any heteroatom (O, N, S, P, Se, and halogens); Al and Ar represent aliphatic and aromatic groups, respectively; = represents double bond; ≡ represents triple bond; — represents aromatic bond as in benzene or delocalized bonds such as the N—O bond in nitro group; ··· represents aromatic single bonds as the C—N bond in pyrrole. [b] Atomic hydrophobicity in the unit of log *P*(octanol-water). [c] The subscript represents hybridization and the superscript its formal oxidation number. The formal oxidation number of a carbon atom = ∑ formal bond orders with electronegative atoms. [d] An α-C may be defined as a C attached through a single bond with —C=X, —C≡X, —C--X. [e] As in nitro, =*N*-oxides. [f] Pyrrole-type structure. [g] Pyridine-type structure. [h] Pyridine *N*-oxide type.

essary to understand this property in different systems such as the self-assembly of globular proteins.[17,18] The octanol–water partition coefficient is often used by medicinal chemists to model this property.[19,20] The second property of interest is the molar fractivity. A brief outline of the theory of atomic refractivity and dispersive force was presented in an earlier work.[12] The main reason in developing a set of molar refractivity parameters is that, to a fair degree of approximation, it is possible to represent the dispersive interaction between the ligand and the receptor as a linear function of the property of the ligand alone, as long as attention is confined to a small local region of the ligand. We showed previously[8] that the correlation coefficient between the atomic refractivities and octanol–water partition coefficient is only 0.432, which suggests that both these properties can be used simultaneously for this purpose. The third property, the formal charge density, can be obtained by molecular orbital calculation. This property determines the electrostatic interaction between the ligand and the receptor. An overall index of a property for a molecule or any part of it may then be obtained as the sum of its constituent atomic properties. Atomic parameters presented here have a clear computational advantage over the more conventional group parameters. The three-dimensional structure of a molecule is represented by the atomic locations,

and we are associating one or more numerical values (properties) with each atomic location.

It should, however, be noted that all physicochemical properties cannot be dissected at the atomic level. Purely atomic properties ought to be scalar, so vector properties, like the dipole moment, cannot be dissected at the atomic level. However, dipolar interaction can be handled by using formal atomic charge density.

METHOD OF CALCULATION

Classification of Atoms. The first step in evaluating atomic contributions to molecular properties is to classify the atomic states of elements into various atomic types. Here "atomic state" implies the oxidation and hybridized state and the influence of immediately bonded atoms. Commonly occurring atomic states of carbon, hydrogen, oxygen, nitrogen, halogens, sulfur, phosphorus, and selenium in organic molecules are represented by 120 atom types as shown in Table I. This includes the earlier set of 110 atom types and an additional set of 10 atom types of which 5 types are for phosphorus. These are ylids, phosphine oxides, phosphates, phosphites, and phosphonates. For selenium, two states were defined: one in which it is attached to two atoms and the other in which it is double bonded with one neighboring atom. Our main interest is, of course, the phosphate type and the double-bonded selenium since such groups are common in many synthetic nucleosides and nucleotides.

Since the constitutive factor of the property has been included by the classification of atoms into numerous types, the total molecular property can be estimated as the sum of individual atomic values in the molecule. This is given by

$$P_{\text{calcd}} = \sum n_i a_i \tag{1}$$

where P_{calcd} refers to either of the properties, molar refractivity (MR_{calcd}) or log of octanol–water partition coefficient (log P).

Preparation of Data. The preparation of data involves (i) collection of molecular physicochemical properties (octanol–water partition coefficients and molar refractivities), (ii) generating the bonding pattern for each structure, and (iii) classification of atoms to different types. The bonding pattern (structural topology) for each molecule is generated by using CHEMSTRUC,[15] a FORTRAN program with a simple interactive command structure similar to CAS ONLINE substructure generation that produces a computer graphic display of the covalent structure. The structural skeleton without the hydrogen atom is generated and the program assigns the necessary hydrogen atoms. Error checking was done by (i) displaying the structure on a graphics terminal, (ii) involving two persons independently in the process, and (iii) comparing the supplied and calculated total number of atoms. The output of the program is fed in as the input of another program, CLASIF,[15] which classifies the atoms in each structure into one of the 120 types (Table I).

Parameter Evaluation. A simple least-squares fitting technique is used to evaluate atomic contributions to hydrophobicity. For evaluating atomic contributions to molar refractivity (MR), we used quadratic programming.[12] Since the molar refractivity is related to the molar volume, the evaluated atomic refractivity should always be positive. A quadratic programming involves an optimization of a quadratic objective function, maintaining some linear constraints. In the present case the objective function is the least-squares function

$$F = \sum [MR_{\text{calcd}} - MR_{\text{obsd}}]^2 \tag{2}$$

where MR_{calcd} is given by eq 1 and the summation is over all molecules in the data set. The linear constraints are

$$a_i \geq l_i;\ i = 1, 2, \ldots, n \tag{3}$$

where l_i's are taken to keep the atomic parameters within physically realistic range. These limits require all atomic contributions to molar refractivity to be positive, and all types of an element in a particular oxidation and hybridized state have comparable magnitude of contributions.

Use of Atomic Parameters for Molecular Superposition. Given a reference structure, the present method gives a quantitative basis of judging the goodness of fit of a geometrically feasible superposition. By the term *reference structure* we mean a fixed conformation of a molecule. The geometrical and computational aspect of molecular superposition is a well-studied subject.[13,21-27] In Appendix I, an algorithm is presented that will allow us to determine the important geometrically feasible molecular superpositions. Each superposition in this approach is represented by a vector, $[j_1, j_2, \ldots, j_n]$, where n is the number of atoms in the reference structure and j_i represents the atom of the test molecule superimposed on the ith atom of the reference. The goodness of fit of a molecular superposition is given by a function, which depends on the physicochemical properties of the reference and superposed atoms

$$F_1 = \sum_k g_k[|x_k| - |x_k - x_{j_k}|] \tag{4}$$

where x_k represents the physicochemical property of the kth atom of the reference structure, j_k is the atom of the test molecule superimposed on the kth atom of the reference structure, and g_k is the weight factor associated with the kth atom as well as the physicochemical property. One immediate use of g_k is to keep selected atoms off the reference during superposition. The corresponding physicochemical property will be zero if no atom is superimposed. The above function assumes that the affinity of the ligand atoms with the receptor site is quadratic in nature and the reference structure lies on the peak. Under such condition the affinity will decrease if the physicochemical property changes on either side of the "ideal" value.

The molecular interactions in many places are linear with the physicochemical properties. Since the relative importance of the physicochemical properties at different regions of the receptor may not be known, it may be assumed that the reference structure experiences attraction from all of its atoms. In other words, the sign of the physicochemical property for attractive interaction is the sign of the corresponding physicochemical property of the reference structure. The goodness of superposition under such condition can be determined by the function

$$F_2 = \sum_k g_k[x_k/|x_k|]x_{j_k} \tag{5}$$

The first term represents the sign of the physicochemical property of the reference atom; the second term is the physicochemical property of the superposed atom.

The maximum possible value of the function F_1 is $\sum g_k|x_k|$, and it corresponds to the matching of the reference itself. The function F_2 does not have any such limit, and the test molecule can have higher value. If it is necessary to use more than one physicochemical property, sum may be taken over other properties as well. However, since an a priori decision about the relative importances of the different physicochemical factors (or the different atoms of the reference molecule) is difficult, they should be given equal weight. In other words, the parameters should be scaled, and g_k values may be considered as unity. For conventional autoscaling, the expression

$$x_{i,\text{new}} = [x_{i,\text{old}} - \mu_{\text{old}}]/s_{\text{old}} \tag{6}$$

is used, where μ_{old} is the mean of the old parameters and s_{old} is their variance. The new parameters have unity variance and zero mean. However, it is expected that the scaling should

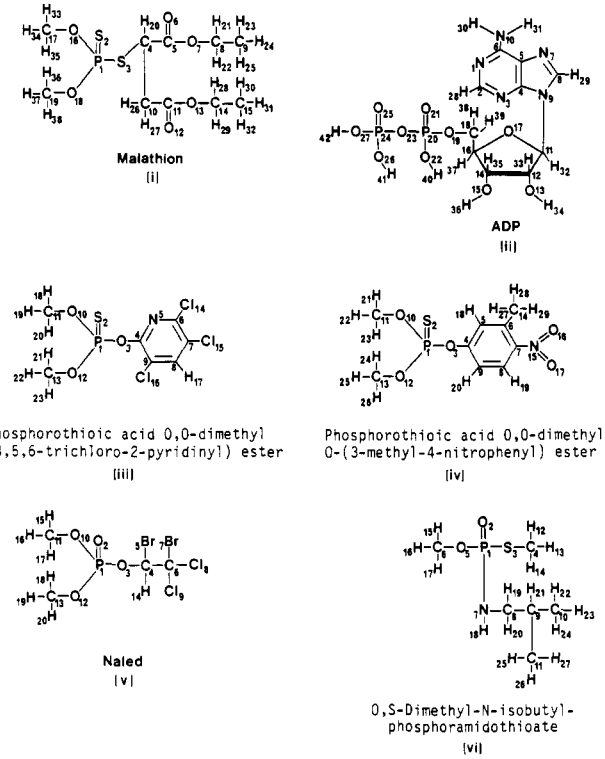

Figure 1. Set of representative molecules to illustrate the atom classification; see Table II.

Table II. Demonstration of Atom Classification Using Some Selected Molecules[a]

structure	atom classification
I, malathion	1(9,15), 2(10), 5(17,19), 6(8,14), 8(4), 40(5,11), 47(21,22,28,29,33-38), 51(20,26,27), 52(23-25,30-32), 58(6,12), 60(7,13,16,18), 107(3), 108(2), 117(1)
II, adenosine diphosphate	6(18), 8(12,14,16), 9(11), 28(5), 29(6), 30(2), 42(8), 43(4), 47(33,35,37,38,39), 48(32), 49(28,29), 50(30,31,34,36,40-42), 56(13,15), 57(22,26,27), 58(21,25), 59(17), 60(19,23), 69(10), 73(9), 75(1,3,7), 117(20,24)
III, methyl chloropyrifos	5(11,13), 24(9), 26(7,8), 29(4,6), 47(17-23), 60(3,10,12), 75(5), 89(15,16), 90(14), 108(2), 117(1)
IV, 3-methyl,dimethyl parathion	1(14), 5(11,13), 24(5,8,9), 25(6), 26(4,7), 46(27-29), 47(18-26), 60(3,10,12), 61(16,17), 76(15), 108(2), 117(1)
V, naled	5(11,13), 9(4), 13(6), 47(15-20), 48(14), 58(2), 60(3,10,12), 88(8,9), 92(5), 93(7), 117(1)
VI, *O,S*-dimethyl *N*-isobutylphosphoramidothioate	1(10,11), 3(9), 5(4,6), 6(8), 46(22-27), 47(12-17,19,20), 50(18), 52(21), 58(2), 60(5), 72(7), 107(3), 117(1)

[a] The data are given in the form $A(B,C,D,...)$, where A denotes atom class and $B,C,D,...$ denote the atom numbers belonging to this class A.

Table III. Statistics of Parameter Fitting in the Two Studies

property	no. of compounds	correln coeff	rms deviation	explained variance
hydrophobicity	893	0.925	0.496	0.838
refractivity	538	0.998	0.774	0.996

be such that it will make a linear transformation of the function expressing the goodness of superposition while considering any particular parameter. Otherwise, the ordering of the binding modes using unscaled and scaled parameters will differ, making the calculation physically unrealistic. The function F_1 in general will maintain the condition only if all the atoms of the reference structure are occupied. The function F_2 in general cannot maintain the condition. In both the cases, the problem is caused by the μ_{old} part of the parameter transformation. Furthermore, autoscaling places the parameters on either side of zero. For charge density and hydrophobicity, positive and negative values have a physical significance. However, as pointed out earlier, molar refractivity is closely related to molar volume and is always positive. Making molar refractivity negative in any scaling is unwanted. Most of these problems can be avoided if a modified autoscaling is used:

$$x_{i,new} = x_{i,old}/s_{old} \quad (7)$$

All parameters in this transformation have identical variance of unity. Use of an extra physicochemical parameter for superposition is justified only if it is linearly independent of the others.

Table IV. Statistics of Predictive Power of the Parameter Sets

property	no. of compounds	correln coeff	rms deviation
hydrophobicity	129	0.876	0.497
refractivity	82	0.996	1.558

Table V. Compounds Showing Large Deviations (1.0 or Greater) of Calculated Values from Observed log $P_{(octanol-water)}$

ID[a]	compound	obsr	calc	dev
360	4,5-dibromo-1,2,3-triazole	2.24	1.07	-1.37
683	ethylene glycol	-1.93	-0.70	1.23
723	acrilonitrile	-0.92	1.05	1.97
731	5-azauracil	-1.87	-0.48	1.39
1156	pyridazine	-0.72	0.28	1.00
1211	cytosine	-1.73	-0.55	1.18
1240	butane-2,3-dione	-1.34	0.85	2.19
1302	naled	1.38	2.57	1.19
1367	2-methyl-2-imidazoline	0.52	-0.55	-1.07
1646	butane-2,3-diol	-0.92	0.12	1.04
1754	2,4,6-trichloropurine	3.90	2.89	-1.01
1869	pyridine *N*-oxide	-1.69	-0.34	1.35
2047	cyclopentane	3.00	1.98	-1.02
2308	hexachlorobenzene	4.13	5.15	1.02
3429	1,3,5-trihydroxyphenol	0.16	1.19	1.03
3557	2-aminopyridine-5-carboxamide	0.70	-0.51	-1.22
3899	sorbose 6-phosphate	0.38	-0.97	-1.35
4962	methyl phenyl selenide	2.87	1.67	-1.20
6647	dimetonthiol	1.93	2.94	1.01
7518	cytidilic acid	0.44	-0.66	-1.06
7539	1,3,6-trimethyl-5-(dimethylamino)uracil	0.99	-1.14	-2.13
7645	2-chloro-1,4-naphthaquinone	2.15	1.07	-1.08
8153	1,2-dimethylindole	2.82	1.60	-1.22
8168	3-(allyloxy)-4-aminobenzoic acid	0.42	1.45	1.03
8209	*N*-cyano-2-(3,3-dimethyl-1-triazeno)-benzamide	0.80	1.91	1.11
8293	inosine	-2.08	-0.74	1.34
8419	fuscaric acid/5-butylpicolinic acid	0.68	2.49	1.81
8669	*O,O*-diethyl *O*-phenyl phosphate	1.34	2.92	1.28
8694	adenosine triphosphate	1.64	0.34	-1.30
8708	*N*-*tert*-pentanoylcyclobutanecarboxamide	0.53	1.90	1.37
8751	4-*tert*-butyl β-glucopyranoside	1.18	-0.31	-1.49
8761	decanol	1.70	3.32	1.62
8769	decylamine	1.92	2.97	1.05
8778	*O,S*-dimethyl *N*-octylphosphoramidothioate	1.51	3.04	1.53
8851	3-*p*-anisoyl-3-bromoacrylic acid	-1.15	1.61	2.76
8869	$2,4\text{-}(NO_2)_2C_6H_3NHN{=}C(CN)COOEt$	4.14	3.05	-1.09
8899	$2\text{-}SO_2Me\text{-}4\text{-}NO_2C_6H_3NHN{=}C(CN)COOMe$	3.21	1.52	-1.69
8936	sulfapyridine	-0.02	1.10	1.12
8993	antipyrene	0.28	1.54	1.26
9410	*p*-hexylpyridine	4.35	3.18	-1.17
9595	dibenzofuran	4.12	2.76	-1.36
9601	1-naphthylmethyl isothiocyanate	4.42	3.30	-1.12
9820	vitavax	2.14	0.09	-2.05
9885	sulfisomidine	-0.30	1.37	1.67
10172	*p*-heptylpyridine	5.00	3.58	-1.42

[a] The number corresponds to the compilation of Hansch and Leo (ref 28) and is given here for easy reference.

Table VI. AM1 Charge Density on Phosphorus in a Few Compounds

compound (molecular formula)	classification (of phosphorus)	charge in au
$(CH_3)_3PS$	P (116)	0.3326
$(CH_3)_3PO$	P (116)	0.6965
$PS(OCH_3)_2Cl$	P (117)	0.6842
$PO(OCH_3)_3$	P (117)	1.1238
$PS(OCH_3)_3$	P (117)	0.7958
$P(OCH_3)_3$	P (118)	0.7927
$P(OCH_3)_2(OH)$	P (118)	0.8180
$P(CH_3)_3$	P (119)	0.2476
$P(CH_3)_2(C_2H_5)$	P (119)	0.2394
$PO(CH_3)(OCH_3)_2$	P (120)	0.9359
$PS(CH_3)(OCH_3)_2$	P (120)	0.6403

Table VII. Compounds Showing Large Differences in Calculated and Observed Values of Molar Refractivity

ID[a]	name	obsr	calc	dev
11366	dichloro(4-ethylphenyl)phosphine	56.02	48.77	−7.25
11367	dichlorophenylphosphine	45.34	43.14	−2.20
11437	triisopropyl phosphite	53.09	55.48	2.39
23568	thionyl chloride	22.12	19.61	−2.50
24937	4-chloro-*N*-methylaniline	29.34	39.12	9.78
25326	antimalarine	91.20	89.02	−2.18
25338	2,5-dimethoxysaffrole	68.27	62.78	−5.49
25453	nonanedioic acid	39.08	45.31	6.23
25465	azobenzene	53.66	57.95	4.29
25510	3,3′-dimethyldiazobenzene	72.51	70.46	−2.05
26103	catechol	32.95	29.47	−3.51
26155	4-iodofluorobenzene	34.96	38.68	3.72
26274	pyrogallol	28.11	31.14	3.03
26356	benzenesulfinyl chloride	25.46	38.50	13.04
26692	3-ethylbenzoic acid	44.84	42.46	−2.38
26710	salicylic acid	31.18	34.51	3.33
26876	3-methylbenzonitrile	34.81	36.84	2.02
27119	2-chlorobenzoxazole	37.33	40.87	3.54
27128	2-methylbenzoxazole	35.01	39.77	4.76
27175	2-phenyl-2-propanol	44.03	41.93	−2.10
28085	2-aminobutane	21.40	23.61	2.21
28178	butyl sulfoxide	54.14	47.69	−7.45
29162	cinnamoyl chloride	49.99	46.99	−3.00
29355	tri-*o*-cresyl phosphate	99.27	102.56	3.29
30195	ethyl 3,5-dinitrobenzoate	59.97	56.98	−2.99
36463	2-benzylpyridine	50.67	53.21	2.54
37974	1-allylthiourea	32.33	35.19	2.86
38430	*N*,*N*-dimethyltrichloroacetamide	40.42	38.32	−2.10
24440	acetyl iodide	26.15	23.39	−2.76
24538	monoethyl adipate	46.04	42.26	−3.78
25065	*o*-nitroanisole	36.89	39.85	2.95
25067	4-nitroanisole	37.38	39.85	2.47
25597	3,3′-dimethylazoxybenzene	78.03	71.71	−6.32

[a] All molecules having ID numbers greater than 24 000 were taken from reference 29. Simply subtract 24 000 to get the *CRC Handbook* number. For the rest of the molecules see reference 12.

RESULTS AND DISCUSSION

Table I lists various atom types defined for this study. The factors important in the present classification are hybridized state, formal charge density, and solvent accessibility. Classification of atoms used in previous studies[8] is retained with the addition of phosphorus and selenium atom types. Figure 1 shows some of the phosphorus-containing molecules used in the study. Table II shows the classification of atoms for these molecules.

As it has been mentioned under Method of Calculation, some constraints are necessary for evaluating atomic contributions to molar refractivity to keep the value in a physically realistic region. One set of constraints in the solution was to ensure that different atom types with the same hybridization do not differ markedly (arbitrarily chosen as 30%) from a base value. This base value was obtained from our previous study.[12] When no constraints were used, both the constrained and unconstrained optimization routines gave identical results. The

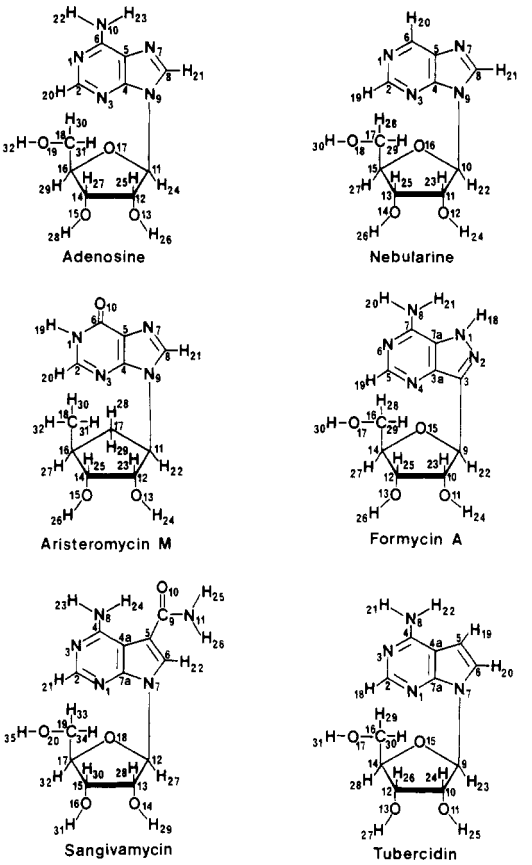

Figure 2. Chemical structures of the antibiotics used for the superposition.

evaluated atomic parameters for hydrophobicity and refractivity are listed in Table I.

Atomic Hydrophobicities. The octanol–water partition coefficients of 893 compounds[28] were used in the evaluation of atomic hydrophobicities. The calculated values showed a standard deviation of 0.496, a correlation coefficient of 0.925, and an explained variance of 0.838 (Table III). These parameters were used to predict the octanol–water coefficient of 129 molecules, which were not included in the parametrization procedure. The calculated values showed a standard deviation of 0.497 and a correlation coefficient of 0.876 (Table IV). Calculated values of 23 compounds deviated by one standard deviation (0.497) or more from their observed values. The two phosphates showed good agreement with the observed values of hydrophobicity. In Table V, the compounds showing a deviation of 1.0 or more of the calculated values of hydrophobicity from their observed values (octanol–water partition coefficient) are presented. These include compounds from both the training set and the test set.

Analysis of the saturated carbon atoms in general showed an increase in hydrophobic contribution when hydrogens are replaced by carbon. In the present as well as previous works,[8] this is generally found to be true, though there are some anomalies, particularly in ethylenic carbons. Acetylenic carbons also satisfied this trend. It is interesting to observe this rule in nitrogen types (compare among types 66, 67, and 68 or among 69, 70, and 71) with better agreement than earlier.[9] However, other unknown factors are bound to complicate the rationalization of the trends.

Validity of the phosphorus atom classification is checked by AM1[10] charge density calculations. In these calculations, two or more phosphorus-containing compounds of each type were generated and atomic positions were optimized. Charges on phosphorus are compared in Table VI. It may be seen that when oxygen is substituted by sulfur, the electron charge

Table VIII. Atom-Based Description of the Molecular Superpositions[a]

compound	atom numbers
sangivamycin	1/3,2/2,3/1,4/7a,5/4a,6/4,7/5,8/6,9/7,10/8,11/12,12/13,13/14,14/15,15/16,16/17,17/18,18/19,19/20,20/21,21/22,22/23,23/24,24/27,25/28,26/29,27/30,28/31,29/32,30/33,31/34,32/35
nebularine	1/1,2/2,3/3,4/4,5/5,6/6,7/7,8/8,9/9,10/20,11/10,12/11,13/12,14/13,15/14,16/15,17/16,18/17,19/18,20/19,21/21,24/22,25/23,26/24,27/25,28/26,29/27,30/28,31/29,32/30
tubercidin	1/3,2/2,3/1,4/7a,5/4a,6/4,7/5,8/6,9/7,10/8,11/9,12/10,13/11,14/12,15/13,16/14,17/15,18/16,19/17,20/18,21/20,22/21,23/22,24/23,25/24,26/25,27/32,28/27,29/28,30/29,31/30,32/31
formycin A	1/6,2/5,3/4,5/7a,6/7,7/1,8/3,9/3a,10/11,11/9,12/10,13/8,14/12,15/13,16/14,17/15,18/16,19/17,20/19,22/20,23/21,24/22,25/23,26/24,27/25,28/26,29/27,30/28,31/29,32/30
aristeromycin M (axial)	1/1,2/2,3/3,4/4,5/5,6/6,7/7,8/8,9/9,10/10,11/11,20/20,21/21
aristeromycin M (equatorial)[b]	1/1,2/2,3/3,4/4,5/5,6/6,7/7,8/8,9/9,10/10,11/11,12/12,14/14,16/16,21/21,27/25,31/31
aristeromycin M (equatorial)[c]	1/1,2/2,3/3,4/4,5/5,6/6,7/7,8/8,9/9,10/10,11/11,12/12,13/13,14/14,17/17,20/20,21/21,24/22,25/23,26/24

[a] The atom numbers of the superposed atoms of each nucleoside antibiotic on the reference compound (adenosine). For each molecule, the atom numbers superposed on adenosine are given in the form X/Y, where X is the atom number of the reference and Y is the corresponding superposed atom of the antibiotic. [b] The numbering refers to the superposition using the function F_1. [c] The numbering refers to the superposition using the function F_2.

Table IX. Results of Molecular Superposition Using the Reference Conformation of Adenosine and the Best Superposed Conformation of the Test Molecules As Obtained by the Present Algorithm

compound	conformational energy[a]	dihedral angle[b]				superpos func	
		w_1	w_2	w_3	w_4	F_1	F_2
adenosine	0.0	60	270	180		44.88	44.88
sangivamycin	0.0	60	270	180	340	41.88	44.06
nebularine	0.0	60	270	180		39.93	39.98
tubercidin	0.0	60	270	180		41.67	43.93
formycin A	0.52	60	260	180		36.82	41.27
aristeromycin M (axial)	0.57	30	0			13.98	18.40
aristeromycin M (equatorial)[c]	1.94	60	240			20.97	
aristeromycin M (equatorial)[d]	0.59	60	260				26.89

[a] Energy values in kilocalories per mole relative to the minimum for each molecule. [b] Descriptions of the dihedral angles are as follows: w_1 = 14-16-18-19 (adenosine); 15-17-19-20 (sangivamycin); 13-15-17-18 (nebularine); 12-14-16-17 (tubercidin and formycin A); 14-16-18-30 (aristeromycin M). w_2 = 8-9-11-12 (adenosine and aristeromycin M); 6-7-12-13 (sangivamycin); 8-9-10-11 (nebularine); 6-7-9-10 (tubercidin); 2-3-9-10 (formycin A). w_3 = 16-18-19-32 (adenosine); 17-19-20-35 (sangivamycin); 15-17-18-30 (nebularine); 14-16-17-31 (tubercidin); 14-16-17-30 (formycin A). w_4 = 5-9-11-25 (sangivamycin). [c] For the superposition based on function F_1. [d] For the superposition based on function F_2.

density on phosphorus increases considerably. Such change ought to affect the hydrophobicity of the atom unless it is counterbalanced by some other factor(s). This effect can be incorporated in two ways: (i) creating more phosphorus atom types or (ii) incorporating charge density in the evaluation of the hydrophobicity.[8] Since increasing the atom types to cover the subtle changes is not very sensible, a basis for incorporating

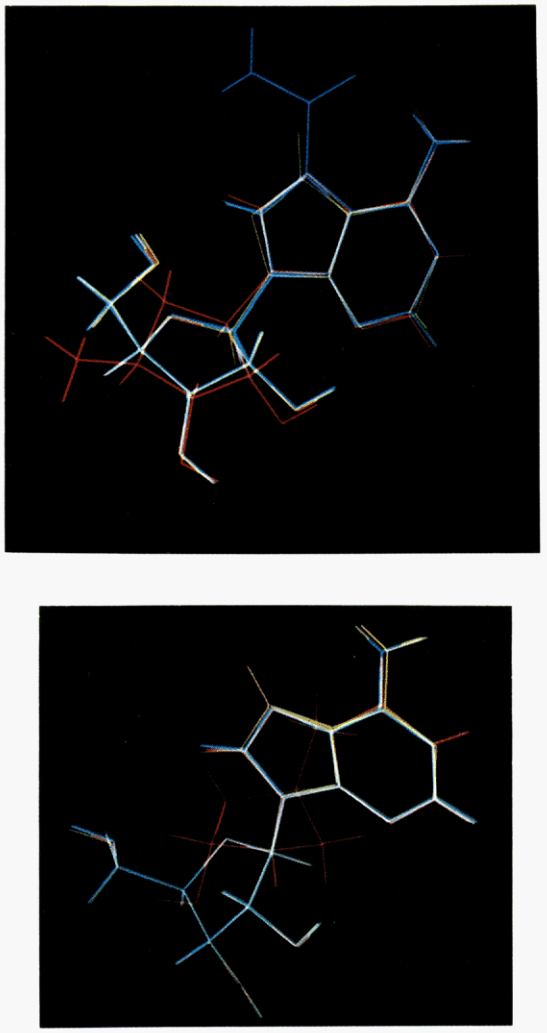

Figure 3. (a, top) Superposition of (i) aristeromycin M in the pseudoequatorial orientation, red, (ii) formycin A, yellow, and (iii) sangivamycic acid, blue, on adenosine in half-bond color. (b, bottom) Superposition of (i) aristeromycin M in the pseudoaxial orientation, red, (ii) tubercidin, yellow, and (iii) nebularine, blue, or adenosine.

charge density and a few other fundamental properties to evaluate the hydrophobicity is under way and will be published in future work.

Atomic Refractivities. The molar refractivities of 538 compounds[12,29,30] were used in obtaining atomic refractivity contributions. Table III shows the statistics of fit. The calculated and the observed values showed a standard deviation of 0.774, a correlation coefficient of 0.998, and an explained variance of 0.996. These parameters were then used to predict the molar refractivity of 82 compounds. The predicted molar refractivities showed a standard deviation of 1.553 and a correlation coefficient of 0.996 with the observed values for these compounds. Among the compounds that were used in the training set, 52 compounds deviated by one standard deviation (0.774) or more from the observed values. Among the compounds used for prediction, 17 molecules showed the same deviation from the observed value. Those molecules showing high deviation (2.0 or more from the observed value) are listed in Table VII, which includes compounds from both the training and the test sets. The density of tri-*o*-cresyl phosphate (*CRC Handbook*,[29] no. 11423) is much higher than other analogous compounds (no. 11422 and 11424), and the calculated value of molar refractivity from the reported density of 1.955 g/cm^3 shows marked deviation, casting doubt on this value.

It is expected that whenever a set of molecules is added to the previous set, the derived parameters show differences. However, any large deviation indicates that the solution is

unstable. In the present study no major deviation of the atomic refractivity values was observed from the previous study.[12] However, when no heteroatom is present, the effect of carbon replacing hydrogen is to decrease the atomic contribution, whereas the reverse situation occurred in the earlier study.[8] When one or two heteroatoms are present, the same situation of decreasing atomic contribution with carbon substituents (atom types 5, 6, 8, and 11 or atom types 7, 9, and 12) develops. The decrease in volume contribution is expected due to substitution by non-hydrogen atoms, since being bulkier than hydrogen the overlapping volume will increase. However, the effect of bond polarity and the overlapping with nonbonded atoms may complicate the situation.

Application of Atomic Physicochemical Parameters for Molecular Superpositions. A set of five nucleoside antibiotics (which are structural analogues of adenosine) is taken for the application of the superposition method described under Method of Calculation. These are nebularine,[31] aristeromycin M,[32] formycin A,[33] sangivamycin,[34] and tubercidin.[35] The covalent structures of these molecules are shown in Figure 2. Aristeromycin M is a carbocyclic nucleoside analogue of adenosine with an imidazolopyrimidine ring attached to a cyclopentane ring with attachments of two hydroxyls and a methyl group similar to the 5-deoxyribose sugar ring. The substituents on the cyclopentane ring can be placed either pseudoaxially or pseudoequatorially, although the pseudoequatorial conformation is expected to be energetically more stable. In the present study both conformations are considered, since our interest is to consider any low-energy conformation and not just the global minimum-energy conformation during the superposition. An objective here would be to check whether the pseudoequatorial conformer superposes better than the pseudoaxial conformer. The three-dimensional structure of each molecule is generated by adding fragments from a crystallographic fragment library, which consists of the 2′-endo,3′-exo (type S) ribose ring and various heterocyclic rings obtained from the literature.[14] Since these molecules are not conformationally rigid, a fixed valence structure molecular mechanics calculation was performed by using a modified version of MM2 CONFOR program.[36] The details of this program are given in Appendix II. For superpositions, the adenosine molecule in its minimum-energy conformation is taken as the reference and a large number of geometrically possible superpositions were evaluated for 100 lowest energy conformations of each molecule by use of the algorithm described in Appendix I. Once the geometrically feasible superpositions were evaluated, eq 4 and 5 were used to estimate their goodness of fit.

The results of these calculations are summarized in Tables VIII and IX. Table VIII represents the atom-based description of the physicochemically best superposition on the reference structure, the minimum energy conformation of adenosine. Table IX gives the conformational details of the superimposed structures and the value of the two superposition functions. For both the functions, the higher the value, the better is the superposition. We can, therefore, conclude that among the various nucleosides tested here sangivamycin resembles adenosine most closely and equatorial aristeromycin is better over the axial one. Except for aristeromycin M, both equations gave the same superposition in all the cases. In Figure 3a, a color picture of the superposition of pseudoequatorial aristeromycin (red), formycin A, (yellow), and sangivamycin (blue) is presented on the reference adenosine molecule (half-bond color) using F_1 function (eq 4). Although tubercidin and nebularine gave excellent superposition, they were omitted to keep the figure clear and the colors legible. Figure 3b presents the superposition of pseudoaxial aristeromycin M (red), tubercidin (yellow), nebularine (blue) on adenosine.

A recent study by Kato et al.[37] also uses a superposition method based on similarity in charge density and conventional least-squares function for molecular fitting. The present development of atomic physicochemical properties and the superposition method described here will aid in selecting meaningful superpositions from many geometrically feasible alternatives and will enable other properties besides charge density to be considered.

CONCLUSIONS

(1) It is possible to dissect many physicochemical properties responsible for molecular interaction at the atomic level. Since most of these properties were not strictly additive, the constitutive part is encoded by an extensive atom classification.

(2) The method is capable of calculating the local physicochemical properties of any part of a molecule and will be of great help for precise empirical receptor mapping.

(3) The electronic and steric effect of the environment is not discrete, so the discretized atom classification and the property assignment should be considered as an oversimplification of the actual state.

(4) The electronic or steric effects persist well beyond the immediately bonded atom. A large number of atom types is necessary to include that effect in the present approach. Alternatively, such effects can be correlated by using topological information and some more fundamental atomic properties.

(5) The atom-based property matching as described here can be a very effective algorithm for the automated superposition of the molecules to avoid any personal bias.

APPENDIX I

Given a reference structure and a finite number of low-energy conformations of a test molecule, the present algorithm will determine the important geometrically feasible superpositions of the test molecule on the reference. First we shall present an algorithm that will determine all geometrically possible superpositions; next we shall show how to save the method from the "combinatorial explosion".

Step 1. Generate the interatomic distance matrix for the reference.

Step 2. Take the lowest energy conformation of the test molecule.

Step 3. Generate the interatomic distance matrix for the test conformation.

Step 4. (i) Loop over all possible combinations of three atoms (r_1, r_2, r_3) from the reference without repetition. (ii) Loop over all permutations of three atoms (t_1, t_2, t_3) from the test molecule. (iii) Check the distances between the superimposed atoms. The condition for superposition is

$$d_t + \delta \geq d_r \geq d_t - \delta \qquad (8)$$

where d_r and d_t represent the distances in the reference and the test molecules, respectively. Accept the "basic superposition" if the distance condition is satisfied. When three atoms of the reference and test molecules are in contact, no more rigid translation or rotation is possible. The other contacts are mere consequence. (iv) Expand the superposition list as follows: Take an unused atom from the reference. Loop over all unused atoms of the test. If the distance condition from the superposed atoms is satisfied, include it in the list. If more than one atom of the test molecule satisfies the condition, the one having minimum absolute deviation per distance from the superimposed atoms is accepted. Restart loop iv until all atoms of the reference are tried. Since distance is invariant to reflection, distance criterion allows the superpositions of enantiomers. To solve the problem, we checked the final

Table X. Torsional Parameters Used outside Regular MM2 (1985) Parameters

torsion type[a]	torsion parameter, kcal/mol V_1	V_2	V_3	ref[b]
9-3-6-1	−1.66	8.98	0.00	MM85 7-3-6-1
9-3-6-20	0.00	0.00	0.00	MM85 7-3-6-20
3-3-6-1	3.53	2.30	−3.53	MM85 2-3-6-1
3-3-6-20	0.00	0.00	0.00	MM85 7-3-6-20
9-3-9-1	0.00	5.00	0.00	MM85 7-3-9-1
9-3-9-23	1.10	5.00	0.00	MM85 7-3-9-28
6-1-9-23	0.00	0.00	0.00	AMBER X-CT-N-X
1-1-9-23	0.00	0.00	0.00	AMBER X-CT-N-X
5-1-9-23	0.00	0.00	0.00	AMBER X-CT-N-X
3-3-9-23	0.00	5.00	0.00	MM85 2-3-9-1
3-3-9-1	0.00	5.00	0.00	MM85 2-3-9-1
6-1-9-3	0.00	0.00	0.00	AMBER X-CT-N-X
2-3-9-23	0.00	5.00	0.00	MM85 2-3-9-1

[a] The number represents the MM2 (1985) atom-type number. [b] For MM85 parameters see reference 38b; for AMBER parameters see Weiner et al. *J. Comput. Chem.* **1986**, *7*, 230.

superposition by using rigid rotation and translation as described by Kenknight.[24] This checking process was done only if it was a new superposition. End loop ii. End loop i. Take the next available conformation of the test molecule and go to step 3.

It is a combinatorial problem, and the number of possibilities becomes astronomical with the increase of the number of atoms in the reference molecule and the test molecule. We have evaluated the basic superposition taking a maximum of 12 important atoms from each molecule. Important atoms are heteroatoms, hydrogen attached to heteroatoms, and carbon multiply bonded to heteroatoms. However, all atoms were used during the expansion of the contact list.

APPENDIX II

Molecular Mechanics Conformational Analysis. The molecular mechanics program uses the 1985 MM2 parameters[38] with some additional torsional parameters obtained from the literature. The unavailability of many stretching and bending parameters did not allow a complete relaxation in the structure. To have good bond length and angle, the molecules were generated from a crystallographic fragment library. This program evaluates the conformational energy as the sum of the nonbonded van der Waals, hydrogen bonding, electrostatic, and torsional interaction

$$E_{\text{total}} = \sum_{i<j} \text{EV}_{ij} + \sum_{\text{H-bond}} \text{EH}_{ij} + \sum_{i<j} q_i q_j / \text{DR}_{ij} + \sum_{\text{dihedral}} [V_1/2(1 + \cos \omega) + V_2/2(1 - \cos 2\omega) + V_3/2(1 + \cos 3\omega)] \quad (9)$$

where

$$\text{EV}_{ij} = \epsilon_{ij}(2.9 \times 10^5 \exp(-12.5/P) - 2.25P^6) \quad (10)$$

if $P \leq 3.311$ and

$$\text{EV}_{ij} = 336.176\epsilon_{ij}P^2 \quad (11)$$

if $P > 3.311$. Here $P = r_{ij}/R_{ij}$. R_{ij} is the internuclear distance; r_{ij} is the sum of the van der Waals radii. $\epsilon_{ij} = \sqrt{(\epsilon_i \epsilon_j)}$, where ϵ_i, etc., are the hardness of the atom. Hydrogen-bonding interaction in MM2 approach may be considered as a special type of van der Waals interaction where the parameters ϵ_{ij} and r_{ij} have been modified.[39] Electrostatic interaction was calculated from the formal atomic charge distribution calculated from the CNDO/2 method[11] and by assuming the effective dielectric constant equal to 2.[40] The energy minimization was done in two steps: First, the pattern search technique[41] was used in which each dihedral angle was rotated sequentially from 0° to 360° at a specified interval and fixed at its minimum-energy value before proceeding to the next dihedral angle. The process was repeated until it converged. In the second step a limited number of important dihedral angles were rotated to generate all combinatorial conformations. The rest of the dihedral angles during this combinatorial search were fixed at the minimum-energy conformation as obtained from the pattern search minimization.

Some of the torsional parameters that were not defined in the MM2 parameter list were either chosen from different force fields or arbitrarily selected by analogy from the same force field. These values are given in Table X. The pyrrole-type nitrogen, aromatic amines, and amides were given the type 9. The program generates the necessary lone pair of electrons.

Registry No. **360**, 15294-81-2; **683**, 107-21-1; **723**, 107-13-1; **731**, 71-33-0; **1156**, 289-80-5; **1211**, 71-30-7; **1240**, 431-03-8; **1302**, 300-76-5; **1367**, 534-26-9; **1646**, 513-85-9; **1869**, 694-59-7; **2047**, 287-92-3; **2308**, 118-74-1; **3429**, 108-73-6; **3557**, 329-89-5; **3899**, 20479-58-7; **4962**, 4346-64-9; **7518**, 63-37-6; **7539**, 38507-32-3; **7645**, 1010-60-2; **8153**, 875-79-6; **8168**, 121498-31-5; **8209**, 66974-92-3; **8293**, 58-63-9; **8419**, 536-69-6; **8669**, 2510-86-3; **8694**, 56-65-5; **8708**, 25031-87-2; **8751**, 29074-04-2; **8761**, 112-30-1; **8769**, 2016-57-1; **8778**, 52067-53-5; **8851**, 16170-76-6; **8869**, 28313-64-6; **8899**, 36865-82-4; **8936**, 144-83-2; **8993**, 60-80-0; **9410**, 27876-24-0; **9595**, 132-64-9; **9601**, 17112-82-2; **9820**, 5234-68-4; **9885**, 515-64-0; **10172**, 40089-90-5; **11366**, 5274-50-0; **11367**, 644-97-3; **11437**, 116-17-6; **23568**, 7719-09-7; **24937**, 932-96-7; **25326**, 551-01-9; **25338**, 121498-32-6; **25453**, 123-99-9; **25465**, 103-33-3; **25510**, 588-04-5; **26103**, 120-80-9; **26155**, 352-34-1; **26274**, 87-66-1; **26356**, 4972-29-6; **26692**, 619-20-5; **26710**, 69-72-7; **26876**, 620-22-4; **27119**, 615-18-9; **27128**, 95-21-6; **27175**, 617-94-7; **28085**, 13952-84-6; **28178**, 2168-93-6; **29162**, 102-92-1; **29355**, 78-30-8; **30195**, 618-71-3; **36463**, 101-82-6; **37974**, 109-57-9; **38430**, 7291-33-0; **24440**, 507-02-8; **24538**, 626-86-8; **25065**, 91-23-6; **25067**, 100-17-4; **25597**, 19618-06-5; I, 121-75-5; II, 58-64-0; III, 121498-30-4; IV, 122-14-5; VI, 52067-51-3; $(CH_3)_3PS$, 2404-55-9; $(CH_3)_3PO$, 676-96-0; $PS(OCH_3)_2Cl$, 2524-03-0; $PO(OCH_3)_3$, 512-56-1; $PS(OCH_3)_3$, 152-18-1; $P(OCH_3)_3$, 121-45-9; $P(OCH_3)_2(OH)$, 96-36-6; $P(CH_3)_3$, 594-09-2; $P(CH_3)_2(C_2H_5)$, 1605-51-2; $PO(CH_3)(OCH_3)_2$, 756-79-6; $PS(CH_3)(OCH_3)_2$, 681-06-1; sangivamycin, 18417-89-5; nebularine, 550-33-4; tubercidin, 69-33-0; formycin A, 6742-12-7; aristeromycin M, 98873-79-1; adenosine, 58-61-7.

REFERENCES

(1) Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. *Science* **1987**, *235*, 574–576.
(2) McCammon, J. A.; Karplus, M. *Acc. Chem. Res.* **1983**, *16*, 187.
(3) Hansch, C. In *Correlation Analysis in Chemistry*; Chapman, N. B., Shorter, J., Eds.; Plenum: New York, 1978.
(4) Ghose, A. K.; Crippen, G. M. *J. Med. Chem.* **1985**, *28*, 333–347.
(5) Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196–7206.
(6) Marshall, G. R. *Annu. Rev. Pharmacol. Toxicol.* **1987**, *27*, 193–213.
(7) Cohen, N. C. *Adv. Drug Res.* **1985**, *14*, 41–145.
(8) Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 80–90.
(9) Hehre, W. J.; Radom, L.; Schleyer, P. v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
(10) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(11) Pople, J. A.; Beveridge, D. L. In *Approximate Molecular Orbital Theory*; McGraw-Hill: New York, 1970.
(12) Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.
(13) Ghose, A. K.; Crippen, G. M. In *Comprehensive Medicinal Chemistry*; Ramsden, C. A., Ed.; Pergamon Press: Oxford, U.K. (in press).
(14) Ghose, A. K.; Crippen, G. M.; Revankar, G. R.; McKernan, P. A.; Smee, D. F.; Robins, R. K. *J. Med. Chem.* **1989**, *32*, 746–756.
(15) Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7*, 565–577.
(16) Ghose, A. K.; Crippen, G. M. *Abstracts of Papers*, 193rd National Meeting of the American Chemical Society, Denver, CO; American Chemical Society: Washington, DC, 1987; COMP 5.
(17) Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H.; Zehfus, M. H. *Science* **1985**, *229*, 834–840.
(18) Viswanadhan, V. N. *Int. J. Biol. Macromol.* **1987**, *9*, 39–48.
(19) Fujita, T.; Iwasa, J.; Hansch, C. *J. Am. Chem. Soc.* **1964**, *86*, 5175.
(20) Rekker, R. F. *The Hydrophobic Fragmental Constants*; Elsevier: New York, 1977.
(21) Nyburg, S. C. *Acta Crystallogr.* **1974**, *B30*, 251.
(22) Yuen, P. S.; Nyburg, S. C. *J. Appl. Crystallogr.* **1979**, *12*, 258.
(23) Barino, L. *Comput. Chem.* **1981**, *5*, 85.

(24) Kenknight, C. E. *Acta Crystallogr.* **1984**, *A40*, 708-712.
(25) Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1985**, *6*, 350-359.
(26) Lejeune, J.; Michael, A.; Vercauteren, D. P. *J. Comput. Chem.* **1986**, *7*, 739-744.
(27) Stouch, T. R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4-12.
(28) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
(29) *CRC Handbook of Chemistry and Physics*, 65th ed.; Weast, R. C., Ed.; CRC: Boca Raton, FL, 1984.
(30) *Dictionary of Organophosphorus Compounds*; Edmandson, R. S., Ed.; Chapman and Hall: London, 1988.
(31) Takeda, T.; Ohashi, Y.; Sasada, Y. *Acta Crystallogr.* **1970**, *B30*, 825-827.
(32) Miyashita, O.; Kasahara, F.; Marumoto, R. *J. Antibiot.* **1985**, *38*, 981-986.
(33) Pruisner, P.; Brennan, T.; Sundaralingam, M. *Biochemistry* **1973**, *12*, 1196-1202.
(34) Loomis, C. R.; Bell, R. M. *J. Biol. Chem.* **1986**, *263*, 1682-1692.
(35) Smulson, M. E.; Suhadolnik, R. J. *J. Biol. Chem.* **1967**, *242*, 2872-2876.
(36) Ghose, A. K.; Crippen, G. M. *J. Med. Chem.* **1984**, *27*, 901.
(37) Kato, Y.; Itai, A.; Iitaka, Y. *Tetrahedron* **1987**, *43*, 5229-5236.
(38) (a) *QCPE No. 395*; Department of Chemistry, Indiana University, Bloomington, IN. (b) Allinger, N. L. Private communication, MMP2 1985, parameters, dated June 17, 1987.
(39) Allinger, N. L.; Kok, R. A.; Imam, M. R. *J. Comput. Chem.* **1988**, *9*, 591-595.
(40) Momany, F. A.; Carruthers, L. M.; McGuire, R. F.; Scheraga, H. A. *J. Phys. Chem.* **1974**, *78*, 1595.
(41) Hooke, R.; Jeeves, T. A. *J. Assoc. Comput. Mach.* **1961**, *8*, 212.

Review of Ring Perception Algorithms for Chemical Graphs

GEOFFREY M. DOWNS, VALERIE J. GILLET, JOHN D. HOLLIDAY, and MICHAEL F. LYNCH*

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received October 18, 1988

Current ring perception algorithms for use on chemical graphs concentrate on processing specific structures. In this review, the various published ring perception algorithms are classified according to the initial ring set obtained, and each algorithm or method of perception is described in detail. The final ring sets obtained are discussed in terms of their suitability for use in representing the ring systems in structurally explicit parts of generic chemical structures.

INTRODUCTION

A necessary component of the generic chemical structure storage and retrieval system currently under development at Sheffield is a method of analysis and representation of ring systems for both specific and generic structures, such as those found in patents. The problems to be addressed are substantial. The representation should ensure consistent description of specific structures and specific ring systems within generic structures for both full and substructure searching. Furthermore, it must permit the description of those parts of generic ring systems for which full structural characterizations are given (**structurally explicit generics**), as well as being capable of extension to the characterization of those parts for which only intensional descriptions are given (**structurally implicit generics**).

Generic ring systems introduce several additional areas of complexity. A variable group may occur within a ring or may substitute onto a ring system to form additional rings, while identical generic structures may be declared different by virtue of the orientation of their partial structures and the partitioning of them into partial structures.

These circumstances call for the production of a searchable representation by means of a ring perception algorithm with the following characteristics:

- It is independent of the projection, orientation, and partitioning of the ring system.
- It is consistent in the selection of rings to ensure maximum recall with minimum false drops.
- It should permit the identification of inter-ring relationships.
- It should enable the overall logic of relationships among ring systems and acyclic parts to be represented faithfully.

A necessary preliminary stage to the production and evaluation of such an algorithm for use on structurally explicit generics is an appraisal of existing algorithms and ring sets used for specific ring systems, together with their potential for extension into a generic environment. This review is a result of such an appraisal, originally given in reference 1. Subsequent papers outline the resultant ring set and algorithm developed from this appraisal.[2-4]

An attempt has been made to standardize the different terminologies used to those of the graph theoretic concepts considered in the paper subsequent to this review. As a result, **vertex** (v) is used instead of atom or node, **edge** (e) instead of bond or arc, **connectivity** instead of degree or valency, and **nullity** (μ) instead of Frerejacque number or cyclomatic number. The sets of all vertices and edges are denoted by V and E, respectively. In addition to the definitions given below, an elementary knowledge of graph theory is assumed.

A **walk** is an alternating sequence of vertices and edges, starting and ending at vertices, where each edge is incident with the vertices on either side of it in the sequence. If the start and end vertices are the same, then the walk is closed and is called a **circuit**; otherwise, it is open. If all edges of a walk occur only once, then it is a **trail**, and if all vertices are also distinct, then it is a **path**. A closed path is a **cycle** ($v \geq 3$). If one or more vertices of a graph occur more than once in a circuit, then it contains **Doppelpunkte**.[5] A circuit without Doppelpunkte is a cycle. If a pair of vertices in a cycle is connected by an edge that does not occur in the cycle, then these vertices are **Nachbarpunkte**. A cycle without Nachbarpunkte is a **simple cycle**.

A graph is **connected** if every pair of vertices in it is joined by a path. The set of all such vertices forms a **component**. If the graph consists of several disconnected sets of vertices, then each set forms a separate component; hence, a disconnected graph has at least two components.

An edge is a **cut edge** if its removal disconnects the graph and increases the number of components; hence, it cannot be part of any cycle. Similarly, a vertex is a **cut vertex** (sometimes called an articulation point) if its removal increases the number

0095-2338/89/1629-0172$01.50/0