The molecular transform as a similarity measure

István Csorvássy, Lajos Tözsér
ALKALOIDA Chemical Company, H-4440 Tiszvasvári, Hungary

Levente Kárpáti and Gábor Náray-Szabó [1]
Department of Theoretical Chemistry, Eötvös University Budapest, P.O. Box 32, H-1518 Budapest 112, Hungary

Received 30 September 1992

We propose to quantify molecular similarity through various forms of molecular transforms directly related to experimental measurements. Various metric distances between molecular transforms are introduced in measuring similarity which can be used in quantitative structure–activity relationships. For simpler classes of compounds like aliphatic alcohols good correlations are obtained between the abstract distance from a lead compound and various physical and pharmalogical properties. For substituted phenols the correlation is worse; however, the predictive power of the descriptors derived from the molecular transform is yet acceptable. For trypsin inhibitors, a class of compounds having very different molecular formulae, the net atomic charge is introduced as a parameter in the generalized form of the molecular transform. Though a poor regression equation is obtained for the differences in the inhibitory power, inactive compounds within a set can be reliably selected.

1. Introduction

Molecular similarity is a useful concept receiving more and more attention in computer-aided molecular design [1]. One of the numerous quantitative measures is based on the molecular transform which is obtainable experimentally, e.g. by electron diffraction [2]. Its application to rational drug design has been first proposed by Soltzberg and Wilkins [3–5]. Somewhat later one of the present authors also advocated the use of abstract molecular distances, defined on the basis of molecular transforms [6,7], as a measure of similarity. Recently, King and co-workers addressed the problem and found 2D and 3D molecular transforms to be useful tools in quantitative structure–property studies [8,9].

It is an important goal to find a definition of molecular similarity which is reliable and simple enough to allow rapid comparisons in large databases. Easy selec-

[1] To whom correspondence should be addressed.

tion of a subset of molecules having a property falling within a certain (prescribed) range is especially useful in the early phase of computer-aided molecular design. This justifies our efforts to develop such a pre-selection method through the measure of similarity by using various forms of molecular transforms.

In the following we provide and compare some similarity definitions based on the molecular transform. These allow to calculate abstract distances defining similarity from a closed form expression; therefore one of the goals, fast comparison, is achieved. The reliability of our method is examined on the prediction of some physical and pharmacological properties of aliphatic alcohols and substituted phenols as well as the estimation of inhibitory potencies of low molecular weight ligands of trypsin.

2. Theory

The molecular transform is derived from the relative intensity of the scattered radiation which is observed in an electron diffraction experiment and is written as a Fourier transform [10],

$$I(s) = K \sum_{i<j}^{N} f_i f_j \int_0^\infty P_{ij}(r) \frac{\sin sr}{sr} \mathrm{d}r, \tag{1}$$

where K is a constant, f_i and f_j are the form factors, $P_{ij}(r)$ is the probability distribution describing the vibrational variation in the distance between atoms i and j. $s = 4\pi\lambda \sin(\theta/2)$, where λ is the wavelength of the electron beam and θ is the scattering angle. Putting $K = 1$, $P_{ij}(r) = \delta(r - r_{ij})$, we get a simplified expression [4],

$$I(s) = \sum_{i<j}^{N} f_i f_j F(x) \tag{2}$$

with $x = sr_{ij}$ and

$$F(x) = \sin(x)/x. \tag{3}$$

r_{ij} may be defined both as the geometric distance and the shortest path in the molecular graph between atoms i and j. The former definition yields a three-dimensional (3D), the latter a topological (2D) representation of the molecule. Another possible choice of F is an exponential form,

$$F(x) = \exp(-x^2/2). \tag{4}$$

In the original definition of the molecular transform f_i is set equal to Z_i, the atomic number; however, other atomic parameters like the net charge or the hydrophobicity increment [11] can also be used yielding generalizations of the original concept.

The abstract distance between two molecular transforms measuring molecular similarity, R_{ab}, may have several definitions. Carbó et al. introduced an index which defines the similarity between two molecules in terms of the overlap of their charge densities [12]. An adaption of their definition to molecular transforms is as follows:

$$R_{ab}^{\mathrm{C}} = \frac{N_{ab}^2}{N_{aa}N_{bb}} \tag{5}$$

with

$$N_{ab}^2 = \int_0^\infty I_a(s) I_b(s)\, \mathrm{d}s, \tag{6}$$

lower indices referring to molecules a and b. Another definition has been proposed by Hodgkin and Richards [13],

$$R_{ab}^{\mathrm{H}} = \frac{2N_{ab}^2}{N_a^2 + N_b^2}. \tag{7}$$

In our previous paper we defined the molecular distance by [6]

$$(r_{ab}^{\mathrm{G}})^2 = \int_0^\infty [I_a(s) - I_b(s)]^2\, \mathrm{d}s, \tag{8}$$

which may be written in a modified form,

$$(R_{ab}^{\mathrm{G}})^2 = \int_0^\infty [I_a(s)/N_{aa} - I_b(s)/N_{bb}]^2\, \mathrm{d}s. \tag{9}$$

It is easy to see that $(R_{ab}^{\mathrm{G}})^2 = 2(1 - R_{ab}^{\mathrm{C}})$. If two molecules are identical, the abstract distances defined in eqs. (5), (7) and (8), (9) become equal to 1 and 0, respectively.

Defining F as in eqs. (3) and (4), N_{ab}^2 can be expressed in a closed form,

$$N_{ab}^2 = \sum_{i<j}^{a} \sum_{k<l}^{b} f_i^a f_j^a f_k^b f_l^b g_{ab}(D_{ij}^a, D_{kl}^b), \tag{10}$$

where

$$g_{ab}(D_{ij}^a, D_{kl}^b) = \tfrac{1}{2}\pi[\max(D_{ij}^a, D_{kl}^b)]^{-1} \tag{11}$$

for eq. (3) and

$$g_{ab}(D_{ij}^a, D_{kl}^b) = \tfrac{1}{2}\sqrt{\pi}[(D_{ij}^a)^2 + (D_{kl}^b)^2]^{-1/2} \tag{12}$$

for eq. (4). Eq. (8) combined with eqs. (3) and (4) involves numerical integration, so the use of r_{ab}^{G} is time consuming.

In order to compare various definitions we calculated R_{ab}^{C} and R_{ab}^{H} for a series of aliphatic alcohols (cf. table 1) and derived the following regression equations:

$$R_{ab,3}^{\mathrm{C}} = 1.2932 R_{ab,4}^{\mathrm{C}} - 0.3157, \quad r = 0.9914;$$

Table 1
Abstract distances of some aliphatic alcohols from methanol as obtained from various definitions.

Molecule	Eqs. (3),(5)	Eqs. (3),(7)	Eqs. (4),(5)	Eqs. (4),(7)
ethanol	0.9850	0.3712	0.9935	0.3683
1-propanol	0.9623	0.2434	0.9809	0.2403
2-propanol	0.9625	0.2360	0.9837	0.2346
1-butanol	0.9397	0.1677	0.9667	0.1655
2-butanol	0.9407	0.1622	0.9703	0.1611
2-methyl-1-propanol	0.9398	0.1626	0.9696	0.1615
2-methyl-2-propanol	0.9418	0.1561	0.9745	0.1565
1-pentanol	0.9170	0.1226	0.9512	0.1212
2-pentanol	0.9217	0.1176	0.9576	0.1170
3-pentanol	0.9195	0.1173	0.9567	0.1169
2-methyl-1-butanol	0.9185	0.1177	0.9559	0.1172
2-methyl-2-butanol	0.9224	0.1136	0.9615	0.1138
3-methyl-1-butanol	0.9195	0.1190	0.9552	0.1182
3-methyl-2-butanol	0.9202	0.1150	0.9591	0.1150
2,2-dimethyl-1-propanol	0.9203	0.1150	0.9591	0.1150
1-hexanol	0.8996	0.0932	0.9348	0.0923
2-hexanol	0.9017	0.0909	0.9422	0.0903
3-hexanol	0.9006	0.0896	0.9429	0.0893
2-methyl-2-pentanol	0.8784	0.0918	0.9274	0.0921
2-methyl-3-pentanol	0.9011	0.0872	0.9460	0.0874
2-ethyl-1-butanol	0.8982	0.0886	0.9426	0.0887
4-methyl-2-pentanol	0.9020	0.0886	0.9448	0.0885
3-methylol-pentane	0.8982	0.0886	0.9426	0.0887
1-heptanol	0.8824	0.0735	0.9255	0.0743
2-heptanol	0.8844	0.0720	0.9287	0.0716
3-heptanol	0.8835	0.0710	0.9296	0.0708
4-heptanol	0.8892	0.0706	0.9297	0.0706
2,4-dimethyl-3-pentanol	0.8845	0.0674	0.9357	0.0680
1-octanol	0.8658	0.0596	0.9124	0.0593
1-methyl-heptanol	0.8698	0.0597	0.9149	0.0592
2-ethyl-hexanol	0.8677	0.0578	0.9168	0.0578
2-methylol-heptane	0.8613	0.0577	0.9180	0.0578
1-nonanol	0.8509	0.0494	0.9004	0.0492
1-decanol	0.8372	0.0417	0.8891	0.0416
1-undecanol	0.8244	0.0357	0.8783	0.0357
1-dodecanol	0.8124	0.0309	0.8681	0.0310
1-tridecanol	0.8012	0.0271	0.8584	0.0272

$$R^{\mathrm{C}}_{ab,3} = 0.1659 \log R^{\mathrm{H}}_{ab,3} + 1.0716, \quad r = 0.9912;$$

$$R^{\mathrm{C}}_{ab,3} = 0.1676 \log R^{\mathrm{H}}_{ab,4} + 1.0731, \quad r = 0.9917, \tag{13}$$

where the lower indices 3 and 4 refer to eqs. (3) and (4), respectively. Owing to the excellent correlation in eq. (13) we apply only one definition of the abstract distance, that of $R^{\mathrm{C}}_{ab,3}$, in all subsequent studies. The definitions in eqs. (8) and (9) are compared for a series of substituted phenols (cf. table 2). The following regression equation holds:

$$r^{\mathrm{G}}_{ab,3} = 0.9930 R^{\mathrm{G}}_{ab,3} + 0.0002, \quad r = 0.9794. \tag{14}$$

According to eqs. (13 and 14) various definitions of distances strongly correlate, therefore all of them have about the same prediction power.

3. Results and discussion

In order to study the feasibility of the molecular transform for the prediction of various properties we considered two possibilities. One is to use the abstract distance of a molecule called i from a lead, R_{0i}, as a descriptor and look for linear regression equations correlating the absolute value of the difference of a certain property for this molecule and the lead, $\Delta P_i = |P_0 - P_i|$, with R_{0i},

Table 2
Abstract distances of some substituted phenols from the 2-*s*-Bu derivative, as obtained from eqs. (8) and (9).

Derivative	Eq. (8)	Eq. (9)
2-*t*-Bu, 4-Me	0.022	0.022
4-*t*-Bu	0.023	0.025
4-*s*-Bu	0.022	0.026
2-Me, 4-*t*-Bu	0.024	0.026
2-*t*-Bu	0.028	0.027
4-Pr	0.041	0.045
4-I	0.068	0.048
2,4-di-Cl	0.055	0.051
4-Ph	0.043	0.054
4-nitro	0.066	0.055
2-nitro	0.073	0.059
4-Br	0.072	0.063
2-Br, 4-Me	0.072	0.068
4-Cl	0.099	0.081
4-Et	0.086	0.084
2-Et	0.090	0.086
3,5-di-Me	0.109	0.092
2,4-di-Me	0.111	0.097
2-Br	0.097	0.103
2-Cl	0.113	0.124
4-F	0.137	0.137
3-OH	0.126	0.140
2-F	0.137	0.146
2-Me	0.152	0.146
H	0.199	0.198

$$\Delta P_i = aR^{\mathrm{C}}_{0i,3} + b. \tag{15}$$

Another, less strict, approach is to study how reliably it can be predicted that ΔP_i exceeds a limit if R_{0i} is beyond another one. In this approach we call successful those predictions for which

$$\Delta P_i > 0.5(\Delta P_{\max} + \Delta P_{\min}) \quad \text{if } R_{0i} < 0.5(R_{0\max} + R_{0\min}) \tag{16}$$

or

$$\Delta P_i < 0.5(\Delta P_{\max} + \Delta P_{\min}) \quad \text{if } R_{0i} > 0.5(R_{0\max} + R_{0\min}) \tag{17}$$

(cf. fig. 1). This means that if R_{0i} is closer to unity than its mean between the extremes obtained for a certain group of molecules ΔP_i exceeds its mean for the same group. In the following we use R_{0i} as defined in eqs. (2), (3) and (5) and call false predictions those points on the ΔP_i versus R_{0i} plot which do not obey the inequalities in eqs. (16) and (17). For a random distribution the number of false predictions is just equal to that of the true ones i.e. 50% of the total number of points.

3.1. ALIPHATIC ALCOHOLS

Some physical properties of aliphatic alcohols as well as their activities obtained in some pharmacological tests are displayed in tables 3 and 4. We defined R_{0i} and property differences considering methanol as the lead compound and displayed regression parameters of eq. (15) and the number of false predictions as defined in eqs. (16) and (17) and in fig. 1, in table 5.

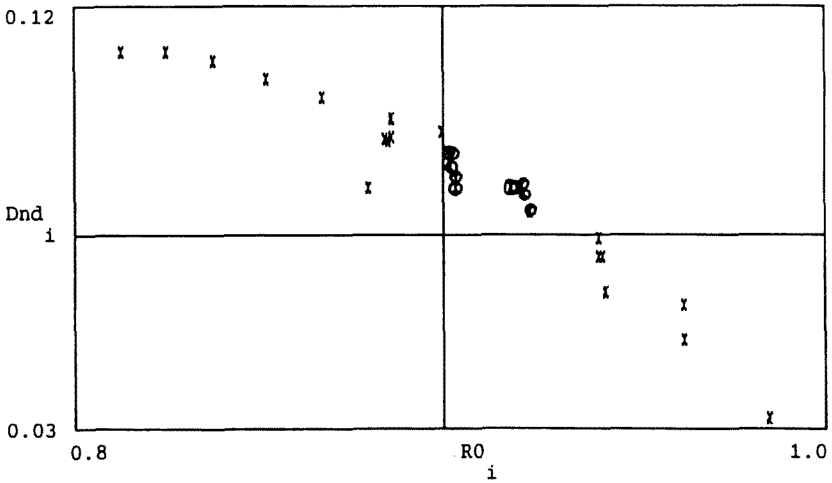


Fig. 1. Δn_{di} versus R_{0i} plot for aliphatic alcohols (cf. table 2). Regions corresponding to eqs. (16) and (17) are the upper left and lower right ones, respectively. False predictions are encircled.

Table 3
Some physical constants of selected aliphatic alcohols. (b.p.: boiling point; m.p.: melting point; n_d: refraction index [14].)

Compound	b.p.	m.p.	Density	n_d
methanol	65.0	−93.9	0.7914	1.3288
ethanol	78.5	−117.3	0.7893	1.3611
1-propanol	97.4	−126.5	0.8035	1.3850
2-propanol	82.4	−89.5	0.7855	1.3776
1-butanol	117.2	−89.5	0.8098	1.3993
2-butanol	99.5	–	0.8080	1.3954
2-methyl-1-propanol	108.0	–	0.8018	1.3955
2-methyl-2-propanol	82.3	25.5	0.7887	1.3878
1-pentanol	137.3	−79.0	0.8144	1.4101
2-pentanol	118.9	–	0.8103	1.4053
3-pentanol	116.1	–	0.8212	1.4104
2-methyl-1-butanol	128.0	–	0.8191	1.4102
3-methyl-2-butanol	112.0	–	0.8225	1.4089
2,2-dimethyl-1-propanol	113.0	52.5	0.812	–
1-hexanol	158.0	−46.7	0.8136	1.4178
2-hexanol	138.0	–	0.8104	1.4126
3-hexanol	132.0	–	0.8213	1.4150
2-methyl-2-pentanol	121.0	−103.0	0.8350	1.4100
2-methyl-3-pentanol	126.7	–	0.8243	1.4175
2-ethyl-1-butanol	146.3	−15.0	0.8326	1.4220
4-methyl-2-pentanol	133.0	–	0.8075	1.4100
1-heptanol	176.0	−34.1	0.8219	1.4249
2-heptanol	161.0	–	0.8190	1.4209
3-heptanol	157.0	−70.0	0.8227	1.4201
4-heptanol	161.0	−42.1	0.8183	1.4205
2,4-dimethyl-3-pentanol	138.7	−70.0	0.8288	1.4250
1-octanol	194.4	−16.7	0.8270	1.4295
1-nonanol	213.5	−5.5	0.8273	1.4333
1-decanol	229.0	–	0.8297	1.4372
1-undecanol	243.0	19.0	0.8298	1.4392
1-dodecanol	257.0	26.0	0.8309	–
1-tridecanol	152.0	32.5	0.8223	–

Good correlations are obtained for boiling points, refraction indices, and pharmacological activities. For melting points and densities where the correlation coefficient is smaller than 0.9 ranking is still predictable to a certain extent. The percentage of false predictions is always smaller than that corresponding to a random distribution. For refractive indices a surprisingly large number of false predictions is observed which is due to the nonlinearity of the Δn_{di} versus R_{0i} curve (cf. fig. 1). A cubic relationship yields higher r and reduces the number of false predictions from 11 to 1 (cf. table 5). The good correlation for pharmacological activities of aliphatic alcohols is not surprising because, due to the simple structure of these

Table 4
Pharmacological activities of selected aliphatic alcohols. (IGC: 50% growth inhibition of *Tetrahymena Pyriformis* [15]; LC: 50% of lethal dose for *Pymephales pyriformis* [16]; PC: toxicity for *Madison 517* fungi [15]; LIP: lipoxygenase inhibition [17]; SHL: sheep liver esterase inhibition [18].)

Compound	IGC	LC	PC	LIP	SHL
methanol	−0.23	−0.06	−0.24	−0.18	3.36
ethanol	−0.58	−0.51	−0.04	0.18	3.85
1-propanol	−1.16	−1.12	0.44	0.68	4.28
2-propanol	–	–	0.24	0.37	–
1-butanol	−1.48	−1.63	0.87	1.15	4.43
2-butanol	–	–	0.60	0.86	–
2-methyl-1-propanol	–	–	0.77	1.13	–
2-methyl-2-propanol	–	–	0.46	0.49	–
1-pentanol	−1.88	−2.27	1.38	1.61	5.05
2-pentanol	–	–	1.08	–	4.40
3-pentanol	–	–	1.01	–	4.26
2-methyl-1-butanol	–	–	1.19	1.34	4.70
2-methyl-2-butanol	–	–	1.44	–	–
3-methyl-1-butanol	–	–	–	–	4.77
2,2-dimethyl-1-propanol	–	–	–	–	4.08
1-hexanol	−2.53	−2.53	1.83	2.10	5.31
2-ethyl-1-butanol	–	–	1.73	–	–
1-heptanol	−3.02	−3.53	2.32	2.60	5.75
1-octanol	−3.50	−3.98	2.86	–	6.05
1-methyl-heptanol	–	–	2.49	–	–
2-ethyl-hexanol	–	–	2.55	–	–
1-nonanol	−3.77	−4.40	3.18	–	6.30
1-decanol	−4.25	−4.82	3.57	–	–
1-undecanol	−4.88	−5.22	–	–	–
1-dodecanol	−5.08	−5.27	–	–	–
1-tridecanol	−5.28	−5.59	–	–	–

Table 5
Coefficients of eq. (15) and false predictions (f.p.) for aliphatic alcohols. Data and notations from tables 1, 3 and 4.

Property	*a*	*b*	*r*	*n*	f.p.	Percent
b.p.	−1092	1063	−0.9473	29	6	21
m.p.	−625.1	610.7	−0.8446	16	5	31
density	−0.2373	0.2397	−0.8230	29	6	21
n_d	−0.4251	0.4664	−0.9458	29	11	38
n_d^3	−0.00839	0.00821	−0.9711	29	1	3
IGC	−26.52	26.27	−0.9947	12	0	0
LC	−28.94	28.86	−0.9956	12	1	8
PC	−25.48	24.98	−0.9862	19	1	5
LIP	−24.06	23.82	−0.9555	11	1	9
SHL	−19.64	19.49	−0.9010	13	1	8

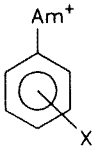
Table 6
Some physical properties and pharmacological activities of substituted phenols. (b.p.: boiling point; m.p.: melting point; n_d: refraction index [14]; AC: acute toxicity; HA: haemolytic activity; AA: antibacterial activity [20].)

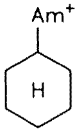
Derivative	b.p.	m.p.	n_d	AC	HA	AA
2-*s*-Bu	227.5	16.0	1.5200	3.37	3.30	3.40
2-*t*-Bu, 4-Me	237.0	55.0	1.4969	3.06	3.70	3.52
4-*t*-Bu	239.5	101.0	1.4787	3.28	3.10	3.19
4-*s*-Bu	241.0	61.5	1.5182	3.35	3.40	3.52
2-Me, 4-*t*-Bu	236.0	27.5	1.5230	3.31	3.52	3.40
2-*t*-Bu	221.0	–	1.5160	3.26	3.40	3.40
4-Pr	232.6	22.0	1.5379	3.23	2.82	3.05
4-I	–	93.5	–	3.15	3.00	3.15
2,4-di-Cl	210.0	45.0	–	3.03	3.00	3.30
4-Ph	–	166.0	–	3.10	3.30	3.05
4-nitro	–	115.0	–	2.60	1.92	2.32
2-nitro	216.0	45.5	1.5723	2.57	1.62	2.40
4-Br	238.0	66.4	–	2.79	2.82	2.82
2-Br, 4-Me	–	–	–	2.98	2.70	2.68
4-Cl	219.7	43.5	1.5579	2.59	2.52	2.70
4-Et	219.0	47.5	1.5239	2.95	2.52	2.92
2-Et	207.0	−18.0	1.5367	2.85	2.52	2.82
3,5-di-Me	219.5	68.0	–	2.89	2.70	2.51
2,4-di-Me	210.0	27.5	1.5420	2.82	2.70	2.52
2-Br	194.5	5.6	1.5890	2.69	2.52	2.52
2-Cl	174.9	9.0	1.5524	2.74	2.22	2.70
4-F	185.5	48.0	–	2.56	2.22	2.10
3-OH	178.0	111.0	–	2.10	0.59	1.52
2-F	151.5	16.1	–	2.32	1.89	2.10
2-Me	–	–	–	2.64	2.00	2.10
H	181.7	43.0	1.5408	2.41	1.24	1.74

molecules, all QSAR methods provide correct predictions. E.g. good correlations have been found between the information content of aliphatic alcohols and their pharmacological activities. For the toxicity of *Madison 517* fungi (PC), inhibition of lipoxygenase (LIP) and sheep liver esterase (SHL), simple linear regression equa-

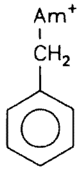
Table 7
Coefficients of eq. (15) and false predictions (f.p.) as defined by eqs. (16) and (17) for substituted phenols. Data and notations from table 6.

Property	*a*	*b*	*r*	*n*	f.p.	Percent
b.p.	337.1	−3.72	0.8120	20	0	0
m.p.	−199.1	59.15	−0.2485	22	7	32
n_d	0.0949	0.0179	0.2365	14	6	43
AC	−5.75	3.31	−0.8136	25	5	20
HA	−15.37	21.90	−0.7934	25	6	24
AA	−16.77	23.54	−0.8391	25	3	12

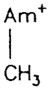




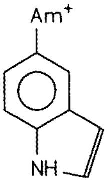
20



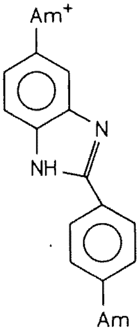
21



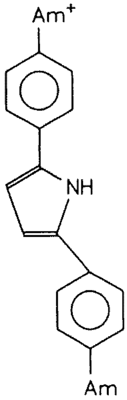
22



23

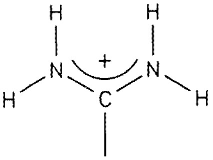


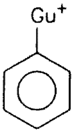
24



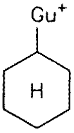
25

Am^+

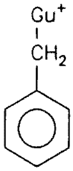




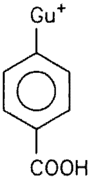
26



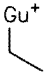
27



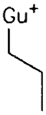
28



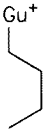
29



30



31



32

Gu^+

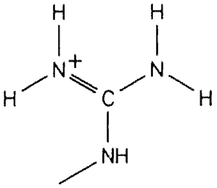


Fig. 2. For caption see next page.

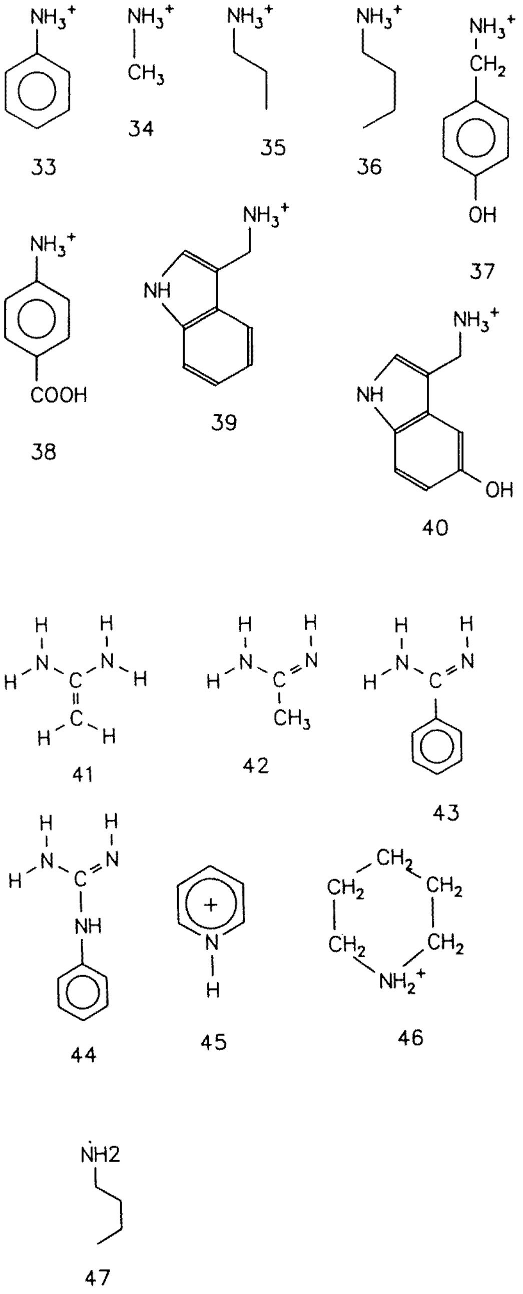


Fig. 2. Structural formulae of benzamidine inhibitors. X = **1**: 3-Me; **2**: 3-OH; **3**: 3-OMe; **4**: 3-OEt; **5**: 3-NO_2; **6**: 3-COOMe; **7**: 3-COOEt; **8**: 3-COMe; **9**: 3-CONHMe; **10**: 4-NH_2; **11**: 4-Me; **12**: 4-OH; **13**: 4-OMe; **14**: 4-OEt; **15**: 4-NO_2; **16**: 4-COOMe; **17**: 4-COOEt; **18**: 4-COMe; **19**: 4-CONHMe.

tions containing the information content yield $r = 0.992$ [19], 0.986 [17] and 0.970 [17], respectively.

3.2. SUBSTITUTED PHENOLS

Since aliphatic molecules are always good targets for QSAR studies the predictive power of molecular transforms has to be tested on more complicated molecules, as well. We selected a family of substituted phenols for that purpose, of which some physical and pharmacological activities are listed in table 6. Correlation coefficients are much smaller than for alcohols but the number of wrong predictions is still acceptable though close to the random value for the molar refractive index. It has to be noticed that, in spite of the very low correlation coefficient for the melting point, a better-than-random classification of the active and inactive derivatives is possible (table 7).

3.3. TRYPSIN INHIBITORS

In the previous examples we put $f_i = Z_i$ in eq. (2), but it is important to notice that mathematically f_i may be replaced by various atomic properties other than Z_i. In this subsection we present a study for low molecular weight trypsin inhibitors (cf. formulas **1–40** in fig. 2) for which we replaced f_i by the atomic charge, q_i and compared the prediction power of both 2D and 3D molecular transforms. Abstract distances, as defined by eqs. (2), (3) and (5) are displayed in table 8 while parameters of various regression equations are given in table 9.

As it is seen, 2D and 3D molecular transforms with $f_i = Z_i$ are in a close correlation which becomes much worse with $f_i = q_i$. This means that while for the former case the use of 3D molecular transforms is not feasible because of the considerably greater computational effort for providing them, if we replace Z_i by q_i the 3D function may give different (better) results in a statistical study. This is apparently not the case (cf. table 9): the correlation coefficient is slightly worse for the 3D than for the 2D molecular transform. The use of atomic net charges is, however, justified; the correlation coefficients considerably increase in both cases even if reaching only very low values. However, the prediction power increases, lowering the number of outliers defined in eqs. (16), (17) from 11 to 9 and 8 if replacing 2DZ by 2DQ and 3DQ (cf. table 8, $n =$ **1–40**). It is interesting to notice that the number of false predictions not obeying eq. (16) is only 2, 2, 1 and 1 for the 2DZ, 3DZ, 2DQ and 3DQ cases, respectively. This means that prediction of a compound being inactive is quite reliable which makes our method feasible for the screening of potentially inactive derivatives as mentioned in the introduction.

We checked the reliability of our predictions for completely ineffective compounds with structures only slightly differing from effective inhibitors (molecules **41–47** in fig. 2 and in table 8). As it is seen, the number of false predictions (cf. eq. (17)) is 5, 5, 1 and 4 for the 2DR, 3DR, 2DQ and 3DQ molecular transforms,

Table 8
Abstract distances, as defined by various forms of eqs. (2), (3) and (5) for trypsin inhibitors. (2DZ: 2D distance, $f_i = Z_i$; 3DZ: 3D distance, $f_i = Z_i$; 2DQ: 2D distance, $f_i = q_i$; 3DQ: 3D distance, $f_i = q_i$, $\Delta pK_i = pK_i(\mathrm{X{=}H}) - pK_i(n)|$, cf. fig. 2.)

Molecule *(n)*	ΔpK_i	2DZ	3DZ	2DQ	3DQ
1	0.286	0.9990	0.9988	0.9180	0.9913
2	0.300	0.9990	0.9989	0.9833	0.9856
3	0.179	0.9965	0.9964	0.9765	0.9869
4	0.212	0.9923	0.9929	0.9794	0.9784
5	0.894	0.9937	0.9938	0.9775	0.9442
6	0.725	0.9890	0.9955	0.9843	0.9738
7	0.594	0.9836	0.9859	0.9800	0.9638
8	0.161	0.9933	0.9933	0.9867	0.9786
9	0.957	0.9888	0.9890	0.7760	0.7958
10	0.209	0.9990	0.9988	0.9742	0.9707
11	0.258	0.9990	0.9988	0.9890	0.9922
12	0.480	0.9990	0.9978	0.9439	0.9550
13	0.286	0.9965	0.9964	0.9890	0.9834
14	0.781	0.9923	0.9928	0.9838	0.9751
15	1.299	0.9938	0.9936	0.9682	0.9442
16	1.258	0.9892	0.9901	0.9784	0.9769
17	1.082	0.9838	0.9857	0.9794	0.9616
18	1.286	0.9934	0.9932	0.9862	0.9810
19	0.927	0.9890	0.9892	0.9659	0.9608
20	1.411	0.9999	0.9976	0.9924	0.9773
21	2.961	0.9981	0.9985	0.9764	0.9796
22	3.191	0.9717	0.9740	0.3191	0.9323
23	0.241	0.9973	0.9961	0.9859	0.9765
24	0.011	0.9597	0.9580	0.9627	0.9415
25	0.029	0.9469	0.9452	0.9488	0.9330
26	0.641	0.9984	0.9988	0.9796	0.9672
27	2.421	0.9976	0.9963	0.9790	0.9627
28	2.651	0.9941	0.9968	0.9818	0.9685
29	1.621	0.9881	0.9892	0.9754	0.9465
30	2.631	0.9967	0.9966	0.9859	0.9692
31	1.501	0.9993	0.9984	0.9902	0.9657
32	1.891	0.9979	0.9958	0.9884	0.9665
33	1.381	0.9988	0.9987	0.6646	0.6480
34	3.571	0.9478	0.9513	0.5266	0.3525
35	2.011	0.9917	0.9926	0.6695	0.5315
36	2.861	0.9976	0.9958	0.7112	0.5694
37	2.301	0.9982	0.9978	0.8482	0.8265
38	1.241	0.9966	0.9960	0.9041	0.8848
39	2.161	0.9968	0.9947	0.5070	0.6572
40	1.561	0.9934	0.9910	0.5905	0.8558
41	–	0.9694	0.9609	0.6348	0.8627
42	–	0.9697	0.9613	0.1263	0.8550
43	–	1.0000	0.9990	−0.0621	0.8016
44	–	0.9985	0.9962	−0.0987	0.8377
45	–	0.9835	0.9906	−0.0409	0.3716
46	–	0.9888	0.9919	0.4777	0.5157
47	–	0.9929	0.9893	0.6950	0.6538
0.5 ($R_{0\max} + R_{0\min}$)		0.9731	0.9721	0.6558	0.6719

Table 9
Coefficients of eq. (15) and false predictions (f.p) as defined by eqs. (16) and (17) for trypsin inhibitors. Data and notations from table 8.

Variable pair	*a*	*b*	*r*	*n*	f.p.	Percent
2DZ–3DZ	0.9672	0.0324	0.9945	40	–	–
2DQ–3DQ	0.6341	0.3341	0.7208	40	–	–
pK_i–2DZ	−5.842	7.051	−0.0759	40	11	28
pK_i–3DZ	−4.209	5.432	−0.0532	40	11	28
pK_i–2DQ	−3.309	4.229	−0.5595	40	9	23
pK_i–3DQ	−3.513	4.433	−0.5226	40	8	20

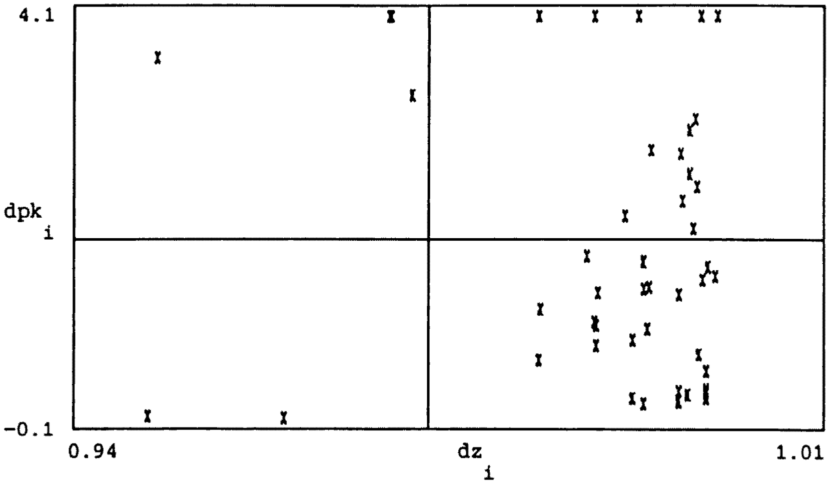


Fig. 3. ΔpK_i versus 2DZ plot for trypsin inhibitors (cf. table 8, $n = 1$–47). Regions corresponding to eqs. (16) and (17) are the upper left and lower right ones, respectively.

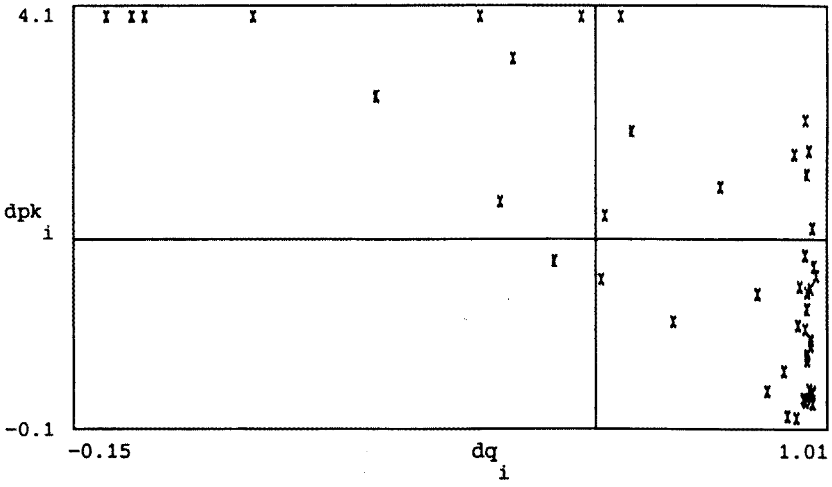


Fig. 4. ΔpK_i versus 2DQ plot for trypsin inhibitors (cf. table 8, $n = 1$–47). Regions corresponding to eqs. (16) and (17) are the upper left and lower right ones, respectively.

respectively. Consequently, the generalization of eq. (2) with $f_i = q_i$ is useful and provides a tool for the distinction between active and inactive inhibitors of trypsin.

4. Conclusions

We propose the abstract distance between molecular transforms of eq. (2) as a measure of molecular similarity. It can be used in quantitative regression equations for simpler classes of compounds like aliphatic alcohols while for more complicated cases a yes/no prediction is possible deciding whether a molecule is active or not. Molecular transforms can be generalized by replacing atomic numbers with net charges, thus making possible a classification between chemically closely related molecules possessing drastically different activities.

References

[1] M.A. Johnson and G.M. Maggiora, eds., *Concepts and Applications of Molecular Similarity* (Wiley, New York, 1990).
[2] H. Lipson and C.A. Taylor, *Fourier Transforms and X-Ray Diffraction* (Bell, London, 1958).
[3] L.J. Soltzberg and C.L. Wilkins, J. Am. Chem. Soc. 98 (1976) 7139.
[4] L.J. Soltzberg and C.L. Wilkins, J. Am. Chem. Soc. 99 (1977) 439.
[5] L.J. Soltzberg and C.L. Wilkins, J. Am. Soc. 99 (1977) 4006.
[6] Z. Gabányi, P.R. Surján and G. Náray-Szabó, Eur. J. Med. Chem. 17 (1982) 1982.
[7] G. Náray-Szabó, J. Mol. Struct. (THEOCHEM) 134 (1986) 401.
[8] J.W. King, R.J. Kassel and B.B. King, Int. J. Quant. Chem. Quant. Biol. Symp. 17 (1990) 27.
[9] J.W. King and R.J. Kassel, Int. J. Quant. Chem. Quant. Biol. Symp. 18 (1991) 289.
[10] R. Wierl, Appl. Phys. (Leipzig) 8 (1931) 521.
[11] P. Broto, G. Moreau and C. Vandycke, Eur. J. Med. Chem. 19 (1984) 71.
[12] R. Carbó, L. Leyda and M. Arnau, Int. J. Quant. Chem. 17 (1980) 1185.
[13] E.E. Hodgkin and W.G. Richards, Int. J. Quant. Chem. Quant. Biol. Symp. 14 (1987) 105.
[14] *Handbook of Chemistry and Physics*, 65th Ed. (CRC Press, Boca Raton, 1984–85) pp. C65–C576.
[15] T.W. Schultz, L.M. Arnold, T.S. Wilke and M.P. Moulton, Ecotox. Environ. 19 (1990) 243.
[16] G.D. Veith, D.J. Can and L.I. Brook, Can. J. Fish. Aquat. Sci. 40 (1983) 743.
[17] L.B. Kier, J. Pharm. Sci. 69 (1980) 807.
[18] D. Glick and C.G. King, J. Biol. Chem. 94 (1932) 497.
[19] V.K. Gombar and S.L. Wadhwa, Arzneimitt.-Forsch. 32 (1982) 715.
[20] G.L. Biagi, J. Med. Chem. 18 (1975) 868.
[21] F. Markwardt, H. Landmann and P. Walsmann, Eur. J. Biochem. 6 (1968) 502.
[22] F. Markwardt, P. Walsmann and H.G. Kazmirowski, Pharmazie 24 (1969) 400.
[23] F. Markwardt, H. Landmann and P. Walsmann, Pharmazie 25 (1970) 551.
[24] F. Markwardt, P. Walsmann, J. Stürzebecher, H. Landmann and G. Wagner, Pharmazie 28 (1973) 5.
[25] T. Inagami, in: *Proteins: Structure and Function*, Vol. 1, eds. M. Funatsu, K. Hiromi, T. Murachi and K. Narita (Kodansha/Wiley, Tokyo/New York, 1972) p. 1.
[26] R.R. Tidwell, J.D. Geratz, O. Dann, G. Volz, D. Zeh and H. Loewe, J. Med. Chem. 21 (1978) 613.