Simulation of Carcinogenic Activity of Chemical Compounds

VLADIMIR K. MUKHOMOROV

University of Naples "Federico II", Naples, ITALY
vmukhomorov@mail.ru

Abstract: - We have established two classification rules that statistically accurate allow to separate carcinogenically active chemical compounds from inactive chemical compounds. The electronic and information properties of molecules are used as molecular descriptors. The threshold values of descriptors that characterize and determine the presence or absence of carcinogenic properties of chemical compounds of various classes are found. Statistical quantitative indicators of the quality of classification rules are given, including the error of model. The proposed classification rules allow one to analyze the carcinogenic properties of different classes of chemical compounds from a unified view. Classification rules were tested for various classes of chemical compounds. We studied the chemical compounds of the following classes: a number of nitroso compounds, halogen-containing organic substances, sulfur-containing organic substances, aromatic amines and related compounds, dyes, oxy compounds, chemical compounds of the mustard type, some medications and polycyclic aromatic hydrocarbons. The model does not use the idea of the existence of reaction regions in molecules. A total of 600 chemical compounds were examined. Carcinogenic properties of a number of known metabolites were analyzed. For the studied series of metabolites, it was found that their molecular descriptors regularly exceed the descriptors of the initial chemical compounds in magnitude.

Keywords: - Carcinogenicity, molecular descriptors, statistical methods, classification rules, threshold values, information function, electronic descriptor, polycyclic hydrocarbons, metabolite, donor, acceptor, intermolecular

1. Introduction

We will present the results of constructing a statistical model that links the carcinogenic activity of chemical compounds with their molecular structure. In connection with the deteriorating ecology, the intensity of the study of carcinogenic agents is not only diminished, but has increased significantly. As reported in [1] motley list of agents that induce malignant tumors shows that the well-documented carcinogenic activity of chemicals does not find common physical or chemical characteristic. The challenge is to understand how can have identical biological effect is so different nature of agents. Here we will return to this issue which has become banal. Using the ideas of condensed matter physics, information theory and statistical methods we try to point out such molecular characteristics that allow probabilistic separate carcinogenic chemical compounds of various classes of chemical compounds that do not possess activity.

2. Problem Formulation

It is now becoming apparent to generate a need to develop new a fast track methodology to identify of carcinogenic chemical compounds. This is particularly important when you consider the laboriousness, high cost and relative duration of experiments with high needs in exploring the vast amount of newly synthesized chemical compounds of various classes. [1]. It is therefore important to establish a relatively simple and unified approach to the evaluation of biological activity of the vast diversity of chemical compounds of different classes on the basis only of knowledge of atomic structure of molecules. This will allow us quickly, simply and statistically reliably to eliminate the potential danger of contact with such substances. In addition, such an approach will allow making prerequisites for scholars to deepen their knowledge about the mechanism of carcinogenesis.

A. Haddow [2] made a review lecture on the VIII International Anticancer Congress. Haddow said the following: “The main issue is to understand, how can such different agents, chemicals ... provide the same final result”. G.M.Badger [3] and R. Schoental [4] has draw attention to this same issue (as so distinguished by its chemical structure of molecules can cause the same end result).

3. Statistical method for constructing classification rules

We have presented a large enough sample of agents in Table 1. Agents belong to different classes of chemical compounds. The experimental data were mostly taken from the reference book [5]. Data have been supplemented with data from the article [6]. Chemical compounds of Table 1, which reliably have carcinogenic activity marked with the symbol "+". Chemical compounds do not possess carcinogenic properties marked with the symbol "-".

The analysis of Table 1 will be carried out using molecular descriptors Z and H [7-9]. The descriptor Z determines the average number of valence electrons in the molecule:

$$Z = \sum_{i=1}^{N} n_i Z_i / N \, . \qquad (1)$$

Here n_i is the number of atoms of the i-th type in the molecule, Z_i is the number of valence electrons of the i-th atom, N is the total number of atoms in the molecule. The descriptor Z was used to evaluate the biological activity of molecules in work [6]. At the same time, assumptions of the pseudopotential theory of the condensed state physics have been used. As is well known, the descriptor Z is the multiplier that fits into the equation of the pseudopotential.

The descriptor H is an information function that is defined by the following equation [10]

$$H = \sum_{i=1}^{N} p_i \log_2 p_i \, , \qquad (2)$$

Here $p_i = n_i / N$ is the fraction of atoms of the n_i species in the molecule. Moreover, for p_i the following relations are satisfied: $0 \le p_i \le 1$, $\sum_{i=1}^{N} p_i = 1$, $\sum_{i=1}^{N} n_i = N$. The information function makes it possible provides insight into the diversity of a multicomponent system. The smaller the information function value, the more diverse (in the relative content of atoms in molecules) a multicomponent system. The choice of the base of logarithm in equation (2) was not a matter of principle importance.

We will use the method of associations (conjugations) to construct a classification rule that will allow us to separate carcinogenic chemicals from non-carcinogenic substances. Preliminarily, we will determine the threshold value of the descriptor $Z^{(\mathrm{th})}$ which separates with some probability the chemical compounds according to their biological activity. Using the data in Table 1, we determine sampling mean of the descriptor Z ($N = 255$). The sample contains N_1 = 141 chemical compounds that have confirmed carcinogenic activity and N_2 = 114 chemical compounds that are not carcinogenic or have no confirmed carcinogenic activity.

An analysis of the Table 1 showed the descriptor Z is not normally distributed. However, if we transform the descriptor of Z then we can lead the sample to a normal distribution form. For transformation, we can select the taking the logarithm to the base 10. The result should be multiplied by 10 so that the numerical values would be higher than 1: $10 \cdot \log_{10}$. This allows us lead to normal view the set of elements that make up the sample [11]. After completion of the operation of taking the logarithm, the sample (Table 1) satisfies the normality condition:

$$\chi^2 = 13.2 < \chi^{2(\mathrm{cr})}_{0.05}(df = 7) = 14.1, \; p = 0.20. \quad (3)$$

The smaller the χ^2 the higher the probability that the random variable is normally distributed; the degrees of freedom is equal to $df = n - l - 1$, l = 2 is the number of intervals of the range of variation of the random variable; l = 2 is the number of parameters for lognormal distribution; p is the level of significance of the criterion, which determines the probability of error in deviating from the hypothesis of normality [12]. We should accept the null hypothesis of a normal distribution because $p >$ 0.05. To test the homogeneity of the normal sample, we can use the criterion τ. For reasons of presentation, we use the following notation: $\langle Z \rangle = 10 \cdot \log(Z)$. Firstly we determine the statistics of the average value $\langle Z \rangle^{(av)}$:

$$N = 255, \quad \langle Z \rangle^{(av)} = 4.87 \pm 0.05, \quad \langle Z \rangle^{(min)} = 2.746, \quad \langle Z \rangle^{(max)} = 8.062, \quad S = 0.718. \qquad (4)$$

Table 1. *The carcinogenic activity, electronic and information descriptors of chemical compounds.*

N	Chemical compound	Gross formula	Activity	*Z*	*H,bits*
1	Hydrazine	N_2H_4	+	2.333	0.919
2	Carbon tetrachloride	CCl_4	+	6.400	0.723
3	Chloroform	$CHCl_3$	+	5.200	1.372
4	Formaldehyde	CH_2O	+	3.000	1.501
5	Thiourea	CH_4N_2S	+	3.000	1.750
6	Semicarbazide	CH_6ClN_3O	+	3.167	1.897
7	Acetaldehyde	C_2H_4O	+	2.572	1.379
8	Ethylene oxide	C_2H_4O	+	2.572	1.379
9	Ethylene sulfide	C_2H_4S	+	2.572	1.379
10	Amitrole	$C_2H_4N_4$	+	3.200	1.522
11	Vinilchloride	C_2H_5Cl	+	2.500	1.299
12	Thioacetamide	C_2H_5NS	+	2.667	1.658
13	Ethylenethiourea	$C_3H_6N_2S$	+	2.833	1.730
14	1,1’-Dimethylhydrazine	$C_2H_8N_2$	+	2.167	1.252
15	1,2-Dimethylhydrazine	$C_2H_8N_2$	+	2.167	1.252
16	Bis(chloromrthyl) ether	$C_2H_4Cl_2O$	+	3.556	1.837
17	Chloromethyl ether	C_2H_5ClO	+	2.889	1.658
18	N-Nitrosodimethylamine	$C_2H_6N_2O$	+	2.727	1.686
19	Acetamide	C_2H_5NO	+	2.667	1.658
20	Methylmethanesulphanate	$C_2H_4O_3S$	+	3.167	1.729
21	Dimethylsulphate	$C_2H_6O_4S$	+	3.385	1.738
22	β-Propiolactone	$C_3H_4O_2$	+	3.111	1.531
23	Propylene oxide	C_3H_6O	+	2.400	1.296
24	Acrylamide	C_3H_5NO	+	2.800	1.686
25	Urethane	$C_3H_7NO_2$	+	2.769	1.669
26	Ethylenethiourea	$C_3H_6N_2S$	+	2.833	1.730
27	Nitrosoethylurea	$C_3H_7N_3O_2$	+	3.067	1.830
28	1,3-Propane sultone	$C_3H_6O_3S$	+	3.231	1.776
29	2-Methylaziridine	C_3H_7N	+	2.182	1.241
30	Ethylmethanesulphonate	$C_3H_8O_3S$	+	2.933	1.673
31	Thiouracil	$C_4H_4N_2OS$	+	3.500	2.085
32	1,2-Diethylhydrazine	$C_4H_{12}N_2$	+	2.111	1.225
33	Tetramethyllead	$Pb(CH_3)_4$	+	1.882	1.087
34	N-Nitrosodiethylamine	$C_4H_{10}N_2O$	+	2.471	1.545
35	Allyl isothiocyanate	C_4H_5NS	+	2.909	1.677
36	Zineb	$C_4H_6N_2S_4Zn$	+	3.412	2.117
37	Gyromitrin	$C_4H_8N_2O$	+	2.667	1.640
38	Methylazoxymethanol acetate	$C_4H_8N_2O_3$	+	3.059	1.808
39	Diethyl sulphate	$C_4H_{10}O_4S$	+	2.947	1.658
40	Mustard gas	$C_4H_8Cl_2S$	+	2.933	1.640
41	Bis-(2-chloroethyl) ether	$C_4H_8Cl_2O$	+	2.933	1.640
42	6-Mercaptopurine	$C_5H_4N_4S$	+	3.571	1.835
43	Methylthiouracil	$C_5H_6N_2OS$	+	3.200	1.966
44	Potassium bis (2-hydroxyethyl) dithiocarbamate	$C_5H_{10}KNO_2S_2$	+	2.857	2.067

Table continuation

45	Nitrogen mustard hydrochloride	$C_5H_{12}Cl_3N$	+	2.762	1.565
46	Azaserine	$C_5H_7N_3O_4$	+	3.474	1.931
47	Niridazole	$C_6H_6N_4O_3S$	+	3.700	2.133
48	2-Amino-5-(nitro-2-furyl)-1,3,4-thiadiazole	$C_6H_4N_4O_3S$	+	4.000	2.155
49	Dichlorbenzene	$C_6H_4Cl_2$	+	3.500	1.459
50	Benzene	C_6H_6	+	2.500	1.000
51	Aniline	C_6H_7N	+	2.572	1.296
52	Thiophosphamide	$C_6H_{12}N_3PS$	+	2.696	1.772
53	1,4-Butanediol dimethan sulphanate	$C_6H_{14}O_6S_2$	+	3.071	1.725
54	Propylthiouracil	$C_7H_{10}N_2OS$	+	2.857	1.780
55	Cyclophosphamide	$C_7H_{15}N_2PO_2Cl_2$	+	2.896	1.953
56	Isophosphamide	$C_7H_{15}N_2PO_2Cl_2$	+	2.896	1.953
57	Styrene	C_8H_8	+	2.500	1.000
58	Styrene oxide	C_8H_8O	+	2.706	1.264
59	Phenelzinesulphate	$C_8H_{12}N_2{\cdot}H_2SO_4$	+	2.966	1.848
60	Allyl isovalerate	$C_8H_{14}O_2$	+	2.417	1.281
61	Streptozoticin	$C_8H_{15}N_3O_7$	+	3.152	1.802
62	Sulfallate*)	$C_8H_{14}ClNS_2$	+	2.692	1.650
63	Cycasin	$C_8H_{16}N_2O_7$	+	3.030	1.722
64	Ethionamide	$C_8H_{10}N_2S$	+	2.762	1.573
65	Bis (1-Aziridinyl) morpholino-phosphine sulphide	$C_8H_{16}N_3OPS$	+	2.667	1.815
66	N-[4-(5-Nitro-2-furyl)-2-thiazolyl] acetomide	$C_9H_7N_3O_4S$	+	3.667	2.046
67	Mirex	$C_{10}Cl_{12}$	+	5.636	0.994
68	Heptachlor	$C_{10}H_5Cl_7$	+	4.273	1.529
69	Dihydrosafrole	$C_{10}H_{12}O_2$	+	2.667	1.325
70	Diallate**)	$C_{10}H_{17}Cl_2NOS$	+	2.750	1.728
71	Safrole	$C_{10}H_{10}O_2$	+	2.818	1.349
72	Benzofluoranthene	$C_{10}H_{12}$	+	2.364	0.994
73	Eugenol	$C_{10}H_{12}O_2$	+	2.667	1.325
74	β - Naphthylamine	$C_{10}H_9N$	+	2.700	1.235
75	2-(2’-Furyl-3-(5-nitro-2-furyl) acrylamide)	$C_{11}H_8N_2O_5$	+	3.539	1.790
76	Polychlorinate biphenyl	$C_{12}Cl_{10}$	+	5.364	0.994
77	Aldrin	$C_{12}H_8Cl_6$	+	3.769	1.526
78	Aramite^{R}	$C_{12}H_{23}ClO_4S$	+	2.634	1.576
79	Dioxin	$C_{12}H_4Cl_4O_2$	+	4.182	1.686
80	4-Aminobiphenyl	$C_{12}H_{11}N$	+	2.667	1.207
81	Chrysoidine	$C_{12}H_{13}N_4Cl$	+	2.933	1.603
82	Benzidine	$C_{12}H_{12}N_2$	+	2.692	1.314
83	Azobenzene	$C_{12}H_{10}N_2$	+	2.833	1.325
84	Aminoazobenzene	$C_{12}H_{11}N_3$	+	2.846	1.400
85	4-Nitrobiphenyl	$C_{12}H_9NO_2$	+	3.083	1.521
86	4-Hydroxyazobezene	$C_{12}H_{10}N_2O$	+	2.960	1.514
87	4,4’-Thidianiline	$C_{12}H_{12}N_2S$	+	2.815	1.494
88	Resorcinol diglycidylether	$C_{12}H_{14}O_4$	+	2.867	1.430
89	Tris(aziridinul)-para-benzoquine	$C_{12}H_{13}N_3O_2$	+	2.933	1.644
90	4,4’- Methylene bis (2-chloraniline)	$C_{13}H_{12}Cl_2N_2$	+	3.035	1.578
91	3-Amino-1,4-dimethyl-5H-pyrido (4,3-b) indole	$C_{13}H_{13}N_3$	+	2.759	1.377
92	4,4’ - Methylenedianiline	$C_{13}H_{14}N_2$	+	2.621	1.292

Table continuation

93	*ortho*-Aminoazotoluene	$C_{14}H_{15}N_3$	+	2.688	1.354
94	DDT	$C_{14}H_9Cl_5$	+	3.571	1.470
95	DDD	$C_{14}H_{10}Cl_4$	+	3.357	1.432
96	3,3'-Dimethylbensidine	$C_{14}H_{16}N_2$	+	2.563	1.272
97	3,3'-Dimethoxy-benzidine	$C_{14}H_{16}N_2O_2$	+	2.765	1.519
98	*para*-Dimethylamino-azobenzene	$C_{14}H_{15}N_3$	+	2.688	1.354
99	Oxazepan	$C_{15}H_{11}ClN_2O_2$	+	3.226	1.707
100	3'-Methoxy-4-dimethyl-aminoazobenzene	$C_{15}H_{17}N_2O$	+	2.657	1.413
101	Sudan Brown RR	$C_{15}H_{14}N_4$	+	2.849	1.411
102	C. I. Disperse yellow 3	$C_{15}H_{15}N_3O_2$	+	2.914	1.588
103	Sudan 1	$C_{16}H_{12}N_2O$	+	2.968	1.438
104	Chlorobenzilate	$C_{16}H_{14}Cl_2O_3$	+	3.143	1.595
105	Methoxychlor	$C_{16}H_{15}Cl_3O_2$	+	3.111	1.577
106	Yellow OB	$C_{17}H_{15}N_3$	+	2.800	1.334
107	Oil orange SS	$C_{17}H_{14}N_2O$	+	2.882	1.717
108	Auromine	$C_{17}H_{22}N_3Cl$	+	2.605	1.418
109	Diacetylaminoazotolune	$C_{18}H_{19}N_3O_2$	+	2.810	1.523
110	Sudan II	$C_{18}H_{16}N_2O$	+	2.811	1.397
111	Citrus red N2	$C_{18}H_{16}N_2O_3$	+	2.974	1.547
112	Ponceau MX	$C_{18}H_{14}N_2Na_2O_7S_2$	+	3.378	2.069
113	Benzanthracene	$C_{18}H_{12}$	+	2.800	0.971
114	Zearalenone	$C_{18}H_{22}O_2$	+	2.524	1.222
115	Ponceau 3R	$C_{19}H_{16}N_2Na_2O_7S_2$	+	3.292	2.036
116	Senkirkine	$C_{19}H_{27}NO_6$	+	2.717	1.490
117	Piperonyl butoxide	$C_{19}H_{30}O_5$	+	2.491	1.426
118	Ethylselenac	$C_{20}H_{40}N_4S_8Se$	+	2.685	1.651
119	7H-Dibenzocarbazole	$C_{20}H_{13}N$	+	2.882	1.130
120	Mestranal	$C_{21}H_{26}O_2$	+	2.490	1.197
121	Hycantone mesilate	$C_{21}H_{28}N_2O_5S_2$	+	2.828	1.678
122	Dibenzacridine	$C_{21}H_{13}N$	+	2.914	1.120
123	Lasiocarpine	$C_{21}H_{33}NO_7$	+	2.645	1.465
124	Dibenzantracene	$C_{22}H_{14}$	+	2.833	0.964
125	Sudan Red 7B	$C_{24}H_{21}N_5$	+	2.840	1.366
126	Searlet Red	$C_{24}H_{20}N_4O$	+	2.898	1.442
127	Dibenzopyrene	$C_{24}H_{14}$	+	2.895	0.950
128	Oestradiol 3-benzoate	$C_{25}H_{28}O_3$	+	2.607	1.246
129	Blue VRS	$C_{27}H_{31}N_2O_6S_2 \cdot Na$	+	2.870	1.739
130	Direct Brown	$C_{31}H_{18}CuN_6Na_2O_9S$	+	3.456	2.048
131	Direct Blue 6	$C_{32}H_{20}N_6Na_4O_{14}S_4$	+	3.625	2.181
132	Direct Black 38	$C_{34}H_{25}N_9Na_2O_7S_2$	+	3.317	1.984
133	Trypan Blue	$C_{34}H_{24}N_6Na_4O_{14}S_4$	+	3.512	2.149
134	Direct Blue 6	$C_{34}H_{24}N_6Na_4O_{14}S_4$	+	3.512	2.149
135	Brilliant Blue FCF	$C_{37}H_{34}N_2O_9S_3 \cdot 2NH_4$	+	2.968	1.722
136	Fast Green FCF	$C_{37}H_{34}N_2O_{10}S_3 \cdot 2Na$	+	3.091	1.827
137	Guinea Green B	$C_{37}H_{35}N_2O_6S_2 \cdot Na$	+	2.916	1.655
138	Light Green SF	$C_{37}H_{34}N_2O_9S_3 \cdot 2Na$	+	3.058	1.811
139	Benzyl Violet 4B	$C_{39}H_{40}N_3O_6S \cdot Na$	+	2.822	1.611
140	Bleomycin A_2	$C_{55}H_{84}N_{17}O_{21}S_3$	+	2.961	1.817
141	Bleomycin B_2	$C_{55}H_{84}N_{20}O_{21}S_2$	+	2.978	1.818
142	Ethylen	C_2H_4	-	2.000	0.919
143	Vinyl bromid	C_2H_3Br	-	3.000	1.459
144	Tetrafluorethylen	C_2F_4	-	5.000	0.920
145	Acrylic acid	$C_3H_4O_2$	-	3.111	1.531
146	2-Amino-5-nitrothiazole	$C_3H_3N_3O_2S$	-	4.000	2.230
147	Allylchloride	C_3H_5Cl	-	2.667	1.325

Table continuation

148	5-Fluorouracil	$C_4H_3FN_2O_2$	-	4.000	2.189
149	γ-Butyrolactone	$C_4H_6O_2$	-	2.833	1.460
150	Dicetone	$C_4H_4O_2$	-	3.200	1.522
151	Succinyl oxide	$C_4H_4O_3$	-	3.455	1.573
152	Allylisothiozyanat	C_4H_5NS	-	2.909	1.677
153	Maneb	$C_4H_6MnN_2S_4$	-	3.412	2.117
154	Alloxan	$C_4H_2N_2O_4$	-	4.333	1.919
155	Maleic hydrazide	$C_4H_4N_2O_2$	-	3.500	1.919
156	Trichlorfon	$C_4H_8Cl_3O_4P$	-	3.700	2.084
157	Dichlorvos	$C_4H_7Cl_2O_4P$	-	3.667	2.078
158	Sodium diethyldithiocarbamate	$C_5H_{10}NNaS_2$	-	2.526	1.783
159	Amonium urate acid	$C_5H_7N_5O_3$	-	3.500	1.941
160	Xantin	$C_5H_4N_4O_2$	-	3.733	1.933
161	5-Nitro-2-furamidoxime	$C_5H_5N_3O_4$	-	3.765	1.972
162	N-Nitrosoproline	$C_5H_8N_2O_3$	-	3.111	1.817
163	N-Nitrosohydroxyproline	$C_5H_8N_2O_4$	-	3.263	1.848
164	Quintezene	$C_6Cl_5NO_2$	-	5.429	1.728
165	5-Nitro-2-furaldehidesemi-carbazone	$C_6H_6N_4O_4$	-	3.700	1.971
166	1,2-Diamino-4-nitrobenzene	$C_6H_7N_3O_2$	-	3.222	1.841
167	N-Vinyl-2-pyrrolidone	C_6H_9NO	-	2.588	1.497
168	Cyclamic acid	$C_6H_{13}NO_3S$	-	2.750	1.736
169	Sodium cyclamate	$C_6H_{12}NNaO_3S$	-	2.750	1.948
170	Phenol	C_6H_6O	-	2.769	1.315
171	Hydroquinone	$C_6H_6O_2$	-	3.000	1.449
172	4-Amino-2-nitrophenol	$C_6H_6N_2O_3$	-	3.412	1.866
173	5-Nitro-2-furaldehyde semicarbazone	$C_6H_6N_4O_4$	-	3.700	1.971
174	Thiram	$C_6H_{12}N_2S_4$	-	2.917	1.730
175	Ledate	$C_6H_{12}N_2S_4Pb$	-	2.960	1.903
176	Ziram	$C_6H_{12}N_2S_4Zn$	-	2.880	1.903
177	Nithiazide	$C_6H_8N_4O_3S$	-	3.455	2.084
178	Treosulphan	$C_6H_{14}O_8S_2$	-	3.267	1.747
179	Trichlorotriethylamine hydrochloride	$C_6H_{12}Cl_3N{\cdot}HCl$	-	3.357	1.964
180	Salicyclic acid	$C_7H_6O_3$	-	3.250	1.505
181	N-methyl-N, 4-dinitroso aniline	$C_7H_7N_3O_2$	-	3.263	1.824
182	Theophyllin	$C_7H_8N_4O_2$	-	3.238	1.838
183	1-[(Nitrofurfurylidine)-amino] hydantion	$C_8H_6N_4O_5$	-	3.826	1.953
184	Alloxantin	$C_8H_6N_4O_8$	-	4.077	1.950
185	Piperonyl	$C_8H_6O_3$	-	3.294	1.484
186	Furazolidone	$C_8H_7N_3O_5$	-	3.652	1.914
187	Coffeinum	$C_8H_{10}N_4O_2$	-	3.083	1.784
188	*para*-Dimethylamino-benzenediazo sodium sulafonate	$C_8H_{10}N_3NaO_3S$	-	3.154	2.134
189	Methyl-parathion	$C_8H_{10}NO_5PS$	-	3.385	2.053
190	Sulfallate	$C_8H_{14}ClNS_2$	-	2.692	1.651
191	Azathioprine	$C_9H_7N_7O_2S$	-	3.692	2.015
192	Ferbam	$C_9H_{18}FeN_2S_6$	-	2.892	1.862
193	Fluometuron	$C_{10}H_{11}F_3N_2O$	-	3.500	1.865
194	Strobane^R	$C_{10}H_9Cl_7$	-	3.769	1.570
195	Sulfametoxazole	$C_{10}H_{11}N_3O_3S$	-	3.214	1.922
196	Chloropropham	$C_{10}H_{12}ClNO_2$	-	3.071	1.843

Table continuation

197	Adenosin	$C_{10}H_{13}N_5O_4$	-	3.188	1.846
198	Malathion	$C_{10}H_{19}O_6PS_2$	-	3.231	2.053
199	Parathion	$C_{10}H_{14}NO_5PS$	-	3.125	1.934
200	1-Naphthylthiourea	$C_{11}H_{10}N_2S$	-	2.917	1.532
201	Sulfafurazole	$C_{11}H_{13}N_3O_3S$	-	3.296	1.942
202	2-(2-Furyl)-3-(5-nitrofuryl) acrylamide	$C_{11}H_8N_2O_5$	-	3.539	1.790
203	Carrageenan	$C_{11}H_{17}O_{12}S$	-	3.390	1.685
204	Fast Yellow C.I.	$C_{12}H_{11}N_3O_6S_2$	-	3.588	2.048
205	Carbaryl	$C_{12}H_{11}NO_2$	-	2.923	1.505
206	Alizarin Yellow R	$C_{12}H_9N_3O_5$	-	3.517	1.827
207	2,4-Diphenyldiamine	$C_{12}H_{12}N_2$	-	2.692	1.315
208	4,4’- Methylenedianiline	$C_{12}H_{14}N_2$	-	2.621	1.292
209	Methyl selenac	$C_{12}H_{24}N_4S_8Se$	-	3.020	1.838
210	Calcium-Cyclamat	$C_{12}H_{24}CaN_2O_6S_2$	-	2.809	1.883
211	Dapsone	$C_{12}H_{12}N_2O_2S$	-	3.035	1.753
212	Dieldrin	$C_{12}H_8Cl_6O$	-	3.852	1.698
213	Alizarin	$C_{14}H_8O_4$	-	3.385	1.419
214	Amido-G-acid	$C_{14}H_8O_4$	-	3.385	1.419
215	Alizarin orange	$C_{14}H_7NO_6$	-	3.714	1.648
216	9-Nitroanthracene	$C_{14}H_9NO_2$	-	3.154	1.476
217	Nitrovin	$C_{14}H_{12}N_8O_6$	-	3.600	1.926
218	Benzoylperoxid	$C_{14}H_{10}O_4$	-	3.214	1.432
219	Kaempherol	$C_{15}H_{10}O_6$	-	3.419	1.492
220	Quercetin	$C_{15}H_{10}O_7$	-	3.500	1.517
221	2’-Trifluoromethylamino-azobenzene	$C_{15}H_{14}N_3F_3$	-	3.143	1.660
222	Disperse Yellow 3	$C_{15}H_{15}N_3O_2$	-	2.914	1.588
223	Methyl Red	$C_{15}H_{15}N_3O_2$	-	2.914	1.588
224	1,8-Dinitropyrene	$C_{16}H_8N_2O_4$	-	3.533	1.640
225	Orange I	$C_{16}H_{11}N_2NaO_4S$	-	3.314	1.928
226	Sunset yellow FCF	$C_{16}H_{10}N_2Na_2O_7S_2$	-	3.590	2.135
227	Pyrene***)	$C_{16}H_{10}$	-	2.846	0.961
228	Cinnamyl antranilate	$C_{16}H_{13}NO_2$	-	2.938	1.434
229	Diazepam	$C_{16}H_{13}ClN_2O$	-	3.030	1.587
230	Orange G	$C_{16}H_{10}N_2Na_2O_7S_2$	-	3.590	2.135
231	Sudan Brown RR	$C_{16}H_{14}N_4$	-	2.882	1.402
232	Sunset Yellow FCF	$C_{16}H_{10}N_2Na_2O_7S_2$	-	3.590	2.135
233	*para*-Anisidine hydrochloride	$C_{17}H_9NO\cdot HCl$	-	3.200	1.484
234	Fusarenon X (105)	$C_{17}H_{22}O_8$	-	2.936	1.478
235	6-Nitrochrisene (112)	$C_{18}H_{11}NO_2$	-	3.125	1.403
236	Ponceau SX	$C_{18}H_{14}N_2Na_2O_7S$	-	3.318	2.015
237	Naphtacene	$C_{18}H_{12}$	-	2.800	0.971
238	2,6-Diamino-3-(phenylazo) pyridine	$C_{18}H_{19}N_3O_2$	-	2.810	1.523
239	Petasitenine	$C_{19}H_{27}NO_7$	-	2.778	1.519
240	Eosin	$C_{20}H_8Br_4O_5$	-	3.946	1.695
241	6-Nitrobenzo(*a*)pyrene	$C_{20}H_{11}NO_2$	-	3.177	1.367
242	Symphytine	$C_{20}H_{31}NO_6$	-	2.621	1.452
243	Ethyl tellurac	$C_{20}H_{40}N_4S_8Te$	-	2.658	1.651
244	Carmoisine	$C_{20}H_{12}N_2Na_2O_7S_2$	-	3.511	2.045
245	Amarant	$C_{20}H_{11}O_{10}Na_3N_2S_3$	-	3.714	2.161
246	Ochratoxin A	$C_{20}H_{18}ClNO_6$	-	3.174	1.676
247	Norgestrel	$C_{21}H_{28}O_2$	-	2.431	1.185
248	Sudan III	$C_{22}H_{16}N_4O$	-	3.023	1.470

End of the table

249	Scharlachrot	$C_{24}H_{20}N_4O$	-	2.898	1.442
250	Sudan Red 7B	$C_{24}H_{21}N_5$	-	2.840	1.366
251	T_2-Trechothecene	$C_{24}H_{34}O_9$	-	2.746	1.416
252	Lauroyl peroxide	$C_{24}H_{46}O_4$	-	2.243	1.181
253	Disulfiram	$C_{30}H_{20}N_2S_4$	-	3.107	1.457
254	6-Nitrobenzo(*a*)pyrene	$C_{20}H_{11}NO_2$	-	3.177	1.367
255	Evans blue	$C_{34}H_{24}N_6Na_4O_{14}S_4$	-	3.512	2.149

[*)] Carcinogen of group 2B according to IARC classification. [**)] Carcinogen of group 3 according to IARC classification. [***)] Carcinogenic activity does not have sufficient evidence [5].

Here $\langle Z\rangle^{(\max)}$ and $\langle Z\rangle^{(\min)}$ refer to chemical compounds of Table 1 under the numbers $N = 2$ and 33, respectively; S is the standard deviation of the sample. We write down the inequality that will allow us to determine the compatibility of the maximal and minimal elements of the sample with other elements of the set:

$$\tau = \left|\left\langle Z^{(\max/\min)}\right\rangle - \left\langle Z^{(\mathrm{av})}\right\rangle\right| / S =$$

$$\begin{cases} 4.46^{(\max)} > \tau_{0.05}^{(\mathrm{cr})}(N = 255) = 3.65, \\ 2.95^{(\min)} < \tau_{0.05}^{(\mathrm{cr})}(N = 255) = 3.65. \end{cases} \qquad (5)$$

Inequality (5) indicates that at the significance level of $\alpha = 0.05$, the maximal value of the characteristic disturbs the homogeneity of the sample. Consequently, we must be weeded out this chemical substance. Using a similar procedure for the remaining elements of the sample, we found that the chemical compounds under numbers 3, 67, 76, 164 are also fall out of the sample. Finally we obtain the following statistics for 250 chemical compounds:

$$N = 250, \quad \langle Z\rangle^{(\mathrm{av})} = 4.81 \pm 0.04, \quad \langle Z\rangle^{(\min)} = 2.746, \ \langle Z\rangle^{(\max)} = 6.99, \quad S = 0.62. \qquad (6)$$

Normality of the sample is confirmed by the inequality:

$$N = 250, \quad \chi^2 = 6.93 < \chi_{0.05}^{2(\mathrm{cr})}(f = 7) = 14.1, \quad p = 0.44. \qquad (7)$$

Now we recalculate the statistics of mean values for a sample of 250 elements. The statistics of the average values will be the following:

$$N = 250, \ Z^{(\mathrm{av})} = 3.06 \pm 0.03, \quad Z^{(\min)} = 1.882, \quad Z^{(\max)} = 5.000, \qquad S = 0.44,$$
$$N_1 = 137, \ Z_1^{(\mathrm{av})} = 2.92 \pm 0.03, \ Z_1^{(\min)} = 1.882, \quad Z_1^{(\max)} = 4.273, \qquad S_1 = 0.40,$$
$$N_2 = 113, \ Z_2^{(\mathrm{av})} = 3.23 \pm 0.04, \ Z_2^{(\min)} = 2.000, \quad Z_2^{(\max)} = 5.000, \qquad S_2 = 0.44. \qquad (8)$$

Let us check whether the average values of $Z_1^{(\mathrm{av})}$ and $Z_2^{(\mathrm{av})}$ are statistically different. It is necessary to compare the variances by using the Fisher test:

$$S_2^2 / S_1^2 = 1.23 < F_{0.05}^{(\mathrm{cr})}(f_2 = 112; f_1 = 136) = 1.36. \qquad (9)$$

Inequality (9) indicates that the variances of the two samples are statistically indistinguishable. Therefore, the difference in mean values we can be verified using the following equation [9]:

$$|Z_1^{(\mathrm{av})} - Z_2^{(\mathrm{av})}| = 0.303 >$$

$$t_{0.05}^{(\mathrm{cr})}(f)\left\{\frac{N[(N_1-1)S_1^2 + (N_2-1)S_2^2]}{N_1 N_2 (N_1 + N_2 - 2)}\right\}^{1/2} \qquad (10)$$

$$= 0.171.$$

Here $f = N_1 + N_2 - 2$; $t_{0.05}^{(\mathrm{cr})}(f)$ is the one-side test. Inequality (10) indicates that the average values for samples N_1 and N_2 are statistically different. That is, active chemical compounds are grouped around of $Z_1^{(\mathrm{av})}$, and inactive chemical compounds are grouped around of $Z_2^{(\text{av})}$.

The value $Z^{(\text{av})}$ = 3.06 is taken as the threshold value: $Z^{(\text{th})} \equiv Z^{(\text{av})}$. If the molecular descriptor Z is below a threshold value, then such chemical compound should hypothetically posses of carcinogenic properties. Chemical compounds that have descriptor $Z >$ $Z^{(\text{th})}$ probably do not have carcinogenic activity. Since the descriptor Z has an alternative variation, we use the method of association of dichotomous features to identify the classification rule. That is, we establish the relationship between the carcinogenic activity of chemical compounds and the value descriptor Z.

First of all, it is necessary to evaluate the statistical interrelation between the two groups (the subsets N_1 and N_2) of chemical compounds. It is convenient to begin analyzing the interrelationship of descriptors by the construction of a contingency table (see Table 2) [13]. Each cell of the table indicates the occurrence frequency q_{ij} of descriptors. Obviously, the classification model the better describes the interrelation of features than the table is closer to the diagonal form. For the method of association of dichotomous features, it is important to know only the occurrence frequency of descriptors. We do not assume the existence of any continuous functional mathematical dependence between the explained variable and the explanatory one.

The corresponding the occurrence frequency q_{ij} of features and the statistics of the relationship of descriptors are presented in the tetrachoric contingency table (Table 2). In Table 2 we also indicate sampling rates: $p_{ij} = q_{ij}/N$. If there is equality for rates of ${}_iP \times P_j = p_{ij}$, then the interrelationship between the carcinogenic activity of chemical compounds and the descriptor Z is absent. If this equality is not fulfilled, then there is an interrelation between the dichotomous features. For example, using Table 2 with data, we have the following inequality: ${}_1P \times P_2 = 0.55 \times 0.44 = 0.24 \neq p_{12} = 0.15$. Therefore, there is an interrelation between molecular descriptor Z and carcinogenic activity of chemical compounds.

Table 2. *Table 2×2 of the association method.*

Separation of chemical compounds based on Z	Separation of chemical compounds based on carcinogenic activity		Total
	Active	Inactive	
$Z \le Z^{(\text{th})} = 3.06$	$q_{11} = 99$ $p_{11} = 0.40$	$q_{21} = 41$ $p_{21} = 0.16$	140 $P_1 = 0.56$
$Z > Z^{(\text{th})} = 3.06$	$q_{12} = 38$ $p_{12} = 0.15$	$q_{22} = 72$ $p_{22} = 0.29$	110 $P_2 = 0.44$
Total	137 ${}_1P = 0.55$	113 ${}_2P = 0.45$	$N = 250$ $\sum_{i=1,2} {}_iP = \sum_{i=1,2} P_i = 1.00$
$Q = 0.64$, $\Phi = 0.28$, $\chi^2 = 32.5 > \chi^{2(\text{cr})}_{0.05}(f = 1) = 3.84$, SE =0.03, $\Omega = 4.58$, $K = 0.34$, $\lvert r_{\text{tet}} \rvert = 0.54$, $\Delta = 0.32$, $SES = 1.25$.			

In the Table 2 we use the following notation: Q is the coefficient of association of Yule, Φ is the coefficient of association of Pearson; the odds ratio is equal to $\Omega = q_{11}q_{22}/(q_{12} \cdot q_{21})$; r_{tet} is the tetrachoric coefficient of association. ***SE*** is the standard association coefficient error; the empirical model error is equal to $\Delta = (q_{12} + q_{21})/N$; SES is the standard odds ratio error; K is Pearson's the mutual association coefficient [13,14]. The standard error (SES) of the odds ratio is determined as follows

$$SES = \Omega\sqrt{1/q_{11} + 1/q_{22} + 1/q_{12} + 1/q_{21}} = 1.25. \tag{11}$$

The Pearson contingency coefficient [10] is determined by the following equation

$$\Phi = \frac{q_{11}q_{22} - q_{12}q_{21}}{[(q_{11}+q_{12})(q_{21}+q_{22})(q_{11}+q_{21})(q_{12}+q_{22})]^{1/2}}$$

$$= 0.28. \tag{12}$$

The statistical significance of the coefficient Φ may be verified with the help of the Student's t-test. A null hypothesis on the independence of features is rejected if the following inequality holds

$$t = \Phi\sqrt{N-2}\,/\sqrt{1-\Phi^2} > t_{0.05}^{(\mathrm{cr})}(f = N-2)\ . \tag{13}$$

Using the value (12) we obtain the following inequality: $t = 4.6 > t_{0.05}^{(\mathrm{cr})}(f = 248) = 1.96$. The standard error ($SE$) of the contingency coefficient may be estimated using equation:

$$\begin{aligned}&SE(\Phi) = 0.5(1-\Phi^2)\times\\&(1/q_{11} + 1/q_{12} + 1/q_{21} + 1/q_{22}) = 0.03.\end{aligned} \tag{14}$$

It is also possible to use the Yule association coefficient [9] to identify the relationship between the factors:

$$Q = \frac{q_{11}q_{22} - q_{12}q_{21}}{q_{11}q_{22} + q_{12}q_{21}} = 0.64\,. \tag{15}$$

The value of the coefficient Q indicates the existence of a relationship between the analyzed factors. Obviously, this coefficient is in the following range: $-1 \le Q \le +1$. It is usually assumed if Q > 0.5, then there is close link between the factors. The tetrachoric association coefficient, allows quantitatively and statistically justified to point out the comparability of the compared factors:

$$r_{\mathrm{tet}} = \cos\begin{bmatrix}\pi(q_{12}q_{21})^{1/2}((q_{11}q_{22})^{1/2}\\ \\ +\,(q_{12}q_{21})^{1/2})^{-1}\end{bmatrix} \tag{16}$$

The chi-criterion [13] is compared with tabulated value of the chi-square distribution function for one degree of freedom (f = 1):

$$\chi^2 = N\Phi^2 =$$

$$\frac{N(q_{11}q_{22} - q_{12}q_{21})^2}{(q_{11}+q_{12})(q_{21}+q_{22})(q_{11}+q_{21})(q_{12}+q_{22})}$$

$$= 32.5 > \chi_{0.05}^{2(\mathrm{cr})}(f = 1)\ = 3.84. \tag{17}$$

That is, at the significance level α = 0.05, the empirical value of Pearson's criterion is much higher than the tabulated value. In this case, the null hypothesis on the independence of factors should be rejected. Since $\chi^2 > \ \chi_{0.05}^{2(\mathrm{cr})}(f = 1)$, we can conclude that there is a statistically significant interrelationship between descriptor of Z and the carcinogenic activity of chemical compounds. The empirical error of the model is equal to: Δ·100% = 32%. In addition, we indicate the frequency relations: q_{11} /(q_{11} + q_{12}) = 0.72 and q_{22} /(q_{22} + q_{21}) = 0.64. These relations are significantly different. It is well known if there is no relationship between factors then these relations should be identically equal.

Now we introduce another descriptor - the information function (2). Table 1 is given numerical values of the information function. We will find out the possibility of constructing the classification rule using the information function. The initial sample satisfies the normal distribution:

$$N = 255,\quad \chi_{0.05}^2 = 5.70\ <\ \chi_{0.05}^{2(\mathrm{cr})}(f = 7) = 14.1,$$

$$df = 7,\ \ p = 0.58\ >\ 0.05. \tag{18}$$

Inequality (18) indicates that the sample satisfies the normality condition. The empirical average value of the information function is equal to

$$N = 255,\qquad H^{(\mathrm{av})}\ = 1.62 \pm 0.02,$$

$$H^{(\min)}\ = 0.724,\ H^{(\max)}\ = 2.23, S = 0.31. \tag{19}$$

The sample at significance level α = 0.05 is homogeneous:

$$\tau = |\,H^{(\max/\min)} - H^{(\mathrm{av})}\,|\,/S\ \ =$$

$$\begin{cases}1.97^{(\max)} < \tau_{0.05}^{(\mathrm{cr})}(N = 255) = 3.65,\\ 2.90^{(\min)} < \tau_{0.05}^{(\mathrm{cr})}(N = 255) = 3.65.\end{cases} \tag{20}$$

That is, all elements of the set are compatible. Inequalities (20) do not allow rejecting the null hypothesis about the homogeneity of the set of elements. For chemical compounds possessing reliably installed carcinogenic activity, the following statistics has been obtained:

$N_1 = 141$, $H_1^{(av)} = 1.56 \pm 0.03$, $H_1^{(min)} = 0.723$, $H_2^{(max)} = 2.181$, $S_1 = 0.31$;
$\chi^2 = 6.16 < \chi_{0.05}^{2(cr)}(df = 9) = 16.9$,
$p = 0.72 > 0.05$. (21)

For non-carcinogenic chemical compounds the statistics will be as follows:

$N_2 = 114$, $H_2^{(av)} = 1.70 \pm 0.03$, $H_2^{(min)} = 0.919$, $H_2^{(max)} = 2.230$, $S_2 = 0.30$;
$\chi^2 = 4.74 < \chi_{0.05}^{2(cr)}(df = 3) = 7.8$, $p = 0.20 > 0.05$. (22)

The statistical discrepancy between values of $H_1^{(av)}$ and $H_2^{(av)}$ is confirmed by the following inequality:

$$H_1^{(av)} - H_2^{(av)} |= 0.143 > t_{0.05}^{(cr)}(f)\times$$

$$\left\{\frac{N[(N_1-1)S_1^2 + (N_2-1)S_2^2]}{N_1 N_2 (N_1 + N_2 - 2)}\right\}^{1/2} = 0.074. \quad (23)$$

Here $f = N_1 + N_2 - 2 = 253$.

Let us check whether descriptor of H can be used to construct a classification rule. The average value of the descriptor $H^{(th)} \equiv H^{(av)} = 1.62$*bits* (19) we will accept for the boundary (threshold) value. Table 3 provides statistics on the application of the association method. Variation of the descriptor $H^{(av)}$ about the average value showed that somewhat higher statistical results can be obtained if we use as the boundary value $H^{(th)} = 1.70$*bits*. Table 3 shows statistics for this threshold value (data are given in parentheses).

***Table 3**. Table 2×2 of the association method.*

Separation of chemical compounds based on H	Separation of chemical compounds based on carcinogenic activity		Total
	Active	Inactive	
$H \le H^{(th)} = 1.62 bits$	$q_{11} = 81(95)$ $p_{11} = 0.32(0.37)$	$q_{21} = 46(56)$ $p_{21} = 0.18(0.22)$	127(151) $P_1 = 0.50(0.59)$
$H > H^{(th)} = 1.62 bits$	$q_{12} = 60(46)$ $p_{12} = 0.23(0.18)$	$q_{22} = 68(58)$ $p_{22} = 0.27(0.23)$	128(104) $P_2 = 0.50(0.41)$
Total	141(141) ${}_1P = 0.55(0.55)$	114(114) ${}_2P = 0.45(0.45)$	$N = 255$ $\sum_{i=1,2} {}_iP = \sum_{i=1,2} P_i = 1.00$
$Q = 0.33(0.36)$, $\Phi = 0.09(0.18)$, $\chi^2 = 7.37(8.70) > \chi_{0.05}^{2(cr)}(f = 1) = 3.84$, $SE = 0.03(0.03)$, $\Omega = 2.00(2.14)$, $K = 0.12(0.21)$, $\lvert r_{tet}\rvert = 0.27(0.29)$, $\Delta = 0.42(0.40)$, $SES = 0.51(0.56)$.			

Since $\chi^2 > \chi_{0.05}^{2(cr)}$, we can agree that there is a statistically significant interrelation between descriptor H and the carcinogenic activity of chemical compounds. This is also indicated by the product of proportions. For example, ${}_2P \cdot P_2 = 0.23(0.19) \neq p_{22} = 0.27(0.23)$. Similarly, the frequency ratio differ significantly: $q_{11}/(q_{11}+q_{12}) = 0.57(0.67) \neq q_{22}/(q_{22}+q_{21}) = 0.60(0.51)$.

Let us now verify the representativeness of a sample of 255 elements. Since the elements of the sample were taken from different literary sources, it is necessary to make sure that they were chosen "with an open mind". For this we use the table of random numbers. Using the table of three-digit random numbers [15], we obtained the following subsample. The random subsample contains 63 elements, which is much smaller than the original sample size (~ 1/4 of the original sample). The random subsample contains the following elements of the set:

148 156 038 020 124 012 250 080 074 001 249 224
102 196 231 191 068 119 120 026 105 240 **144**
137 070 013 203 187 245 249 184 179 088 254
154 209 069 144 034 122 213 230 171 008 146 238
230 130 **164** 162 **002** 219 168 042 192 175 127
233 045 005 163 033 204.

(24)

These numbers correspond to the numbering of the Table 1. Elements 002, 144 and 164 of the subset should be excluded from the subsample, since they are incompatible with the rest of the elements of the subset by descriptor *Z*.

Using the random subsample (24) we obtained the following statistics for the descriptor of Z:

$N = 60$, $Z^{(\mathrm{av})} = 3.13 \pm 0.06$, $Z^{(\mathrm{min})} = 1.882$, $Z^{(\mathrm{max})} = 4.333$, $S = 0.49$,
$N_1 = 28$, $Z_1^{(\mathrm{av})} = 2.86 \pm 0.081$, $Z_1^{(\mathrm{min})} = 1.882$, $Z_1^{(\mathrm{max})} = 4.273$, $S_1 = 0.43$,
$N_2 = 32$, $Z_2^{(\mathrm{av})} = 3.36 \pm 0.08$, $Z_2^{(\mathrm{min})} = 2.750$, $Z_2^{(\mathrm{max})} = 4.333$, $S_2 = 0.43$. (25)

The average values of the descriptor (25) turned out to be close to the average values of (8). Therefore, the sample for Table 1 can be considered as the representative sample.

Let us now verify the representativeness of the sample (Table 1) on the basis of H. On grounds of the information function the random sample (24) is normally distributed:

$N = 60$, $\chi^2 = 2.84 < \chi^{2(\mathrm{cr})}_{0.05}(df = 5) = 11.07$, $p = 0.72 \gg 0.05$. (26)

Using the random sample (24), the following statistics were obtained for the information function:

$N = 60$, $H^{(\mathrm{av})} = (1.63 \pm 0.04)$*bits*, $H^{(\mathrm{min})} = 0.919$*bits*, $H^{(\mathrm{max})} = 2.236$*bits*, $S = 0.34$,
$N_1 = 28$, $H_1^{(\mathrm{av})} = (1.45 \pm 0.06)$*bits*, $H_1^{(\mathrm{min})} = 0.919$*bits*, $H_1^{(\mathrm{max})} = 2.048$*bits*, $S_1 = 0.31$,
$N_2 = 32$, $H_2^{(\mathrm{av})} = (1.79 \pm 0.05)$*bits*, $H_2^{(\mathrm{min})} = 1.204$*bits*, $H_2^{(\mathrm{max})} = 2.230$*bits*, $S_2 = 0.28$. (27)

The average values the descriptor of $H^{(\mathrm{av})}$ (27) turned out to be close to the average values (16).

Using the threshold value $Z^{(\mathrm{th})} \equiv Z^{(\mathrm{av})} = 3.05$ (variation within the confidence interval (25)) we have obtained the following association statistics for the random sub-sample (24):

$N = 60$, $q_{11} = 22$, $q_{12} = 6$, $q_{22} = 22$, $q_{21} = 10$;
$Q = 0.78$, $\Phi = 0.39$,
$\chi^2 = 13.4 > \chi^{2(\mathrm{cr})}_{0.05}(f = 1) = 3.84$, $SE = 0.15$,
$\Omega = 8.07$, $K = 0.44$, $|r_{\mathrm{tet}}| = 0.68$, $\Delta = 0.27$, $SES = 4.82$. (28)

Similarly, we can obtain statistic on the interrelation between the carcinogenic activity of chemical compounds and the information function. The threshold value of the information function is equal to $H^{(\mathrm{th})} \equiv H^{(\mathrm{av})} = 1.66$*bits* (variation of the threshold value within the confidence interval (23)). The association statistics for molecular descriptor of *H* and carcinogenic activity of chemical compounds will be as follows:

$N = 60$, $q_{11} = 20$, $q_{12} = 8$, $q_{22} = 19$, $q_{21} = 13$; $Q = 0.57$, $\Phi = 0.26$,
$\chi^2 = 5.74 > \chi^{2(\mathrm{cr})}_{0.05}(f = 1) = 3.84$, $SE = 0.14$,
$\Omega = 3.65$, $K = 0.33$, $|r_{\mathrm{tet}}| = 0.47$, $\Delta = 0.35$, $SES = 2.02$. (29)

Thus, classification rules are performed for random sampling.

We have found that there is a statistically significant interrelation between the molecular descriptors. Using the random sample (24), we obtained the following correlation equation:

$N = 60$, $H(Z) = b_0 + b_1 \cdot Z$, $b_0 = 0.27 \pm 0.21$, $b_1 = 0.43 \pm 0.07$, $t(b_0) = 1.28$, $t(b_1) = 6.47$, $R = 0.65$,
$F = 41.8 > F^{(\mathrm{cr})}_{0.05}(f_1 = 1; f_2 = 58) = 4.00$,
Std. Err. = 0.25. (30)

At the same time, for the original sample (Table 1) we obtained the following regression equation

$N = 250$, $H(Z) = b_0 + b_1 \cdot Z$, $b_0 = 0.22 \pm 0.10$, $b_1 = 0.47 \pm 0.03$, $t(b_0) = 2.13$, $t(b_1) = 13.8$, $R = 0.66$, $F = 190 > F_{0.05}^{(cr)}(f_1 = 1; f_2 = 248) = 3.89$, *Sdt. Err.* $= 0.22$. (31)

Obviously, equations (30) and (31) are essential identical. This result also indicates the identity of the two samples and the representativeness of the original sample. We will also write out the correlation equations for the carcinogenic active chemical compounds of the initial sample:

$N_1 = 137$, $H(Z) = b_0 + b_1 \cdot Z$, $b_0 = 0.176 \pm 0.143$, $b_1 = 0.478 \pm 0.048$, $t(b_0) = 1.23$, $t(b_1) = 9.88$, $R = 0.65$, $F = 97.8 > F_{0.05}^{(cr)}(f_1 = 1; f_2 = 135) = 3.92$, *Std. Err.* = 0.223. (32)

For inactive chemical compounds we have obtained the following correlation equation:

$N_2 = 113$, $H(Z) = b_0 + b_1 \cdot Z$, $b_0 = 0.280 \pm 0.172$, $b_1 = 0.445 \pm 0.053$, $t(b_0) = 1.63$, $t(b_1) = 8.36$, $R = 0.62$, $F = 69.8 > F_{0.05}^{(cr)}(f_1 = 1; f_2 = 111) = 3.94$, *Std.Err.* $= 0.227$. (33)

We use the Chow test for *F*-statistics to determine whether the equations (32) and (33) are statistically dissimilar:

$$F = \frac{(\Sigma_0 - \Sigma_1 - \Sigma_2)(N - 2m - 2)}{(\Sigma_1 + \Sigma_2)(m + 1)}, \qquad (34)$$

Here $\Sigma_1 = 6.714$ and $\Sigma_2 = 5.732$ are the sum of the squared deviations of the actual values from the regression lines for the first equation (32) and the second equation (33). $\Sigma_0 = 12.463$ is the sum of the squared deviations for the combined sample ($N = N_1 + N_2 = 250$). For the combined sample, the regression equation has the form (34); the number of characteristic factors is equal to: $m = 1$. From the equation (31) we obtain the inequality for *F*-statistics (degrees of freedom $f_1 = m + 1$, $f_2 = N_1 + N_2 - 2m - 2$):

$$F = 0.34 << F_{0.05}^{(cr)}(f_1 = 2, f_2 = 246) = 4.71. \qquad (35)$$

Therefore, it is impossible to reject the null hypothesis at the significance level α = 0.05. That is, regressions (32) and (33) are not statistically distinguishable.

Let us check the classification rules using chemical compounds that are not included in the original sample. For example, we calculated molecular descriptors for theophylline ($C_7H_8N_4O_2$; $Z = 3.24$, $H = 1.84$*bits*), azathiprine ($C_9H_7N_7O_2S$; $Z = 3.69$, $H = 2.02$*bits*), adenosine ($C_{10}H_{13}N_5O_4$; $Z = 3.19$, $H = 1.85$*bits*), endrin ($C_{12}H_8Cl_6O$; $Z = 3.85$, $H = 1.69$*bits*) and dirimal ($C_{12}H_{18}N_4O_6S$; $Z = 3.12$, $H = 1.90$*bits*), E331 ($C_6H_5O_7Na_3$; $Z = 3.52$, $H = 1.94$*bits*). These chemicals are not carcinogenic. Obviously, the descriptors of *Z* and *H* are greater than the threshold values. That is, the results of observations do not contradict the formulated classification rules. Below we will dwell in more detail on the predictive capabilities of the model for chemical compounds of various classes.

Using the Table 1 we will compile new subsample. A subsample includes all chemical compounds that contain only carbon, hydrogen, nitrogen, and oxygen atoms. Using this sub-sample, we will verify the validity of the formulated classification rules (Tables 2 and 3). In total, the sub-sample contains 63 elements. The number of active chemical compounds and the number of inactive chemical compounds are $N_1 = 26$ and $N_2 = 37$, respectively. The sub-sample satisfies the normality condition. The sub-sample statistics will be the following:

$$N = 63, \quad Z^{(av)} = 3.15 \pm 0.05, \quad Z^{(min)} = 2.471, \quad Z^{(max)} = 4.333, \quad S = 0.40, \qquad (36)$$

The compatibility conditions for the elements of the set are satisfied:

$$\tau = | Z^{(max/min)} - Z^{(av)} | / S = \begin{cases} 2.96^{(max)} < \tau_{0.05}^{(cr)}(f = 63) = 3.22, \\ 1.70^{(min)} < \tau_{0.05}^{(cr)}(f = 63) = 3.22. \end{cases} \qquad (37)$$

Statistics of average values for carcinogenic chemical compounds and for inactive chemical compounds will be the following:

$N_1 = 26$, $Z_1^{(\mathrm{av})} = 2.90 \pm 0.05$, $Z_1^{(\mathrm{min})} = 2.471$, $Z_1^{(\mathrm{max})} = 3.539$, $S_1 = 0.24$,
$N_2 = 37$, $Z_2^{(\mathrm{av})} = 3.32 \pm 0.07$, $Z_2^{(\mathrm{min})} = 2.588$, $Z_2^{(\mathrm{max})} = 4.333$, $S_2 = 0.40$. (38)

Let us verify the reliability of the statistically significant difference between the average values of $Z_1^{(\mathrm{av})}$ and $Z_2^{(\mathrm{av})}$. Using the F distribution, we determine the difference between the variances of these two subsamples: $F = S_2^2 / S_1^2 = 2.78 > F_{0.05}^{(\mathrm{cr})}(f_2 = 36; f_1 = 25) = = 1.90$. From this inequality it follows that two sample variances must be recognized as different at the significance level $\alpha = 0.05$ Therefore, to compare the average values of two clusters, it is necessary to use the approximate T-test [16]:

$$| Z_1^{(\mathrm{av})} - Z_2^{(\mathrm{av})} | = 0.42 > T = \frac{\mathrm{v}_1 t_{0.05}^{(\mathrm{cr})}(f_1) + \mathrm{v}_2 t_{0.05}^{(\mathrm{cr})}(f_2)}{(\mathrm{v}_1 + \mathrm{v}_2)^{0.5}} = 0.17, \quad (39)$$

Here $\mathrm{v}_1 = S_1^2 / N_1 = 2.215 \cdot 10^{-3}$, $\mathrm{v}_2 = S_2^2 / N_2 = 4.33 \cdot 10^{-3}$. Inequality (39) indicates that the difference in the mean values is statistically significant at a significance level of 5%. Thus, the null hypothesis on the equality of mean values must be rejected. Therefore, the difference between the average values should be considered statistically significant. The inequality (39) is even persisted. If we take a very stringent value 0.001 for the significance level the inequality (39) is still persisted. That is, we are immune from the error of the so-called first kind [16], namely the possibility of accepting the hypothesis of equality of mean values, whereas they actually differ. Thus, it can be assumed that the active carcinogens are grouped around the average value of $Z_1^{(\mathrm{av})}$, while inactive compounds are grouped around the average value of $Z_2^{(\mathrm{av})}$. In the framework of a method that using tetrachoric contingency tables we obtained the following statistics:

$N = 63$, $q_{11} = 23$, $q_{12} = 3$, $q_{22} = 23$, $q_{21} = 14$; $Q = 0.85$, $\Phi = 0.48$, $SE = 0.19$, $\chi^2 = 16.14 > \chi_{0.05}^{2(\mathrm{cr})}(f = 1) = 3.84$, $\Omega = 12.6$, $K = 0.49$, $|r_{\mathrm{tet}}| = 0.77$, $\Delta = 0.27$, $SES = 8.83$. (40)

It follows that there is a statistically significant relationship between the value of Z and the carcinogenic activity of chemical compounds.

Similarly, you can obtain statistics for the information function:

$N = 63$, $H^{(\mathrm{av})} = 1.69 \pm 0.03$,
$H^{(\mathrm{min})} = 1.204$, $H^{(\mathrm{max})} = 1.972$, $S = 0.20$,

$$\tau = | H^{(\mathrm{max/min})} - H^{(\mathrm{av})} | / S \ = \begin{cases} 1.46^{(\mathrm{max})} < \tau_{0.05}^{(\mathrm{cr})}(f = 63) = 3.22, \\ 2.38^{(\mathrm{min})} < \tau_{0.05}^{(\mathrm{cr})}(f = 63) = 3.22. \end{cases}$$

$N_1 = 26$, $H_1^{(\mathrm{av})} = 1.62 \pm 0.03$, $H_1^{(\mathrm{min})} = 1.397$, $H_1^{(\mathrm{max})} = 1.931$, $S_1 = 0.15$,
$N_2 = 37$, $H_2^{(\mathrm{av})} = 1.71 \pm 0.04$, $H_2^{(\mathrm{min})} = 1.204$, $H_2^{(\mathrm{max})} = 1.972$, $S_2 = 0.22$,

$$F = S_2^2 / S_1^2 = 2.24 > F_{0.05}^{(\mathrm{cr})}(f_2 = 36; f_1 = 25) = 1.90,$$
$$| H_1^{(\mathrm{av})} - H_2^{(\mathrm{av})} | = 0.097 > T = \frac{\mathrm{v}_1 t_{0.05}^{(\mathrm{cr})}(f_1) + \mathrm{v}_2 t_{0.05}^{(\mathrm{cr})}(f_2)}{(\mathrm{v}_1 + \mathrm{v}_2)^{0.5}} = 0.004. \quad (41)$$

Assuming that the threshold value is equal to $H^{(\mathrm{th})}$ = 1.69*bits*, we obtain the following statistics of the association method:

$N = 63$, $q_{11} = 19$, $q_{12} = 7$, $q_{22} = 21$, $q_{21} = 16$; $Q = 0.56$, $\Phi = 0.26$, $SE = 0.14$, $\chi^2 = 5.50 > \chi_{0.05}^{2(\mathrm{cr})}(f = 1) = 3.84$, $\Omega = 3.56$, $K = 0.37$, $|r_{\mathrm{tet}}| = 0.46$, $\Delta = 0.37$, $SES = 1.97$. (42)

It follows from the statistics (36), (38), and (41) the average values of the descriptors Z and H found are close to the average values (8), (16), (25), and (27). This indicates the stability of statistical results. The descriptors of Z and H are interrelated (Fig. 1).

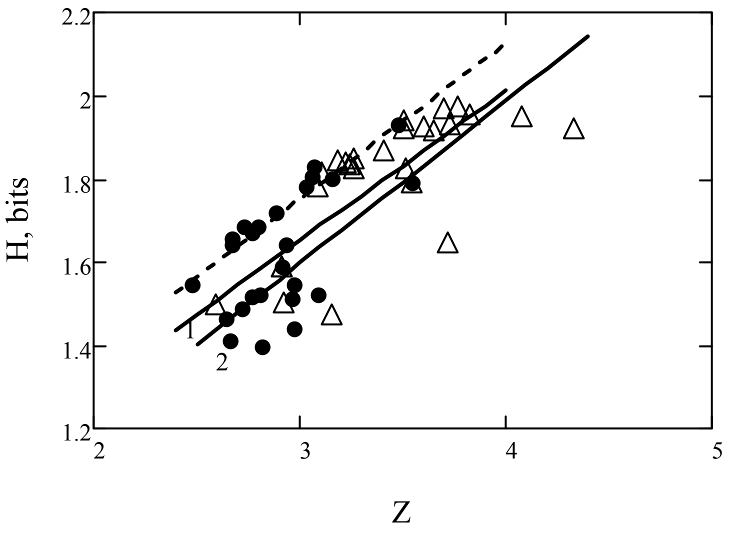

Fig. 1. *Diagram of scattering the descriptors for the sub-sample* (36). *Active chemical compounds are marked with dots* (•). (1) *Regression has the following form*: $H(Z) = (0.57 \pm 0.03\ ) + (0.36 \pm 0.10)\cdot Z$, $R = 0.59$, $S_1 = 0.12$, $F = 13.06 > F_{0.05}^{(\mathrm{cr})}(1;24) = 4.26$, $N_1 = 26$. *Inactive chemical compounds are marked by triangles* (Δ). (2) *Regression has the following form*: $H(Z) = (0.43 \pm 0.20) + (0.39 \pm 0.06)\cdot Z$, $R = 0.74$, $S_2 = 0.14$, $F = 42.9 > F_{0.05}^{(\mathrm{cr})}(1;35) = 4.11$, $N_2 = 37$. *If the correlation coefficient is within 0.25 ≤ R ≤ 0.75, then the correlation is viewed as moderate* [17].

We will use the Chow test for *F*-statistics to check whether regressions 1 and 2 in Figure 1 are statistically different:

$$F = \frac{(\Sigma_0 - \Sigma_1 - \Sigma_2)(N - 2m - 2)}{(\Sigma_1 + \Sigma_2)(m + 1)}. \qquad (43)$$

Here $\Sigma_1 = 0.341$ and $\Sigma_2 = 0.787$ are the sum of the squared deviations of the empirical values of the descriptor from the regression lines for the first (N_1 = 26) and the second equation (N_2 = 37). $\Sigma_0 = 1.146$ is the sum of the squared deviations for the combined sample ($N = N_1 + N_2$). The regression equation has the following form: $H(Z) = (0.58 \pm 0.14) + (0.33 \pm 0.04)\cdot Z$, $N = 63$, $R = 0.72$, *Std.Err.of Estimate* = 0.13. Number of descriptor-factor is equal to $m = 1$. From the equation (43) we obtain the inequality for *F*-statistics (degree of freedom $f_1 = m + 1, f_2 = N_1 + N_2 - 2m - 2$):

$$F = 0.36 < F_{0.05}^{(\mathrm{cr})}(f_1 = 2, f_2 = 59) = 3.15\,. \qquad (44)$$

Therefore, we can't reject the null hypothesis. That is regressions (1) and (2) in Figure 1 are not statistically dissimilar.

Significant scattering around regression lines 1 and 2 (Fig.1), presumably due to the fact that the sub-sample contains compounds of different chemical classes. This is confirmed by a more detailed analysis of the correlation. For related chemical compounds under the numbers (Table 1):

18, 19, 24, 25, 27, 34, 37, 38, 46, 61, 63, 107. (45)

Interrelation can be approximated by the following linear equation (dashed line in Fig. 1)

$H(Z) = A + B\cdot Z$, $A = 0.64 \pm 0.06$, $B = 0.38 \pm 0.02$, $R = 0.98$, $N = 12$, $t(B) = 17.3 > t(A) = 10.0 > t_{0.05}^{(\mathrm{cr})}(f = 10) = 1.81$, $F = 299.8 \gg F_{0.05}^{(\mathrm{cr})}(f_1 = 1, f_2 = 10) = 4.96$, *Std.Err. of Estimate* = 0.0197. (46)

Approximation error is equal to:

$$\delta H(Z) = 100\% \sum_{i=1}^{N} | H_i - H(Z_i) | / N / H_i \; = 1.16\%. \qquad (47)$$

Here H_i is the actual value of the information function of the series (45).

Using the Table 1, we will also compile a sub-sample of chemical compounds containing a sulfur atom. The sub-sample contains N_1 = 44 carcinogens active chemical and N_2 = 34 chemical compounds that do not have carcinogenic activity. The combined sample has the following statistics for the average values:

$$N = 78, \quad Z^{(av)} = 3.15 \pm 0.04, \quad S = 0.35,$$
$$H^{(av)} = 1.87 \pm 0.02, \quad S = 0.21. \qquad (48)$$

Using the data [5, 19] we compiled a sub-sample that contain halogen atoms. We took into account only those chemical compounds whose biological activity is reliably established. After the sifting out the descriptors that violate the homogeneity of the set we obtained a subsample which is presented in Table 4. After the sifting out of chemical compounds whose descriptors violate the homogeneity of the set of elements, we obtained a sub-sample, which is presented in Table 4.

Table 4. *The sub-sample of halogen-containing chemical compounds*

N	Chemical compounds	Gross formula	*Z*	*H,bits*
Chemical compounds with confirmed carcinogenic activity				
1	Ethylene dibromide	$C_2H_4Br_2$	3.25	1.50
2	Epichlor hydrin	C_3H_5ClO	3.00	1.69
3	1,2-Dibromo-3-chloropripane	$C_3H_5ClBr_2$	3.46	1.79
4	2-Fluoroethylnitrosourea	$C_3H_4N_3O_2F$	3.85	2.20
5	1,3-Dichloropropene	$C_3H_4Cl_2$	3.33	1.53
6	Glycerol iodinated	$C_3H_7O_3J$	3.14	1.73
7	Isobutenyl chloride	C_4H_6Cl	2.64	1.32
8	Tris (2-chloroethyl) phosphate	$C_6H_{12}C_{13}PO$	2.96	1.77
9	3-(Chloromethyl) pyridinehydrochloride	$C_6H_7NCl_2$	3.13	1.68
10	Mannitol dibromone	$C_6H_{12}O_6Br_2$	3.31	1.78
11	Isophosphamide	$C_7H_{15}Cl_2N_2O_2P$	2.90	1.95
12	Benzyl chloride *)	C_7H_7Cl	2.80	1.29
13	Benzal chloride *)	$C_7H_6Cl_2$	3.20	1.43
14	Benzotrixhloride	$C_7H_5Cl_3$	3.60	1.51
15	Sulfallate	$C_8H_{14}ClNS_2$	2.69	1.65
16	Tris-2,4-dichlorophenoxyethyl-phosphate	$C_9H_{15}Br_6O_4P$	3.49	1.97
17	Chlordimeform	$C_{10}H_{13}ClN_2$	2.69	1.50
18	Chlorobenzilate	$C_{10}H_{14}Cl_2O_3$	2.97	1.64
19	Aramite^R	$C_{12}H_{23}ClO_4S$	2.63	1.58
20	Melphalan	$C_{13}H_{18}Cl_2N_2O_2$	2.87	1.72
21	DDT	$C_{14}H_9Cl_5$	3.57	1.47
22	Dicofol	$C_{14}H_9Cl_5O$	3.66	1.63
23	Chloambucil	$C_{14}H_{19}Cl_2NO_2$	2.79	1.62
24	Nitrogen mustad	$C_{15}H_{11}Cl_2N$	2.63	1.53
25	Phenoxybezamine-hydrochloride	$C_{17}H_{12}NOCl_2$	2.67	1.48
26	Prednimustine	$C_{35}H_{45}Cl_2NO_6$	2.70	1.49
27	Methyl iodide	CH_3J	2.80	1.37
28	Epichlorohydrin	C_3H_5ClO	3.00	1.69
29	Dimethylcarbamoyl chloride	C_3H_6ClNO	3.00	1.90
30	Manuron	$C_9H_{11}ClN_2O$	2.92	1.73
31	Prometalon hydrochloride	$C_{15}H_{19}NO \cdot HCl$	2.58	1.43
32	Griseofulvin	$C_{17}H_{17}ClO_6$	3.17	1.71
33	N,N’-Bis (2-Chloroethyl)-2-napthylamine	$C_{14}H_{15}NCl_2$	2,81	1.44
34	Chrysoidine	$C_{12}H_{13}N_4Cl$	2.93	1.60
35	Melphalan	$C_{13}H_{18}N_2O_2Cl_2$	2.87	1.72

36	Mustard gas	$C_5H_{12}NCl_3$	2.93	1.64
37	Nitrogen mustard hydrochloride	$C_5H_{12}NCl_3$	2.76	1.57
38	Oestradiol mustard	$C_{42}H_{50}N_2O_4Cl_4$	2.75	1.51
Chemical compounds that do not have confirmed carcinogenic activity				
39	1,1-Dichloroethane	$C_2H_4Cl_2$	3.25	1.50
40	Iodoacetamide	C_2H_4JNO	3.33	2.06
41	Ethyl chloride	C_2H_5Cl	2.50	1.30
42	Chloracetic acid	$C_2H_3ClO_2$	3.75	1.91
43	Propylene dichloride	$C_3H_4Cl_2$	3.33	1.53
44	1-Chlorobutane	C_4H_9Cl	2.29	1.20
45	5-Fluorouracil	$C_4H_3FN_2O_2$	4.00	2.19
46	Trichlorfon	$C_4H_8Cl_3O_4P$	3.70	2.08
47	Chlorocholine chloride	$C_5H_{17}NCl_2$	2.24	1.32
48	Dibromneopentyl glycol	$C_5H_{10}Br_2O_2$	2.95	1.68
49	2-(Chloromethyl) pyridine hydrochloride	$C_6H_7NCl_2$	3.13	1.68
50	Heptachlor	$C_{10}H_5Cl_7$	4.27	1.53
51	Bis-chloroisopropyl ether	$C_6H_{13}ClO$	2.38	1.36
52	Carbromalum	$C_7H_{13}BrN_2O_2$	2.80	1.77
53	Phenacyl chloride	C_8H_7OCl	3.06	1.52
54	Fluometuron	$C_{10}H_{11}F_3N_2O$	3.26	1.87
55	Strobane^R	$C_{10}H_9Cl_7$	3.77	1.57
56	Chloramphenicol	$C_{11}H_{12}Cl_2N_2O_5$	3.44	1.98
57	Dieldrin	$C_{12}H_8Cl_6O$	3.86	1.70
58	Photodihydrin	$C_{12}H_6Cl_6O$	4.08	1.68
59	Endrin	$C_{12}H_8Cl_6O$	3.86	1.70
60	Aldrin	$C_{12}H_8Cl_6$	3.77	1.53
61	Trifluraline	$C_{13}H_{16}F_3N_3O_4$	3.51	1.96
62	Coumaphos	$C_{14}H_{16}ClO_5PS$	3.16	1.86
63	Dicofol	$C_{14}H_9Cl_5O$	3.66	1.64
64	Methoxychlor	$C_{16}H_{15}Cl_3O_2$	3.11	1.58
65	Flecainide acetate	$C_{17}H_{20}F_6N_2O_3$	3.29	1.87
66	*p,p*'-Ethyl-DDD	$C_{18}H_{20}Cl_2$	2.65	1.24
67	Trichlorotriethylamine hydrochloride	$C_6H_{12}NCl_3{\cdot}HCl$	3.36	1.96
68	Eosin disodium salt	$C_{20}H_6Br_4Na_2O_5$	3.94	1.87
69	Hexachlorophene	$C_{13}H_6O_2Cl_6$	4.15	1.75
70	2,4,6-Trichlorophenol	$C_6H_3OCl_3$	4.15	1.78
71	*para*-Anizidine hydrochloride	$C_{17}H_9NO{\cdot}HCl$	3.20	1.48
72	4-Chloro-*ortho*-phenylendiamine	$C_6H_7N_2Cl$	3.00	1.68
73	Fluometuron	$C_{10}H_{11}F_3N_2O$	3.26	1.87

*) There is insufficient data on carcinogenicity of chemical compound [5].

Statistics of average values of molecular descriptors (Table 4).

1. The descriptor Z:

$N = 73$, $Z^{(\mathrm{av})} = 3.17 \pm 0.05$, $Z^{(\mathrm{min})} = 2.24$, $Z^{(\mathrm{max})} = 4.15$, $S = 0.47$,

$\chi^2 = 6.85 < \chi^{2(\mathrm{cr})}_{0.05}(f = 6) = 12.6$, $p = 0.34$,

$q_{11} = 28$, $q_{12} = 10$, $q_{22} = 23$, $q_{21} = 12$; $Q = 0.69$, $\Phi = 0.31$, $\chi^2 = 11.42 > \chi^{2(\mathrm{cr})}_{0.05}(f = 1) = 3.84$, $SE = 0.12$, $\Omega = 5.4$, $K = 0.37$, $|r_{\mathrm{tet}}| = 0.58$, $\Delta = 0.30$, $SES = 2.75$. Error of model is equal to 30%.

$$N_1 = 38,\quad Z_1^{(\mathrm{av})} = 3.01 \pm 0.05,\quad Z_1^{(\mathrm{min})} = 2.58,\quad Z_1^{(\mathrm{max})} = 3.85,\quad S_{z1} = 0.33,$$
$$N_2 = 35,\quad Z_2^{(\mathrm{av})} = 3.35 \pm 0.09,\quad Z_2^{(\mathrm{min})} = 2.24,\quad Z_2^{(\mathrm{max})} = 4.15,\quad S_{z2} = 0.53. \qquad (49)$$

Let us check whether active and inactive chemical compounds really belong to different subsets and are primarily grouped around $Z_1^{(\mathrm{av})}$ and $Z_2^{(\mathrm{av})}$. Preliminarily, we compare the ratio of the larger variance to the smaller variance with the critical value of the Fisher distribution: $F = S_{z2}^2 / S_{z1}^2 = 2.57 > F_{0.05}^{(\mathrm{cr})}(f_2 = N_2 - 1 = 34;$

$f_1 = N_1 - 1 = 37)$ = 1.72. Obviously, the distinction in variances turns out to be statistically significant. Therefore, to verify the distinction in the average values we use the following inequality:

$$|Z_1^{(\mathrm{av})} - Z_2^{(\mathrm{av})}| = 0.34 > T = \frac{\nu_1 t_{0.05}^{(\mathrm{cr})}(f_1) + \nu_2 t_{0.05}^{(\mathrm{cr})}(f_2)}{(\nu_1 + \nu_2)^{0.5}} = 0.18\,, \qquad (50)$$

where $\nu_1 = S_1^2 / N_1 = 2.86 \cdot 10^{-3}$, $\nu_2 = S_2^2 / N_2 = 8.03 \cdot 10^{-3}$. Inequality (50) indicates that the distinction between the average values is statistically significant and the null hypothesis can be rejected.

2. The descriptor *H*:

N =73, $H^{(\mathrm{av})}$ = 1.66 ± 0.03, $H^{(\mathrm{min})}$ = 1.20, $H^{(\mathrm{max})}$ = 2.20, S = 0.22,

$\chi^2 = 1.82 < \chi_{0.05}^{2(\mathrm{cr})}(f = 4) = 9.5$, $p = 0.77$,

q_{11} = 23, q_{12} = 15, q_{22} = 21, q_{21} = 14; Q = 0.39, Φ = 0.13,

$\chi^2 = 3.07 < \chi_{0.05}^{2(\mathrm{cr})}(f = 1) = 3.84$, SE = 0.11, Ω = 2.3, K = 0.19, $|r_{\mathrm{tet}}|$ = 0.32, Δ = 0.40, SES = 1.10. Error of model is equal to: 40%.

N_1 =38, $H_1^{(\mathrm{av})}$ = 1.63 ± 0.03, $H_1^{(\mathrm{min})}$ = 1.29, $H_1^{(\mathrm{max})}$ = 2.20, S_{H1} = 0.19,

N_2 =35, $H_2^{(\mathrm{av})}$ = 1.69 ± 0.04, $H_2^{(\mathrm{min})}$ = 1.20, $H_2^{(\mathrm{max})}$ = 2.19, S_{H2} = 0.25. (51)

Let's check whether the average values of $H_1^{(\mathrm{av})}$ and $H_2^{(\mathrm{av})}$ are statistically different. Let us compare the variances of two subsets: $F = S_{H2}^2 / S_{H1}^2$ = 1.73 > $F_{0.05}^{(\mathrm{cr})}(f_2 = 34; f_1 = 37) = 1.71$. Therefore, a comparison of average values can be made using the following relationship:

$$|H_1^{(\mathrm{av})} - H_2^{(\mathrm{av})}| = 0.06 < T = \frac{\nu_1 t_{0.05}^{(\mathrm{cr})}(f_1) + \nu_2 t_{0.05}^{(\mathrm{cr})}(f_2)}{(\nu_1 + \nu_2)^{0.5}} = 0.10\,, \qquad (52)$$

where $\nu_1 = S_1^2 / N_1 = 9.5 \cdot 10^{-4}$, $\nu_2 = S_2^2 / N_2 = 1.84 \cdot 10^{-3}$. Inequality (52) allows us to reject the null hypothesis.

Thus, the application of classification rules to the table (4) makes it possible to separate active chemical compounds from inactive agents. To verify the impartiality of these results (49) and (51) we composed a random sub-sample using each and every data of handbook [5]. In the handbook [5] the total number of organohalogen compounds is 132. Using the table of random numbers [15] we obtained the subsample (see Table 5). After eliminating the incompatible elements, the statistics of the average values of the molecular descriptors will be as follows:

$$N = 36,\ Z^{(\mathrm{av})} = 3.15 \pm 0.07,\ H^{(\mathrm{av})} = 1.67 \pm 0.03. \qquad (53)$$

These average values are very close to the results (49) and (51). Such precision of mean values indicates stability and nonrandomness of results.

Table 5. *Random sampling of halogen containing chemical compounds*

N	Chemical compounds	Gross formula	*Z*	*H,bits*
Active chemical compounds				
1	1,2-Bis (chlormethoxy) ethan	$C_4H_8O_2Cl_2$	3.13	1.75
2	Diallate	$C_{10}H_{17}NOSCl_2$	2.75	1.73
3	Chlorobenzilate	$C_{16}H_{14}O_3Cl_2$	3.14	1.59
4	Cyclophosphamid	$C_7H_{17}Cl_2N_2O_3P$	2.88	1.94
5	Chlordimeform	$C_{10}H_{13}N_2Cl$	2.69	1.50
6	Chlorobenzilate	$C_{10}H_{14}O_3Cl_2$	2.97	1.64
7	Griseofulvin	$C_{17}H_{17}O_6Cl$	3.17	1.71
8	Mirex	$C_{10}Cl_{12}$	5.64	0.99
9	Chrysoidine	$C_{12}H_{13}N_4Cl$	2.93	1.60
10	Oxazepam	$C_{15}H_{11}N_2O_2Cl$	3.23	1.71
11	Tetrachlorvinphos	$C_{10}H_9Cl_4O_4P$	3.87	2.25

12	DDT	$C_{14}H_9Cl_5$	3.57	1.47
13	Chlorothanil	$C_8Cl_4N_2$	5.00	1.38
14	4,4'-Methelene bis (2-chloroline)	$C_{13}H_{12}Cl_2N_2$	3.03	1.58
15	Chlorobenilate	$C_{10}H_{14}Cl_2O_3$	2.97	1.64
16	Nitrofen	$C_{12}H_7Cl_2NO_3$	3.68	1.87
17	Ethylene dibromide	$C_2H_4Br_2$	3.25	1.50
18	Chlormethyl methyl ether	C_2H_5ClO	2.89	1.66
19	1,1,2-Trichloroethan	$C_2H_3Cl_3$	4.20	1.97
20	Isophosphamide	$C_7H_{15}Cl_2N_2O_2P$	2.90	1.96
21	Sulfallate	$C_8H_{14}ClNS_2$	2.69	1.65
22	Melphalan	$C_{13}H_{18}Cl_2N_2O_2$	2.87	1.72
23	Nitrofen	$C_{12}H_7Cl_2NO_3$	3.68	1.87
24	3,3'-Dichloro bezidine	$C_{12}H_{14}Cl_2$	3.00	1.59
25	Benzyl chloride	C_7H_7Cl	2.80	1.29
26	Benzidine hydrochloride	$C_{12}H_{12}N_2 \cdot HCl$	2.79	1.48
27	Dimethylcarbamoyl chloride	C_3H_6ClNO	3.00	1.90
Inactive chemical compounds				
28	Nitrogen mustard N-oxide	$C_5H_{11}Cl_2NO$	2.80	1.74
29	1,1,2,2-Tetrachloroethane$^{*)}$	$C_2H_2Cl_4$	4.75	1.50
30	2,4,6-Trichlorophenol	$C_6H_3Cl_3O$	4.25	1.56
31	Magenta	$C_{20}H_{20}N_3Cl$	2.77	1.42
32	Dichlorvos	$C_4H_7Cl_2O_4P$	3.67	2.08
33	2,4,6-Trichlorophenol	$C_6H_3Cl_3O$	4.25	1.56
34	Chlorotrianisene	$C_{23}H_{21}ClO_3$	2.88	1.40
35	2,4,5-Trichlorophenoxyacetic acid$^{*)}$	$C_8H_5Cl_3O_3$	4.00	1.87
36	Diazepam	$C_{11}6H_{13}ClN_2O$	3.03	1.59
37	Clomiphene$^{*)}$	$C_{26}H_{28}ClNO$	2.63	1.33
38	Trichlorotriethylamine	$C_6H_{12}Cl_3N \cdot HCl$	3.36	1.96
39	Benzoyl chloride$^{*)}$	C_7H_5ClO	3.29	1.57
40	*para*-Dichlorobenzene$^{*)}$	$C_6H_4Cl_2$	3.50	1.46
41	*ortho*-Dichlorobenzene$^{*)}$	$C_6H_4Cl_2$	3.50	1.46

$^{*)}$ There is insufficient data on carcinogenicity of chemical compound [5].

Let's check the classification rules (8), (16), (25), (27), (36) and (38). For this purpose we will compile a random sample from the data of the handbook [5]. We previously numbered sequentially throughout of almost all chemical compounds of the handbook (Chapters: 1,5-12,14-21,23,25,26,28-32,35). The total number of numbered chemical compounds is equal to 541. Using the table of three-digit random numbers [15] we obtained the sub-sample, which contains 85 random chemical compounds of different classes. We give the numbering of chemical compounds that form a random sub-sample.

489 156 038 460 420 522 020 379 124 487 477 349 012 250 080 074 001
249 224 368 303 371 196 231 380 438 351 323 374 191 464 529 068 119
350 120 026 304 428 447 503 336 534 148 105 473 240 435 422 144 137
070 345 456 277 316 013 203 187 245 352 184 179 088 254 154 209 069
275 034 122 213 230 341 171 284 008 146 291 354 377 415 358 238 402.

(54)

Using random sub-sample (54), we obtained the following statistics for the average values of the molecular descriptors. Active chemical compounds:

$N_1 = 58$, $Z_1^{(av)} = 3.04 \pm 0.09$, $Z_1^{\min} = 2.250$, $Z_1^{\max} = 5.999$, $S_{1Z} = 0.68$,

$N_1 = 58$, $H_1^{(av)} = 1.53 \pm 0.04$, $H_1^{\min} = 1.532$, $H_2^{\max} = 2.069$, $S_{1H} = 0.29$.

Inactive chemical compounds:

$N_2 = 27$, $Z_2^{(av)} = 3.20 \pm 0.11$, $Z_2^{\min} = 2.316$, $Z_2^{\max} = 5.430$, $S_{2Z} = 0.58$,

$N_2 = 27$, $H_2^{(av)} = 1.63 \pm 0.06$, $H_2^{\min} = 0.979$, $H_2^{\max} = 2.135$, $S_{2H} = 0.31$. (55)

For the random sub-sample (54) the average values will be as follows:

$$N = 85, \qquad Z^{(av)} = 3.09 \pm 0.07, \qquad H^{(av)} = 1.56 \pm 0.03. \qquad (56)$$

The average values (55) and (56) do not differ substantially (within the width of the confidence interval) from the average values that were obtained for other samples (see (8), (16), (25), (27) and (38)). That is, the threshold values of molecular descriptors, as well as the average values of descriptors $Z_{1,2}$ and $H_{1,2}$ approximately retain their values for different samples. Hence, the samples formed on the basis of various assumptions yield similar results, thus that the results are stable (Table 6).

Table 6. *A summary table of threshold and average descriptors values for different subsamples.*

Original Table 1		
$N = 250$ $N = 255$	$H^{(th)} \equiv H^{(av)} = 1.62 \pm 0.02$ (16)	$Z^{(th)} \equiv Z^{(av)} = 3.06 \pm 0.03$ (10)
$N_1 = 137$ $N_1 = 141$	$H_1^{(av)} = 1.56 \pm 0.03$	$Z_1^{(av)} = 2.92 \pm 0.03$
$N_2 = 113$ $N_2 = 114$	$H_2^{(av)} = 1.70 \pm 0.03$	$Z_2^{(av)} = 3.23 \pm 0.04$
The sub-sample (25)		
$N = 60$	$H^{(th)} \equiv H^{(av)} = 1.63 \pm 0.04$	$Z^{(th)} \equiv Z^{(av)} = 3.13 \pm 0.06$
$N_1 = 28$	$H_1^{(av)} = 1.45 \pm 0.06$	$Z_1^{(av)} = 2.86 \pm 0.08$
$N_2 = 32$	$H_2^{(av)} = 1.79 \pm 0.05$	$Z_2^{(av)} = 3.36 \pm 0.08$
The sub-sample (54)		
$N = 85$	$H^{(th)} \equiv H^{(av)} = 1.56 \pm 0.03$	$Z^{(th)} \equiv Z^{(av)} = 3.09 \pm 0.07$
$N_1 = 58$	$H_1^{(av)} = 1.53 \pm 0.04$	$Z_1^{(av)} = 3.04 \pm 0.09$
$N_2 = 27$	$H_2^{(av)} = 1.63 \pm 0.06$	$Z_2^{(av)} = 3.20 \pm 0.11$
The sub-sample (41)		
$N = 63$	$H^{(th)} \equiv H^{(av)} = 1.68 \pm 0.03$	$Z^{(th)} \equiv Z^{(av)} = 3.15 \pm 0.05$
$N_1 = 26$	$H_1^{(av)} = 1.62 \pm 0.03$	$Z_1^{(av)} = 2.91 \pm 0.05$
$N_2 = 37$	$H_2^{(av)} = 1.71 \pm 0.04$	$Z_2^{(av)} = 3.41 \pm 0.08$
The sub-sample (48)		
$N = 78$	$H^{(th)} \equiv H^{(av)} = 1.87 \pm 0.02$	$Z^{(th)} \equiv Z^{(av)} = 3.15 \pm 0.04$
$N_1 = 44$	$H_1^{(av)} = 1.83 \pm 0.03$	$Z_1^{(av)} = 3.09 \pm 0.05$
$N_2 = 34$	$H_2^{(av)} = 1.93 \pm 0.03$	$Z_2^{(av)} = 3.23 \pm 0.06$
The sub-sample (Table 4)		
$N = 73$	$H^{(th)} \equiv H^{(av)} = 1.87 \pm 0.02$	$Z^{(th)} \equiv Z^{(av)} = 3.17 \pm 0.05$
$N_1 = 38$	$H_1^{(av)} = 1.63 \pm 0.03$	$Z_1^{(av)} = 3.01 \pm 0.05$
$N_2 = 35$	$H_2^{(av)} = 1.69 \pm 0.04$	$Z_2^{(av)} = 3.35 \pm 0.09$

It should be noted that the assessment of the biological activity of certain chemical compounds in the handbook [5] contains an uncertainty that is associated with lack of knowledge of chemicals carcinogenic activity. The purpose of the classification model is to help the researcher quickly assess the likely presence or absence of carcinogenic properties of a new substance or poorly explored chemicals. In this case, the molecular descriptors Z and H complement each other. The easily calculated molecular descriptors offered here make it easier for the researcher to identify probabilistically the biological activity of substances that have not been thoroughly studied. Thus, carcinogenically active chemical compounds are preferably located in the region below of the threshold values $Z^{(\mathrm{th})}$ and $H^{(\mathrm{th})}$. Inactive chemical compounds are preferably located above these threshold values. The possibility of a preliminary probabilistic evaluation of the biological activity of the agent may be useful in the synthesis of new chemical compounds. In addition, the classification rules allow researchers to pay attention to chemical compounds that have already been included in the reference books on carcinogenic activity but they are characterized as "insufficiently studied" or "experimental data are inadequate", "evidence is limited", "impossible to estimate the carcinogenic activity" [5].

4. Comparison with Monitoring

Let us check the possibilities of the proposed here classification model to evaluate the carcinogenic activity of chemicals. We will analyze the following series of chemical compounds: organonitroso-compounds using the data of the handbook (each and every chemical compounds from Chapter 16 of the handbook [5]), oxy-compounds (each and every chemical compounds from ([3] Appendix II, Russian edition), mustard (each and every chemical compounds from Chapter 8 of the handbook [5]), drugs (each and every chemical compounds from Chapter 23 of the handbook [5]), as well as aromatic amines and related compounds (each and every chemical compounds from Chapter 4 of the handbook [5]).

The table 7 below shows the aromatic amines and chemically related compounds. All chemical compounds without exception from the handbook ([5] Chapter 4, Subsection "Some aromatic amines, hydrazine and related chemical compounds") are included in this sub-sample. We apply the classification rules (8) and (16). Only in one case (4-nitrobiphenyl) the descriptor of Z slightly exceeds the threshold value $Z^{(\mathrm{th})} = 3.06$. And this exceeding fits into the confidence limits of the threshold value. There are no chemical compounds that violate classification rule on the basis of descriptor H ($H^{(\mathrm{th})} = 1.62$ *bits*).

Table 7. *Carcinogenic properties of aromatic amines and related compounds*

N	Chemical compounds	Gross formula	Activity	*Z*	*H,bits*
1	3,3'-Dimethoxy benzidine	$C_{14}H_{16}O_2N_2$	+	$2.77 < Z^{(\mathrm{th})}$	$1.52 < H^{(\mathrm{th})}$
2	Magenta	$C_{20}H_{19}N_3 \cdot HCl$	+	$2.77 < Z^{(\mathrm{th})}$	$1.43 < H^{(\mathrm{th})}$
3	4,4'-Methylene bis(2-chloraniline)	$C_{13}H_{12}N_2Cl_2$	+	$2.86 < Z^{(\mathrm{th})}$	$1.58 < H^{(\mathrm{th})}$
4	4,4'-Methylene bis(2-methylaniline)	$C_{15}H_{18}N_2$	+	$2.51 < Z^{(\mathrm{th})}$	$1.25 < H^{(\mathrm{th})}$
5	4,4'-Methylene bis (2-dianiline)	$C_5H_{11}Cl_2N$	+	$2.62 < Z^{(\mathrm{th})}$	$1.29 < H^{(\mathrm{th})}$
6	1-Naphthylamine	$C_{10}H_9N$	+	$2.70 < Z^{(\mathrm{th})}$	$1.23 < H^{(\mathrm{th})}$
7	2-Naphthylamine	$C_{10}H_9N$	+	$2.70 < Z^{(\mathrm{th})}$	$1.23 < H^{(\mathrm{th})}$
8	4-Nitrobiphenyl	$C_{12}H_9NO_2$	+	$3.08 \approx Z^{(\mathrm{th})}$	$1.52 < H^{(\mathrm{th})}$
9	N,N-Bis(2-Chloroethyl)-2-naphthylamine	$C_{14}H_{15}Cl_2N$	+	$2.81 < Z^{(\mathrm{th})}$	$1.44 < H^{(\mathrm{th})}$
10	Hydrazine	N_2H_4	+	$2.33 < Z^{(\mathrm{th})}$	$0.92 < H^{(\mathrm{th})}$
11	1,1-Dimethylhydrazine	$C_2H_8N_2$	+	$2.17 < Z^{(\mathrm{th})}$	$1.25 < H^{(\mathrm{th})}$
12	1,2-Dimethylhydrazine	$C_2H_8N_2$	+	$2.17 < Z^{(\mathrm{th})}$	$1.25 < H^{(\mathrm{th})}$
13	1,2-Diethylhydrazine	$C_4H_{12}N_2$	+	$1.88 < Z^{(\mathrm{th})}$	$1.06 < H^{(\mathrm{th})}$
14	Izonicotinic acid hydrazide	$C_6H_7N_3O$	-	$3.06 \approx Z^{(\mathrm{th})}$	$1.74 > H^{(\mathrm{th})}$
15	Maleic hydrazide	$C_4H_4N_2O_2$	-	$3.50 > Z^{(\mathrm{th})}$	$1.92 > H^{(\mathrm{th})}$

As analysis has shows the descriptors for chemical compounds (Table 7) are interrelated (Fig. 2):

$$H(Z) = A + B \cdot Z, \quad A = -0.06 \pm 0.24, B = 0.54 \pm 0.09,$$
$$t(B) = 6.048 > t^{(cr)}_{0.05}(f = 13) = 1.77 > |t(A)| = 0.240, \quad N = 15, \quad R = 0.86,$$
$$F = 36.44 > F^{(cr)}_{0.05}(f_1 = 1; f_2 = 13) = 4.67,$$
$$Std.Err.\ of\ Estimate = 0.137, \ \delta H(Z) = 9.08\%. \qquad (57)$$

It is not difficult to see (Fig.2) that hydrazine gives the greatest deviation from the regression line. According to the formal definition of the group, hydrazine is referred to the aromatic amines [5]. Nevertheless, its molecular descriptors differ significantly from the other fourteen chemical compounds. The relationship between molecular descriptors can be used for the purposes of quantifying chemical compounds as related compounds. To quantify of the likeness of chemical compounds it can be used statistical criteria, for example, an approximation error or *RMSE* value. From this point of view, hydrazine falls out of the class of aromatic compounds. Eliminating hydrazine from the subsample (Table 7) reduces the magnitude of the approximation error to 6.5%.

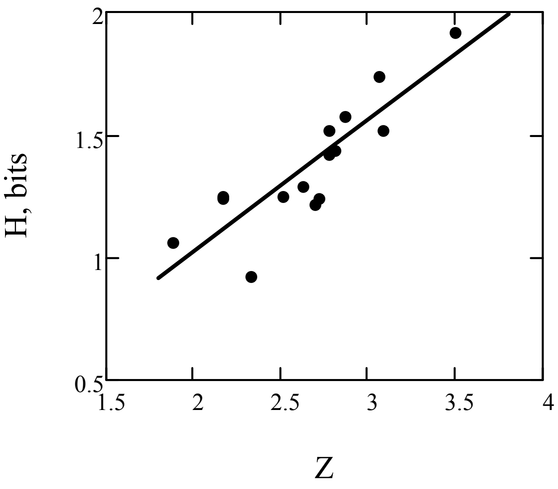

Fig. 2. *The correlation field of the information function and the electronic descriptor for aromatic amines and relative compounds. Points are the data Table 7. Regression line is approximated by the function* (57).

An important criterion for the quality of the classification model is the determination of the magnitude of the error in the prediction of biological activity for chemical compounds that were not included in original sample. The resulting classification rules allow us specifying only probabilistically the presence or absence of carcinogenic activity of chemicals. At the same time, the classification rules do not allow us to establish any monotonous relationship between the change in carcinogenic activity and the change in descriptors.

It is well known to change the carcinogenic activity of 4-aminobiphenyl [5, 19] by varying the substituents of the molecule. Now, let's compare the change in the carcinogenic activity of 4-aminobiphenyl (Z = 2.67, H = 1.21*bits*) at changes in the electronic and information descriptors of the molecule. For example, the replacement of the hydrogen atom by amino group -N(OH)-OCCH$_3$ leads to an increase in the carcinogenic activity of the molecule and simultaneously to an increase of the descriptor values: Z = 2.88, H = 1.54*bits*. At the same time, the addition of two methyl radicals at positions 3 and 3' is accompanied by an increase in the carcinogenic properties of the chemical compound. Nevertheless, the descriptor values are decreased: Z = 2.48, H = 1.18*bits*. The replacement of hydrogen atom at the 4' position with atomic groups NH$_2$ (Z = 2.69, H = 1.31*bits*), NO$_2$ (Z = 3.07, H = 1.62*bits*), NOHOCCH$_3$ (Z = 2.88, H = 1.54*bits*) leads to an increase in the values descriptors and, at the same time, enhancing the carcinogenic activity of the molecule. A similar situation is observed by varying the molecular structure of 2-acetylaminofluorene which has a carcinogenic activity (Z = 2.80, H = 1.35*bits*). The following chemical compounds have been tested whose molecular structures are close to the structure of the molecule 2-acetylaminofluorene: 2-diacetylaminofluorene (Z = 2.91, H = 1.42*bits*), 2-methylaminofluorene (Z = 2.70, H = 1.19*bits*), 2-dimethylaminofluorene (Z = 2.63, H = 1.18*bits*), 7-fluoro-2-acetylaminofluorene (Z = 3.00, H = 1.52*bits*). These chemical compounds also have carcinogenic activity. However, they are markedly weaker than 2-acetylaminofluorene. It is important to note molecular descriptors can be either higher or lower than the descriptors of the initial compound. Apparently, the probabilistic model does not allow us to reveal such subtle variations of molecule structures.

In addition, heterocyclic derivatives of 2-acetylaminofluorene were investigated. For example, 3-acetylaminodibenzothiophene (Z = 2.87, H = 1.53*bits*), has a higher carcinogenic activity than 2-acetylaminofluorene. At the same time the descriptor values is closer to the threshold values. That is, there is a multidirectional change in molecular descriptors and carcinogenic activities of molecules. Such results justify the application of the method of associations (conjugation) in constructing a mathematical model. In this case, the main factor is the threshold effect. Similar non-monotonic relationships exist for other classes of chemical compounds.

The class of aromatic hydrocarbons includes aminostilbenes. The carcinogenic activity of the following chemical compounds was investigated [3,19]: 4-aminostilbene (Z = 2.59, H = 1.18*bits*), 4-methylaminostilbene (Z = 2.58, H = 1.17*bits*), 4-diethylaminostilbene (Z = 2.45, H = 1.14*bits*), 4-dimethylamino-2'-methylstilbene (Z = 2.44, H = 1.15*bits*). All these chemicals have carcinogenic activity. Molecular descriptors are in the region below the threshold values. This does not contradict the classification rules. However, for stilbene (Z = 2.61, H = 0.99*bits*), carcinogenic activity was not detected, as well as for 4'-fluoro-4-aminostilbene (Z = 2.86, H = 1.37*bits*). Apparently, it is necessary to study the electronic structure of molecules much more thoroughly, using more rigorous quantum mechanical methods. Quite subtle differences in the molecular structure can affect the carcinogenicity of chemical compounds [19]. However, the calculation *ab initio* of the electronic structure of large molecules is not a simple task even with the current state of computer technology. This is due to the need to optimize the geometry of polyatomic molecules, for example, such as dyes.

Table 8 shows the carcinogenic activity and molecular descriptors of a series of dyes. We used each and every the data from Chapter 15 (the "Dyes" section) of the handbook [5]. These data were supplemented by data from the monograph ([3] Appendix II, Russian edition). We used the threshold values $Z^{(\mathrm{th})}$ = 3.15**,** $H^{(\mathrm{th})}$ =1.87*bits* (see Table 6). Table 8 demonstrates the descriptors do not contradict the classification rules for all chemical compounds without exception.

Table 8. *Carcinogenic properties of dyes*

N	Chemical compounds	Gross formula	Activity	*Z*	*H, bits*
1	Acredine orange	$C_{17}H_{19}N_3$	+	$2.62 < Z^{(\mathrm{th})}$	$1.31 < H^{(\mathrm{th})}$
2	Benzyl violet 4B	$C_{39}H_{40}N_3O_6S_2{\cdot}Na$	+	$2.82 < Z^{(\mathrm{th})}$	$1.61 < H^{(\mathrm{th})}$
3	Brilliant blue FCF	$C_{37}H_{34}N_2O_9S_3{\cdot}2NH_2$	+	$2.97 < Z^{(\mathrm{th})}$	$1.72 < H^{(\mathrm{th})}$
4	Disodium salt (sour celestial blue)	$C_{37}H_{34}N_2O_9S_3{\cdot}2Na$	+	$3.05 < Z^{(\mathrm{th})}$	$1.81 < H^{(\mathrm{th})}$
5	Fast green FCF	$C_{37}H_{34}N_2O_{10}S_3{\cdot}2Na$	+	$3.09 < Z^{(\mathrm{th})}$	$1.83 < H^{(\mathrm{th})}$
6	Guinea green B	$C_{37}H_{35}N_2O_6S_2{\cdot}Na$	+	$2.92 < Z^{(\mathrm{th})}$	$1.66 < H^{(\mathrm{th})}$
7	Rhodamine B	$C_{28}H_{31}N_2O_3{\cdot}Cl$	+	$2.74 < Z^{(\mathrm{th})}$	$1.49 < H^{(\mathrm{th})}$
8	Rhodamine 6G	$C_{28}H_{30}N_2O_3{\cdot}HCl$	+	$2.74 < Z^{(\mathrm{th})}$	$1.49 < H^{(\mathrm{th})}$
9	Light green SF	$C_{37}H_{34}N_2O_9S_3{\cdot}2Na$	+	$3.06 < Z^{(\mathrm{th})}$	$1.81 < H^{(\mathrm{th})}$
10	Blue VRS	$C_{27}H_{31}N_2O_6S_2{\cdot}Na$	+	$2.87 < Z^{(\mathrm{th})}$	$1.74 < H^{(\mathrm{th})}$
11	Acid green	$C_{36}H_{34}N_2S_3O_9Na_2$	+	$3.09 < Z^{(\mathrm{th})}$	$1.88 \approx H^{(\mathrm{th})}$
12	Acid red C	$C_{20}H_{12}N_2O_7S_2Na_2$	-	$3.59 > Z^{(\mathrm{th})}$	$2.15 > H^{(\mathrm{th})}$
13	Indigo carmine	$C_{16}H_8N_2O_8S_2Na_2$	-	$3.79 > Z^{(\mathrm{th})}$	$2.14 > H^{(\mathrm{th})}$
14	Tartrazine	$C_{16}H_9N_4O_9S_2Na_3$	-	$3.78 > Z^{(\mathrm{th})}$	$2.27 > H^{(\mathrm{th})}$

Table 9 presents a group of chemical compounds belonging to the class of nitroso-compounds. These chemical compounds contain only carbon, hydrogen, nitrogen and oxygen atoms. As threshold values, we took the values (36) and (41): $Z^{(\mathrm{th})} \equiv Z^{(\mathrm{av})} = 3.15$, $H^{(\mathrm{th})} \equiv H^{(\mathrm{av})} = 1.69$*bits*.

From the Table 9 it follows that the use of classification rules on the basis of Z led to an error in biological activity in four cases (the error is equal to 20%) and in eight cases (the error is equal to 40%) when using descriptor of H. These results correspond to the models (36) and (41), which involves empirical errors: 27% and 37%. It is suggested [19] that the majority of nitroso-compounds belong to the "indirect" carcinogens, which demonstrate a blastomogenic effect. Activation is apparently associated with the disintegration of the molecule of nitroso-compounds and the formation of alkylating agents, which, in turn, are the "final" carcinogens. At the same time, carcinogen *N*-nitroso-*N*-methylurethane ($Z = 3.06 < Z^{(\mathrm{th})}$) do not need to be pre-metabolized and cause neoplasms with local application.

Table 9. *Carcinogenic activity of organonitroso compounds* *)

N	Chemical compounds	Gross formula	Activity	*Z* **)	*H,bits* **)
1	*N*-Nitrosodi-*n*-butylamine	$C_8H_{18}N_2O$	+	$2.28 < Z^{(\mathrm{th})}$	$1.37 < H^{(\mathrm{th})}$
2	*N*-Nitrosodimethylamine	$C_2H_6N_2O$	+	$2.73 < Z^{(\mathrm{th})}$	$1.69 = H^{(\mathrm{th})}$
3	N-Nitrosoidi-n-propylamine	$C_6H_{14}N_2O$	+	$2.35 < Z^{(\mathrm{th})}$	$1.45 < H^{(\mathrm{th})}$
4	*N*-Nitrosodi-*n*-propylamine	$C_4H_{10}N_2O_3$	+	$2.84 < Z^{(\mathrm{th})}$	$\mathbf{1.72} > H^{(\mathrm{th})}$
5	*N*-Nitrosodiethylamine	$C_4H_{10}N_2O$	+	$2.47 < Z^{(\mathrm{th})}$	$1.54 < H^{(\mathrm{th})}$
6	*N*-Nitrosomethylvinyl-amine	$C_3H_6N_2O$	+	$2.83 < Z^{(\mathrm{th})}$	$1.55 < H^{(\mathrm{th})}$
7	*N*-Nitroso-*N*-methylurea	$C_2H_5N_3O_2$	+	$\mathbf{3.33} > Z^{(\mathrm{th})}$	$\mathbf{1.89} > H^{(\mathrm{th})}$
8	*N*-Nitrosomethylethyl-amine	$C_3H_8N_2O$	+	$2.57 < Z^{(\mathrm{th})}$	$1.61 < H^{(\mathrm{th})}$
9	*N*-Nitrosomorpholine	$C_4H_8N_2O_2$	+	$2.88 < Z^{(\mathrm{th})}$	$\mathbf{1.75} > H^{(\mathrm{th})}$
10	*N*'-Nitrosonicotine	$C_9H_{14}N_3O$	+	$2.63 < Z^{(\mathrm{th})}$	$1.55 < H^{(\mathrm{th})}$
11	*N*-Nitrosopiperidine	$C_5H_{10}N_2O$	+	$2.56 < Z^{(\mathrm{th})}$	$1.57 < H^{(\mathrm{th})}$
12	*N*-Nitrosopyrrolidine	$C_4H_8N_2O$	+	$2.67 < Z^{(\mathrm{th})}$	$1.62 < H^{(\mathrm{th})}$
13	*N*-Nitrososarcosine	$C_3H_6N_2O_3$	+	$\mathbf{3.29} > Z^{(\mathrm{th})}$	$\mathbf{1.88} > H^{(\mathrm{th})}$
14	N-Methyl-N'-nitro-N-nitrosoguanidine	$C_3H_7N_3O_2$	+	$\mathbf{3.73} > Z^{(\mathrm{th})}$	$\mathbf{1.91} > H^{(\mathrm{th})}$
15	*N*-Nitroso-*N*'-methylurethane	$C_4H_8N_2O_3$	+	$3.06 < Z^{(\mathrm{th})}$	$\mathbf{1.81} > H^{(\mathrm{th})}$
16	Streptozotocin	$C_8H_{15}N_3O_7$	+	$3.15 = Z^{(\mathrm{th})}$	$\mathbf{1.80} > H^{(\mathrm{th})}$
17	*N*-Nitroso-*N*-ethylurea	$C_3H_7N_3O_2$	+	$3.07 < Z^{(\mathrm{th})}$	$\mathbf{1.83} > H^{(\mathrm{th})}$
18	*N*-Nitrosoproline	$C_5H_8N_2O_3$	-	$\mathbf{3.11} < Z^{(\mathrm{th})}$	$1.81 > H^{(\mathrm{th})}$
19	*N*-Nitrosohydrooxyproline	$C_5H_8N_2O_4$	-	$3.26 > Z^{(\mathrm{th})}$	$1.85 > H^{(\mathrm{th})}$
20	*N*-Nitrosofolie acid	$C_{19}H_{18}N_8O_7$	-	$3.39 > Z^{(\mathrm{th})}$	$1.87 > H^{(\mathrm{th})}$

*) Chemical compounds that violate the classification rule are indicated in bold type.

**) Descriptors of chemical compounds, for which the classification rules are satisfied with a confidence interval, are italicized.

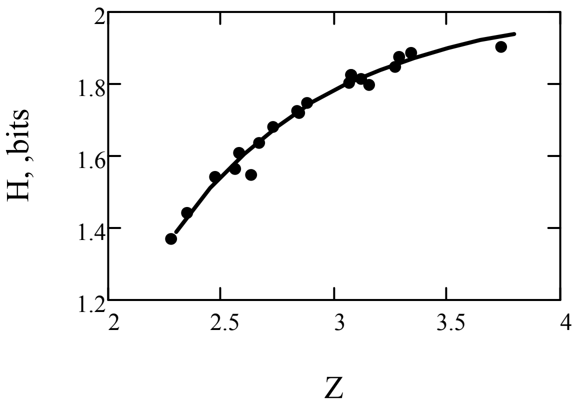

Fig. 3. *The correlation field of the information function H and the electronic characteristic Z for nitroso compounds. Points are the Table* 9. *Regression line is approximated by the following function*: $H(Z) = A + B\cdot\exp(-C\cdot Z)$, $A = 2.01 \pm 0.05$, $B = -16.6 \pm 7.27$, $C = 1.43 \pm 0.21$, $N = 20$, $RMSE = 0.025$, $\delta H(Z) = 2.01\%$.

For the homologous series of chemical compounds the interrelation between the descriptors has such a small *RMSE* value that the statistical interrelation approaches the functional dependence (Fig.3). At the same time, if the sample contains chemical compounds belonging to different classes, then the scattering around the regression lines becomes more noticeable (see Fig. 1). However, a statistically significant interrelation between descriptors is remained intimately.

Now we will apply the classification rules (36) and (41) to the series of oxy-compounds (Table 10). The table 10 provides a summary of oxy-compounds from the following references: [3 Supplement 2, Russian edition], [5] and [20].

Table 10. Carcinogenic properties and molecular descriptors of oxy-compounds

N	Chemical compounds	Gross formula	Activity	*Z*	*H,bits*
1	Patulin	$C_7H_6O_4$	+	**3.41** $> Z^{(th)}$	1.55 $< H^{(th)}$
2	Sarcomodil	$C_7H_8O_3$	+	3.00 $< Z^{(th)}$	1.48 $< H^{(th)}$
3	Methyl protoanemonin	$C_6H_6O_2$	+	3.00 $< Z^{(th)}$	1.45 $< H^{(th)}$
4	β -Angelikolakton	$C_5H_5O_2$	+	3.08 $< Z^{(th)}$	1.48 $< H^{(th)}$
5	Penicillic acid	$C_8H_{10}O_4$	+	3.00 $< Z^{(th)}$	1.50 $< H^{(th)}$
6	Aflatoxins B_1	$C_{17}H_{12}O_6$	+	**3.31** $> Z^{(th)}$	1.47 $< H^{(th)}$
7	Parasorbic acid	$C_6H_7O_2$	+	2.87 $< Z^{(th)}$	1.43 $< H^{(th)}$
8	Lactone-4-oxyhexenic acid	$C_6H_7O_2$	+	2.87 $< Z^{(th)}$	1.43 $< H^{(th)}$
9	Aflatoxins M_1	$C_{17}H_{12}O_7$	+	**3.39** $> Z^{(th)}$	1.50 $< H^{(th)}$
10	1,2,3,4-Diepoxybutane	$C_4H_6O_2$	+	2.83 $< Z^{(th)}$	1.46 $< H^{(th)}$
11	β-Propiolactone	$C_3H_7O_2$	+	3.11 $< Z^{(th)}$	1.53 $< H^{(th)}$
12	1-Ethylene-oxy-3,4-Epoxycyclohexane	$C_8H_{11}O_2$	+	2.62 $< Z^{(th)}$	1.34 $< H^{(th)}$
13	1-Ethylene-oxy-3,4-epoxycyclohexane	$C_8H_{11}O_2$	+	2.62 $< Z^{(th)}$	1.34 $< H^{(th)}$
14	1,2-Epoxybutane	$C_4H_6O_2$	-	***2.83*** $< Z^{(th)}$	***1.46*** $< H^{(th)}$
15	d1-Diepoxybutane	$C_4H_6O_2$	-	***2.83*** $< Z^{(th)}$	***1.46*** $< H^{(th)}$
16	Styrene oxide	C_8H_8O	-	***2.71*** $< Z^{(th)}$	***1.26*** $< H^{(th)}$
17	9,10- Epoxystearic acid	$C_{18}H_{34}O_3$	-	***2.25*** $< Z^{(th)}$	***1.19*** $< H^{(th)}$
18	6,7,9,10- Epoxystearic acid	$C_{18}H_{32}O_4$	-	***2.37*** $< Z^{(th)}$	***1.25*** $< H^{(th)}$

19	Hexaepoxysvalol	$C_{30}H_{48}O_6$	-	***2.48*** $< Z^{(\mathrm{th})}$	***1.26*** $< H^{(\mathrm{th})}$
		Hydroperoxides			
20	1-Vinyl-1-hydroperoxide of cyclohexane	$C_8H_{11}O_2$	+	2.62 $< Z^{(\mathrm{th})}$	1.53 $< H^{(\mathrm{th})}$
21	1-Vinilcyclohexane-3	C_8H_{11}	+	2.26 $< Z^{(\mathrm{th})}$	0.98 $< H^{(\mathrm{th})}$
22	Benzene peroxide	$C_{14}H_{10}O_4$	-	3.21 $> Z^{(\mathrm{th})}$	**1.43** $< H^{(\mathrm{th})}$
23	Lauroyl peroxide	$C_{24}H_{46}O_4$	-	***2.24*** $< Z^{(\mathrm{th})}$	***1.18*** $< H^{(\mathrm{th})}$

The verification of the applicability of the classification rule (36) demonstrated that the descriptor of Z in ten cases gives an incorrect valuation of the carcinogenicity (the error is equal to 43%). At the same time, descriptor of H incorrectly estimates the carcinogenic properties for eight agents (the error is 35%). That is, for this sub-sample the descriptor of H turned out to be more informative than the descriptor of Z. However, it is necessary to make the following important remark. For example, the chemical compound at number 19 (Table 10) is potentially active agent ($Z < Z^{(\mathrm{th})}$). We assume that the hypothesis [8] on the role of hydrophobicity is acceptable not only for radioprotectors, but also for carcinogenic activity. Then the absence of carcinogenic activity of chemical compounds with numbers from 18 up to 20 (the descriptors for these molecules are marked in italics in Table 10) is due to a change in the hydrophobic properties of the molecules. Such chemical compounds seem to be potentially active carcinogens, but do not show activity, since they have a large number of CH_2 and CH (m >14) atomic groups. We note an analogous situation for the molecule of lauroyl peroxide in the series of hydroperoxides (chemical compound at number 23; Table 10). Chemical compounds at numbers 14 and 15 are also potentially active. However, they do not have confirmed carcinogenic activity. According to the data of [20], the agent at number 14 has very weak carcinogenic activity. For these chemicals, the index m is less than 5. That is, the index m lies outside the permissible region defined in [7]. At the same time, for 1-ethylene-hydroxy-3,4-epoxycyclohexane and 1-vinyl-1-hydroperoxide of cyclohexane-3, the number of such atomic groups is equal to 9 and 10, respectively. This practically coincides with the area of maximum bioactivity [7]. The index of carcinogenic activity of these chemical compounds on a five-point scale is 3 and 5 [20]. This hypothesis does not contradict the carcinogenic activity of triethylene glycol diglycidyl ether [5]. This compound belongs to the same class of agents. The molecular descriptors for this molecule are below threshold values. The index m falls into the range of values: $7 > m > 5$. This range of index m [7] does not prevent the manifestation of biological activity. Thus, if we take into account the influence of the length of carbon chains on carcinogenic activity, this increases the accuracy of the model by a factor of two. The paper [21] indicates that the elongation of an alkyl radical chain reduces the carcinogenic activity of chemical compounds, until it completely disappears. In this connection, it can be noted that the accumulation of methyl groups in the azo dye molecule of 2,5,4',6'-tetramethyl-4-aminoazobenzene also leads to a decrease in carcinogenic activity [22]. A similar situation exists for piperonyl butoxide ($C_{19}H_{30}O_5$), for which the molecular descriptors satisfy the classification rules (36) and (41). However, it should be noted that the carcinogenicity for this insecticide has not been proven [5]. This preparation has long hydrocarbon chains, for which the index m is greater than 10. According to [7-9], this probably does not contribute to the manifestation of the biological activity of the chemical compound.

A similar situation is noted [23] for a number of nitrosomethylamines. At first, increasing the length of the hydrocarbon chain is accompanied by an increase in the carcinogenic activity (in scope of 4-ball scale). Then it has been noted a decrease in carcinogenic activity. (Table 11, Fig.4). The hydrophobicity of the molecules of the homologous series $ON\text{-}N\text{-}CH_3(CH_2)_mCH_3$ was determined by the additive increment method [24]. The index m ranges from 1 up to 12. The contribution to the hydrophobicity of one atomic group CH_2 is equal to: $\pi = \log(P) = 0.52$, here P is the hydrophobicity.

Table 11. *Comparative carcinogenic activity of the homologous series of nitrosomethylalkylamines.*

N	Chemical compound ON-N-$CH_3(CH_2)_mCH_3$	Gross formula	Activity (A)	*Z*	*H,bits*	*π*
1	*m* = 2	$C_4H_{10}N_2O$	++++	2.47	1.55	1.04
2	*m* = 3	$C_5H_{12}N_2O$	++++	2.40	1.49	1.56
3	*m* = 4	$C_6H_{14}N_2O$	++++	2.35	1.45	2.08
4	*m* = 5	$C_7H_{16}N_2O$	++++	2.31	1.41	2.60
5	*m* = 1	$C_3H_8N_2O$	+++	2.57	1.61	0.52
6	*m* = 6	$C_8H_{18}N_2O$	+++	2.28	1.37	3.12
7	*m* = 7	$C_9H_{20}N_2O$	+++	2.25	1.34	3.64
8	*m* = 8	$C_{10}H_{22}N_2O$	+++	2.23	1.32	4.16
9	*m* = 9	$C_{11}H_{24}N_2O$	++	2.21	1.30	4.68
10	*m* = 10	$C_{12}H_{26}N_2O$	++	2.20	1.28	5.20
11	*m* = 11	$C_{13}H_{28}N_2O$	++	2.18	1.26	5.72
12	*m* = 12	$C_{14}H_{30}N_2O$	++	2.17	1.25	6.24

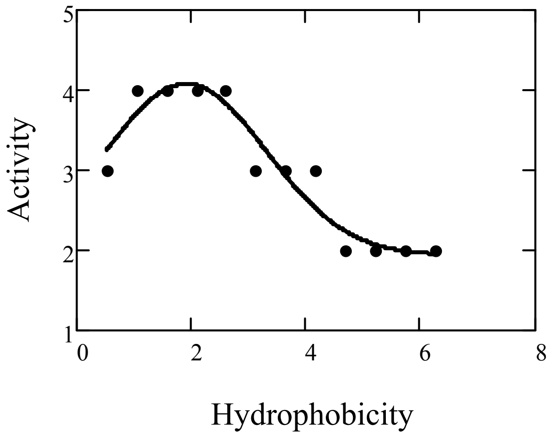

Fig. 4. *Interrelation nitrosomethylalkylamines carcinogenic activity (Table 11) with their hydrophobicity. Points are the comparative carcinogenic activity (Table 11). The envelope of the regression line is defined by the following equation*: A(π) = $D + A\cdot\exp(-(\pi-B)^2/C^2)$, A = 2.14 ± 0.23, B = 1.91 ± 0.15, C = -1.98 ± 0.33, D = 1.94 ± 0.21, *RMSE* = 0.28.

Figures 5A and 5B have demonstrated the close interrelation of molecular descriptors, as well as the interrelation of the information function to the hydrophobic contribution of CH_2 atomic groups to the total hydrophobicity of homologous series molecules.

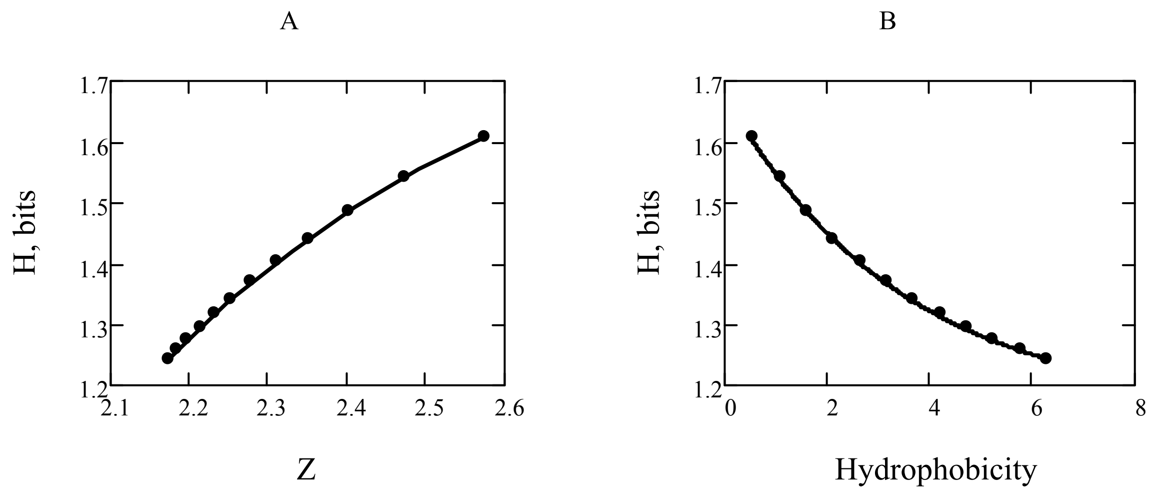

Fig.5. (A) *Interrelation of molecular descriptors for a series of nitrosomethylalkylamines. Points are the Table 11. The regression line is determined by the equation: H(Z) = B + A·exp(-C·Z), A* = -58.2 ± 0.96*, B* = 1.89 ± 0.002*, C* = 2.07 ± 5.59 , *RMSE* = 0.0002. (B) *Interrelation of information function with the hydrophobicity. Points are the Table 5. The regression line is determined by the equation: H(π) = B + A·exp(-C·π), A* = 0.55 ± 0.003*, B* = 1.16 ± 0.005*, C* = 0.29 ± 0.006*, RMSE* = 0.002.

The figure 6 shows the nonlinear association of descriptors *Z* and *H* for oxy-compounds. The approximation parameter *C* determines the curvature of a curve. For nitrosocompounds and hydroxy compounds, the parameter *C* has similar values. It is interesting to see *RMSE* so small ($\approx 10^{-3}$-10^{-4}) that the interrelations are practically functional for closely related chemical compounds.

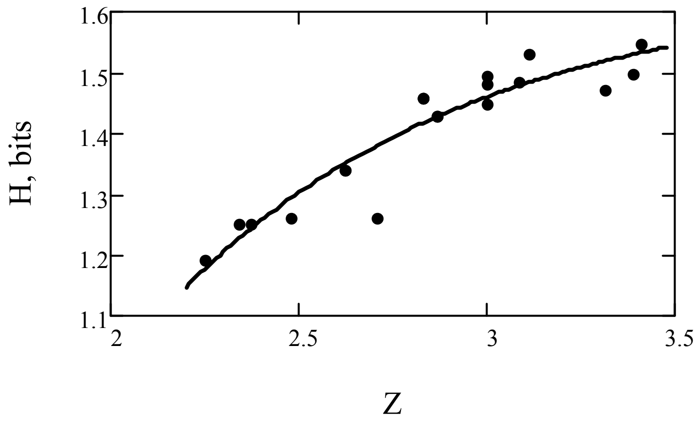

Fig. 6. *Oxy-compounds from Table 10. The regression line is determined by the equation: H(Z) = A + B·exp(-C·Z), A* = 1.64 ± 0.11*, B* = -9.01 ± 10.40*, C* = 1.32 ± 0.59*, RMSE* = 0.043, $\delta H(Z)$ = 2.21%.

We will check how effectively the application of classification rules in the analysis of the carcinogenic properties of chemical compounds such as mustard gas. The Table 12 bellow shows the quantitative values of chemical compounds molecular descriptors, as well as their carcinogenic activity. We analyze each and every chemical compounds of the type mustard from references [3,5]. For this series of chemical compounds that contain sulfur and chlorine, we will use the threshold values of molecular descriptors (8) and (19).

Table 12. *Carcinogenic properties of chemical compounds such as mustard gas*

N	Chemical compounds	Gross formula	Activity	*Z*	*H,bits*
1	Bis(2-chloroethyl) ether	$C_4H_8OCl_2$	+	$2.93 < Z^{(th)}$	$1.64 \approx H^{(th)}$
2	Mannomustine dihydrochloride	$C_{10}H_{24}Cl_4N_2O_4$	+	$2.86 < Z^{(th)}$	$\mathbf{1.80} > H^{(th)}$
3	Melphan	$C_{13}H_{18}Cl_2N_2O_2$	+	$2.86 < Z^{(th)}$	$\mathbf{1.72} > H^{(th)}$
4	Mustard gas	$C_4H_8Cl_2S$	+	$2.93 < Z^{(th)}$	$1.64 \approx H^{(th)}$
5	Nitrogen mustard	$C_5H_{11}Cl_2N$	+	$2.63 < Z^{(th)}$	$1.53 < H^{(th)}$
6	Nitrogen mustard hydrochloride	$C_5H_{12}Cl_3N$	+	$2.76 < Z^{(th)}$	$1.57 < H^{(th)}$
7	N-Nitrogen mustad	$C_5H_{11}Cl_2NO$	+	$2.80 < Z^{(th)}$	$\mathbf{1.74} > H^{(th)}$
8	Nitrogen mustard *N*-oxide	$C_5H_{12}Cl_3NO$	+	$2.91 < Z^{(th)}$	$\mathbf{1.76} > H^{(th)}$
9	Oestradiol mustard	$C_{42}H_{50}Cl_4N_2O_4$	+	$2.75 < Z^{(th)}$	$1.52 < H^{(th)}$
10	Uracil mustard	$C_8H_{11}Cl_2N_3O_2$	+	$\mathbf{3.23} > Z^{(th)}$	$\mathbf{1.98} > H^{(th)}$
11	Methyl-di-(2-chloroethyl)-amine	$C_5H_{11}Cl_2N$	+	$2.63 < Z^{(th)}$	$1.53 < H^{(th)}$
12	Phenyl-di-(2-chloroethyl)-amine	$C_{10}H_{13}Cl_2N$	+	$2.77 < Z^{(th)}$	$1.50 < H^{(th)}$
13	Trichlorotriethylamine hydrochloride	$C_6H_{12}Cl_3N{\cdot}HCl$	-	$3.36 > Z^{(th)}$	$1.96 > H^{(th)}$

For agents of the mustard gas type (Table 12) the situation is reversed in comparison with the oxy-compounds. The use of the information function gives an erroneous result in five cases (≈39%), whereas the classification rule using the descriptor *Z* gives only one erroneous result (≈ 8%). That is, the descriptor *Z* is more telling in this case. There is also a statistically significant interrelationship between the electronic and information descriptors for chemical compounds of the mustard type (Figure 7).

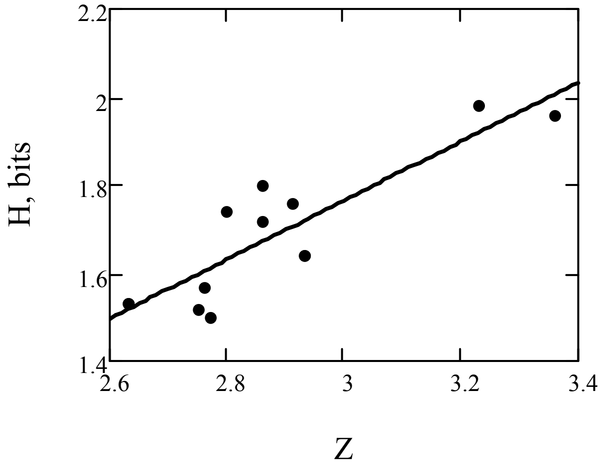

Fig.7. *Interrelation of the descriptors (Table 12). The regression equation has the following form: H*(*Z*) = *A* + *B·Z*, *R* = 0.88, *A* = -(0.25 ± 0.32), *B* = 0.67 ± 0.11, *F* = 38.2 > $F_{0.05}^{(cr)}(f_1 = 1; f_2 = 11)$ = 5.59, $\delta H(Z) = 3.94\%$.

Now we will examine the applicability of the classification rules using a sub-sample of medicines. We have compiled the Table 13, taking into account the threshold values (8) and (16) for molecular descriptors. The table includes each and every without exception the medicines from Chapter 23 of the handbook [5].

Only four chemical compounds (3, 6, 21, 22) breaking the classification rules of Table 13. However, the information function is less than the threshold value for a chemical compound at number of 6. That is, this chemical compound is the carcinogen on the grounds of *H*. Thus, the classification rules lead to an error of ≈ 14%.

Now let's analyze the carcinogenic activity of the metabolites of such initial chemical compounds that are carcinogenically inactive or weakly active. For example, cholesterol is a weak carcinogen ($Z = 2.16$, $H = 1.04$*bits*). There are differ opinions on the carcinogenicity of cholesterol [5]. According to [3], the product of cholesterol transformation is 6*β*-peroxy-Δ^4-cholesten-3-on. This metabolite is a strong carcinogen, whereas its descriptors bigger than cholesterol: $Z = 2.27$, $H = 1.15$*bits*. That is, the increase of descriptors has been accompanied by an increased in the carcinogenic effect. The situation is similar with tryptophan ($Z = 2.89$, $H = 1.60$*bits*), which has no evidence of carcinogenic activity. As a result of tryptophan metabolism three carcinogenic chemical compounds are formed: 3-hydroxykinurenine ($Z = 2.96$, $H = 1.68$*bits*), 2-amino-3-hydroxyacetophenol ($Z = 2.90$, $H = 1.60$*bits*), 3-hydroxyanthranilic acid ($Z = 3.22$, $H = 1.72$*bits*) ([3] Appendix II, Russian edition). For all metabolites molecular descriptors exceed the descriptors of the tryptophan molecule.

Apparently, such an increase in the descriptors is commonplace for metabolites. For example, the transformation of aflatoxin B1 ($Z = 3.31$, $H = 1.47$*bits*) into carcinogenic aflatoxin M1 ($Z = 3.39$, $H = 1.50$*bits*), 4-aminobiphenyl ($Z = 2.67$, $H = 1.21$*bits*) and into carcinogenic N-hydroxy-4-aminobiphenyl ($Z = 2.92$, $H = 1.53$*bits*) [23], is also accompanied by an increase in descriptors. It is assumed that the "final" carcinogen of aflatoxin B_1 is its metabolite 2,3-epoxyaflatoxin B1 ($Z = 3.33$, $H = 1.52$*bits*) [23]. The disappearance of the carcinogenic activity of aflatoxin B1 in hypophysectomy argues in favor of the indirect carcinogenic effect of the initial compound, as the metabolism significantly changes. For all metabolites analyzed (Table 14) the descriptors are closer to the threshold values than for the initial chemical compounds. At the same time, the carcinogenic activity of metabolites is higher than for the initial chemical compounds. However, it is important to note that the descriptors *Z* and *H* remain below the threshold.

Studies of the carcinogenic activity of aromatic amines have shown that 2-acetylaminofluorene is a strong carcinogen. Chemical compounds similar in molecular structure to 2-acetylaminofluorene were tested [3]. These agents include heterocyclic derivatives: 3-acetylaminodibenzothiophene, 3-acetylaminodibenzothiophene-5-oxide, 3-acetylaminodibenzofuran. And all of them turned out to be carcinogens, and 3-acetylaminodibenzothiophene is even more active carcinogen than 2-acetylaminofluorene.

Table 13. *Carcinogenic properties of some drugs*

N	Chemical compounds	Gross formula	Activity	*Z*	*H,bits*
1	Clofibrate	$C_{12}H_{15}O_3Cl$	+	$2.84 < Z^{(\text{th})}$	$1.52 \approx H^{(\text{th})}$
2	Dapsone	$C_{12}H_{12}N_2O_2S$	+	$3.03 < Z^{(\text{th})}$	$1.75 < H^{(\text{th})}$
3	Dihydroxymethyl-furatrizine$^{*)}$	$C_{11}H_{11}N_5O_5$	+	$\mathbf{3.44} > Z^{(\text{th})}$	$\mathbf{1.90} > H^{(\text{th})}$
4	Hydralazine	$C_8H_8N_4$	+	$3.00 < Z^{(\text{th})}$	$1.52 < H^{(\text{th})}$
5	Hydralazine hydrochloride$^{*)}$	$C_8H_8N_4{\cdot}HCl$	+	$3.09 > Z^{(\text{th})}$	$\mathbf{1.71} > H^{(\text{th})}$
6	Methoxsalen$^{*)}$	$C_{12}H_8O$	+	$\mathbf{3.33} > Z^{(\text{th})}$	$1.46 < H^{(\text{th})}$
7	Nafenopin	$C_{20}H_{22}O_3$	+	$2.67 < Z^{(\text{th})}$	$1.29 < H^{(\text{th})}$
8	Phenacetin	$C_{10}H_{13}NO_2$	+	$2.69 < Z^{(\text{th})}$	$1.50 > H^{(\text{th})}$
9	Phenazopyridine	$C_{11}H_{11}N_5$	+	$2.96 < Z^{(\text{th})}$	$1.51 < H^{(\text{th})}$
10	Phenazopyridine Hydrochloride	$C_{11}H_{125}$	+	$3.03 < Z^{(\text{th})}$	$\mathbf{1.66} > H^{(\text{th})}$

11	Phenelzine	$C_8H_{12}N_2$	+	$2.46 < Z^{(\mathrm{th})}$	$1.32 < H^{(\mathrm{th})}$
12	Phenelzine sulphate	$C_8H_{12}N_2{\cdot}H_2SO_4$	+	$2.97 < Z^{(\mathrm{th})}$	$1.50 < H^{(\mathrm{th})}$
13	Phenoxybenzamine hydrochloride	$C_{18}H_{23}Cl_2NO$	+	$2.67 < Z^{(\mathrm{th})}$	$1.47 > H^{(\mathrm{th})}$
14	Proflavine	$C_{13}H_{11}N_3$	+	$2.67 < Z^{(\mathrm{th})}$	$1.39 < H^{(\mathrm{th})}$
15	Proflavine dihydrochloride	$C_{13}H_{11}N_3{\cdot}2HCl$	+	$3.03 < Z^{(\mathrm{th})}$	$1.63 \approx H^{(\mathrm{th})}$
16	Proflavine monohydrochloride	$C_{13}H_{11}N_3{\cdot}HCl$	+	$2.97 < Z^{(\mathrm{th})}$	$1.55 < H^{(\mathrm{th})}$
17	Reserpine	$C_{33}H_{40}N_2O_9$	+	$2.81 < Z^{(\mathrm{th})}$	$1.51 < H^{(\mathrm{th})}$
18	Rifampicin	$C_{43}H_{58}N_4O_{12}$	+	$2.75 < Z^{(\mathrm{th})}$	$1.54 < H^{(\mathrm{th})}$
19	Spironolactone	$C_{24}H_{32}O_4S$	+	$2.59 < Z^{(\mathrm{th})}$	$1.37 < H^{(\mathrm{th})}$
20	Doxorubicin	$C_{27}H_{29}NO_{11}$	+	$3.06 < Z^{(\mathrm{th})}$	$1.57 < H^{(\mathrm{th})}$
21	Azacitidine	$C_8H_{12}N_4O_5$	+	$\mathbf{3.24} > Z^{(\mathrm{th})}$	$\mathbf{1.87} > H^{(\mathrm{th})}$
22	Bergaptene	$C_{12}H_7O_4$	+	$\mathbf{3.44} > Z^{(\mathrm{th})}$	$1.45 < H^{(\mathrm{th})}$
23	Procarbazine	$C_{12}H_{19}N_3O$	+	$2.51 < Z^{(\mathrm{th})}$	$1.46 < H^{(\mathrm{th})}$
24	Phenacetin	$C_{10}H_{13}NO_2$	+	$2.69 < Z^{(\mathrm{th})}$	$1.50 < H^{(\mathrm{th})}$
25	Sulfafurazole	$C_{11}H_{13}N_3SO_2$	-	$3.21 > Z^{(\mathrm{th})}$	$1.92 > H^{(\mathrm{th})}$
26	Sulfamethoxazole	$C_{10}H_{11}N_3SO_3$	-	$3.30 > Z^{(\mathrm{th})}$	$1.94 > H^{(\mathrm{th})}$
27	Profilvinum hemisulphate **)	$C_{26}H_{22}N_6SO_4$	-	$3.16 > Z^{(\mathrm{th})}$	$\mathbf{1.75} < H^{(\mathrm{th})}$
28	Chloramphenicol**)	$C_{11}H_{12}Cl_2N_2O_5$	-	$3.44 > Z^{(\mathrm{th})}$	$1.98 > H^{(\mathrm{th})}$

*) There is insufficient data on carcinogenicity of chemical compound [5].**) Carcinogenicity assessment is inadequate [5].

We introduce an additional molecular descriptor, namely, the information function of redundancy. This descriptor will allow us to trace the change in the activity of chemical compounds as molecular structures change. The dimensionless redundancy information function is defined as follows:

$$D = 1 - H / H_{\max} \, . \qquad (58)$$

Here $H_{\max} = \log_2(n)$, n is the number of different atoms in the molecule. Table 15 shows a number of aromatic amines similar in structure to 2-acetylaminofluorene.

Table 14. *Carcinogenic activity, electronic descriptors and information functions for the initial chemical compounds and their metabolites.*

N	The original chemical compound	Carcinogenic activity	Metabolite	Activity
1	Cholesterol $Z = 2.16$, $H = 1.04$*bits*	Data is not adequate [7]	6-β-Peroxy-Δ4-cholesten-3-on $Z = 2.27$, $H = 1.1$*bits*	Very active
2	Tryptophan $Z = 2.89$, $H = 1.60$*bits*	Inactive	3-Oxykinurenine $Z = 2.96$, $H = 1.68$*bits*	Active
			2-Amino-3-hydroxyacetophenol $Z = 2.90$, $H = 1.60$*bits*	Active
			3-Oxyanthranilic acid $Z = 3.22$, $H = 1.72$*bits*	Active
3	Aflatoxin B_1 $Z = 3.31$, $H = 1.47$*bits*	Slightly active	Aflatoxin M_1 $Z = 3.39$, $H = 1.50$*bits*	Active
			2,3- Epoxiflatoxin B_1 $Z = 3.33$, $H = 1.52$*bits*	Active
4	Safrole $Z = 2.82$, $H = 1.35$*bits*	Slightly active	1'-Hydroxisafrole $Z = 2.96$, $H = 1.43$*bits*	Active
			$C_{12}H_{12}O_4$ *) [9] $Z = 3.00$, $H = 1.45$*bits*	Active
5	4-Aminobiphenyl $Z = 2.67$, $H = 1.21$*bits*	Slightly active	N-Oxy-4-aminobiphenyl $Z = 2.92$, $H = 1.53$*bits*	Active
6	2-Naphtylamine	Slightly active	2-Amino-1-naphthol $Z = 2.77$, $H = 1.65$*bits*	Very active
			2-Naphthylhydroxyamine $Z = 2.86$, $H = 1.45$*bits*	Very active

	$Z = 2.70$, $H = 1.23$*bits*		Bis (2-hydroxylamine)-1-naphthyl phosphate $Z = 3.17$, $H = 1.75$*bits*	Very active
			Bis (2-amino-1-naphthyl) phosphate $Z = 3.17$, $H = 1.75$*bits*	Very active
7	1,2-Benzo(a)pyrene $Z = 2.13$, $H = 0.96$*bits*	Carcinogenic Nonmutagenic	7,8-Dihydroxy-9,10-epoxy-7,8,9,10-tetrahydro-1,2-benzopyrene $Z = 2.46$, $H = 1.30$*bits*	Carcinogenic Very active Mutagenic
8	Herbicide AAP $Z = 2.88$, $H = 0.95$*bits*	Carcinogenic	7,8-Dihydroxy-9,10-epoxy-7,8,9,10-tetrahydro-1,2-benzopyrene $Z = 2.82$, $H = 1.42$*bits*	Active
9	4-Acetylaminostilbene $Z = 2.73$, $H = 1.33$*bits*	Slightly active	*N*-Oxy-4-acetylaminostilbene $Z = 2.82$, $H = 1.42$*bits*	Very active
10	4-Aminobiphenyl $Z = 2.67$, $H = 1.21$*bits*	Carcinogenic	3-Oxy-4-aminobiphenyl $Z = 2.80$, $H = 1.59$*bits*	Very active
			N-Oxy-2-acetylaminofluorene $Z = 2.80$, $H = 1.40$*bits*	Very active
			N-Oxy-4-aminobiphenyl $Z = 2.80$, $H = 1.40$*bits*	Very active
			4-Amino-3-hydroxy-diphenylsulfate $Z = 3.24$, $H = 1.79$*bits*	Very active
11	Methylurea $Z = 2.73$, $H = 1.69$*bits*	Inactive	*N*-Nitroso-*N*-methylurea $Z = 3.33$, $H = 1.89$*bits*	Very active
12	*N*-Nitrozodimethyl amine $Z = 2.73$, $H = 1.69$*bits*	Carcinogenic Nonmutagenic	Methyl diazohydroxide $Z = 3.00$, $H = 1.79$*bits*	Carcinogenic Very active
13	Cycasin $Z = 3.03$, $H = 1.72$*bits*	Active	Diazomethane $Z = 3.21$, $H = 1.52$*bits*	Very active
14	4-Amino-4-acetylamino-biphenyl $Z = 2.76$, $H = 1.36$*bits*	Active	*N*-Oxy-4-acetylaminobiphenyl $Z = 2.87$, $H = 1.46$*bits*	Active
15	2-Acetylaminofluorene $Z = 2.80$, $H = 1.35$*bits*	Active	*N*-Oxy-2-acetylaminofluorene $Z = 2.90$, $H = 1.45$*bits*	Very active
			2-Oxy-2-acetylaminofluorene $Z = 2.90$, $H = 1.45$*bits*	Active
16	3- Acetylaminofluorene $Z = 2.80$, $H = 1.35$*bits*	Inactive	2-Oxy-2-acetylaminofluorene $Z = 2.90$, $H = 1.45$*bits*	Active
17	Benz-(1,2)-anthracene $Z = 2.80$, $H = 0.97$*bits*	Active	4'-Oxybenz-(1,2)-anthracene $Z = 2.90$, $H = 1.15$*bits*	Active
18	9,10-Dimethylbenz-(1,2)-anthracene $Z = 2.67$, $H = 0.99$*bits*	Active	4'-Oxy-9,10- dimethylbenz-(1,2)-anthracene $Z = 2.76$, $H = 1.14$*bits*	Active
19	Chrysene $Z = 2.80$, $H = 0.98$*bits*	Active	1'-Oxychrysene $Z = 2.90$, $H = 1.15$*bits*	Active
20	*N*,*N*'-Dimethyl-4-aminoazobenzene $Z = 2.59$, $H = 1.34$*bits*	Active	*N*-Hydroxy-4-monomethylaminoazo-benzene $Z = 2.79$, $H = 1.47$*bits*	Active
21	7,12-Dimethylbenz (a) anthracene $Z = 2.67$, $H = 0.99$*bits*	Very active	7-Oxymethyl-12-methylbenz(a) anthracene $Z = 2.76$, $H = 1.14$*bits*	Inactive, toxic
22	4-Aminostilbene $Z = 2.64$, $H = 1.19$*bits*	Active	*N*-Oxy-4-acetylaminostilbene $Z = 2.82$, $H = 1.42$*bits*	Active
23	Bensidine $Z = 2.69$, $H = 1.32$*bits*	Active	3,3'- Oxybensidine $Z = 2.93$, $H = 1.59$*bits*	Active
			4'-Acetylamino-4-aminobiphenyl $Z = 2.77$, $H = 1.45$*bits*	Active
			4'-Acetylamino-4-amino-3-oxybiphenyl $Z = 2.88$, $H = 1.54$*bits*	Active
			3-Oxybensidine $Z = 2.82$, $H = 1.49$*bits*	Active

*) Presumptive metabolite [23]. **) Metabolite after hydrolysis. ***) It is formed in the body with nitrosation.

Table 15 shows the interrelation of molecular descriptors with the level of carcinogenic activity for related chemical compounds. The increase in the molecular descriptor *D* is accompanied by an increase in the level of carcinogenic activity of the chemical compound. The use of the molecular descriptor *D* establishes a relatively monotonous interrelation. However, the values of the molecular descriptors are less than the threshold values. An increase in the level of carcinogenic activity of chemical compound in a number of related compounds is also accompanied by a tendency to increase the descriptor of *H*. We note a similar trend for the descriptors of the anthracene molecule (H = 0.98*bits*, Z = 2.75) and its methyl derivatives. Anthracene itself is not a carcinogen. However, the 2-methyl derivative of anthracene (H = 0.99*bits*, Z = 2.67) has a carcinogenic activity. The addition of the second methyl group to the anthracene molecule leads to increase the carcinogenic activity. Thus, for example, the carcinogenic activity of the 2,6-dimethyl derivative of anthracene (H = 1.00*bits*, Z = 2.60) is increased in fourfold ([3] Appendix I, Russian edition).

Table 15. *A number of aromatic amines close to 2-acetylaminofluorene*

N	Chemical compounds	Gross formula	Activity	*D*	*Z*	*H,bits*
1	3-Acetylaminodibenzothiophene	$C_{14}H_{13}NOS$	+++	0.35	2.87	1.53
2	2-Acetylaminofluorene	$C_{15}H_{13}NO$	++	0.33	2.80	1.35
3	3-Acetylaminophenanthrene	$C_{16}H_{13}NO$	+	0.33	2.84	1.34
4	2-Acetylaminophenanthrene	$C_{14}H_{13}NO$	+	0.32	2.76	1.36
5	3-Acetylaminodibenzothiophene-5-oxide	$C_{14}H_{13}NO_2S$	+	0.30	2.97	1.62
6	3-Acetylaminodibenzfuran	$C_{14}H_{13}NO_2$	+	0.27	2.87	1.46
7	2-Aminofluorene	$C_{13}H_{10}N$	+	0.25	2.79	1.20
8	2-Aminoanthracene	$C_{14}H_{11}N$	+	0.25	2.77	1.19

Table 16.

N	Chemical compound	Gross formula	Activity	*D*	*Z*	*H,bits*
1	1-Ethyloxy-3,4-epoxycyclohexane	$C_8H_{12}O_2$	3	0.17	2.55	1.32
2	1,2-Epoxybutane	C_4H_6O	1	0.17	2.55	1.32
3	d1-Diepoxybutane	$C_4H_6O_2$	0	0.08	2.83	1.46
4	Mesodiepoxy butane	$C_4H_6O_2$	0	0.08	2.83	1.46
5	Styrene oxide	C_8H_8O	0	0.20	2.70	1.26
6	9,10-Epoxystearic acid	$C_{18}H_{33}O_3$	0	0.25	2.27	1.19
7	6,7,9,10-Epoxystearic acid	$C_{18}H_{32}O_4$	0	0.21	2.37	1.25
8	Hexoepoxysvalol	$C_{30}H_{54}O_4$	0	0.27	2.25	1.10

In the book [3] information is given on carcinogenic activity (on a five-point scale) 1-vinyl-1-hydroperoxide cyclohexane-3 (activity 5 points) and 1-vinylcyclohexane-3 (activity 1 point). For these chemical compounds, the following values of descriptors were obtained: Z = 2.62, H = 1.34*bits*, D = 0.15 and Z = 2.20, H = 0.97*bits*, D = 0.03 respectively. Thus, the classification rule is satisfied for substances that are close in molecular structure. The analysis has shown that descriptor *D* is more reliable than descriptors *Z* and *H* for determining the variability of carcinogenic activity of closely related chemical compounds. Table 16 shows the activity of chemical compounds on a five-point scale [3].

Chemical compounds under the number from 6 up to 8, have high values of descriptor *D*. However they did not show carcinogenic activity. As discussed above, this is possible due to the large number of hydrocarbon chains. In this case, their number is $m \geq 17$.

An interesting situation is created. The range for bioactivity of chemical compounds that belong to different classes is determined by the inequality $H \leq H^{(\mathrm{th})}$ (or $Z \leq Z^{(\mathrm{th})}$). That is, the descriptor is less than the threshold value. At the same time, the level of carcinogenic activity increases with increasing of *H* (or *Z*, *D*) when the molecular structure of the related molecules is varied. Apparently, such interrelation of molecular descriptors with the level of carcinogenic activity of chemicals is most effective for closely related agents. It should be pointed out that the classification rules are not sensitive to the difference between the isomer molecules, and also between iso-atomic molecules (or with substituent that have the same number of valence electrons, for example, chlorine, fluorine, etc.). The handbook [5] states that the available knowledge of the carcinogenicity of certain chemical compounds is not sufficient (or contradictory) and requires additional research. Classification rules allow probabilistically indication of chemical agents, which first of all require the attention of researchers.

Analogues of kinetin possessing anticarcinogenic properties and selectively acting on malignant cells have also been studied [25]. The following chemical compounds have been analyzed: furfuryl-6-aminopurine (Z = 3.20), $\mathrm{N}^{\mathrm{NH}_2}$-puryl-6-tryptamine ($Z$ = 3.20), $\mathrm{N}^{\mathrm{NH}_2}$-puryl-6-tyramine ($Z$ = 3.00), $\mathrm{N}^{\mathrm{NH}_2}$-puryl-6-histamine ($Z$ = 3.15), N^{ε}-puryl-6-lysine (Z = 2.91), N, N'-dipuryl-6-ethylenediamine (Z = 3.24). It is important to note that all these agents are characterized by a rather high *Z* descriptor value, which is significantly higher than the average value of the descriptor for active carcinogens (8) and (21).

Using the data of [8], it can be noted that descriptors of sulfur-containing radioprotective agents and descriptors of carcinogenic chemical compounds overlap. Indeed, sulfur-containing chemical compounds against ionizing radiation have the descriptor of $Z^{(\mathrm{th})}_{\mathrm{prot.}} = 2.83$ is less than threshold descriptor of carcinogenic agents (Table 12). A similar situation occurs for information function. It is important to note that the electronic and information descriptors of molecules are determined from different principles, but they lead to the same consequences. Thus, it is possible that radioprotectors can be carcinogenic [7,9].

5 Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons

Polycyclic aromatic hydrocarbons have several features which distinguish them from many of the more discovered carcinogens. They act the site of application. The effective dose is minute, of the order of micrograms, and they have been found to induce tumours in almost every tissue and animal species in which they have been tested [4]. Polycyclic aromatic hydrocarbons are characterized by the presence in the chemical structure of two or more condensed benzene rings.

Polycyclic aromatic hydrocarbons are involved in intermolecular interactions in the body. This is confirmed by the discovery of hydrocarbon-cancer tissue complexes. Detailed studies using chemical compounds labeled with radioactive carbon showed that the formation of such complexes plays a major role in the development of tumors. There is a parallelism between the incidence of tumors and the size of the complex [29]. Brookes and Lawley [30] has been established with high accuracy that polycyclic aromatic compounds are attached to DNA, and carcinogenic activity is proportional to the number of molecules of hydrocarbons involved in complex formation. For example, it is known that polycyclic hydrocarbons are capable to install into DNA. Moreover, the carcinogenic effect is proportional to the amount of bound hydrocarbon.

We can get some additional knowledge about the physical interpretation of the molecular descriptor *Z*. For this purpose, we will analyze the additive components of the total energy of the pair intermolecular interaction of polycyclic hydrocarbons with model molecular systems.

The fact that osmium tetroxide is attached at position 9,10 of the phenanthrene molecule (*K*-bond [29]) served as the basis for allowing the possibility of binding cyclic hydrocarbons to the active center of the organism precisely by this region of the molecule. The book [31] shows the model studies of the additive components of the intermolecular interactions of polycyclic aromatic hydrocarbons with tetramethyl-uric acid. The table 17 demonstrates the following additive contributions to the pair interaction energy: E_{elect} is the electrostatic energy, E_{pol} is the polarization interaction energy, E_{disp} is the dispersion energy, and E_{rep} is the energy of the short-range exchange repulsion.

***Table 17**. Additive components of the pair interaction tnergy of tetramethyl-uric acid with polycyclic hydrocarbons.*

N	Chemical compound	Brutto formula	Contributions to the energy of pair interaction, kcal/mol [32]					*Z**)	*H**), *bits*
			E_{elect}	E_{pol}	E_{disp}	E_{rep}	E_{total}		
1	Dibenz-1,2,3,4-pyrene	$C_{24}H_{14}$	-1.93	-1.11	-10.37	4.95	-8.46	2.90	0.95
2	Anthanthren	$C_{22}H_{12}$	-1.34	-0.93	-8.72	4.79	-6.20	2.94	0.94
3	Perylene	$C_{20}H_{12}$	-1.35	-0.86	-7.80	4.25	-5.76	2.88	0.96
4	Benz-1,2-pyrene	$C_{20}H_{12}$	-1.77	-0.96	-9.23	4.51	-7.45	2.88	0.96
5	Benz-3,4-pyrene	$C_{20}H_{12}$	-1.38	-0.86	-8.64	4.48	-6.40	2.88	0.96
6	Pyrene	$C_{16}H_{10}$	-1.46	-0.78	-7.76	4.12	-5.88	2.85	0.96
7	Dibenz-1,2,5,6-anthracene	$C_{22}H_{14}$	-1.09	-0.79	-8.23	3.88	-6.23	2.83	0.96
8	Chrysene	$C_{18}H_{12}$	-1.18	-0.80	-7.92	4.15	-5.75	2.80	0.97
9	Benz-1,2- anthracene	$C_{18}H_{12}$	-1.18	-0.80	-7.92	4.15	-5.75	2.80	0.97
10	Phenantrene	$C_{14}H_{10}$	-1.57	-0.65	-7.13	3.77	-5.58	2.75	0.98
11	Anthracene	$C_{14}H_{10}$	-1.27	-0.79	-6.84	3.67	-5.22	2.75	0.98

*) The elements of the sets *Z* and *H* satisfy the normal distribution. Statistics of Wilk-Shapiro: $W_Z = 0.94 > W_H = 0.92 > W_{0.95}^{(\text{cr})}(f = 11) = 0.85$.

Statistical analysis has shown that there is a reliable relationship between the descriptor *Z* with the contribution to the energy of short-range repulsion, the polarization contribution, and the dispersion interaction. We have obtained the following correlation equation for the relationship of the descriptor *Z* with the energy E_{rep}:

$$E_{\text{rep}}(Z) = a_0 + a_1 Z, \quad N = 11, \quad R = 0.88,$$
$$a_0 = -12.28 \pm 3{,}00, \quad a_1 = 5.82 \pm 1.06,$$
$$F = 30.3 > F_{0.05}^{(\text{cr})}(f_1 = 1; f_2 = 9) = 5.1;$$
$$W_{E\text{rep}} = 0.96 > W_{0.95}^{(\text{cr})}(f = 11) = 0.85. \qquad (59)$$

Here *f* is the number of freedom degrees. The statistical significance of the correlation coefficient is confirmed by the inequality: $t = 0.5\cdot\ln[(1+|R|)/(1-|R|)]\cdot(N - 3)^{1/2} = 3.89 > t_{0.95}^{(\text{cr})}(f = N - 2) = 2.26$. The relationship between the contributions E_{pol} and E_{disp} with the molecular descriptor *Z* has the following statistics:

$$E_{\text{pol}}(Z) = a_0 + a_1 Z, \quad N = 11, \quad R = -\,0.77,$$
$$a_0 = 3.48 \pm 1{,}21, \quad a_1 = -1.52 \pm 0.43,$$
$$F = 12.7 > F_{0.05}^{(\text{cr})}(f_1 = 1; f_2 = 9) = 5.1; \quad W_{E\text{pol}} = 0.93 > W_{0.95}^{(\text{cr})}(f = 11) = 0.85, \qquad (60)$$

the statistical significance of the correlation coefficient *R*: $t = 2.85 > t_{0.95}^{(\text{cr})}(f = 9) = 2.26$.

$$E_{\text{disp}}(Z) = a_0 + a_1 Z, \quad N = 11, \quad R = -\,0.74, \quad a_0 = 25.3 \pm 10{,}3, \quad a_1 = -\,11.8 \pm 3.62,$$
$$F = 10.7 > F_{0.05}^{(\text{cr})}(f_1 = 1; f_2 = 9) = 5.1; \quad W_{E\text{disp}} = 0.95 > W_{0.95}^{(\text{cr})}(f = 11) = 0.85, \qquad (61)$$

the statistical significance of the correlation coefficient *R*: $t = 2.66 > t_{0.95}^{(\text{cr})}(f = 9) = 2.26$.

Apparently, the molecular descriptor *Z* is allowed intermolecular interaction to be identified and described in accordance with the specification short-range components of intermolecular interactions that determine the formation of complexes of organic substances with molecular systems that are simulating biosystems. It is important to emphasize that the relationship exists for the dispersion interaction, which gives the dominant contribution to the total interaction energy. As the analysis showed, a linear relationship is absent for long-range electrostatic interactions. However, it is necessary to note the following important fact. The detection of the interrelations (59) - (61) is not unexpected. Previously, we found similar relationships for other classes of chemical compounds. For example, there is a close interrelation between *Z* descriptor (and also *H* descriptor) and the absolute value of the dispersion interaction energy for halogenated compounds [9]:

$$E_{\mathrm{disp}}(Z) = a_0 + a_1 Z, \quad N = 12, \quad R = 0.96,$$

$$a_0 = -1.10 \pm 0{,}95, \quad a_1 = 2.52 \pm 0.24,$$

$$F = 108.50 >> F^{(\mathrm{cr})}_{0.05}(f_1 = 1; f_2 = 10) = 5.00\ ;$$

$$W_{\mathrm{Edisp}} = 0.95 > W^{(\mathrm{cr})}_{0.95}(f = 12) = 0.86, \qquad (62)$$

the statistical significance of the correlation coefficient: $t = 5.83 > t^{(\mathrm{cr})}_{0.95}(f = 10) = 2.23$.

Statistical analysis showed a close relationship between the molecular descriptors *Z* and *H*. However, polycyclic hydrocarbons have peculiarity in comparison with other chemical compounds of different classes (see Sect. 4). For polycyclic aromatic hydrocarbons, the relationship has a statistically significant opposite direction:

$$H(Z) = a_0 + a_1 \cdot Z, \quad N = 11, \quad R = -0.96,$$

$$a_0 = 1.49 \pm 0.05, \quad a_1 = -0.186 \pm 0.02,$$

$$RMSE = 0.0035, \quad t(a_0) = 29.6 > |t(a_1)| = 10.48 > t^{(\mathrm{cr})}_{0.05}(f = 9) = 2.26,$$

$$F = 109.98 > F^{(\mathrm{cr})}_{0.05}(f_1 = 1, f = 9) = 5.1, \qquad (63)$$

the statistical significance of the correlation coefficient: $t = 5.83 > t^{(\mathrm{cr})}_{0.95}(f = 10) = 2.23$.

Statistical analysis confirms the opposite trend in the interrelationship between the descriptors, even if a wider range of polycyclic hydrocarbons are used (*N* = 41 [29], *R* = -0.996, *RMSE* = 0.0012). Set of elements *Z* and *H* satisfy the normal distribution. An increase the sample size leads to the appearance of a weak nonlinearity (Fig. 8A). In this case, the *RMSE* decreases to the value 0.0005. That is, the interrelationship is approaching to a functional dependence. The nonlinearity of the interrelationship occurs for most classes of chemical compounds [9].

It follows from the statistics (57) - (61) that the molecular descriptors *Z* and *H* are statistically significantly informative in order to characterize the pair intermolecular interaction. We emphasize that the descriptors have been derived of various principles. We will use these molecular descriptors to evaluate the carcinogenicity of polycyclic hydrocarbons.

The most active carcinogens are molecules that have the greatest value of the molecular descriptor *Z*. For example, dibenz-3,4,9,10-pyrene and dibenz-3,4,8,9-pyrene, for which descriptor *Z* is equal to 2.895. Taking into account the Table 17 and the relationships (57) - (60), it can be assumed that molecules with a high value of *Z* are preferred in paired dispersion interactions. Intermolecular interactions can only be a prerequisite for the appearance of complexes. Strong covalent chemical interactions might arise at the final stage of complex formation. This situation was indeed observed in the complexation of dibenz-1,2,5,6-anthracene with the cellular receptor [29].

Here we will analyze the experimental data that were used by Pullman A. and Pullman B. [29, 33] for interpretation of carcinogenic activity of polycyclic aromatic hydrocarbons. Pullmans [33] highlight the following group of polycyclic hydrocarbons, which contains from four to five condensed benzene rings: A very strong carcinogen – benz-3,4-pyrene (++++); moderately carcinogenic – dibenz-1,2,5,6-anthracene (++) and weak carcinogens – dibenz-1,2,3,4-phenanthrene (+), dibenz-1,2,7,8-anthracene (+); inactive – triphenylene (–) [5].

The sequence of values molecular descriptor *Z* will be as follows: 2.875; for the three penultimate compounds – 2.830; for the last triphenylene – 2.800. There is a parallelism between the carcinogenic activity of polycyclic aromatic hydrocarbons and the values of their molecular descriptor *Z*.

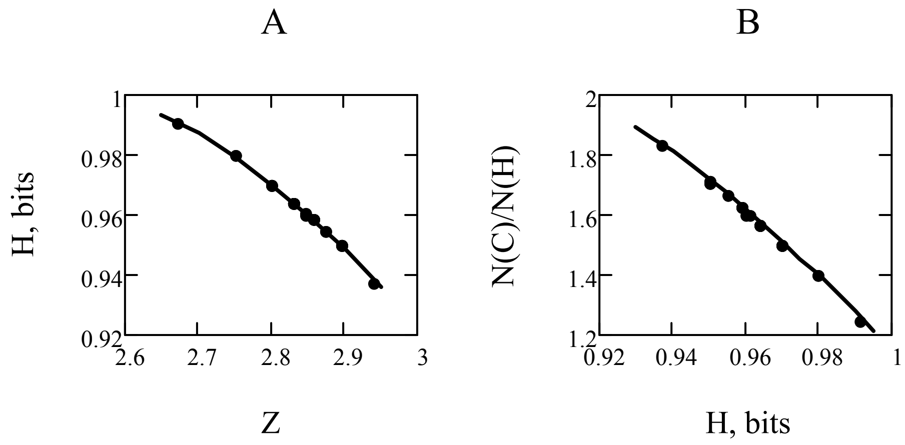

Fig.8. The polycyclic aromatic hydrocarbons. (A) *Interrelation of molecular descriptors H and Z.* (B) *Interrelation of molecular descriptors of* $\gamma = N(\mathrm{C})/N(\mathrm{H})$ *and H.*

Pullman proposed [29] distinguish three groups of carcinogenic substances according to their relative activity. The first group (inactive or weakly active) is benz-3,4-phenanthrene (Z = 2.80, H = 0.97*bits*), dibenz-1,2,5,6-phenanthrene (Z = 2.83, H = 0.96*bits*), dibenz-1,2,3,4-phenanthrene (Z = 2.83, H = 0.96*bits*); the second group (active) is dibenz-1,2,5,6-anthracene (Z = 2.83, H = 0.96*bits*), dibenz-1,2,7,8-anthracene (Z = 2.83, H = 0.96*bits*); the third group (the most active carcinogens) is benz-3,4-pyrene (Z = 2.875, H = 0.955*bits*), dibenz-1,2,3,4-pyrene (Z = 2.895, H = 0.95*bits*), dibenz-3,4,8,9-pyrene Z = 2.895, H = 0.95*bits*), dibenz-3,4,9,10-pyrene (Z = 2.895, H = 0.95*bits*). Thus, in this case we can also make a comparative assessment of the carcinogenicity for each cyclic hydrocarbon group using the molecular descriptor of Z (as well as the descriptor H; Fig. 8A). According to the level of carcinogenicity, three groups of compounds maintain their consistency. The greater the value of the descriptor Z, the more likely the intermolecular interaction is stronger.

Thus, the use of molecular descriptors Z and H does not contradict the criteria of carcinogenicity suggested by Pullmans, which were called the K and L regions of the molecules [33]. For example, dibenz-1,2,3,4-pyrene, dibenz-3,4,8,9-pyrene and dibenz-3,4,9,10-pyrene do not have the L region, but they have a suitable index of K region and therefore they are carcinogens. The molecular descriptors Z = 2.895 and H = 0.95*bits* also indicate the carcinogenicity of these compounds. Descriptors not just indicate, but the descriptors correctly determine the sequence in the relative activity of this series of compounds. Pullman also pointed out [29] that the criteria have exceptions for polycyclic hydrocarbons. First, benz-3,4-phenanthrene, dibenz-1,2,5,6-phenanthrene and dibenzo-1,2,3,4-phenanthren do not subject to the criteria of Pullmans [33]. Perhaps this is due to the inaccuracy of the established threshold values for the criteria of the L and K regions. As shown above, the use as a quantitative measure of the molecular descriptor Z = 2.80-2.83 does not deny the weak carcinogenicity of these compounds. This is also confirmed by the test for the promotive activity of mice for benz-3,4-phenanthrene. The test was positive [5]. Secondly, according to Pullmans, anthanthrene should have carcinogenic activity. The value of descriptor Z = 2.94 to point to this. According to modern data [5], anthanthrene is indeed of carcinogenic activity. The strong carcinogen the dibenz-3,4,6,7-pyrene does not have K region. At the same time, the molecular descriptor for this polycyclic hydrocarbon is very high (Z = 2.895) and hence this substance must be a carcinogen. This result does not contradict observations.

Thus, two different models based on different principles do not lead to contradictory results. At the same time, it is not necessary to perform complex and labor-intensive quantum mechanical calculations to determine the descriptors Z and H. It is enough to know only the gross formula of a polycyclic hydrocarbon. To determine the quantitative characteristics of the K and L regions, knowledge of the various quantum chemical parameters of the molecules is required.

We note further that the most studied addition reactions occurring in the *L* region of the molecule are (according to Pullmans [33]): *a*) fixation of maleic anhydride and *b*) photochemical oxidation. The descriptor *Z* (or the descriptor *H*) enable tracking the facility of fixing maleic anhydride. Pullmans [33] indicated that the facility of fixation is consistently decreasing in the following series of compounds:

Anthracene (Z = 2.75) > benz-1,2-anthracene (Z = 2.80) > dibenz-1,2,5,6-anthracene (Z = 2.83) > benz-3,4-pyrene (Z = 2.875). (64)

Obviously, this sequence of fixations correlates with the value of descriptor *Z*. This gradation of the ease of fixation is confirmed by experience.

Photochemical oxidation very easily occurs in non-carcinogenic anthracene molecules (Z = 2.75) and naphthacene (Z = 2.80). Photooxidation becomes the more difficult with an increase of the number of side rings. Dibenz-1,2,5,6-anthracene molecule (Z = 2.83) weakly react, and benz-3,4-pyrene molecule (Z = 2.875) is completely inactive in the photooxidation reaction. Methyl substitution of the hydrocarbon increases the reactivity of the chemical compound. For example, 9,10-dimethylbenzanthracene (Z = 2.67, H = 0.99*bits*) is much more active than the initial hydrocarbon benzanthracene (Z = 2.80, H = 0.97*bits*), in reactions with maleic anhydride and reaction of photooxidation. These experimental facts also correlate with the variation of the molecular descriptor *Z* (or *H*). Indeed, benz-1,2-pyrene molecule (Z = 2.875) possesses a high carcinogenic activity [34], as well as the following polycyclic hydrocarbons (in order of increasing activity): dibenz-1,2,3,4-pyrene (Z = 2.895), dibenz-3,4,8,9-pyrene (Z = 2.895), dibenz-3,4,9,10-pyrene (Z = 2.895), and also ovalene (Z = 3.087). The ovalene has pronounced carcinogenic, mutagenic and teratogenic properties. As demonstrated in the article [35], the dibenz-3,4,6,7-pyrene is very active (no less active than the benz-3,4-pyrene or the dibenz-3,4,8,9-pyrene). The dibenz-3,4,6,7-pyrene has the descriptor Z = 2.895. At the same time it is important to note that dibenz-3,4,6,7-pyrene does not have a sufficiently active *K* region.

The addition reaction of osmium tetraoxide was investigated in detail to study the reactivity of the *K* region. The following sequence of polycyclic compounds has been drawn up [29] in order of increasing their reactivity with respect to this agent:

Phenanthrene (Z = 2.75; I =7.75 eV; (–)) < benz-1,2-anthracene (Z = 2.80; I =7.50 ev; (+/–)) < dibenz-1,2,5,6-anthracene (Z = 2.83; I = 7.09-7.80 eV; (++) < benz-3,4-pyren (Z = 2.875; I = 7.19 eV; (++++)). (65)

It is not difficult to see that in this case there is also a parallelism of the reactivity with the magnitude of the descriptor *Z*. The relative carcinogenic activity of molecules increases in the same sequence. Here we give the values of the first ionization potentials (*I*) of the molecules [36]. The increase in descriptor *Z* is attended by a decrease in the ionization potential of the molecule for a given series (65) of polycyclic hydrocarbons. There is the interrelation (65) between the magnitude of descriptor *Z* and the magnitude of the first ionization potential of polycyclic aromatic hydrocarbons. It can be assumed that the lower the ionization energy of the molecule, the better the donor properties of the molecule. The increase in the descriptor *Z* often coincides with the dynamics of the increase in the donor ability of molecules. Obviously, for the sequence of observations (65), the descriptor *Z* and the associated signs of *H* and *I* are reliably informative. The variation of a number of chemical compounds (65) indicates the existence of a relationship between the reactivity of molecules and the magnitude of molecular features. The descriptor value Z = 2.80-2.83 we can approximately taken as the boundary value in accordance with the sequence (65). This value approximately separates carcinogens from inactive polycyclic hydrocarbons. In this connection, we note that two dibenzphenanthrenes XIII and XIV (the numbering of Pullmans [33]) does not contain the reaction *K* region. Consequently, they cannot be attributed to the carcinogens. However, the molecular descriptor (Z = 2.83) indicates that these chemical compounds should have a weak carcinogenic activity. This is confirmed by observations [5]. Molecular descriptors are also useful for quantitatively assessing the reactivity of the *L* region of molecules. The order of decreasing reactivity towards $Pb(OAc)_4$ is :

Anthracene ($Z = 2.75$, $H = 0.98$*bits*; (–)) > benz-1,2-anthracene ($Z = 2.80$, $H = 0.97$*bits* (–/+)) > dibenz-1,2,5,6-anthracene ($Z = 2.83$, $H = 0.96$*bits* (+)) = 0. (66)

Molecular descriptors indicate their parallelism with the reactivity of the *L* region and in this case.

Another molecular descriptor can be proposed for the purpose of rapid assessment of the carcinogenic activity of polycyclic aromatic hydrocarbons. This descriptor is defined by the ratio of the number of carbon atoms to the number of hydrogen atoms: $\gamma = N(\mathrm{C})/N(\mathrm{H})$. As shown by statistical analysis, the descriptor γ is closely related to the molecular descriptors *Z* and *H* (Fig. 8B). There is a statistically significant close relationship between the descriptors for the forty-one polycyclic aromatic hydrocarbons represented by Pullmans [33]. For example, the relationship between the information function *H* and the descriptor γ can be approximated by a linear form:

$$\gamma(H) = a_0 + a_1 \cdot H, \qquad R = -0.997, \qquad N = 41,$$
$$RMSE = 0.0079, \qquad a_0 = 11.86 \pm 0.13,$$
$$a_1 = -10.67 \pm 0.14, \qquad |t(a_1)| = 78.6 \quad >>$$
$$t_{0.05}^{(\mathrm{cr})}(f = 34) = 2.03. \qquad (67)$$

The increase in the sample size leads to the appearance of a statistically significant weak nonlinearity for the relationship between the signs of γ and *H* (Fig. 8B). A similar close relationship exists between the descriptors *Z* and γ. It is important to note that the molecular descriptors that were obtained on the basis of different baseline principles are closely interrelated. Their application leads to identical results for polycyclic aromatic hydrocarbons.

The above experimental facts can also be interpreted using the molecular descriptor γ, which is in no way connected with the *K* and *L* regions of the polycyclic molecule. For example, the carcinogenicity of chemical compounds increases in the following sequence:

Phenanthrene ($\gamma = 1.40$) < benz-1,2-anthracene ($\gamma = 1.50$) < dibenz-1,2,5,6-anthracene ($\gamma = 1.57$) < benz-3,4-pyrene ($\gamma = 1.67$) < dibenz-3,4,6,7-pyrene ($\gamma = 1.71$) < valen ($\gamma = 2.29$). (68)

At the same time, the molecular descriptor γ is obviously related to the amount of carcinogenic activity: an increase in the carcinogenicity of the substance is accompanied by an increase in the value of γ (as well as the sign of *Z*).

It is of fundamental importance that we do not use the idea of special local regions of molecules (that is, the *K* and *L* regions). On the contrary, we use molecular descriptors that characterize the properties of molecules in general. These descriptors make it possible to successfully interpret the observations that underlie the Pullmans model. Obviously, within the framework of our alternative model, it is possible to remove the following important remark made by Ladik J. : "... why just hydrocarbons that have joined the *K* region are carcinogenic, while the hydrocarbons that have joined in position *L* are inactive" [37]. From point of view of Pullmans model it is extremely puzzling [33] to observe that, whereas benz-3,4-pyrene is cancerogen, the $2'$ and $3'$ methyl derivatives are total inactive. However, the descriptors γ and *Z* for benz-3,4-pyrene are 1.67 and 2.88, and for methyl derivatives are 1.50 and 2.80, respectively. That is, the magnitude of the descriptors decreases. A similar decrease in descriptors occurs for inactive 6,7-dimethyl-3,4-benzphenanthrene ($\gamma = 1.25$, $Z = 2.67$) and 4,5-dimethylchrysene ($\gamma = 1.13$, $Z = 2.59$).

Some polycyclic hydrocarbons may have a high *Z* descriptor value, but they do not have carcinogenic activity. This may be due to the limiting factors of the manifestation of biological activity [8,9]. For example, the molecular descriptors *Z*, *H* and γ do not distinguish between isomer molecules. The benz-3,4-pyrene ($Z = 2.875$) is very active, and benz(e)pyrene ($Z = 2.875$) is weakly active. However, the affinity energy (*A*) for these molecules is markedly different. The affinity energy for benz(e)pyrene is equal to 0.07eV - 0.40 eV (comparison of different data). At the same time, the affinity energy for benz-3,4-pyrene, according to various sources, is in the range 0.70eV - 0.77 eV [36].

Pullman singled out [29], the following six polycyclic hydrocarbons, the carcinogenic whose activity significantly changes at little structural changes:

Benz-3,4-pyrene (Z = 2.875; A = 0.77 eV; ++++); dibenz-1,2,5,6-anthracene (Z = 2.833; A = 0.68 eV; ++); dibenz-1,2,5,6-phenanthrene (Z = 2.833; A = 0.31 eV; +); dibenz-1,2,3,4-phenanthrene (Z = 2.833; A = 0.40-0.43 eV; +); dibenz-1,2,7,8-anthracene (Z = 2.833; A = 0.23-0.69 eV; –/+); benz-1,2-pyrene (Z = 2.875; A = 0.49-0.61eV; –); dibenz-1,2,3,4-anthracene (Z = 2.833; A = 0.22-0.54 eV; –). (69)

The experimental values of the affinity energy have been taken from the reference book [36]. To this series of compounds, the following polycyclic hydrocarbon can be added: dibenz-3,4,6,7-pyren (Z = 2.895, A = 0.82 eV; ++++). The lower bounds of the affinity energy indicate their parallelism with the carcinogenic activity of the molecules. It is possible that activity pyrene (Z = 2.846; +/–) is limited by the value of the affinity energy of the molecule ($A \geq 0.09$ eV).

The affinity energy gives a measure of the oxidative ability of a molecule. The transfer of an electron is usually accompanied by endoenergetic or exoenergetic processes that can affect the state of the objects with which the polycyclic hydrocarbon interacts. In the first case, for the transfer of an electron, energy is required to be at least two times greater than for benzo-3,4-pyrene. Since the carcinogen molecule interacts with molecular structures, in particular RNA and DNA [38], the locally released energy during electron transfer can be directed to the destruction of these ordered molecular structures. The concentration of hydrocarbons [30] significantly enhances the effect of the released energy in the interaction region.

The limiting effect on the carcinogenicity of a substance can provide the addition of an alkyl group to a polycyclic hydrocarbon [39]. First, this leads to a decrease in the molecular descriptors Z and γ. Secondly, an increase in the length of the alkyl group is accompanied by a change in the hydrophobic properties of the substance. According to Pullmans, steric hindrances can also be limiting factors. Shape, size and steric factors of the compounds are of importance for cancerogenic activity, but in the last resort this depends on the recipient tissue. It depends on the species, the particular strain of the test animal, its age, sex, its nutritional and hormonal state, and the phase of mitotic activity of a particular cell [4].

In accordance with the model of Pullmans, the chemical compounds LVIII and LIX (the numbering of Pullmans [33]) have a large spatial size and should be carcinogens. The same conclusion follows from calculations of molecular descriptors, which are equal to Z = 2.96 and 3.00, γ = 1.88 and 2.00, respectively. However, it was not possible to find the results of the experimental verification of these substances.

It is also useful to emphasize the high information content of the Z descriptor. For example, Table 18 shows the experimental values of the energy ΔE of the most intense electronic transition [40] for several aromatic hydrocarbons.

Table 18. The energies of the most intense electronic transition ΔE and the molecular descriptors of polyacenes

N	Chemical compound	ΔE, eV [40]	Z	H, *bits*	γ
1	Benzene	4.70	2.50	1.00	1.00
2	Naphthalene	3.94	2.67	0.99	1.25
3	Anthracene	3.27	2.75	0.98	1.40
4	Naphthacene	2.63	2.80	0.97	1.50
5	Pentacene	2.16	2.83	0.96	1.57

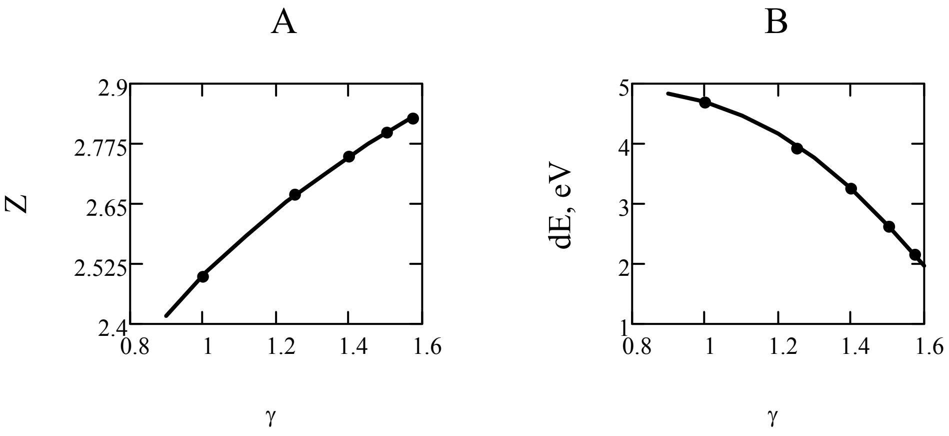

Fig. 9. (A) *The relationship between descriptors Z and γ. Statistics of the interrelation*: $Z(\gamma) = B + A\cdot\exp(-C\cdot\gamma)$, $B = 3.23 \pm 0.03$, $A = -2.09 \pm 0.04$, $C = 1.05 \pm 0.06$, $RMSE = 0.0012$. (B) *The relationship between descriptors ΔE and γ. Statistics of the interrelation*: $\Delta E(\gamma) = A\cdot\gamma^2 + B\cdot\gamma + C$, $B = 7.85 \pm 0.99$, $A = -4.78 \pm 0.39$, $C = 1.63 \pm 0.63$, $RMSE = 0.028$. *Here we use the following notation*: d$E \equiv \Delta E$.

Since the descriptor γ is closely related to the descriptor Z (Fig. 9A), the energy of the electronic transition ΔE also obviously correlates (Fig.9B) with the value of the molecular descriptor Z.

We can also note the high information content of the Z descriptor. This descriptor correlates with electron donor activity not only for polycyclic hydrocarbons, but also for other classes of compounds. The quadrupole splitting of the gamma-resonance line of tin in the Mössbauer spectrum of tin dibutyl chloride was studied [41]. The following order was obtained for a decrease in the donor properties of solvents towards to the organometallic compound $(C_4H_9)_2SnCl$ for seven solvents: CH_3SOCH_3 (Z = 2.60) > $HCON(CH_3)_2$ (Z = 2.50) > $[(CH_3)_2N]_3PO$ (Z = 2.35) > $CH_3OCH_2CH_2OCH_3$ (Z = 2.57) > tetrahydrofuran (Z = 2.30) > $H_5C_2OCH_2CH_2OC_2H_5$ (Z = 2.27) > $(C_2H_5)_2O$ (Z = 2.14). The donor ability reaches a minimum when the descriptor value Z for $(C_2H_5)_2O$ practically coincides with the value Z = 2.17 for the organometallic compound $(C_4H_9)_2SnCl$.

6. Conclusion

Classification rules allow us to identify the relationship between the biological response and the molecular structure of a chemical. The rules can be practically useful in the preliminary projection of the carcinogenic activity of new chemical compounds. It is important to emphasize that simple calculations of molecular descriptors require only knowledge of the chemical structural formula of a molecule. This approach makes it possible to considerably facilitate the search for new carcinogens, and also to draw the attention of researchers to poorly studied chemical compounds. However, it should be noted that the determination of the molecule descriptor is not sensitive to the study of iso-electronic molecular systems, and also when comparing the bioactivity of isomer molecules.

The ability of the Z descriptor to separate potentially carcinogenic compounds from non-carcinogenic substances is apparently not accidental but reflected the action of the real electrostatic molecular potential. This potential is generated by a set of charged particles (nuclei and electrons). The magnitude of the potential varies from molecule to molecule. It is very difficult to use the total molecular potential, which includes the Coulomb potential of nuclei and electrons. However, from analytic formulas for the pseudopotential [6] it follows that the general characteristic feature for the pseudopotential is the factor Z.

The change in the carcinogenic properties of molecules with a change in the molecular potential does not contradict the known notions of the mechanism of chemical carcinogenesis. The known data [26], as well as quantum-chemical calculations [27], allow us to conclude that, at least in a series of close congener chemical compounds, their carcinogenicity is directly dependent on the ability to electrophylic attack.

It is suggested [26] that carcinogens induce DNA single-strand breaks. In this case, purine bases (especially guanine) are the target for them. In this regard, it should be noted that the molecular descriptor Z for all purine bases is larger than the threshold values (Table 6). The descriptors maximum values are achieved for guanine (Z = 3.50, H = 1.82*bits*). For other purine bases the value of the molecular descriptor Z is also higher than the threshold values: adenine (Z = 3.33), guanine (Z = 3.50), thymine (Z = 3.36), cytosine (Z = 3.23), uracil (Z = 3.5).

We used descriptors Z and H to study the radioprotective properties of sulfur-containing chemical compounds in the articles [7,8]. This opens the possibility to compare the preferred areas characterizing the radio-protective effect of sulfur-containing chemicals and their carcinogenic properties. Effective radioprotectors (dose $\leq$ 1 mM/kg; survival > 50%) are preferably in the following areas Z(protect.) $\leq$ 2.80 и H(protect.) $\leq$ 1.80*bits* [8]. At the same time, sulfur-containing carcinogens are likely to be in the following areas of Z(cancer) $\leq$ 3.15 and H(cancer) $\leq$ 1.87*bits*. Obviously, these areas for the descriptor of Z (or H) are overlap. Therefore, it is possible that effective radioprotectors may have carcinogenic properties. As well known from the literature there is the chemical compounds that have been tested for both antiradiation protection [28] and their carcinogenic activity [5]. For example, thiourea (Z = 3.00, H = 1.75*bits*) has radioprotective properties and at the same time is a carcinogen.

References

[1] Pliss G.B. In: Proceedings of the VIII International of Anti-Cancer Congress. 1963. vol. 2. Moscow. pp. 312-314.

[2] Haddow A. In: Proceedings of the VIII International of Anti-Cancer Congress. 1963. vol. 2. Moscow. pp.267-272

[3] Badger G.M. The Chemical Basis of Carcinogenic Activity. Charles C. Thomas – Publisher: Springfield-Illinois-USA. 1962.

[4] Schoental R. In: Clar E. “Polycyclic Hydrocarbons”. V.1. Ch.18. London-N.Y., Berlin-Göttingen-Heidelberg. 1964.

[5] Carcinogenic substances. Handbook. Ed. Turusov V.S. (IARC Monographs on the Evaluation the Carcinonogenic Risk of Chemicals to Humans). Moscow. 1987.

[6] Mukhomorov V.K. “Simulation of Carcinogenic Activity of Polycyclic Hydrocarbons”. In: Proceedings of the IV International Conference "Actual Problems of the Development of World Science". Part 1. pp. 86-97. Kiev. Ukraine. February 28, 2018. (in Russian).

[7] Mukhomorov V.K. *Adv. in Biological Chem*., 1, 2011, pp.1.

[8] Mukhomorov V.K. *Biomedical Statistics and Informatics*, 1 (2016) 24.

[9] Mukhomorov V.K. Modeling of Chemical Compounds Bioactivity. Relationships of Structure - Bioactivity. Lambert Academic Publisher. Saarbrűcken. Germany. 2012. (in Russian).

[10] Quastler G. In: Theory of Information in Biology. Ed. Blumenfeld LA. Moscow.1960.

[11] L’vovsky E.N. Statistical Methods for Constructing Empirical Formulas. Moscow. High School. 1988. (in Russian).

[12] Khalafyan A.A. Textbook. Statistica 6. Statistical Analysis of Data. Moscow. Publisher: Binomial. 2007. (in Russian).

[13] Fleiss J.L. Statistical Methods for Rates and Proportions. New York – Chichester – Brisbane – Toronto – Singapore. John Wiley & Sons, Inc. 1981.

[14] Forster E., Ronz B. Methoden der Korrelations- und Regressonanalise. Verlag Die Wirtschaft Berlin. 1979.

[15] Urbach V. Yu. Statistical Analysis in Biological and Medical Research. Moscow. Medicine. 1975. (in Russian).

[16] Pustylnik E.I. Statistical Methods of Analysis and Processing of Observations. Moscow. 1968. 288 p. (in Russian).

[17] Hollander M., Wolfe A. Nonparametric Statistical Methods. New York-London. John Wiley and Sons. Inc. 1973.

[18] http://potency.berkeley.edu/cpdb/html

[19] Rubenchik B.L. Biochemistry of Carcinogenesis. Kiev. 1977. (in Russian).

[20] Orris A., Yan Duuren B.L., Nelsen N. In: Proceedings of the VIII International of Anti-Cancer Congress. 1963. vol. 2, p. 305.

[21] Buu-Hoi N.P. *Cancer Res*., 24, 1964, pp.1511.

[22] Fukui J.A. et al. *Nippon Kaguki Zasshi*, 82, 1961, pp.474 (*Chem Abstr*. 57 (1962) 12378-b).

[23] Rubenchik B.L. Formation of Carcinogens from Nitrogen Compounds. Kiev. 1990.

[24] Leo A., Hansch C. *Chem. Rev*., 71, 1971, pp.525.

[25] Hidvedy T.Y., Arky I., Antoni F., Kõteles G. In: Proceedings of the VIII International of Anti-Cancer Congress. 1963. Moscow. vol. 2. p.225.

[26] Vilenchik MM Regularities of the Molecular-Genetic effect of Chemical Carcinogens. Moscow. 1977.

[27] Pullman B., Pullman A. Quantum Biochemistry. 1965.

[28] Doherty D. Radiation Protection and Recovery. New York. 1964.

[29] Pullman B. La Biochimie Electronique. Press Universitaires de France. 1963.

[30] Brookes P., Lawley P.D. *Nature*, 202, No.1964 4934, pp.781.

[31] Molecular Associations in Biology. Ed. by Pullman B. Academic Press, 1968. 217 p.

[32] Intermolecular Interactions: From Diatomics to Biopolymers. Ed. by Pullman B. John Wiley and Sons. Chichester-New-York. 1980.

[33] Pullman A., Pullman B. In: Advances in Cancer Research. Eds. by Greenstein J., Haddow A. Vol. 3. 1955. Acad. Press. Inc. Publisher. New-York.

[34] Badger G.M. The Chemical Basis of Carcinogenic Activity. Illinois, USA: Thomas-Publisher: Springfield, 1962. 98 p.

[35] Vylegzhanin N.I. *Oncology Affairs*. Vol. 21, Issue 5, 1952, pp.135. (in Russian).

[36] Gurvich L.V., Karachevtsev G.V., Kondratiev Yu. A. et al. The Energy of Chemical Linkages Crushing. Potentials of Ionization and Electron Affinity M. Nauka, 1974. (in Russian).

[37] Ladik J. Quantenbiochemie für Chemiker und Biologen. Akadimiai Kiado. Budapest. 1972.

[38] Heidelberger C., Moldenhauer M.G. *Cancer Research*, 16, 1956, pp.442.

[39] Fieser L.F., Putnam S.T. *J. Am. Chem. Soc*., 69. 1947. 1041.

[40] Peacock T.E. Electronic Properties of Aromatic and Heterocyclic Molecules. Acad. Press. London and New York. 1965.

[41] Goldanskii, V.I. *Angew. Chem.*, 79, 1967, pp.844.