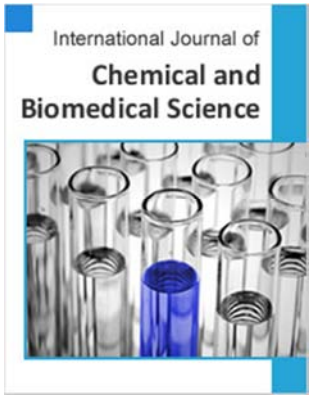
International Journal of Chemical and Biomedical Science
2017; 3(4): 39-63
http://www.aascit.org/journal/ijcbs
ISSN: 2472-9574 (Print); ISSN: 2472-9590 (Online)

American Association for Science and Technology



Keywords
Carcinogenicity,
Statistical Methods,
Classification Rules,
Information Function,
Electronic

Received: April 20, 2017
Accepted: June 6, 2017
Published: August 21, 2017

Statistical Modeling and Prediction of Carcinogenic Activity of Chemical Compounds

Vladimir Mukhomorov

Universita degli Studi di Napoli "Federico II" Via Cintia, Napoli, Italy

Email address
vmukhomorov@mail.ru

Citation
Vladimir Mukhomorov. Statistical Modeling and Prediction of Carcinogenic Activity of Chemical Compounds. *International Journal of Chemical and Biomedical Science.* Vol. 3, No. 4, 2017, pp. 39-63.

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, and some medications. A total of 541 chemical compounds were examined.

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.

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. A. 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).

2. 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 "-".

Table 1. *The carcinogenic activity, electronic and information factors 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	N-Nitroso-N-ethylurea	$C_3H_7N_3O_2$	+	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

N	**Chemical compound**	**Gross-formula**	**Activity**	*Z*	*H, bits*
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	$C_{20}H_{12}O_2$	+	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- Oxyazobenzene	$C_{12}H_{10}N_2O$	+	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
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	$C_{16}H_{14}N_4$	+	2.882	1.402
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

N	**Chemical compound**	**Gross-formula**	**Activity**	*Z*	*H, bits*
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	1.417
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
148	5-Fluorouracil	$C_4H_3FN_2O_2$	-	4.000	2.189
149	γ-Butyrolactone	$C_4H_6O_2$	-	2.833	1.460
150	Dicetene	$C_4H_4O_2$	-	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

N	**Chemical compound**	**Gross-formula**	**Activity**	*Z*	*H, bits*
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
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

N	Chemical compound	Gross-formula	Activity	*Z*	*H*, *bits*
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	Indigo carmine	$C_{16}H_8N_2Na_2O_8S_2$	-	3.790	2.143
231	Sudan Brown RR	$C_{16}H_{14}N_4$	-	2.882	1.402
232	Tartrazine	$C_{16}H_9N_4Na_3O_9S_2$	-	3.767	2.268
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
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].

The analysis of Table 1 will be carried out using molecular descriptors Z and H [6-8]. The descriptor Z determines the average number of valence electrons in the molecule:

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

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 close in meaning was used in Ref. [9]. 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]

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

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 $Z^{(th)}$ 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:

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

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 $\langle Z \rangle^{(\mathrm{av})}$:

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

Here $\langle Z \rangle^{(\mathrm{max})}$ and $\langle Z \rangle^{(\mathrm{min})}$ refer to chemical compounds in 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:

$$\tau = \left| \left\langle Z^{(\mathrm{max/min})} \right\rangle - \left\langle Z^{(\mathrm{av})} \right\rangle \right| / S = \begin{cases} 4.46^{(\mathrm{max})} > \tau^{(\mathrm{cr})}_{0.05}(N = 255) = 3.65, \\ 2.95^{(\mathrm{min})} < \tau^{(\mathrm{cr})}_{0.05}(N = 255) = 3.65. \end{cases} \quad (5)$$

Inequality (5) indicates that at the significance level of α = 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:

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

Normality of the sample is confirmed by the inequality:

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

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

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

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:

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

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 [13]:

$$| Z_1^{(\mathrm{av})} - Z_2^{(\mathrm{av})} | = 0.303 > t^{(\mathrm{cr})}_{0.05}(f) \left\{ \frac{N[(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2]}{N_1 N_2 (N_1 + N_2 - 2)} \right\}^{1/2} = 0.171. \quad (10)$$

Here $f = N_1 + N_2 - 2$. 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 $Z_1^{(av)}$, and inactive chemical compounds are grouped around of $Z_2^{(av)}$.

The value $Z^{(av)} = 3.06$ is taken as the threshold value: $Z^{(th)} \equiv Z^{(av)}$. If the molecular descriptor Z is below a threshold value, then such chemical compound should hypothetically posses of carcinogenic properties. The smaller molecular descriptor Z, the higher the electrophilicity of the chemical compound. It is well known that most carcinogens are electrophiles. Chemical compounds that have descriptor $Z > Z^{(th)}$ probably do not have carcinogenic activity. Since descriptor of 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) [14]. 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 of 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 of 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**	
$Z \le Z^{(th)} = 3.06$	$q_{11} = 99$ $p_{11} = 0.40$	$q_{21} = 41$ $p_{21} = 0.16$	140 $P_1 = 0.56$
$Z > Z^{(th)} = 3.06$	$q_{12} = 38$ $p_{12} = 0.15$	$q_{22} = 72$ $p_{22} = 0.29$	110 $P_2 = 0.44$
	137	113	$N = 250$
Total	${}_1P = 0.55$	${}_2P = 0.45$	$\sum_i {}_iP = \sum_i P_i = 1.00$

$Q = 0.64$, $\Phi = 0.28$, $\chi^2 = 32.5 > \chi^{2(cr)}_{0.05}(f = 1) = 3.84$, $SE = 0.03$, $\Omega = 4.58$, $K = 0.34$, $|r_{tet}| = 0.54$, $\Delta = 0.32$, $SES = 1.25$.

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 [14]. The standard error (*SES*) of the odds ratio is determined as follows

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

The Pearson contingency coefficient [15] is determined by the following equation

$$\Phi = \frac{q_{11}q_{22} - q_{12}q_{21}}{[(q_{11} + q_{12})(q_{21} + q_{22})(q_{11} + q_{21})(q_{12} + q_{22})]^{1/2}} = 0.28. \quad (12)$$

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

$$t = \Phi\sqrt{N-2}/\sqrt{1-\Phi^2} > t^{(cr)}_{0.05}(f = N-2). \quad (13)$$

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

$$SE(\Phi) = 0.5(1-\Phi^2)(1/q_{11} + 1/q_{12} + 1/q_{21} + 1/q_{22}) = 0.03. \quad (14)$$

It is also possible to use the Yule association coefficient [15] to identify the relationship between the factors:

$$Q = \frac{q_{11}q_{22} - q_{12}q_{21}}{q_{11}q_{22} + q_{12}q_{21}} = 0.64. \quad (15)$$

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 that 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:

$$r_{\text{tet}} = \cos\left[\pi \cdot (q_{12}q_{21})^{1/2}((q_{11}q_{22})^{1/2} + (q_{12}q_{21})^{1/2})^{-1}\right]. \quad (16)$$

The chi-criterion [15] is compared with critical value of the chi-square distribution function for one degree of freedom ($f = 1$):

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

$$= 32.5 > \chi^{2(\text{cr})}_{0.05}(f = 1) = 3.84, \quad (17)$$

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 $\chi^2 > \chi^{2(\text{cr})}_{0.05}(f = 1)$, 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 that if there is no relationship between factors then these relations should be identically equal.

Now we will use 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:

$$N = 255, \quad \chi^2_{0.05} = 5.70 < \chi^{2(\text{cr})}_{0.05}(f = 7) = 14.1, \quad df = 7, \quad p = 0.58 >> 0.05. \quad (18)$$

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

$$N = 255, \quad H^{(\text{av})} = 1.62 \pm 0.02, \quad H^{(\text{min})} = 0.724, \quad H^{(\text{max})} = 2.23, \quad S = 0.31. \quad (19)$$

The sample at significance level α = 0.05 is homogeneous:

$$\tau = | H^{(\text{max/min})} - H^{(\text{av})} | / S = \begin{cases} 1.97^{(\text{max})} < \tau^{(\text{cr})}_{0.05}(N = 255) = 3.65, \\ 2.90^{(\text{min})} < \tau^{(\text{cr})}_{0.05}(N = 255) = 3.65. \end{cases} \quad (20)$$

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:

$$N_1 = 141, \; H_1^{(\text{av})} = 1.56 \pm 0.03, \; H_1^{(\text{min})} = 0.723, \; H_2^{(\text{max})} = 2.181, S_1 = 0.31;$$

$$\chi^2 = 6.16 < \chi^{2(\text{cr})}_{0.05}(df = 9) = 16.9, p = 0.72 >> 0.05. \quad (21)$$

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

$$N_2 = 114, \; H_2^{(\text{av})} = 1.70 \pm 0.03, \; H_2^{(\text{min})} = 0.919, \; H_2^{(\text{max})} = 2.230, S_2 = 0.30;$$

$$\chi^2 = 4.74 < \chi^{2(\text{cr})}_{0.05}(df = 3) = 7.8, p = 0.20 > 0.05. \quad (22)$$

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

$$| H_1^{(\text{av})} - H_2^{(\text{av})} | = 0.143 > t^{(\text{cr})}_{0.05}(f)\left\{\frac{N[(N_1-1)S_1^2 + (N_2-1)S_2^2]}{N_1N_2(N_1+N_2-2)}\right\}^{1/2} = 0.074. \quad (23)$$

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 $H^{(\text{th})} \equiv H^{(\text{av})} = 1.62bits$ (17) we will accept for the boundary (threshold) value. Table 3 provides statistics on the application of the association method. Variation the descriptor $H^{(\text{av})}$ about the average value showed that somewhat higher statistical results can be obtained if we use as the boundary value $H^{(\text{th})} = 1.70bits$. Table 3 shows statistics for this threshold value (data are given in parentheses). Since $\chi^2 > \chi^{2(\text{cr})}_{0.05}$ 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: $q_{11}/(q_{11}+q_{12}) = 0.57(0.67) \neq q_{22}/(q_{22}+q_{21}) = 0.60(0.51)$.

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	
$H \le H^{(\text{th})} = 1.62 bits$	$q_{11} = 81(95)$ $p_{11} = 0.32(0.37)$	$q_{21} = 46(56)$ $p_{21} = 0.18(0.22)$	127(151) $P_1 = 0.50(0.59)$
$H > H^{(\text{th})} = 1.62 bits$	$q_{12} = 60(46)$ $p_{12} = 0.23(0.18)$	$q_{22} = 68(58)$ $p_{22} = 0.27(0.23)$	128(104) $P_2 = 0.50(0.41)$
	141(141)	114(114)	$N = 255$
Total	${}_1P = 0.55(0.55)$	${}_2P = 0.45(0.45)$	$\sum_i {}_iP = \sum_i P_i = 1.00$

$Q = 0.33(0.36)$, $\Phi = 0.09(0.18)$, $\chi^2 = 7.37(8.70) > \chi^{2(\text{cr})}_{0.05}(f = 1) = 3.84$, $SE = 0.03(0.03)$, $\Omega = 2.00(2.14)$, $K = 0.12(0.21)$, $|r_{\text{tet}}| = 0.27(0.29)$, $\Delta = 0.42(0.40)$, $SES = 0.51(0.56)$.

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 [16], we obtained the following sub-sample. The random subsample contains 63 elements, which is much smaller than the original sample size (~ 1/4 of the original sample). The random sub-sample 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 (bold font) of the subset should be excluded from the sub-sample, 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 :

$$N = 60,\ Z^{(\text{av})} = 3.13 \pm 0.06,\ Z^{(\text{min})} = 1.882,\ Z^{(\text{max})} = 4.333,\ S = 0.49,$$

$$N_1 = 28,\ Z_1^{(\text{av})} = 2.86 \pm 0.081,\ Z_1^{(\text{min})} = 1.882,\ Z_1^{(\text{max})} = 4.273,\ S_1 = 0.43,$$

$$N_2 = 32,\ Z_2^{(\text{av})} = 3.36 \pm 0.08,\ Z_2^{(\text{min})} = 2.750,\ Z_2^{(\text{max})} = 4.333,\ S_2 = 0.43. \quad (25)$$

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:

$$N = 60,\ \chi^2 = 2.84 < \chi^{2(\text{cr})}_{0.05}(df = 5) = 11.07,$$

$$p = 0.72 >> 0.05. \quad (26)$$

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

$$N = 60,\ H^{(\text{av})} = (1.63 \pm 0.04) bits,\ H^{(\text{min})} = 0.919 bits,\ H^{(\text{max})} = 2.236 bits,\ S = 0.34,$$

$$N_1 = 28,\ H_1^{(\text{av})} = (1.45 \pm 0.06) bits,\ H_1^{(\text{min})} = 0.919 bits,\ H_1^{(\text{max})} = 2.048 bits,\ S_1 = 0.31,$$

$$N_2 = 32,\ H_2^{(\text{av})} = (1.79 \pm 0.05) bits,\ H_2^{(\text{min})} = 1.204 bits,\ H_2^{(\text{max})} = 2.230 bits,\ S_2 = 0.28. \quad (27)$$

The average values the descriptor of $H^{(\text{av})}$ (27) turned out to be close to the average values (19).

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

$$N = 60,\ q_{11} = 22,\ q_{12} = 6,\ q_{22} = 22,\ q_{21} = 10;\ Q = 0.78,\ \Phi = 0.39,$$

$$\chi^2 = 13.4 > \chi^{2(\text{cr})}_{0.05}(f = 1) = 3.84,\ SE = 0.15,\ \Omega = 8.07,$$

$$K = 0.44,\ |r_{\text{tet}}| = 0.68,\ \Delta = 0.27,\ SES = 4.82. \quad (28)$$

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 $H^{(\text{th})} \equiv H^{(\text{av})} = 1.66 bits$ (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:

$$N = 60,\ q_{11} = 20,\ q_{12} = 8,\ q_{22} = 19,\ q_{21} = 13;\ Q = 0.57,$$

$$\Phi = 0.26,$$

$$\chi^2 = 5.74 > \chi^{2(\mathrm{cr})}_{0.05}(f = 1) = 3.84,\ SE = 0.14,\ \Omega = 3.65,$$

$$K = 0.33,\ |r_{\mathrm{tet}}| = 0.47,\ \Delta = 0.35,\ SES = 2.02. \quad (29)$$

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:

$$N = 60,\ H(Z) = b_0 + b_1 \cdot Z,\ b_0 = 0.27 \pm 0.21,\ b_1 = 0.43 \pm 0.07,$$

$$t(b_0) = 1.28,\ t(b_1) = 6.47,\ R = 0.65,$$

$$F = 41.8 > F^{(\mathrm{cr})}_{0.05}(f_1 = 1; f_2 = 58) = 4.00,$$

$$Std.\ Err. = 0.25. \quad (30)$$

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

$$N = 250,\ H(Z) = b_0 + b_1 \cdot Z,\ b_0 = 0.22 \pm 0.10,\ b_1 = 0.47 \pm 0.03,$$

$$t(b_0) = 2.13,\ t(b_1) = 13.8,\ R = 0.66,$$

$$F = 190 > F^{(\mathrm{cr})}_{0.05}(f_1 = 1; f_2 = 248) = 3.89,$$

$$Sdt.\ Err. = 0.22. \quad (31)$$

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:

$$N_1 = 137,\ H(Z) = b_0 + b_1 \cdot Z,\ b_0 = 0.176 \pm 0.143,$$

$$b_1 = 0.478 \pm 0.048,\ t(b_0) = 1.23,\ t(b_1) = 9.88,\ R = 0.65,$$

$$F = 97.8 > F^{(\mathrm{cr})}_{0.05}(f_1 = 1; f_2 = 135) = 3.92,$$

$$Std.\ Err. = 0.223. \quad (32)$$

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

$$N_2 = 113,\ H(Z) = b_0 + b_1 \cdot Z,\ b_0 = 0.280 \pm 0.172,$$

$$b_1 = 0.445 \pm 0.053,\ t(b_0) = 1.63,\ t(b_1) = 8.36,\ R = 0.62,$$

$$F = 69.8 > F^{(\mathrm{cr})}_{0.05}(f_1 = 1; f_2 = 111) = 3.94,$$

$$Std.Err. = 0.227. \quad (33)$$

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

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

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 following form (31); the number of characteristic factors is equal to: $m = 1$. From the equation (34) we obtain the inequality for *F*-statistics (degrees of freedom: $f_1 = m + 1$, $f_2 = N_1 + N_2 - 2m - 2$):

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

Therefore, it is impossible to reject the null hypothesis at the significance level $\alpha = 0.05$. 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 sub-sample. A sub-sample 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:

$$N = 63,\ Z^{(\mathrm{av})} = 3.15 \pm 0.05,\ Z^{(\mathrm{min})} = 2.471,\ Z^{(\mathrm{av})} = 4.333,$$

$$S = 0.40. \quad (36)$$

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

$$\tau = |Z^{\left(\frac{\max}{\min}\right)} - Z^{(\mathrm{av})}| / S = \begin{cases} 2.96^{(\max)} < \tau^{(\mathrm{cr})}_{0.05}(f = 63) = 3.22, \\ 1.70^{(\min)} < \tau^{(\mathrm{cr})}_{0.05}(f = 63) = 3.22. \end{cases} \quad (37)$$

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

$$N_1 = 26,\ Z_1^{(\mathrm{av})} = 2.90 \pm 0.05,\ Z_1^{(\mathrm{min})} = 2.471,\ Z_1^{(\max)} = 3.539,\ S_1 = 0.24,$$

$$N_2 = 37,\ Z_2^{(\mathrm{av})} = 3.32 \pm 0.07,\ Z_2^{(\mathrm{min})} = 2.588,\ Z_2^{(\mathrm{max})} = 4.333,\ S_2 = 0.40. \quad (38)$$

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 α = 0.05 Therefore, to compare the average values of two clusters, it is necessary to use the approximate T-test [13]:

$$| Z_1^{(\mathrm{av})} - Z_2^{(\mathrm{av})} | = 0.42 > T = \frac{\nu_1 t_{0.05}^{(\mathrm{cr})}(f_1) + \nu_2 t_{0.05}^{(\mathrm{cr})}(f_2)}{(\nu_1 + \nu_2)^{0.5}} = 0.17, \quad (39)$$

here $\nu_1 = S_1^2 / N_1 = 2.215 \cdot 10^{-3}$ $\nu_2 = S_2^2 / N_2 = 4.33 \cdot 10^{-3}$. 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. 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 [13], 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:

$$N = 63,\ q_{11} = 23,\ q_{12} = 3,\ q_{22} = 23,\ q_{21} = 14;\ Q = 0.85,$$

$$\Phi = 0.48,$$

$$\chi^2 = 16.14 > \chi_{0.05}^{2(\mathrm{cr})}(f = 1) = 3.84,\ SE = 0.19,\ \Omega = 12.6,$$

$$K = 0.49,\ |r_{\mathrm{tet}}| = 0.77,\ \Delta = 0.27,\ SES = 8.83. \quad (40)$$

It follows that there is a statistically significant relationship between the value of Z and the carcinogenic activity of chemical compounds. Similarly, we will obtain statistics for the information function:

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

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

$$N_1 = 26,\quad H_1^{(\mathrm{av})} = 1.62 \pm 0.03,\quad H_1^{(\mathrm{min})} = 1.397,\quad H_1^{(\mathrm{max})} = 1.931,\quad S_1 = 0.15,$$

$$N_2 = 37,\quad H_2^{(\mathrm{av})} = 1.71 \pm 0.04,\quad H_2^{(\mathrm{min})} = 1.204,\quad H_2^{(\mathrm{max})} = 1.972,\quad S_2 = 0.22,$$

$$F = S_2^2 / S_1^2 = 2.24 > F_{0.05}^{(\mathrm{cr})}(f_2 = 36; f_1 = 25) = 1.90,$$

$$| H_1^{(\mathrm{av})} - H_2^{(\mathrm{av})} | = 0.097 > T = \frac{\nu_1 t_{0.05}^{(\mathrm{cr})}(f_1) + \nu_2 t_{0.05}^{(\mathrm{cr})}(f_2)}{(\nu_1 + \nu_2)^{0.5}} = 0.004. \quad (41)$$

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

$$N = 63,\quad q_{11} = 19,\quad q_{12} = 7,\quad q_{22} = 21,\quad q_{21} = 16;\ Q = 0.56,$$

$$\Phi = 0.26,$$

$$\chi^2 = 5.50 > \chi_{0.05}^{2(\mathrm{cr})}(f = 1) = 3.84,\quad SE = 0.14,\quad \Omega = 3.56,$$

$$K = 0.37,\quad |r_{\mathrm{tet}}| = 0.46,\quad \Delta = 0.37,\quad SES = 1.97. \quad (42)$$

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), (19), (25), and (27). This indicates the stability of statistical results. The descriptors Z and H are interrelated (Figure 1).

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

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

Here Σ_1= 0.341 and Σ_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). Σ_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 ± 0.14) + (0.33 ± 0.04)·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 ($f_1 = m + 1, f_2 = N_1 + N_2 - 2m - 2$):

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

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

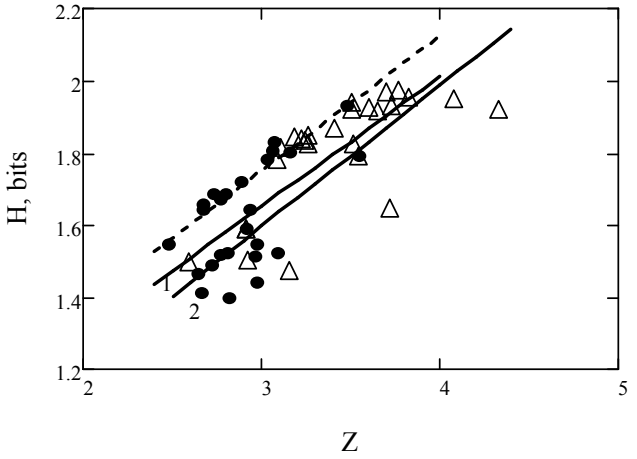


Figure 1. *Diagram of scattering the descriptors for the sub-sample (33). Active chemical compounds are marked with dots (•). (1) Regression has the following form: H(Z) = (0.57 ± 0.03) + (0.36 ± 0.10)·Z, R = 0.59, S_1 = 0.12, $F = 13.06 > F_{0.05}^{(cr)}(1;24) = 4.26$, N_1 = 26. Inactive chemical compounds are marked by triangles (Δ). (2) Regression has the following form: H(Z) = (0.43 ± 0.20) + (0.39 ± 0.06)·Z, R = 0.74, S_2 = 0.14, $F = 42.9 > F_{0.05}^{(cr)}(1;35) = 4.11$, N_2 = 37. If the correlation coefficient is within 0.25 ≤ R ≤ 0.75, then the correlation is viewed as moderate [17].*

Significant scattering around regression lines 1 and 2 (Figure 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):

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

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

$$H(Z) = A + B \cdot Z,\ A = 0.64 \pm 0.06,\ B = 0.38 \pm 0.02,\ R = 0.98,$$

$$N = 12,\ t(B) = 17.3 > t(A) = 10.0 > t_{0.05}^{(cr)}(f = 10) = 1.81,$$

$$F = 299.8 >> F_{0.05}^{(cr)}(f_1 = 1, f_2 = 10) = 4.96,$$

$$Std.Err.\ of\ Estimate = 0.0197. \qquad (46)$$

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, 18] 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	$C_6H_{12}C_{l3}PO$	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

N	Chemical compounds	Gross-formula	*Z*	*H,bits*
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 of *Z*:

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

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

$$q_{11} = 28,\ q_{12} = 10,\ q_{22} = 23,\ q_{21} = 12;\ Q = 0.69,\ \Phi = 0.31,$$

$$\chi^2 = 11.42 > \chi^{2(cr)}_{0.05}(f = 1) = 3.84,\ SE = 0.12,$$

$\Omega = 5.4$, $K = 0.37$, $|r_{tet}| = 0.58$, $\Delta = 0.30$, $SES = 2.75$.

Error of model is equal to 30%**.**

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

Let us check whether active and inactive chemical compounds really belong to different subsets and are primarily grouped around $Z_1^{(av)}$ and $Z_2^{(av)}$. Preliminarily, we compare the ratio of the larger variance to the smaller variance with the critical value of the Fisher distribution: $F = S_{z2}^2 / S_{z1}^2 = 2.57 > F_{0.05}^{(cr)}(f_2 = N_2 - 1 = 34; f_1 = N_1 - 1 = 37) = 1.72$. 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:

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

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

2. The descriptor of *H*:

$$N = 73,\ H^{(av)} = 1.66 \pm 0.03,\ H^{(min)} = 1.20,\ H^{(max)} = 2.20,$$

$$S = 0.22,$$

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

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

$$\chi^2 = 3.07 < \chi_{0.05}^{2(cr)}(f = 1) = 3.84 ,\ SE = 0.11,$$

$\Omega = 2.3$, $K = 0.19$, $|r_{tet}| = 0.32$, $\Delta = 0.40$, $SES = 1.10$. Error of model is equal to: 40%.

$$N_1 = 38,\ H_1^{(av)} = 1.63 \pm 0.03,\ H_1^{(min)} = 1.29,\ H_1^{(max)} = 2.20,$$

$$S_{H1} = 0.19,$$

$$N_2 = 35,\ H_2^{(av)} = 1.69 \pm 0.04,\ H_2^{(min)} = 1.20,\ H_2^{(max)} = 2.19,$$

$$S_{H2} = 0.25. \quad (51)$$

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

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

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 [16] 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:

$$N = 36,\ Z(av) = 3.15 \pm 0.07,\ H(av) = 1.67 \pm 0.03. \quad (53)$$

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

N	Chemical compounds	Gross-formula	*Z*	*H,bits*
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	Chlomiphene*)	$C_{26}H_{28}ClNO$	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), (19), (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 this handbook (all compounds from 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 [16] 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:

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

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

For the random sub-sample (51) the average values will be as follows:

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

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), (19), (25), (27), (36) 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 sub-samples.*

Original Table 1				
$N = 250$ $N = 255$	$H^{(th)} \equiv H^{(av)} = 1.62 \pm 0.02$	Eq.(19)	$Z^{(th)} \equiv Z^{(av)} = 3.06 \pm 0.03$	Eq. (10)
The sub-sample. Eq. (25) $N = 60$	$H^{(th)} \equiv H^{(av)} = 1.63 \pm 0.04$		$Z^{(th)} \equiv Z^{(av)} = 3.13 \pm 0.06$	
The sub-sample. Eq. (51) $N = 85$	$H^{(th)} \equiv H^{(av)} = 1.56 \pm 0.03$		$Z^{(th)} \equiv Z^{(av)} = 3.09 \pm 0.07$	
The sub-sample. Eq. (41) $N = 63$	$H^{(th)} \equiv H^{(av)} = 1.68 \pm 0.03$		$Z^{(th)} \equiv Z^{(av)} = 3.15 \pm 0.05$	
The sub-sample. Eq. (45) $N = 78$	$H^{(th)} \equiv H^{(av)} = 1.87 \pm 0.02$		$Z^{(th)} \equiv Z^{(av)} = 3.15 \pm 0.04$	
The sub-sample (Table 4) $N = 73$	$H^{(th)} \equiv H^{(av)} = 1.87 \pm 0.02$		$Z^{(th)} \equiv Z^{(av)} = 3.17 \pm 0.05$	

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 $Z^{(th)}$ and $H^{(th)}$. The descriptors of 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].

3. Discussion and Comparison with Monitoring

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 (19). Only in one case (4-Nitrobiphenyl) the descriptor of Z slightly exceeds the threshold value $Z^{(th)} = 3.06$. 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 H ($H^{(th)} = 1.62 bits$). As analysis has shows the descriptors for chemical compounds (Table 7) are interrelated (Figure 2):

$$H(Z) = A + B \cdot Z,\ A = -0.0565 \pm 0.2397,\ B = 0.5376 \pm 0.0891,$$

$$t(B) = 6.048 > t_{0.05}^{(cr)}(f = 13) = 1.77 > |t(A)| = 0.240,\ N = 15,$$

$$R = 0.86,$$

$$F = 36.44 > F_{0.05}^{(cr)}(f_1 = 1; f_2 = 13) = 4.67,\ \textit{Std.Err. of Estimate} = 0.137,\ \delta H(Z) = 9.08\%. \quad (57)$$

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

N	Chemical compounds	Gross-formula	Activity	Z	H,bits
1	3,3'-Dimethoxy benzidine	C14H16O2N2	+	$2.77 < Z^{(th)}$	$1.52 < H^{(th)}$
2	Magenta	C20H19N3·HCl	+	$2.77 < Z^{(th)}$	$1.43 < H^{(th)}$
3	4,4'-Methylene bis(2-chloraniline)	C13H12N2Cl2	+	$2.86 < Z^{(th)}$	$1.58 < H^{(th)}$
4	4,4'-Methylene bis(2-methylaniline)	C15H18N2	+	$2.51 < Z^{(th)}$	$1.25 < H^{(th)}$
5	4,4'-Methylene bis (2-dianiline)	C5H11Cl2N	+	$2.62 < Z^{(th)}$	$1.29 < H^{(th)}$
6	1-Naphthylamine	C10H9N	+	$2.70 < Z^{(th)}$	$1.23 < H^{(th)}$
7	2-Naphthylamine	C10H9N	+	$2.70 < Z^{(th)}$	$1.23 < H^{(th)}$
8	4-Nitrobiphenyl	C12H9NO2	+	$3.08 \approx Z^{(th)}$	$1.52 < H^{(th)}$
9	N,N-Bis(2-Chloroethyl)-2-naphthylamine	C14H15Cl2N	+	$2.81 < Z^{(th)}$	$1.44 < H^{(th)}$
10	Hydrazine	N2H4	+	$2.33 < Z^{(th)}$	$0.92 < H^{(th)}$
11	1,1-Dimethylhydrazine	C2H8N2	+	$2.17 < Z^{(th)}$	$1.25 < H^{(th)}$
12	1,2-Dimethylhydrazine	C2H8N2	+	$2.17 < Z^{(th)}$	$1.25 < H^{(th)}$
13	1,2-Diethylhydrazine	C4H12N2	+	$1.88 < Z^{(th)}$	$1.06 < H^{(th)}$
14	Izonicotinic acid hydrazide	C6H7N3O	-	$3.06 \approx Z^{(th)}$	$1.74 > H^{(th)}$
15	Maleic hydrazide	C4H4N2O2	-	$3.50 > Z^{(th)}$	$1.92 > H^{(th)}$

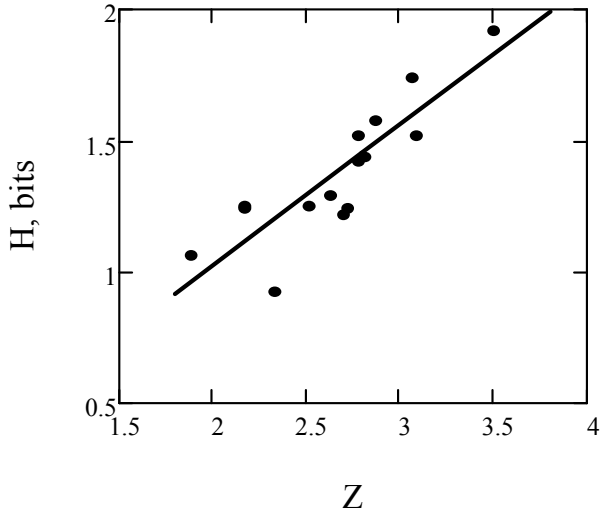


Figure 2. *The correlation field of the information function and the electronic descriptor for aromatic amines and relative compounds. Points are the data Table 7. Regression line is approximated by the function (57).*

It is not difficult to see (Figure 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 model error to 6.5%.

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-acetylaminodibenzo-thiophene (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-amino-stilbene (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 section "Dyes") of the handbook [5]. These data were supplemented by data from the book [3 (Appendix II, Russian edition)]. We used the threshold values $Z^{(th)}$ = 3.15, $H^{(th)}$ =1.87*bits* (see Table 6). Table 8 bellow

demonstrates the descriptors do not contradict the classification rules for all dyes without exception.

Table 8. *Carcinogenic properties of dyes.*

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

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): $Z^{(th)} \equiv Z^{(av)} = 3.15$, $H^{(th)} \equiv H^{(av)} = 1.69 bits$. 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 of model is equal to 20%) and in eight cases (the error of model 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%. 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 (Figure 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 Figure 1). However, a statistically significant interrelation between descriptors is remained intimately.

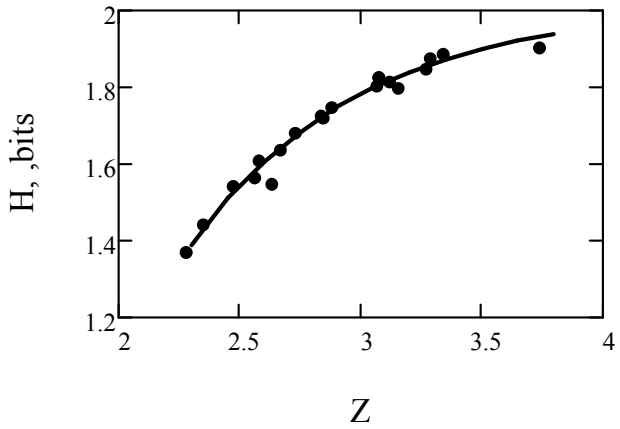


Figure 3. *The correlation field of the information function H and the electronic characteristic Z for nitroso-compounds. Points are the data Table 9. Regression line is approximated by the following function: H(Z) = A + B·exp(-C·Z), A = 2.01 ± 0.05, B = -16.60 ± 7.27, C = 1.43 ± 0.21, N = 20, RMSE = 0.025, δH(Z) = 2.01%.*

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

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

[*)] 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.

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 (Appendix II, Russian edition)], [5] and [20]. 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 of model is equal to 43%). At the same time, descriptor of *H* incorrectly estimates the carcinogenic properties for eight agents (the error of model is equal to 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 18 (Table 10) is *potentially active* agent ($Z < Z^{(th)}$). We assume that the hypothesis [6-8] 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 16 up to 18 (the descriptors for these molecules are marked in italics) 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 *m* of CH_2 and CH ($m > 14$) atomic groups [6-8].

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	1,2-Epoxybutane	$C_4H_6O_2$	-	$2.83 < Z^{(th)}$	$1.46 < H^{(th)}$
14	d1-Diepoxybutane	$C_4H_6O_2$	-	$2.83 < Z^{(th)}$	$1.46 < H^{(th)}$
15	Styrene oxide	C_8H_8O	-	$2.71 < Z^{(th)}$	$1.26 < H^{(th)}$
16	9,10- Epoxystearic acid	$C_{18}H_{34}O_3$	-	$2.25 < Z^{(th)}$	$1.19 < H^{(th)}$
17	6,7,9,10- Epoxystearic acid	$C_{18}H_{32}O_4$	-	$2.37 < Z^{(th)}$	$1.25 < H^{(th)}$
18	Hexaepoxysvalol	$C_{30}H_{48}O_6$	-	$2.48 < Z^{(th)}$	$1.26 < H^{(th)}$
Hydroperoxides					
19	1-Vinyl-1-hydroperoxide of cyclohexane	$C_8H_{11}O_2$	+	$2.62 < Z^{(th)}$	$1.53 < H^{(th)}$
20	1-Vinilcyclohexane-3	C_8H_{11}	+	$2.26 < Z^{(th)}$	$0.98 < H^{(th)}$
21	Benzene peroxide	$C_{14}H_{10}O_4$	-	$3.21 > Z^{(th)}$	$1.43 < H^{(th)}$
22	Lauroyl peroxide	$C_{24}H_{46}O_4$	-	$2.34 < Z^{(th)}$	$1.18 < H^{(th)}$

We note an analogous situation for the molecule of lauroyl peroxide in the series of hydroperoxides (chemical compound at number 22). Chemical compounds at numbers 13 and 14 are also potentially active. However, they do not have confirmed carcinogenic activity. According to the data of [20], the agent at number 13 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 works [6-8]. 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 [6-8]. 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: $7 \geq m \geq 5$. This range of index *m* 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 [7-8], 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, Figure 4). The hydrophobicity of the homologous series ON-N-CH_3 $(CH_2)_m$ CH_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.

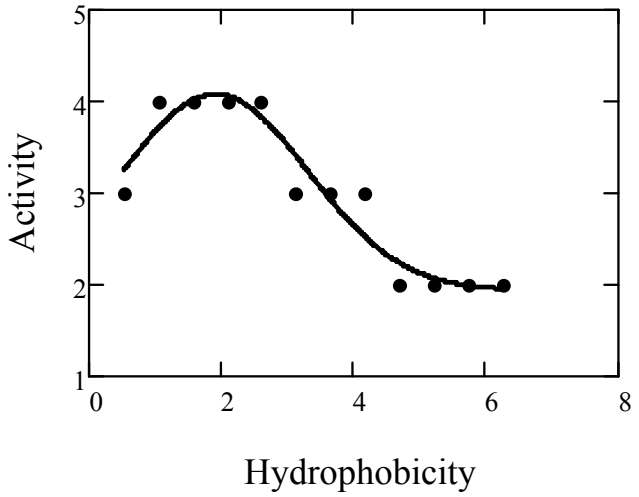


Figure 4. *Interrelation nitrosomethylalkylamines carcinogenic activity A(π) with their hydrophobicity. Points are the comparative carcinogenic activity (Table 11). The envelope of the regression line is defined by the following equation:* $A(\pi) = D + A \cdot \exp(-(\pi-B)^2/C^2)$, $A = 2.14 \pm 0.23$,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

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

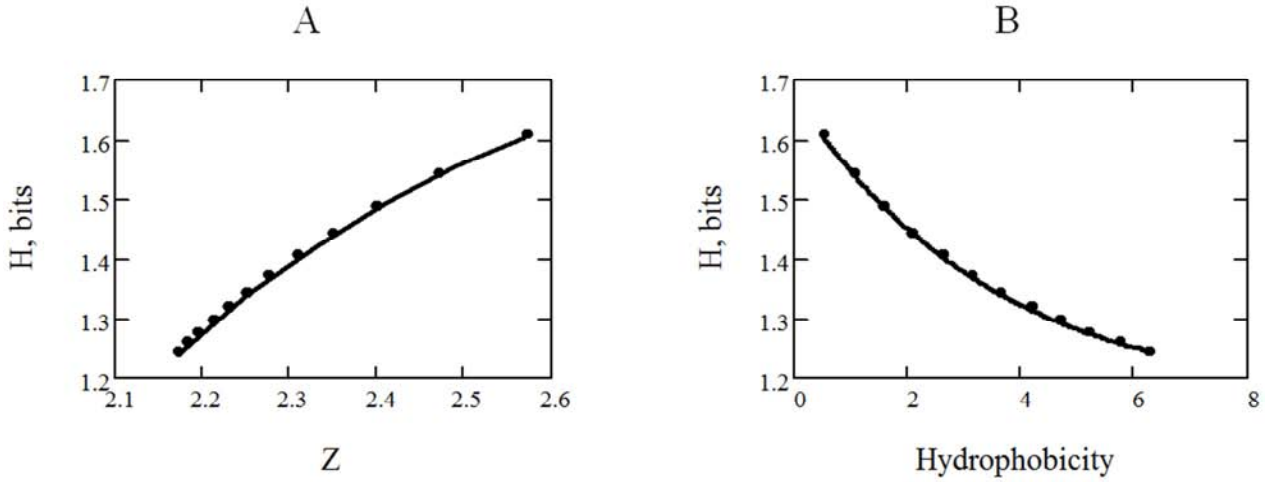


Figure 5. *(A) Interrelation of molecular descriptors for a series of nitrosomethylalkylamines. Points are the data of Table 11. The regression line is determined by the equation:* $H(Z) = B + A \cdot \exp(-C \cdot Z)$, $A = -58.2 \pm 0.96$, $B = 1.89 \pm 0.01$, $C = 2.07 \pm 5.59$, $RMSE = 0.0002$. *(B) Interrelation of information function with hydrophobicity. Points are the data of Table 11. The regression line is determined by the equation:* $H(\pi) = B + A \cdot \exp(-C \cdot \pi)$, $A = 0.55 \pm 0.01$, $B = 1.16 \pm 0.01$, $C = 0.29 \pm 0.01$, $RMSE = 0.002$.

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. 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.

Next, 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 (8) and (19).

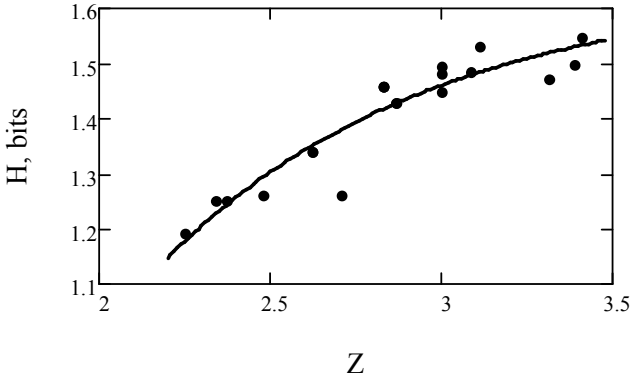


Figure 6. *For oxy-compounds there is a close relationship between the information function and electronic descriptor (Table 10). The regression line is determined by the equation: $H(Z) = A + B \cdot \exp(-C \cdot Z)$, $A = 1.64 \pm 0.11$, $B = -9.01 \pm 0.40$, $C = 1.32 \pm 0.59$, $RMSE = 0.043$, $\delta H(Z) = 2.21\%$.*

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	$C_{10}H_{24}Cl_4N_2O_4$	+	$2.86 < Z^{(th)}$	$1.80 > H^{(th)}$
3	Melphan	$C_{13}H_{18}Cl_2N_2O_2$	+	$2.86 < Z^{(th)}$	$1.72 > H^{(th)}$
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	$C_5H_{11}Cl_2NO$	+	$2.80 < Z^{(th)}$	$1.74 > H^{(th)}$
8	Nitrogen mustard N-oxide	$C_5H_{12}Cl_3NO$	+	$2.91 < Z^{(th)}$	$1.76 > H^{(th)}$
9	Oestradiol mustard	$C_{42}H_{50}Cl_4N_2O_4$	+	$2.75 < Z^{(th)}$	$1.52 < H^{(th)}$
10	Uracil mustard Uracil	$C_8H_{11}Cl_2N_3O_2$	+	$3.23 > Z^{(th)}$	$1.98 > H^{(th)}$
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 ($\approx$ 39%), whereas the classification rule using the descriptor Z gives only one erroneous result ($\approx$ 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).

Now we will examine the applicability of the classification rules using a sub-sample of drags. We have compiled Table 13, taking into account the threshold values (8) and (19) for molecular descriptors. Table 13 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 $\approx$ 14%.

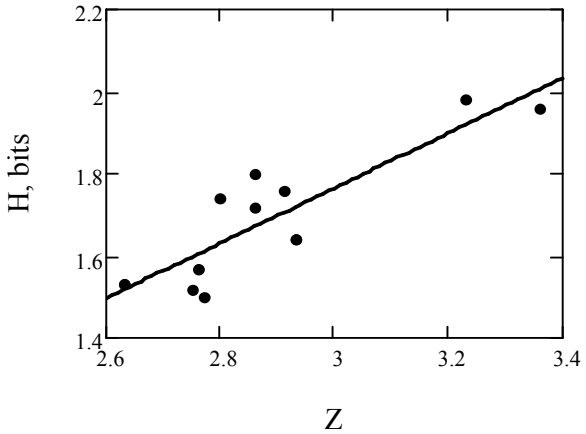


Figure 7. *Interrelation of the descriptors (Table 12). The regression equation has the following form: $H(Z) = A + B \cdot Z$, $R = 0.88$, $A = -(0.25 \pm 0.32)$, $B = 0.67 \pm 0.11$, $F = 38.2 > F_{0.05}^{(cr)}(f_1 = 1; f_2 = 11) = 5.59$, $\delta H(Z) = 3.94\%$.*

Table 13. *Carcinogenic properties of some drugs*

N	Chemical compounds	Gross-formula	Activity	*Z*	*H,bits*
1	Clofibrate	$C_{12}H_{15}O_3Cl$	+	$2.84 < Z^{(th)}$	$1.52 \approx H^{(th)}$
2	Dapsone	$C_{12}H_{12}N_2O_2S$	+	$3.03 < Z^{(th)}$	$1.75 < H^{(th)}$
3	Dihydroxymethylfuratrizine*)	$C_{11}H_{11}N_5O_5$	+	$3.44 > Z^{(th)}$	$1.90 > H^{(th)}$
4	Hydralazine	$C_8H_8N_4$	+	$3.00 < Z^{(th)}$	$1.52 < H^{(th)}$
5	Hydralazine hydrochloride*)	$C_8H_8N_4 \cdot HCl$	+	$3.09 > Z^{(th)}$	$1.71 > H^{(th)}$
6	Methoxsalen*)	$C_{12}H_8O$	+	$3.33 > Z^{(th)}$	$1.46 < H^{(th)}$
7	Nafenopin	$C_{20}H_{22}O_3$	+	$2.67 < Z^{(th)}$	$1.29 < H^{(th)}$
8	Phenacetin	$C_{10}H_{13}NO_2$	+	$2.69 < Z^{(th)}$	$1.50 > H^{(th)}$
9	Phenazopyridine	$C_{11}H_{11}N_5$	+	$2.96 < Z^{(th)}$	$1.51 < H^{(th)}$
10	Phenazopyridine hydrochloride	$C_{11}H_{12}ClN_5$	+	$3.03 < Z^{(th)}$	$1.66 > H^{(th)}$
11	Phenelzine	$C_8H_{12}N_2$	+	$2.46 < Z^{(th)}$	$1.32 < H^{(th)}$
12	Phenelzine sulphate	$C_8H_{12}N_2 \cdot H_2SO_4$	+	$2.97 < Z^{(th)}$	$1.50 < H^{(th)}$
13	Phenoxybenzamine hydrochloride	$C_{18}H_{23}Cl_2NO$	+	$2.67 < Z^{(th)}$	$1.47 > H^{(th)}$
14	Proflavine	$C_{13}H_{11}N_3$	+	$2.67 < Z^{(th)}$	$1.39 < H^{(th)}$
15	Proflavine dihydrochloride	$C_{13}H_{11}N_3 \cdot 2HCl$	+	$3.03 < Z^{(th)}$	$1.63 \approx H^{(th)}$
16	Proflavine monohydrochloride	$C_{13}H_{11}N_3 \cdot HCl$	+	$2.97 < Z^{(th)}$	$1.55 < H^{(th)}$
17	Reserpine	$C_{33}H_{40}N_2O_9$	+	$2.81 < Z^{(th)}$	$1.51 < H^{(th)}$
18	Rifampicin	$C_{43}H_{58}N_4O_{12}$	+	$2.75 < Z^{(th)}$	$1.54 < H^{(th)}$
19	Spironolactone	$C_{24}H_{32}O_4S$	+	$2.59 < Z^{(th)}$	$1.37 < H^{(th)}$
20	Doxorubicin	$C_{27}H_{29}NO_{11}$	+	$3.06 < Z^{(th)}$	$1.57 < H^{(th)}$
21	Azacitidine	$C_8H_{12}N_4O_5$	+	$3.24 > Z^{(th)}$	$1.87 > H^{(th)}$
22	Bergaptene	$C_{12}H_7O_4$	+	$3.44 > Z^{(th)}$	$1.45 < H^{(th)}$
23	Procarbazine	$C_{12}H_{19}N_3O$	+	$2.51 < Z^{(th)}$	$1.46 < H^{(th)}$
24	Phenacetin	$C_{10}H_{13}NO_2$	+	$2.69 < Z^{(th)}$	$1.50 < H^{(th)}$
25	Sulfafurazole	$C_{11}H_{13}N_3SO_2$	-	$3.21 > Z^{(th)}$	$1.92 > H^{(th)}$
26	Sulfamethoxazole	$C_{10}H_{11}N_3SO_3$	-	$3.30 > Z^{(th)}$	$1.94 > H^{(th)}$
27	Profilvinum hemisulphate **)	$C_{26}H_{22}N_6SO_4$	-	$3.16 > Z^{(th)}$	$1.75 < H^{(th)}$
28	Chloramphenicol**)	$C_{11}H_{12}Cl_2N_2O_5$	-	$3.44 > Z^{(th)}$	$1.98 > H^{(th)}$

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

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-acetylaminodibenzo-thiophene-5-oxide, 3-acetylaminodibenzo-furan. And all of them turned out to be carcinogens, and 3-acetylaminodibenzothiophene is even more active carcinogen than 2-acetylaminofluorene.

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 9 shows a number of aromatic amines similar in structure to 2-acetylaminofluorene.

Table 14 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 14. *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

Using the data of [7], 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 $Z_{\text{prot.}}^{(\text{th})} = 2.83$ is less than threshold descriptor of carcinogenic agents (Table 6). 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, 8]. For example, thiourea (Z = 3.00, H = 1.75*bits*) has radioprotective properties and at the same time is a carcinogen. Anticarcinogenic properties and selectively action agents on malignant cells have been studied [25]. The following agents have been analyzed: Furfuryl-6-aminopurine (Z =3.20), N^{NH_2} -puryl-6-tryptamine (Z = 3.20), N^{NH_2} -puryl-6-tyramine (Z = 3.00), N^{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 mean values of the characteristic for active carcinogens.

4. 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 [9] 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 electrophilic 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).

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". 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. *Adv. in Biological Chem*., 1, (2011) 1.

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

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

[9] Veljkoviĉ V., Laloviĉ D. *Experientia*, 33 (1977) 1228.

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

[11] L'vovsky E. N. Statistical Methods for Constructing Empirical Formulas. Moscow. High School. 1988.

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

[13] Pustylnik E. I. Statistical Methods of Analysis and Processing of Observations. Moscow. 1968. 288.

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

[15] Förster E., Rönz B. Methoden der Korrelations-und Regressonanalise. Verlag Die Wirtschaft Berlin. 1979.

[16] Urbach V. Yu. Statistical Analysis in Biological and Medical Research. Moscow. Medicine. 1975.

[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.

[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) 1511.

[22] Fukui J. A. et al. *Nippon Kaguki Zasshi*, 82 (1961) 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) 525.

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

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

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