CROATICA CHEMICA ACTA CCACAA **69** (3) 1159–1173 (1996)

ISSN-0011-1643
CCA–2366

Original Scientific Paper

Estimation of the Normal Boiling Points of Haloalkanes Using Molecular Similarity

Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald

Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway, Duluth, MN 55811, USA

Received April 3, 1995; revised November 3, 1995; accepted November 4, 1995

A molecular similarity measure has been used to estimate the normal boiling points of a set of 267 haloalkanes with 1–4 carbon atoms. Molecular similarity/dissimilarity was quantified in terms of Euclidean distances of molecules in the eight dimensional principal component space derived from fifty-nine topological indices. Correlation coefficients between the experimental and estimated boiling points ranged from 0.854 to 0.943 in the *K*-nearest neighbor estimation of boiling points using a different number of nearest neighbors (K = 1–10, 15, 20, 25).

INTRODUCTION

The use of structural analogy as a tool to classify chemicals, as well as predict the behaviour of chemical species, is as old as chemistry. In 1819, Mitscherlich[1] described the phenomenon of isomorphism, in which substitution of one atom by another leads to similar lattice structures. At the turn of this century, Langmuir[2] observed that isosteric chemical species, those which contain the same total number of atoms and electrons, have very similar properties. Members of isosteric pairs, like N_2-CO and N_2O-CO_2, have many similar physical constants.[3] The structural similarity of the isosteric amino acids valine and threonine poses some interesting problems in the protein synthesis mechanism of cells. Being sterically similar, valine and threonine may be charged to the same tRNA. The incorrectly formed aminoacyl adenylate and aminoacyl tRNA are discriminated and destroyed *via* a »double sieve«, involving steric exclusion and ineffective binding, before they are used in protein synthesis.[4]

Similarity plays an important role in biological activity. The enzyme dihydrofolate reductase normally facilitates the reduction of dihydrofolate to tetrahydrofolate. Methotrexate, a compound whose structure is similar to dihydrofolate, inhibits the action of the reductase.[5] Competitive inhibition of enzymes can also result from interaction of the enzyme with transition state analogs of the substrate. For example, proline racemase from *Clostridium sticklandii* preferentially binds the transition state of proline. As a result, the racemase is subject to inhibition by compounds which are structural analogs of the transition state of proline, such as pyrrole-2-carboxylate and pyrroline-2-carboxylate, which bind to the enzyme with a much greater affinity than does proline.[6] Furthermore, the structural similarity between a macromolecular biotarget and its antiidiotypic antibody is believed to be the reason for the use of such antibodies as model receptors in the screening of chemicals for drug discovery.[7]

The last decade has seen an upsurge of interest in the development of similarity measures and their applications in chemical research, drug design, and toxicology.[8–25] Such methods are based on different representations of chemical species, viz., topological, geometrical, quantum chemical, etc. In drug design, similarity searching of databases is used to identify potential leads. Also, dissimilarity based methods are used to select chemicals for screening in the drug discovery process.[11] In toxicology, structural and functional analogy are used to assess the ecological and human health risk of the new and existing chemicals.[26–28]

In the United States, the majority of chemicals submitted to the Environmental Protection Agency (USEPA) for registration do not have any test data.[27] One of the methods used by regulators for the hazard assessment of such chemicals is to select their analogs and, subsequently, estimate the hazard of the chemical of interest from the hazard of the analogs. Such selection of analogs is often done subjectively by individual experts on the basis of an intuitive notion of similarity.[27] In USEPA's approach to ecological risk assessment, class specific QSARs are preferred over the use of analogs, although in human health hazard assessment, analog-based estimation of toxic potential is still the most important factor.[28]

Rapid selection of analogs for drug design and hazard assessment requires automated methods that are computationally feasible. Similarity methods based on parameters that can be calculated directly from molecular structure fall into this category.[8–25] Topological indices derived from a molecular graph comprise a set of parameters which can be computed for any chemical structure.[29]

In some of our recent studies, we have developed novel methods of quantifying molecular similarity using topological indices and substructural features like atom pairs.[13–22] We have applied similarity techniques in the selection of analogs and in the estimation of molecular properties such as

boiling point, lipophilicity, and mutagenicity for different sets of chemicals. In this paper, we have carried out a similarity based estimation of the normal boiling point for a set of 267 chlorofluorocarbons (CFCs) using a similarity method based on topological indices.

DATABASES

The data analyzed in this study consist of the normal boiling points for 267 CFCs with 1–4 carbon atoms. These data were originally collected from Beilstein's *Handbuch der Organischen Chemie*, the *CRC Handbook of Chemistry and Physics*, Heilbron's *Dictionary of Organic Compounds* and Smith and Srivastava's *Thermodynamic Data for Pure Compounds, Part B* for use in several studies by Balaban *et al.*[30,31] For our purposes, the subset of 276 CFC's[30] was further reduced to 267, to remove outliers. Nine compounds whose normal boiling points were more than two standard deviations from the mean boiling point of the group were removed. This was done to enhance the estimation by removing compounds that had only one or two neighbours which would give reasonable estimates of boiling point. Table I is a listing of the compounds used in this study and their normal boiling points.

METHODS

Calculation of Topological Indices

The fifty-nine topological indices used in this study were calculated using POLLY 2.3 which uses the SMILES line notation input of chemical structures.[32] The TIs calculated are listed in Table II and include the Wiener index calculated by the method of Wiener,[33] connectivity indices as calculated by Randić[34] and by Kier and Hall,[35] information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić[36] as well as those of Raychaudhury *et al.*,[37] parameters derived on the neighbourhood complexity of vertices in hydrogen-filled molecular graphs,[38–41] path lengths, and Balaban's J indices.[42–44]

Data Reduction

Initially, all TIs were transformed by the natural log of the TI plus one. The natural logarithm transformation was done because some TIs may be several orders of magnitude greater than others. One was added before the log transformation since many of the TIs may be zero. Principal component analysis (PCA) was used to reduce the dimensionality of the set of 59 topological indices (TIs). With PCA, linear combinations of the TIs, called prin-

TABLE I

Normal boiling points of 267 haloalkanes with 1–4 carbon atoms

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
1	carbon tetrachloride	76.7	33.3	43.4
2	trichloromethane	61.2	28.6	32.6
3	dichloromethane	39.8	7.1	32.7
4	trichlorofluoromethane	23.7	1.3	22.4
5	dichlorofluoromethane	8.9	4.3	4.6
6	chlorofluoromethane	–9.1	3.5	–12.6
7	chloromethane	–24.3	–9.3	–15.0
8	dichlorodifluoromethane	–29.8	6.9	–36.7
9	chlorodifluoromethane	–40.8	–8.4	–32.4
10	difluoromethane	–51.7	12.0	–63.7
11	hexachloroethane	184.4	146.6	37.8
12	1,1,1,2,2-pentachloro-2-fluoroethane	137.9	136.2	1.7
13	1,1,1,2-tetrachloro-2-fluoroethane	117.0	106.9	10.1
14	1,1,2,2-tetrachloro-1-fluroethane	116.6	97.7	18.9
15	1,1,2-trichloroethane	113.7	78.7	35.0
16	1,1,2-trichloro-2-fluoroethane	102.4	68.8	33.6
17	1,1,2,2-tetrachloro-1,2-difluoroethane	92.7	55.0	37.7
18	1,1,1-trichloroethane	74.0	25.2	48.8
19	1,2-dichloro-1-fluoroethane	73.8	71.4	2.4
20	1,1,2-trichloro-1,2-difluoroethane	72.5	60.4	12.1
21	1,2-dichloro-1,2-difluoroethane	58.5	33.9	24.6
22	1,1-dichloroethane	57.2	4.8	52.4
23	1,1,1-trichloro-2,2,2-trifluoroethane	45.8	57.5	–11.7
24	1,2-dichloro-1,1-difluoroethane	46.6	43.8	2.8
25	2-chloro-1,1-difluoroethane	35.1	57.6	–22.5
26	1,1-dichloro-1-fluoroethane	32.0	15.2	16.8
27	2,2-dichloro-1,1,1-trifluoroethane	28.7	42.3	–13.6
28	1-chloro-1-fluoroethane	16.1	32.6	–16.5
29	chloroethane	12.3	7.6	4.7
30	1-chloro-1,1,2-trifluoroethane	12.0	20.0	–8.0
31	2-chloro-1,1,1-trifluoroethane	6.9	21.1	–14.2
32	1,1,2-trifluoroethane	5.0	27.7	–22.7
33	1,2-dichloro-1,1,2,2-tetrafluoroethane	3.6	40.6	–37.0
34	2,2-dichloro-1,1,1,2-tetrafluoroethane	3.6	29.1	–25.5
35	1-chloro-1,1,2,2-tetrafluoroethane	–12.0	24.1	–36.1
36	1,1,2,2-tetrafluoroethane	–22.8	63.6	–86.4
37	1,1-difluoroethane	–25.8	27.8	–53.6
38	1,1,1,2-tetrafluoroethane	–26.1	–1.1	–25.0
39	fluoroethane	–37.8	17.6	–55.4
40	1,1,1-trifluoroethane	–47.3	–8.5	–38.8
41	1,1,1,2,2-pentafluoroethane	–48.3	–14.4	–33.9
42	1,1,2,2,3,3-hexachloropropane	218.5	201.4	17.1
43	1,1,1,2,2,3-hexachloropropane	218.0	199.9	18.1
44	1,1,1,2,3,3-hexachloro-2,3-difluoropropane	196.0	183.8	12.2
45	1,1,1,2,2,3-hexachloro-3,3-difluoropropane	193.4	197.4	–4.0

TABLE I

(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
46	1,1,1,2,2-pentachloro-3,3-difluoropropane	175.0	178.0	–3.0
47	1,1,1,3,3-pentachloro-2,2-difluoropropane	174.0	147.8	26.2
48	1,1,2,2-tetrachloropropane	165.5	173.1	–7.6
49	1,1,3,3-tetrachloropropane	161.9	170.6	–8.7
50	1,2,3-trichloropropane	156.8	153.4	3.4
51	1,1,2,3,3-pentachloro-1,2,3-trifluoropropane	154.7	128.6	26.1
52	1,1,2,2-tetrachloropropane	153.0	151.6	1.4
53	1,1,2,2,3-pentachloro-1,3,3-trifluoropropane	152.3	156.8	–4.5
54	1,1,1,3-tetrachloro-2,2-difluoropropane	151.2	130.7	20.5
55	1,1,1,2-tetrachloropropane	150.4	152.1	–1.7
56	1,1,2,2-tetrachloro-3,3-difluoropropane	147.6	132.1	15.5
57	1,1,3-trichloropropane	145.5	140.9	4.6
58	1,1,2,2-tetrachloro-1,3,3-trifluoropropane	134.6	126.9	7.7
59	1,2,3-trichloro-2-fluoropropane	130.8	111.9	18.9
60	1,1,2,3-tetrachloro-2,3,3-trifluoropropane	129.8	106.8	23.0
61	1,1,3-trichloro-2,2-difluoropropane	127.3	99.2	28.1
62	1,1,2,2-tetrachloro-3,3,3-trifluoropropane	126.2	132.3	–6.1
63	1,1,2-trichloro-2-fluoropropane	116.7	99.2	17.5
64	1,1,3,3-tetrachloro-1,2,2,3-tetrafluoropropane	114.0	105.0	9.0
65	1,1,1,3-tetrachloro-2,2,3,3-tetrafluoropropane	113.9	104.3	9.6
66	1,1,2-trichloro-1-fluoropropane	113.5	99.9	13.6
67	1,1,1,2-tetrachloro-2,3,3,3-tetrafluoropropane	112.5	121.2	–8.7
68	1,1,2,2-tetrachloro-1,3,3,3-tetrafluoropropane	112.3	125.2	–12.9
69	1,2,2,3-tetrachloro-1,1,3,3-tetrafluoropropane	112.2	113.4	–1.2
70	1,1,3-trichloro-1,2,2-trifluoropropane	109.5	87.2	22.3
71	1,1,1-trichloropropane	108.0	105.9	2.1
72	1,2,2-trichloro-3,3,3-trifluoropropane	104.5	120.8	–16.3
73	1,3-dichloro-2,2-difluoropropane	96.7	93.6	3.1
74	1,3,3-trichloro-1,1,2,2-tetrafluoropropane	91.8	90.2	1.6
75	1,2,2-trichloro-1,1-difluoropropane	90.2	86.5	3.7
76	1,2,3-trichloro-1,1,2,3-tetrafluoropropane	90.0	84.5	5.5
77	2,3-dichloro-1,1,2,3-tetrafluoropropane	89.8	72.2	17.6
78	1,2,3-trichloro-1,1,3,3-tetrafluoropropane	88.0	76.7	11.3
79	1-chloro-3-fluoropropane	81.0	101.7	–20.7
80	1,2,3-trichloro-1,1,2,3,3-pentafluoropropane	73.7	65.4	8.3
81	2,3,3-trichloro-1,1,1,2,3-pentafluoropropane	73.4	65.4	8.0
82	1,3,3-trichloro-1,1,2,2,3-pentafluoropropane	73.0	65.5	7.5
83	2,2,3-trichloro-1,1,1,3,3-pentafluoropropane	72.0	80.6	–8.6
84	1,2-dichloro-1,1-difluoropropane	70.0	82.1	–12.1
85	2,2-dichloropropane	69.3	34.3	35.0
86	1-chloro-2-fluoropropane	68.5	70.7	–2.2
87	1,1-dichloro-1-fluoropropane	66.6	77.6	–11.0
88	1-chloro-2,2-difluoropropane	55.1	44.0	11.1
89	1-chloro-1,2-difluoropropane	52.9	72.8	–19.9
90	2,2-dichloro-1,1,1-trifluoropropane	48.8	72.0	–23.2

TABLE I

(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
91	1-chloropropane	46.6	43.8	2.8
92	3,3-dichloro-1,1,1,2,2-pentafluoropropane	45.5	58.8	−13.3
93	1,3-difluoropropane	41.6	88.8	−47.2
94	2-chloropropane	35.7	31.3	4.4
95	1,3-dichloro-1,1,2,2,3,3-hexafluoropropane	35.7	33.8	1.9
96	2-chloro-2-fluoropropane	35.2	25.5	9.7
97	3,3-dichloro-1,1,1,2,2,3-hexafluoropropane	35.0	50.5	−15.5
98	3-chloro-1,1,1,2,2-pentafluoropropane	27.6	42.1	−14.5
99	1-chloro-1,1-difluoropropane	25.4	50.0	−24.6
100	1-chloro-1,1,2,2,3,3-hexafluoropropane	21.0	32.2	−11.2
101	1,1,1,2,3-pentafluoropropane	20.0	20.6	−0.6
102	1,1,2,2,3,3-hexafluoropropane	10.5	11.2	−0.7
103	1,1-difluoropropane	7.5	61.1	−53.6
104	1,1,1,2,3,3-hexafluoropropane	5.0	9.0	−4.0
105	1,1,1,2,2,3-hexafluoropropane	1.2	5.3	−4.1
106	2-chloro-1,1,1,2,3,3,3-heptafluoropropane	−2.0	12.6	−14.6
107	2-fluoropropane	−9.7	18.1	−27.8
108	1,1,1-trifluoropropane	−12.5	7.1	−19.6
109	1,1,1,2,3,3,3-heptafluoropropane	−19.0	20.6	−39.6
110	1-chloro-2-fluoroethane	53.0	48.7	4.3
111	1-chloro-1,1-difluoroethane	−9.8	−1.9	−7.9
112	1-chloro-1,1,2,2,2-pentafluoroethane	−38.0	4.8	−42.8
113	1,2-dichloroethane	83.5	50.5	33.0
114	1,1-dichloro-2,2-difluoroethane	60.0	57.1	2.9
115	1,1-dichloro-1,2,2-trifluoroethane	30.2	41.3	−11.1
116	1,2-dichloro-1,1,2-trifluoroethane	28.2	41.7	−13.5
117	1,1,2-trichloro-1-fluoroethane	88.5	56.5	32.0
118	1,1,1-trichloro-2,2-difluoroethane	73.0	68.1	4.9
119	1,1,2-trichloro-2,2-difluoroethane	71.2	62.3	8.9
120	1,1,2-trichloro-1,2,2-trifluoroethane	47.6	47.4	0.2
121	1,1,1,2-tetrachloroethane	130.5	98.3	32.2
122	1,1,2,2-tetrachloroethane	146.3	61.7	84.6
123	1,1,1,2-tetrachloro-2,2-difluoroethane	91.6	102.0	−10.4
124	1,1,1,2,2-pentachloroethane	161.9	149.5	12.4
125	1,1,2,2,3-pentachloropropane	196.0	193.0	3.0
126	1,1,2,3,3-pentachloropropane	199.0	184.4	14.6
127	1,1,2,2,3-pentachloro-3,3-difluoropropane	168.4	148.0	20.4
128	1,1,2,3,3-pentachloro-1,3-difluoropropane	167.4	130.5	36.9
129	1,1,1,2,2-pentachloro-3,3,3-trifluoropropane	153.0	172.7	−19.7
130	1,1,1,2,3-pentachloro-2,3,3-trifluoropropane	153.3	145.2	8.1
131	1,1,1,3,3-pentachloro-2,2,3-trifluoropropane	153.0	137.3	15.7
132	1,1,1,2,3,3-hexachloropropane	217.0	207.2	9.8
133	1,1,1,3,3,3-hexachloropropane	206.0	151.7	54.3
134	1,1,1,2,2,3-hexachloro-3-fluoropropane	210.0	201.5	8.5
135	1,1,1,2,3,3-hexachloro-3-fluoropropane	207.0	178.7	28.3

TABLE I

(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
136	1,1,2,2,3,3-hexachloro-1-fluoropropane	210.0	185.3	24.7
137	1,1,1,3,3,3-hexachloro-2,2-difluoropropane	194.2	148.5	45.7
138	1,1,2,2,3,3-hexachloro-1,3-difluoropropane	194.2	181.4	12.8
139	1,2-dichloro-1,1,2,3,3-pentafluoropropane	56.3	65.1	–8.8
140	2,3-dichloro-1,1,1,2,3-pentafluoropropane	56.0	65.6	–9.6
141	1,1,2-trichloropropane	133.0	94.7	38.3
142	1,2,2-trichloropropane	122.0	103.1	18.9
143	1,1,1-trichloro-2,2-difluoropropane	102.0	75.0	27.0
144	1,2,2-trichloro-1,1,3,3-tetrafluoropropane	92.0	99.4	–7.4
145	3,3,3-trichloro-1,1,1,2,2-pentafluoropropane	70.5	89.4	–18.9
146	1,1,2-trichloro-1,2-difluoropropane	97.7	74.0	23.7
147	1,1,3-trichloro-3,3-difluoropropane	107.8	70.3	37.5
148	3-chloro-1,1,1,3,3-pentafluoropropane	28.4	63.7	–35.3
149	2-chloro-1,1,1,3,3,3-hexafluoropropane	15.5	20.3	–4.8
150	3-chloro-1,1,1,2,2,3,3-heptafluoropropane	–2.5	15.0	–17.5
151	3-chloro-1,1,1,2,2,3-hexafluoropropane	20.0	22.2	–2.2
152	1,1-dichloropropane	88.1	91.8	–3.7
153	1,2-dichloropropane	96.0	90.3	5.7
154	1,3-dichloropropane	120.8	105.4	15.4
155	1,2-dichloro-2-fluoropropane	88.6	57.5	31.1
156	1,2-dichloro-1-fluoropropane	93.0	64.8	28.2
157	1,1-dichloro-2,2-difluoropropane	79.0	80.3	–1.3
158	1,3-dichloro-1,1-difluoropropane	80.8	79.6	1.2
159	1,1-dichloro-1,2,2-trifluoropropane	60.2	62.4	–2.2
160	3,3-dichloro-1,1,1-trifluoropropane	72.4	81.3	–8.9
161	1,2-dichloro-1,1,2-trifluoropropane	55.6	63.1	–7.5
162	2,3-dichloro-l,1,1-trifluoropropane	76.7	99.6	–22.9
163	1,3-dichloro-1,1,2,2-tetrafluoropropane	68.2	71.8	–3.6
164	2,3-dichloro-1,1,1,3,3-pentafluoropropane	50.4	70.3	–20.0
165	2,3-dichloro-1,1,1,2,3,3-hexafluoropropane	34.7	40.0	–5.3
166	1,2,3-trichloro-1,1-difluoropropane	114.3	102.0	12.3
167	1,1,1-trichloro-3,3,3-trifluoropropane	95.1	109.6	–14.5
168	1,1,2-trichloro-3,3,3-trifluoropropane	106.8	103.0	3.8
169	2,3,3-trichloro-1,1,1,3-tetrafluoropropane	87.2	91.4	–4.2
170	1,1,3-trichloro-1,2,2,3-tetrafluoropropane	90.5	109.6	–19.1
171	1,1,1,3-tetrachloropropane	158.0	142.6	15.4
172	1,1,2,3-tetrachloropropane	180.0	156.6	23.4
173	1,1,1,2-tetrachloro-2-fluoropropane	139.6	113.5	26.1
174	1,1,2,2-tetrachloro-1-fluoropropane	135.0	114.4	20.6
175	1,1,1,3-tetrachloro-3,3-difluoropropane	132.0	112.1	19.9
176	1,1,1,2-tetrachloro-3,3,3-trifluoropropane	125.1	141.4	–16.3
177	1,1,2,3-tetrachloro-1,3,3-trifluoropropane	128.7	111.9	16.8
178	1,1,3,3-tetrachloro-2,2,3-trifluoropropane	127.0	105.6	21.4
179	1,1,2,3-tetrachloro-1,2,3,3-tetrafluoropropane	112.5	97.0	15.5
180	1-fluoropropane	–2.3	24.7	–27.0

TABLE I

(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
181	octafluoropropane	–38.0	–0.5	–37.5
182	2,2-difluoropropane	–0.5	11.3	–11.8
183	1,1,1,3-tetrafluoropropane	29.4	23.3	6.1
184	1,1,1,3,3,3-hexafluoropropane	0.8	18.2	–17.4
185	1,1,1,2,2,3,3-heptafluoropropane	–17.0	11.7	–28.7
186	1-chloro-1-fluoropropane	48.0	74.8	–26.8
187	3-chloro-1,1,1-trifluoropropane	45.1	65.1	–19.0
188	2-chloro-1,1-difluoropropane	52.0	74.7	–22.7
189	2-chloro-1,1,1-trifluoropropane	30.0	47.7	–17.7
190	1-fluorobutane	32.2	66.7	–34.5
191	2-fluorobutane	24.7	46.3	–21.6
192	1,1,1,2,2,4,4,4-octafluorobutane	18.0	59.5	–41.5
193	1,1,2,2,3,3,4,4-octafluorobutane	43.0	38.1	4.9
194	1,1,1,2,2,3,3,4,4-nonafluorobutane	14.0	55.4	–41.4
195	decafluorobutane	–2.0	43.0	–45.0
196	1-chlorobutane	78.5	90.1	–11.6
197	2-chlorobutane	68.5	58.6	9.9
198	1-chloro-4-fluorobutane	115.0	109.7	5.3
199	1-chloro-l,1-difluorobutane	55.5	56.1	–0.6
200	3-chloro-1,1,1-trifluorobutane	66.0	72.3	–6.3
201	1-chloro-1,1,3,3-tetrafluorobutane	70.5	78.0	–7.5
202	2-chloro-1,1,1,3,3,3-hexafluorobutane	51.0	74.2	–23.2
203	4-chloro-1,1,1,2,2,3,3-heptafluorobutane	54.0	49.3	4.7
204	4-chloro-1,1,1,2,2,3,3,4,4-nonafluorobutane	30.0	45.0	–15.0
205	1,1-dichlorobutane	115.0	145.0	–30.0
206	1,2-dichlorobutane	123.5	150.2	–26.7
207	1,3-dichlorobutane	133.0	140.8	–7.8
208	1,4-dichlorobutane	155.0	130.8	24.2
209	1,3-dichloro-1,1,3-trifluorobutane	129.0	80.9	48.1
210	3,4-dichloro-1,1,1,2,2,3-hexafluorobutane	72.0	78.5	–6.5
211	1,4-dichloro-1,1,3-trifluorobutane	118.5	97.5	21.0
212	2,3-dichloro-1,1,1,4,4,4-hexafluorobutane	78.0	71.9	6.1
213	4,4-dichloro-1,1,1,2,2,3,3-heptafluorobutane	76.5	72.0	4.5
214	4,4-dichloro-1,1,1,2,2,3,3,4-octafluorobutane	62.8	71.1	–8.3
215	3,4-dichloro-1,1,1,2,2,3,4,4-octafluorobutane	66.0	70.6	–4.6
216	1,4-dichloro-1,1,2,2,3,3,4,4-octafluorobutane	64.0	71.3	–7.3
217	2,2-dichloro-1,1,1,3,3,4,4,4-octafluorobutane	64.0	70.5	–6.5
218	2,3-dichloro-1,1,1,2,3,4,4,4-octafluorobutane	64.0	76.3	–12.3
219	1,1,1-trichlorobutane	133.5	137.9	–4.4
220	1,1,2-trichlorobutane	156.8	143.5	13.3
221	1,1,3-trichlorobutane	153.8	143.3	10.5
222	1,1,4-trichlorobutane	183.8	133.3	50.5
223	2,2,3-trichloro-1,1,1,4,4,4-hexafluorobutane	104.0	107.7	–3.7
224	4,4,4-trichloro-1,1,1,2,2,3,3-heptafluorobutane	96.5	85.4	11.1
225	1,3,4-trichloro-1,1,2,2,3,4,4-heptafluorobutane	99.0	91.3	7.7

TABLE I

(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
226	2,2,3-trichloro-1,1,1,3,4,4,4-heptafluorobutane	97.4	77.7	19.7
227	1,1,4,4-tetrachlorobutane	200.0	148.7	51.3
228	1,2,4,4-tetrachloro-1,1,2,3,3,4-hexafluorobutane	134.0	92.6	41.4
229	1,2,3,4-tetrachloro-1,1,2,3,4,4-hexafluorobutane	134.0	85.2	48.8
230	1,1,2,3,4,4-hexachloro-1,2,3,4-tetrafluorobutane	208.0	113.2	94.8
231	1-chloroisobutane	68.3	60.5	7.8
232	2-chloroisobutane	50.7	38.5	12.2
233	1-chloro-1-fluoroisobutane	82.5	109.8	–27.3
234	1,1-dichloroisobutane	105.0	107.4	–2.4
235	1,2-dichloroisobutane	106.5	99.1	7.4
236	1,3-dichloroisobutane	136.0	134.6	1.4
237	1,1-dichloro-1-fluoroisobutane	107.0	116.1	–9.1
238	1,2,3-trichloroisobutane	163.0	146.0	17.0
239	1,1,2,3-tetrachloroisobutane	191.0	185.2	5.8
240	1,2,3-trichloro-2-chloromethylpropane	211.0	183.3	27.7
241	1,1,2,3-tetrachloro-2-chloromethylpropane	227.0	204.3	22.7
242	1-fluoroisobutane	16.0	56.1	–40.1
243	2-fluoroisobutane	12.0	38.1	–26.1
244	1,1,1,3,3,3-hexafluoroisobutane	21.5	25.5	–4.0
245	1,1,1,3,3,3-hexafluoro-2-fluoromethylpropane	40.0	18.9	21.1
246	1,1,1,3,33-hexafluoro-2-difluoromethylpropane	33.0	20.3	12.7
247	1,1,1,3,3,3-hexafluoro-2-trifluoromethylpropane	12.0	6.0	6.0
248	decafluoroisobutane	–0.3	3.6	–3.9
249	3-chloro-1,1,1,3,3-pentafluoroisobutane	59.0	68.6	–9.6
250	1,1,1,3,3,3-hexafluoro-2-chloromethylpropane	58.0	39.6	18.4
251	2,3-dichloro-1,1,1-trifluoroisobutane	93.5	101.0	–7.5
252	2,3-dichloro-1,1,1,3,3-pentafluoroisobutane	75.3	91.2	–15.9
253	2,3-dichloro-1,1,1,3,3-pentafluoro-2-trifluoromethylpropane	65.0	65.8	–0.8
254	1,1,2-trichloroisobutane	163.0	143.1	19.9
255	1,2,3-trichloro-1,1-difluoroisobutane	132.0	114.7	17.3
256	2,3,3-trichloro-1,1,1-trifluoroisobutane	123.7	124.6	–0.9
267	1,1,1,3,3,3-hexafluoro-2-trichloromethylpropane	107.0	106.7	0.3
258	1,1,1,2-tetrachloro-3,3,3-trifluoroisobutane	148.5	148.7	–0.2
259	1,1,1,2,3-pentachloroisobutane	211.0	201.3	9.7
260	1-chloro-1,1,2,2-tetrafluoropropane	19.9	27.5	–7.6
261	1,1,1-trichloropropane	104.0	99.6	4.4
262	2,3-dichlorobutane	116.0	105.2	10.8
263	2,2,3-trichlorobutane	143.0	152.3	–9.3
264	1,2,3-trichlorobutane	166.0	141.7	24.3
265	1,4-difluorobutane	77.8	122.0	–44.2
266	2,2-difluorobutane	30.9	40.0	–9.1
267	1,2-difluoroethane	26.0	40.7	–14.7

TABLE II

Topological index symbols and definitions

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I_D^W}$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by the frequency of occurrences of distance h
O	Order of neighbourhood when IC_r reaches its maximum value for the hydrogen-filled graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighbourhood of vertices
O_{ORB}	Maximum order of neighbourhood of vertices for I_{ORB} within the hydrogen-suppressed graph
M_1	A Zagreb group parameter = sum of the square of degree over all vertices
M_2	A Zagreb group parameter = sum of the cross-product of degrees over all neighbouring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0–3) order neighbourhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for the r^{th} (r = 0–3) order neighbourhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for the r^{th} (r = 0–3) order neighbourhood of vertices in a hydrogen-filled graph
${}^h\chi$	Path connectivity index of the order h = 0–5
${}^h\chi_C$	Cluster connectivity index of the order h = 3–6
${}^h\chi_{PC}$	Path-cluster connectivity index of the order h = 4–6
${}^h\chi^v$	Valence path connectivity index of the order h = 0–5
${}^h\chi_C^v$	Valence cluster connectivity index of the order h = 3–6
${}^h\chi_{PC}^v$	Valence path-cluster connectivity index of the order h = 4–6
P_h	Number of paths of length h = 0–5
J	Balaban's J index based on distance
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

cipal components (PCs) are derived from the correlation matrix. The first PC has the largest variance, or eigenvalue, of the linear combination of TIs. Each subsequent PC explains the maximal index variance orthogonal to previous PCs. With 59 TIs available, 59 PCs can be generated. For this study, PCs with an eigenvalue greater than one were retained. The PCA analysis and selection of PCs was accomplished using the SAS procedure PRINCOMP.[45] Basak *et al.*[13] provide more detail on this approach.

Computation of Similarity

Intermolecular similarity was measured by the Euclidean distance (ED) within an n-dimensional space. This n-dimensional space consisted of orthogonal variables (PCs) derived from the TIs. ED between the molecule's i and j is defined as:

$$ED_{ij} = \left[\sum_{k=1}^{n} (D_{ik} - D_{jk})^2 \right]^{1/2}$$

where n equals the number of dimensions retained from PCA. D_{ik} and D_{jk} are the data values of the k^{th} dimension for chemicals i and j, respectively.

K-nearest Neighbour Selection and Boiling Point Estimation

Following the quantification of the intermolecular similarity of the CFCs, the K-nearest neighbours (K = 1–10, 15, 20, 25) were determined on the basis of ED. The mean observed boiling point of the K-nearest neighbours for a compound was used as the estimated boiling point and the standard error (s.e.) of the estimates were used to assess the efficacy of this similarity method.

RESULTS

From the PCA of 59 TIs for 267 CFCs, eight PCs with eigenvalues greater than one were retained. These eight PCs explained, cumulatively, 95.0% of the total variation within the TI data. Table III lists the eigenvalues of the eight PCs, the proportion of variance explained by each PC, and the cumulative variance explained. In addition, Table III lists the two TIs most correlated with each PC. The first PC is strongly correlated with the parameters that characterize the size of the molecular graphs and the increasing number of chlorofluoro substitutions, *viz.* P_0 (number of atoms) and P_1 (number of bonds). The second PC is highly correlated with higher order complexity indices including SIC_2 and CIC_2. For the third PC, the highest

TABLE III

Summary of the principal components of 59 TIs for the 267 haloalkanes and the correlation coefficients of the two most correlated with each principal component

PC	Eigenvalue	Percent of variance	Cumulative percent	First correlated TI		Second correlated TI	
1	31.9	54.0	54.0	P_0	0.982	P_1	0.982
2	8.7	14.8	68.8	SIC_2	0.949	CIC_2	–0.922
3	5.2	8.8	77.6	$^4\chi^v_C$	–0.668	$^3\chi^v_C$	–0.637
4	3.6	6.1	83.7	$^1\chi^v$	0.475	$^3\chi^v$	0.440
5	2.1	3.6	87.3	$^1\chi^v$	0.495	$^4\chi^v$	0.482
6	1.9	3.2	90.5	P_5	0.579	$^5\chi$	0.574
7	1.5	2.5	93.0	$^2\chi^v$	0.282	$^3\chi^v_C$	0.280
8	1.2	2.0	95.0	$^4\chi_C$	0.324	H^v	–0.313

correlations occur with the valence cluster connectivity TIs such as $^4\chi^v_C$ and $^3\chi^v_C$. The fourth PC was characterized by lower order valence path connectivity indices such as $^1\chi^v$ and $^3\chi^v$ and the fifth PC by the higher order valence path connectivity indices such as $^5\chi^v$ and $^4\chi^v$. Interpretation beyond the fifth level PC becomes more difficult, as it can be seen in Table III. These PC/TI correlations agree with our expectations based on previous research.[16,17,19,20] Generally, PCs and TIs correlate as follows: PC_1 with the size of the molecular graph, PC_2 with higher order complexity indices, PC_3 with cluster and path-cluster connectivity, and PC_4 with low order information theoretic indices.

TABLE IV

Summary of the *K*-nearest neighbour normal boiling point estimation for 267 chlorofluorohydrocarbons

K	*r*	s.e. (°C)
1	0.854	33.2
2	0.908	26.4
3	0.923	24.5
4	0.927	24.2
5	0.933	23.7
6	0.934	24.3
7	0.934	24.3
8	0.936	24.3
9	0.939	24.4
10	0.939	24.7
15	0.936	26.2
20	0.936	27.7
25	0.943	28.0

Table IV reports the correlation and standard errors of boiling point estimates obtained by the *K*-nearest neighour estimation with the observed boiling point values. Each line of the table represents a different *K* level. The standard error for estimation was at its minimum of 23.7 °C for $K = 5$. The correlation, however, continued an upward trend as *K* increased.

DISCUSSION

The goal of this paper was to investigate the usefulness of general similarity methods based on graph invariants in the prediction of the boiling points of a set of 267 chlorofluorocarbons. To this end, we used Euclidean distance in an eight dimensional PC-space as the measure of structural similarity/dissimilarity of CFCs. The results in Table IV show that the best estimates of the normal boiling point are obtained at $K = 5$. Our previous studies on similarity-based prediction of properties like lipophilicity,[17], boiling point,[16, 19] and mutagenicity[16, 19, 20] have shown that a small number of neighbours ($K = 5–10$) will usually give the best results in property estimation.

Comparison of the K-nearest neighbour estimates reported in this paper with previous studies on the same set of CFCs shows that similarity-based estimates are inferior to predictions derived by neural net models.[30] In the neural net model, parametrization was done with an eye to specific structural features of CFCs. In contrast, the PC-based similarity approach used a set of general structural parameters which quantify such structural features of chemical graphs as size, shape, degree of branching, *etc*. Yet, similarity methods based on such graph theoretic parameters give a reasonably good estimate of the normal boiling point of CFCs analyzed in this paper. The usefulness of the similarity approach depends on the context, *i.e*. what level of accuracy is required.

In risk assessment, molecular similarity is used in the selection of analogs of chemicals for hazard estimation. Very often, one has to do rapid estimation of a large number of properties. Such estimations should be based on parameters that can be algorithmically derived, *i.e*., can be computed for any chemical species directly from structure. The graph invariants used in this paper fall into this category. The results reported here show that such methods can be used as a first order estimation of properties.

The parameters used in this paper did not include any stereoelectronic property that might influence the normal boiling points of CFCs. It would be interesting to see whether similarity methods give better estimates of boiling points when stereoelectronic variables are included in the set of parameters. Such studies are in progress and will be reported subsequently.

Acknowledgments. – This paper is the contribution number 150 from the Center for Water and the Environment of the Natural Resources Research Institue. Research reported in this paper was supported, in part, by grant F-49620-94-1-0401 from the United States Air Force, Exxon Biomedical Sciences, Inc. and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute at the University of Minnesota. The authors would also like to extend their thanks to Professor A. T. Balaban of the Polytechnic University of Bucharest, Romania, for his helpful discussions.

REFERENCES

1. E. Mitscherlich, *Abhandl. Akad. Wiss. Berlin*, (1819) 427.
2. I. Langmuir, *J. Am. Chem. Soc.* **41** (1919) 1543.
3. T. Moeller, *Inorganic Chemistry*, John Wiley & Sons, New York 1852.
4. A. Cornish-Bowden and C. W. Wharton, *Enzyme Kinetics*, D. Rickwood (Ed.), IRL Press, Oxford, UK, 1988.
5. P. Calabresi and B. A. Chabner, *Antineoplastic Agents*, in: *The Pharmacological Basis of Therapeutics*, A. G. Gilman, T. W. Rall, A. S. Nies, and P. Taylor (Eds.), Eighth Edition, Pergamon Press, New York, 1990.
6. D. Voet and J. G. Voet, *Biochemistry*, John Wiley & Sons, New York, 1990.
7. J. Couraud, E. Escher, D. Regoli, V. Imhoff, B. Rossignol, and P. Pradelles, *J. Bio. Chem.* **260** (1985) 9461.
8. R. E. Carhart, D. H. Smith, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **25** (1985) 64.
9. C. L. Wilkins and M. Randić, *Theor. Chim. Acta (Berl.)* **58** (1980) 45.
10. D. H. Rouvray, *The evolution of the concept of molecular similarity*, in: *Concepts and Applications of Molecular Similarity*, M. A. Johnson and G. M. Maggiora (Eds.), John Wiley & Sons, New York, 15–42 (1990).
11. M. S. Lajiness, *Molecular similarity-based methods for selecting compounds for screening*, in: *Computational Chemical Graph Theory*, D. H. Rouvray (Ed.), Nova, New York, 299–316 (1990).
12. *Concepts and Applications of Molecular Similarity*, M. A. Johnson and G. M. Maggiora (Eds.), John Wiley & Sons, New York, 1990.
13. S. C. Basak, V. R. Magnuson, G. J. Niemi, and R. R. Regal, *Discrete Appl. Math.* **19** (1988) 17.
14. M. Johnson, S. C. Basak, and G. Maggiora, *Math. and Comp. Modelling* **II** (1988) 630.
15. S. C. Basak, S. Bertelsen, and G. Grunwald, *J. Chem. Inf. Comput. Sci.* **34** (1994) 270.
16. S. C. Basak and G. D. Grunwald, *SAR and QSAR Environ. Res.* **2** (1994) 289.
17. S. C. Basak and G. D. Grunwald, *New. J. Chem.* **19** (1995) 231.
18. S. C. Basak and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **35** (1995) 366.
19. S. C. Basak and G. D. Grunwald, *SAR and QSAR Environ. Res.* **3** (1995) 265.
20. S. C. Basak and G. D. Grunwald, *Chemosphere* **31** (1995) 2529.
21. S. C. Basak, S. Bertelsen and G. D. Grunwald, *Toxicol. Lett.* **79** (1995) 239.
22. S. C. Basak and G. D. Grunwald, *Math. Modelling and Sci. Comput.*, in press.
23. P. Willet and V. Winterman, *Quant. Struct.-Act. Relat.* **5** (1986) 18.
24. W. Fisanick, K. P. Cross, and A. Rusinko, III, *J. Chem. Inf. Comput. Sci.* **32** (1992) 664.
25. P. E. Bowen-Jenkins, D. L. Cooper, and W. G. Richards, *J. Phys. Chem.* **89** (1985) 2195.
26. J. C. Arcos, *Environ. Sci. Tech.* **21** (1987) 743.
27. C. M. Auer, J. V. Nabholz, and K. P. Baetcke, *Environ. Health Perspect.* **87** (1990) 183.
28. C. M. Auer, M. Zeeman, J. V. Nabholz, and R. G. Clements, *SAR and QSAR Environ. Res.* **2** (1994) 29.
29. N. Trinajstić, *Chemical Graph Theory*, Second Edition, CRC Press, Inc., Boca Raton, Florida, 1992.
30. A. T. Balaban, S. C. Basak, T. Colburn, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1118.
31. A. T. Balaban, N. Joshi, L. B. Kier, and L. H. Hall, *J. Chem. Inf. Comput. Sci.* **32** (1992) 233.

32. S. C. Basak, D. K. Harriss, and V. R. Magnuson, POLLY: Copyright of the University of Minnesota, 1988.
33. H. J. Wiener, *J. Am. Chem. Soc.* **69** (1947) 17.
34. M. Randić, *J. Am. Chem. Soc.* **97** (1975) 6609.
35. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth, Hertfordshire, UK 1986.
36. D. Bonchev and N. Trinajstić, *J. Chem. Phys.* **67** (1977) 4517.
37. C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, and S. C. Basak, *J. Comput. Chem.* **5** (1984) 581.
38. A. B. Roy, S. C. Basak, D. K. Harriss, and V. R. Magnuson, *Neighborhood complexities and symmetry of chemical graphs, and their biological applications*, in. *Math. Modelling in Sci. and Tech.*, X. J. R. Avula, R. E. Kalman, A. I. Liapis, and E. Y. Rodin (Eds.), Pergamon Press, New York, 745–750 (1984).
39. S. C. Basak, A. B. Roy, and J. J. Ghosh, *Proc. IInd. Int. Conf. on Math. Modelling* **II** (1980) 851.
40. S. C. Basak and V. R. Magnuson, *Arzneim.-Forsch.* **33** (1983) 501.
41. R. Sarkar, A. B. Roy, and P. K. Sarkar, *Math Biosci.* **39** (1978) 299.
42. A. T. Balaban, *Chem. Phys. Lett.* **89** (1982) 399.
43. A. T. Balaban, *Pure & Appl. Chem.* **55** (1983) 199.
44. A. T. Balaban, *MATCH* **21** (1986) 115.
45. SAS Institute Inc., In *SAS/STAT User's Guide, Release 6.03 Edition*, SAS Institute Inc., Cary, NC, Chapter 34 (1988) 949–965.

SAŽETAK

Procjena normalnih vrelišta haloalkana na osnovi molekulske sličnosti

Subhash C. Basak, Brian D. Gute i Gregory D. Grunwald

Molekulska sličnost upotrijebljena je za procjenu normalnih vrelišta skupa od 276 haloalkana s 1 do 4 ugljikova atoma. Molekulska sličnost/različitost kvantificirana je Euklidovom udaljenošću molekula u osmerodimenzijskom prostoru glavnih komponenti izvedenih iz 59 topoloških indeksa. Koeficijent korelacije između eksperimentalnih i procijenjenih vrelišta iznosi između 0.854 i 0.943 za procjene vrelišta pomoću K najbližih susjeda, uz različite brojeve najbližih susjeda (K = 1, ..., 10, 15, 20, 25).