Performance of Kier-Hall E-states descriptors in QSAR of multi-functional molecules

Darko Butina
ChemoMine Consultancy

Kier-Hall E-state descriptors

- Pharm.Res. **1990**, 7, 801-807
- JCICS 1991, 31, 76-82
- $I=(\delta^V+1)/\delta$
  - δ^V and δ are counts of valence and sigma electrons of atoms associated with the molecular skeleton
- $S_i=I_i+\Delta I_i$
  - E-state value, S_i, for skeletal atom I
- ΔIi, is given as $\Sigma(I_i-I_j)/r_{ij}^2$

Intrinsic-State Values

Table I. Intrinsic-State Values

atom (skeletal hydride group)	intrinsic-state value[a]
)C(	1.250
)CH–	1.333
$-CH_2-$	1.500
)C=	1.667
$-CH_3$, =CH–,)N–	2.000
≡C–, –NH–	2.500
$=CH_2$, =N–	3.000
–O–	3.500
≡CH, $-NH_2$	4.000
=NH	5.000
≡N, –OH	6.000
=O	7.000
–F	8.000
–Cl	4.111
–Br	2.750
–I	2.120
=S	3.667
–SH	3.222
–S–	1.833

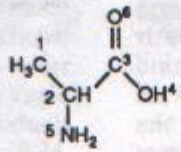
[a] Calculated from eq 2.

Kier-Hall Atom Types

RowNo	atom-types-Kier-Hal	RowNo	atom-types-Kier-Hal
1	sOH	19	ddssS
2	dO	20	sF
3	ssO	21	sCl
4	aaO	22	sBr
5	sNH2	23	sI
6	dNH	24	sCH3
7	ssNH	25	ssCH2
8	aaNH	26	dCH2
9	tN	27	sssCH1
10	dsN	28	dsCH1
11	aaN	29	tCH
12	sssN	30	aaCH
13	ddsN	31	aasC
14	ssssN+	32	ddC
15	sSH	33	tsC
16	dS	34	dssC
17	ssS	35	sssC
18	aaS		

Kier-Hall Algorithm

Table II. Electrotopological-State Calculations for Alanine

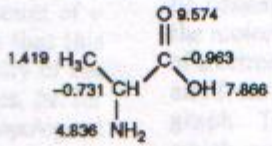


intrinsic values	intrinsic values	intrinsic values
$I(1) = 2.000$ $I(2) = 1.333$	$I(3) = 1.667$ $I(4) = 6.000$	$I(5) = 4.000$ $I(6) = 7.000$

$(I_i - I_j)/r_{ij}^2$ Matrix

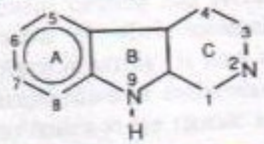
i \ j	1	2	3	4	5	6	ΔI = row sum
1	0.0	0.1667	0.0370	−0.2500	−0.2222	−0.3125	−0.5810
2	−0.1667	0.0	−0.083	−0.5185	−0.6667	−0.6296	−2.0648
3	−0.0370	0.0833	0.0	−1.0833	−0.2593	−1.3333	−2.6296
4	0.2500	0.5185	1.0833	0.0	0.1250	−0.1111	1.8657
5	0.2222	0.6667	0.2593	−0.1250	0.0	−0.1875	0.8356
6	0.3125	0.6296	1.3333	0.1111	0.1875	0.0	2.5741
							0.0000

$$S_i = I_i + \Delta I_i$$



QSAR example 1

Table 6. Binding of beta-carbolines to the benzodiazepine receptor and electrotopological state values

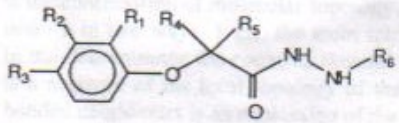


Obs	R_1	R_3	R_4[a]	R_5	R_6	R_7	S(CO)[b]	S(NN)	$S(C_{14})$	pIC_{50}[c]	Calc[d]	Res[e]
1	CH_3	H	H	oH	H	H	0.930	3.828	1.555	−0.450	−1.166	0.716
2	CH_3	H	H	H	H	OCH_3	0.921	3.831	1.523	−1.980	−1.228	−0.752
3	CH_3	H	H	H	H	OH	0.902	3.748	1.480	−1.900	−2.505	0.605
4	H	CH_2OH	H	H	H	H	4.867	3.712	1.845	1.591	1.972	−0.381
5	H	$COOCH_3$	H	H	H	H	5.872	3.658	1.702	2.097	1.698	0.399
6	H	$COOC_2H_5$	H	H	H	H	5.997	3.688	1.719	2.155	2.277	−0.122
7	H	$COOC_3H_7$	H	H	H	H	6.082	3.711	1.731	1.921	2.702	−0.781
8	H	$COOCH_3$	C_2H_5	H	OCH_3	OCH_3	6.162	3.787	1.253	2.400	2.146	0.254
9	H	$COOC_2H_5$	CH_2OCH_3	$OCH_2C_6H_5$	H	H	6.360	3.850	1.170	2.959	2.875	0.084
10	H	$COOC_2H_5$	CH_3	O-iC_3H_7	H	H	6.225	3.800	1.236	2.960	2.319	0.641
11	H	$COONHCH_3$	H	H	H	H	5.934	3.661	1.685	1.780	1.737	0.043
12	CH_3	H_2	H_2	H	H	OCH_3	0.451	3.947	1.068	−2.813	−1.708	−1.105
13	CH_3	H_2	H_2	H	H	OH	0.434	3.864	1.023	−2.748	−2.969	0.221
14	H	$COOC_2H_5$	CH_2OCH_3	H	$OCH_2C_6H_5$	H	6.335	3.836	1.181	2.959	2.703	0.256
15	H	H	H	H	H	H	0.911	3.696	1.946	−1.000	−1.587	0.587
16	H	H	OOCH	H	H	H	5.448	3.530	1.574	−1.491	−0.825	−0.666

a. In compounds 12 and 13 the C ring is not aromatic; all others are.
b. $S(CO) = (S(C_3) + S(O))/2$; S(O) is the E-state value for the oxygen attached to the carbon atom in the substituent on position 3. See text. $S(NN) = (S(N_2) + S(N_9))/2$. See text. $S(C_{14}) = (S(C_1) + S(C_4))/2$. See text.
c. The negative log of the IC_{50} value for binding.
d. Calc is the value computed for pIC_{50} from Eq. 9.
e. Res = Calc − pIC_{50}.

QSAR example 2

Table 5. Hydrazide monoamine oxidase inhibitors and their electrotopological state values



Obs	R_1	R_2	R_3	R_4	R_5	R_6	$S(CH_{14})^a$	$S(NH_{11})^b$	pIC_{50}^c	Calc[d]	Res[e]
1	H	H	H	H	H	$CH(CH_3)_2$	1.943	2.538	5.42	5.42	0.00
2	Cl	H	H	H	H	$CH(CH_3)_2$	1.926	2.515	5.60	5.66	−0.06
3	H	Cl	H	H	H	$CH(CH_3)_2$	1.929	2.520	5.40	5.61	−0.21
4	H	H	Cl	H	H	$CH(CH_3)_2$	1.931	2.524	5.96	5.57	0.39
5	CH_3	H	H	H	H	$CH(CH_3)_2$	1.947	2.558	5.54	5.22	0.32
6	H	CH_3	H	H	H	$CH(CH_3)_2$	1.946	2.553	5.05	5.27	−0.22
7	H	H	CH_3	H	H	$CH(CH_3)_2$	1.946	2.550	5.40	5.30	0.10
8	OCH_3	H	H	H	H	$CH(CH_3)_2$	1.932	2.535	5.62	5.47	0.15
9	H	OCH_3	H	H	H	$CH(CH_3)_2$	1.934	2.536	5.42	5.45	−0.03
10	H	H	OCH_3	H	H	$CH(CH_3)_2$	1.935	2.536	5.52	5.45	0.07
11	H	H	H	H	CH_3	$CH(CH_3)_2$	1.947	2.588	5.00	4.94	0.06
12	Cl	H	H	H	CH_3	$CH(CH_3)_2$	1.930	2.565	5.16	5.19	−0.03
13	H	Cl	H	H	CH_3	$CH(CH_3)_2$	1.933	2.570	4.96	5.13	−0.17
14	H	H	Cl	H	CH_3	$CH(CH_3)_2$	1.936	2.573	5.00	5.10	−0.10
15	H	H	H	CH_3	CH_3	$CH(CH_3)_2$	1.950	2.629	4.34	4.55	−0.21
16	H	H	Cl	CH_3	CH_3	$CH(CH_3)_2$	1.938	2.614	4.80	4.71	0.09
17	H	CH_3	H	H	CH_3	$CH(CH_3)_2$	1.951	2.603	4.90	4.79	0.11
18	H	H	H	H	H	C_2H_5	1.901	2.489	5.82	5.95	−0.13
19	H	H	Cl	H	H	C_2H_5	1.889	2.474	6.00	6.12	−0.12
20	H	H	H	H	H	$CH_2C_6H_5$	1.103	2.587	6.14	6.44	−0.30
21	H	H	H	H	H	$CH(CH_3)C_6H_5$	1.108	2.637	5.70	5.96	−0.26
22	H	H	CH_3	H	H	$CH(CH_3)C_6H_5$	1.109	2.649	6.05	5.85	0.20
23	H	H	OCH_3	H	H	$CH(CH_3)C_6H_5$	1.096	2.635	6.00	6.00	−0.00
24	H	H	Cl	H	H	$CH_2C_6H_5$	1.089	2.573	6.96	6.60	0.36

a. $S(CH_{14})$ is the E-state value for the first carbon in the substituent R_6.
b. $S(NH_{11})$ is the E-state value for the NH adjacent to the carbonyl group of the hydrazide funtional group.
c. pIC_{50} is the negative log of the inhibitory power $IC_{50}(\mu M)$.
d. Calc is the calculated activity (pIC_{50}) from Eq. 8.
e. Res. = pIC_{50} − Calc.

What to assign as E-sate value of the atom type not present?

- E-state value of ‘0’ is valid result so reporting value of ‘0’ for missing atom type should not be used (as in C2 – Accelyrs)
- Use of –999 as E-state value for missing atom types as input for QSAR

What are the issues with E-states and multi-functional molecules?

- 35 atom types that are the bases for calculating K-H E-sates are too general
- When dealing with QSAR for datasets where atom-by-atom matching is not possible and any given atom type hit more than once → the result is ambiguity that no statistical tool will resolve

More on ambiguity

- For example: ssNH could be part of
  - Sulphonamide, RNHSO2R and
  - Amine, RNHR
  - Same atom type, both part of the same molecule – but in very different chemical environment
- What to calculate?
  - An average
  - Sum or
  - Both – the sum and the average

Testing hypothesis that simple counts should do at least as good as information rich K-H E-states

- Develop the program that will read in the same atom types and do the counts
- Choose several datasets that from QSAR area that feature multi functional type of molecules
- Use the same statistical approach to compare the performance of two sets of descriptors

Protocol used for comparison

- Descriptors:
  - E-state
    - 35 descriptors based on average E-state values
    - 35 descriptors based on sum of E-states
  - Counts
    - 35 based on the counts of K-H –state atom types
- Datasets
  - logP*, aqueous solubility, Human Intestinal Absorption and Blood Brain Barrier
- Statistical Tools
  - PCA/PLS in Simca (Umetrics)

Smarts Definitions for Kier-Hall Atom Types

RowNo	smarts-definitions	estates-atom-types-KH	RowNo	smarts-definitions	estates-atom-types-KH
1	[OH1][*]	sOH	19	S(=[*])(=[*])([*])[*]	ddssS
2	O=[*]	dO	20	[F][*]	sF
3	[OH0]([*])[*]	ssO	21	[Cl][*]	sCl
4	[o]	aaO	22	[Br][*]	sBr
5	[NH2][*]	sNH2	23	[I][*]	sI
6	[NH1]=[*]	dNH	24	[CH3][*]	sCH3
7	[NH1]([*])[*]	ssNH	25	[CH2]([*])[*]	ssCH2
8	[nH1]	aaNH	26	[CH2]=[*]	dCH2
9	N#[*]	tN	27	[CH1]([*])([*])[*]	sssCH1
10	[ND2](=[*])[*]	dsN	28	[CH1](=[*])[*]	dsCH1
11	[nH0]	aaN	29	[CH1]#[*]	tCH
12	N([*])([*])[*]	sssN	30	[cH]	aaCH
13	N(=[*])(=[*])[*]	ddsN	31	[cH0]	aasC
14	[N;+]([*])([*])([*])[*]	ssssN+	32	C(=[*])=[*]	ddC
15	[SH1][*]	sSH	33	C(#[*])[*]	tsC
16	S=[*]	dS	34	C(=[*])([*])[*]	dssC
17	[SX2]([*])[*]	ssS	35	C([*])([*])([*])[*]	ssssC
18	[s]	aaS			

Calculating E-state Descriptors

Name	%F (HIA)	sOH-sum	sOH-av	dO-sum	dO-av	ssO-sum	ssO-av	aaO-sum
raffinose	0.3	108.94	9.9	-999	-999	26.46	5.29	-999
lactulose	0.6	76.52	9.56	-999	-999	15.31	5.1	-999
aztreonam	1	18.06	9.03	57.84	11.57	4.92	4.92	-999
ceftriaxone	1	9.89	9.89	62.29	12.46	4.74	4.74	-999
cefuroxime	1	9.5	9.5	47.57	11.89	9.3	4.65	5.13
kanamycin	1	70.91	10.13	-999	-999	22.2	5.55	-999

Counts of Kier-Hall Atom Types

Name	%F (HIA)	sOH	dO	ssO	aaO	sNH2	dNH	ssNH	aaNH
raffinose	0.3	11	0	5	0	0	0	0	0
lactulose	0.6	8	0	3	0	0	0	0	0
aztreonam	1	2	5	1	0	1	0	1	0
ceftriaxone	1	1	5	1	0	1	0	1	1
cefuroxime	1	1	4	2	1	1	0	1	0
kanamycin	1	7	0	4	0	4	0	0	0

Objectives

- Compare quality of the models (R^2), based on training set alone and using in-built cross-validation Q^2 (LMO) within Simca
- Each of the datasets used has been analysed in the literature using similar approaches but with different descriptors
- NOT designed to build best models for those datasets

Performance of E-states vs Counts using Simca and PLS

Data Set (size)	e-states (ES)	counts of ES at-type	Performance
	R^2	R^2	$(R^2(ES)-R^2(Counts))*100$
Aq Sol (n=3000)	0.655	0.659	-0.4
HIA (n=300)	0.306	0.49	-18.4
BBB (n=145)	0.611	0.59	2.1
logP (n=10,000)	0.42	0.718	-29.8

Conclusions

- Simple counts of the same atom types that Kier-Hall Estate descriptors are built on work at least as good in building the models for BBB and solubility, and outperform E-states when building models for HIA and logP, 18% and 30% respectively

Acknowledgment

- Thanks to Daylight for supplying programming toolkits for coding E-states algorithm and development of software for counting atom types based on smarts definitions