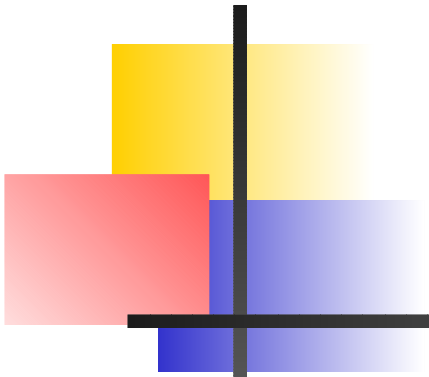


Atom Pairs as Descriptors

Rajarshi Guha

Penn State University

Atom Pair Descriptors

- **Aim**: [a] AP's were devised as a means of measuring structural similarity.
- **Problem**: Since a given molecule can have more than 1 atom pair fragment, how can we combine them and their numbers to describe a molecule?
- **Solution**: Combine them with trivial operations!

[a] Carhart et al. *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 64-73

Calculation Method

- First evaluate the atom property value:

$$64 * (\text{atom type}) + 16 * (\text{no. of pi electrons}) + (\text{no. of neighbors})$$

- Calculate the atom pair key:

$$\min[ap(j), ap(k)] + 1024 * \max[ap(j), ap(k)] + 1024 * D(k)$$

where $ap(j)$ & $ap(k)$ are the atom property values & $D(K)$ is the shortest distance between atoms j & k

- Use atom pair keys to generate descriptors.
- Currrently the atomic weight is used as atom type value.

The Descriptors

- 5 descriptors were devised:
  - *apnum*: Number of atom pair fragments.
  - *apmax*: Maximum atom pair key value.
  - *apmin*: Minimum atom pair key value.
  - *apsum*: Sum of all atom pair key values.
  - *apavg*: Average of all atom pair key values.

Dataset

- Glass T_g dataset [a].
- 251 molecules.
- TSET = 201, PSET = 26, CVSET = 24.

[a] Mattioni et al, *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 232-240

Descriptors

- Descriptor reduction gave a reduced pool of 20 descriptors.
- Reduction used a correlation cutoff = 0.8 & identical cutoff = 0.9.
- The reduced pool contained *apsum, apmin & apavg*
- Reducing the cutoffs generated reduced pools which still contained 2 or 3 of the AP descriptors.
- They do seem to be information rich!

Correlation Matrix

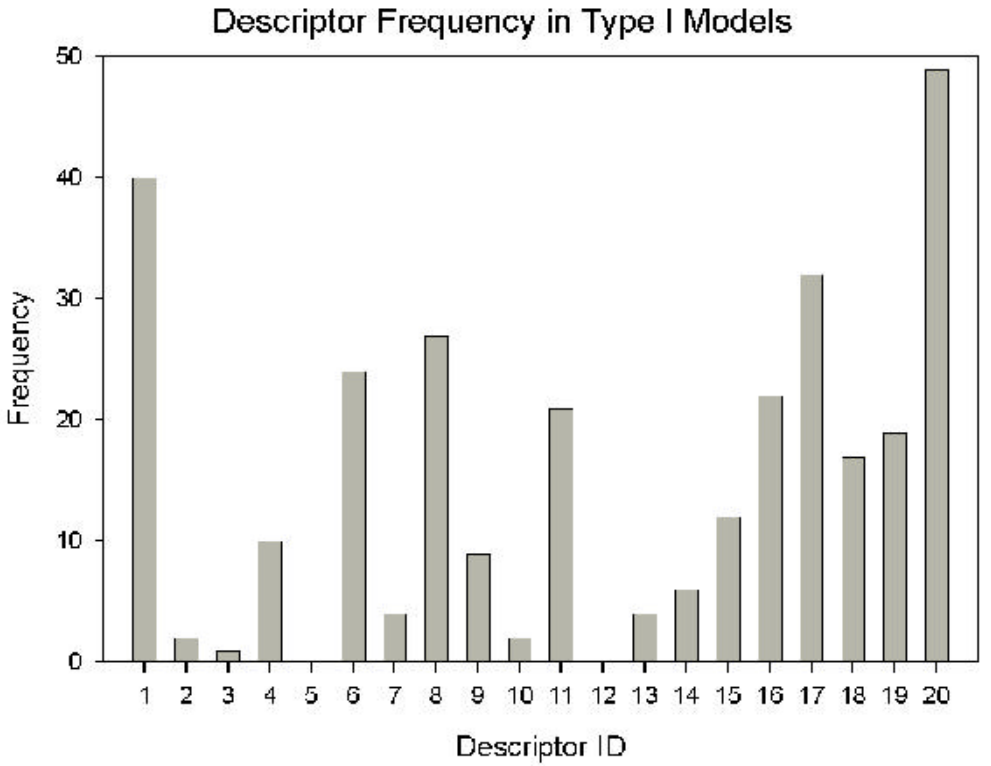
	NUM	SUM	MAX	MIN	AVG
NUM	0				
SUM	0.97	0			
MAX	0.66	0.69	0		
MIN	0.21	0.30	0.36	0	
AVG	0.56	0.63	0.96	0.39	0

- The correlation matrix for the 5 AP descriptors.
- Uses the entire dataset.

Type I Models

- A total of 58 models were generated.

1	KAPA	10	MDE 11
2	V4C	11	MDE 12
3	V6C	12	MDE 14
4	MOLC	13	MDE 22
5	NCl	14	MDE 24
6	NSB	15	SYMM
7	WTPT	16	ESUM
8	WTPT	17	EDIF
9	3SP3	18	apsum
10	MDE	19	apmin
11	MDE	20	apavg



Type I Models

Statistics for the Best Model

		Beta	S.E.	t	P
X( )	CONS	3.282496E+02	2.3357E+01	+14.0534	1.0014E-05
X(1)	KAPA	-1.446837E+01	1.8605E+00	-7.7768	1.0014E-05
X(8)	WTPT	1.506828E+01	1.9907E+00	+7.5692	1.0014E-05
X(9)	3SP3	2.527958E+01	6.6662E+00	+3.7922	2.3911E-01
X(13)	MDE	4.613150E+00	1.2894E+00	+3.5779	2.7855E-01
X(14)	MDE	-4.664483E+00	1.1227E+00	-4.1547	1.8148E-01
X(15)	SYMM	-1.064462E+02	2.5916E+01	-4.1073	1.8837E-01
X(16)	ESUM	-1.026281E+00	1.7201E-01	-5.9664	1.0014E-05
X(17)	EDIF	7.814406E+00	9.9928E-01	+7.8201	1.0014E-05
X(18)	ap_sum	1.445844E-05	2.5490E-06	+5.6721	1.0014E-05
X(20)	ap_avg	3.659924E-03	3.5660E-04	+10.2635	1.0014E-05

RSS = 479925.7812 SD = 47.3566 RSQ = 0.8059 Adj. RSQ = 0.7969

Overall F(10,214) = 80.4075 P = 3.3889E-08 N = 225

Multiple R = 0.8977

Type II Models

Architecture	TSET	CVSET	PSET
9-5-1	28.88	27.62	44.83
8-5-1	33.53	28.16	37.43
10-5-1	26.87	31.80	36.83
10-5-1 $*$	15.67	15.08	21.76

Descriptors selected:

8-5-1 : KAPA NSB WTPT MDE12 MDE24 ESUM EDIF apavg

9-5-1 : KAPA NSB MDE12 MDE24 ESUM EDIF apsum apmin apavg

10-5-1: KAPA WTPT4 WTPT5 MDE22 MDE24 SYMM ESUM EDIF apsum apmin

10-5-1$*$: V6CH N6PC MDE44 GRAV4 QSUM1 RPCS 1SP2 MOLC8 MOLC9 NSB

Descriptor Issues

- Correlation between the AP descriptors themselves.
- Is there a better way to combine the atom pair values for a given molecule?
- Should the atom pair fragment code play a role?

AP Value Issues

- The hashing method needs to be improved to take into account any atom type.
- The original method is restrictive since it tries to pack the AP value into a 32 bit integer.