*Molecules* **2004**, *9*, 1004-1009

*molecules*

**ISSN 1420-3049**
http://www.mdpi.org

# Performance of Kier-Hall E-state Descriptors in Quantitative Structure Activity Relationship (QSAR) Studies of Multifunctional Molecules

**Darko Butina** *

ChemoMine Consultancy, 201 Icknield Way, Letchworth Garden City, Herts SG6 4TT, U.K.
Telephone: (+44) (0)1462 634167

* To whom correspondence should be addressed; e-mail: darko.butina@chemomine.co.uk

*Received: 2 June 2004 / Accepted: 15 July 2004 / Published: 31 December 2004*

**Abstract:** Performance of the E-state descriptors was tested against simple counts of the 35 atom types that the Kier-Hall E-states are based upon, by building PLS models for clogP, aqueous solubility, human intestinal absorption (HIA) and blood brain barrier (BBB). The results indicate that the simple counts work at least as well as E-state descriptors in building models for solubility and BBB, while surprisingly, simple counts have outperformed E-states by 18% and 30%, respectively, when building the models for HIA and clogP.

**Keywords**: E-state descriptors, Kier-Hall, atom types, QSAR, 2D descriptors.

## Introduction

Kier and Hall [1,2] have developed the concept of E-states, an electrotopological-state index for atoms in a molecule. Examples of uses of the E-states in QSAR in the areas of NMR chemical shifts, inhibition of monoamine oxidase and receptor binding affinities of beta carbolines have been published [3]. An intrinsic atom value is assigned to each atom as $I=(\delta^V+1)/\delta$, in which $\delta^V$ and $\delta$ are counts of valence and sigma electrons of atoms associated with the molecular skeleton. The E-state value, $S_i$, for skeletal atom I is defined as $S_i=I_i+ \Delta I_i$, where the influence of other atoms on atom *i*, $\Delta Ii$, is given as $\Sigma(I_i-I_j)/r_{ij}^2$ in which $r_{ij}$ is the graph separation between atoms *i* and *j*, counted as the number of atoms, including *i* and *j*.

If one looks into drug like molecules and uses C, N, O, S and the halogens as the main building blocks, Kier and Hall use 35 atom types to calculate E-states (Table 1):

**Table 1**

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

The symbols associated with the atom types are *s* for single bond, *d* for double, *t* for triple and *a* for aromatic. The attraction and proposed advantage of E-states over simple counts of the equivalent atom types is that E-states values for each atom in a given molecule 'reflect' the steric and electronic effects of the surrounding atoms and as such, could be best described as information rich atomic descriptors. Thus, for example, if two different molecules have one phenol group, simple phenolic OH counts would not differentiate between two different substitution patterns that the phenolic group might have, while E-states would.

An example of the power of this approach in the field of QSAR was exemplified in the case of receptor binding of a series of beta-carbolines [3] where each atom of the 6-5-6 ring system could be uniquely mapped and E-states calculated. However, in the majority of QSAR applications in drug discovery, one is dealing with varying levels of structural diversity and multifunctional environments where the individual atom type used as a basis for calculating E-states will occur more than once and most likely in different chemical environments. For example, one can easily have a drug molecule where ssNH would be part of sulphonamide, like $RNHSO_2R$ and basic amine RNHR, both part of the same molecule. There are the following potential problems with using E-state values in QSAR problems where any atom type is present in a given molecule more than once and the molecules could not be matched using atom-by-atom overlap:

1. An average of two or more E-states is calculated for each atom type
2. The sum of two or more E-states is calculated for each atom type
3. Both an average and the sum are reported for each atom type
4. What E-state value to report if a given atom type is NOT present in a molecule?

The problem of using in QSAR applications E-state values that for a given atom type are the result of either an average or a sum is in the ambiguity of the resulting values. Two very different molecules could have an identical average E-state value for the same atom type but in a very different chemical environment, which in turn would reflect in poor performance in QSAR terms and regardless of statistical approach used to build the putative model.

To test that hypothesis, comparison was made in building PLS based models for clogP [4], aqueous solubility [5], Blood Brain Barrier (BBB) [6] and Human Intestinal Absorption (HIA) [7] using E-state descriptors and simple counts for 35 atom types that the Kier-Hall algorithm is based upon (Table 1). The issue of reporting E-state value for atom types that are not present in the molecule is discussed in the next section.

## Methodology

The E-states algorithm is relatively easy to implement, and this author has used the Daylight software toolkit [8,9] to code in the original algorithm. E-state atom types were coded using smarts (substructure features within Daylight) and Table 1 shows the coding details. The atom types used are based on the following atoms: C, N, O, S and the halogens.

One interesting problem occurs when reporting E-state values for atom types that are not present in the molecule. Reporting the E-state value of '0' should be only used if the average value for a given atom type is calculated as '0' (there are no intrinsic values for any atom type that are '0'), which is a real possibility. The author in this paper is using -999 as the E-state value for any atom type that is not present in the molecule. In addition to calculating E-states, another software program was written that counts the presence of each of the 35 atom types listed in Table 1.

For each data set, two comma separated output files were produced, one with E-states descriptors with both sum and the average value reported (-999 used if atom type is not present in the molecule), and another one with the simple counts for the same atom types.

## Statistical Analysis

The principal component analysis (PCA) and partial least squares discriminant analysis (PLS) [10] are chemometric tools for extracting and rationalizing the information from any multivariate description of a biological system. Complexity reduction and data simplification are two of the most important features of such tools. PCA and PLS condense the overall information into two smaller matrixes, namely the score plot (which shows the pattern of compounds) and the loading plot (which shows the pattern of descriptors). Because the chemical interpretation of score and loading plots is simple and straightforward, PCA and PLS are usually preferred to other nonlinear methods. PLS analysis was implemented using SIMCA software, version 9, supplied by Umetrics (Umeå, Sweden). Within Simca's implementation of PLS, all descriptors are normalised and use of the LMO (leave many out) controls the fitting procedure and avoids over-fitted models.

## Results and Discussion

For each of the four sets used: logP (10,000 molecules), aqueous solubility (3,000 molecules), BBB (145 molecules) and HIA (300 molecules), two sets of descriptors have been calculated:

- E-states (sum and the average) for each of 35 descriptors as in Table 1 (70 descriptors for each molecule).
- Simple counts for each of 35 descriptors as in Table 1.

As discussed earlier, for any descriptor that was not present in a given molecule, –999 was used for both E-state sum and the average value. The resulting output file was read into Simca and resulting $R^2$ and the difference between the two approaches is reported in Table 2:

**Table 2**

| **Data Set** | **e-states (ES) $R^2$** | **counts of ES at-type $R^2$** | **Difference between the models $(R^2(ES)-R^2(Counts))*100$** |
|---|---|---|---|
| aqueous solubility | 0.655 | 0.659 | -0.4 |
| HIA | 0.306 | 0.49 | -18.4 |
| BBB | 0.611 | 0.59 | 2.1 |
| logP | 0.42 | 0.718 | -29.8 |

The statistical measure of how well a regression line approximates real data points is reported as $R^2$, with a value of 1 indicating perfect fit. What one would expect to see is that the richer descriptors, like E-states, would systematically outperform simple counts of the identical type of descriptors, irrespective of overall quality of the models and therefore give large and positive difference between the two $R^2$ (last column in Table 2). However, as one can see, models for aqueous solubility and BBB are for all practical purposes almost identical, while models for HIA and logP are 18% and 30% better, respectively, when using simple counts of 35 descriptors that E-states are built on. It is important to bear in mind that the objective of this work was NOT to build the best models for the four datasets, but to make comparisons of the performance of the two sets of descriptors using identical datasets and the same statistical method.

As indicated earlier, the most likely explanation for the observed results could be found in the specificity, or rather lack of it, of the basic set of 35 atomic types that the E-states are based upon. Most of the drug discovery related problems have multifunctional and chemically diverse structures that could not be matched using atom-by-atom overlap, very much like the sets used in this paper. Ambiguity resulting from calculating either an average or the sum or both E-state values for the same atom type that could produce an identical E-state value for the atoms that could be in very different chemical environment will result in 'degradation' of the quality of E-state based descriptors and poor quality QSAR models.

## Conclusions

Performance of the E-state descriptors was tested against the simple counts of the 35 atom types that the E-state descriptors are based on, using multifunctional data sets from drug discovery projects, like logP, solubility, HIA and BBB. The statistical method used was PLS within Simca software, a standard in pharma industry when dealing with linear regression based models. The results indicate that the simple counts work at least as well as E-state descriptors in two datasets, solubility and BBB, while the count based descriptors have outperformed E-states in HIA and logP datasets. Possible explanation for the lack of expected superior performance by information-rich descriptors like E-states in the multifunctional type of molecules is most likely due to the ambiguity that arises when the same atom type is present more than once in the given molecule and in a very different chemical environment. The only values, in the case of E-states, that can be calculated for that atom type are an average value, the sum, or both, which will result in an ambiguous value that cannot be properly resolved by standard statistical approaches, like PLS. While this paper is based on only four datasets and use of the single statistical approach, this author has been using the same descriptors, different statistical approaches (decision trees, NN, kNN) and different activity/property datasets in the past and has observed very similar behaviour.

## Acknowledgments

My thanks to Daylight and Yosi Taitz for supplying programming toolkits that were instrumental in coding e-states algorithm and development of the software to do the counts for the same atom types.

## References and Notes

1. Kier, L.B.; Hall, L.H. *Molecular Structure Descriptors: The Electrotopological State*. Academic Press: New York, **1999**
2. Hall, L.H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem Inf. Comput. Sci.* **1991**, *31*, 76-82
3. Hall, L.H.; Mohney, B.; Kier, L. B. An Electrotopological-State: An Atom Index for QSAR. *Quant. Struct. –Act. Relat.* **1991**, *10*, 43-51
4. Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. Estimation of Molecular LFER Descriptors by a Group Contribution Approach. 2. Prediction of Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71-80
5. Butina, D.; Gola, J. M. R. Modelling Aqueous Solubility. *J. Chem Inf. Comput. Sci.* **2003**, *43*, 837-841
6. Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A.; Zhao, Y. H.; Ijaz, L. Correlation and prediction of large blood-brain distribution data set – an LFER study. *Eur. J. Med. Chem.* **2001**, *36*, 719-730
7. Zhao, Y. H.; Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. Evaluation of Human Intestinal Absorption Data and Subsequent derivation of QSAR with Abraham Descriptors. *J. Pharm. Res.* **2000**, *90*, 749-784
8. Daylight software, v 4.72 was used for this work, implemented on a Linux v.8 single processor laptop
9. For all details on smarts and toolkits see Daylight's home page: www.daylight.com
10. Dunn, W.J.; Wold, S. Pattern Recognition Techniques in Drug Design. In *Comprehensive Medicinal Chemistry*; Pergamon Press: Oxford, **1990**; Vol.4, pp. 691-714