http://dx.doi.org/10.5012/bkcs.2013.34.4.1043

Prediction of Physicochemical Properties of Organic Molecules Using Semi-Empirical Methods

Chan Kyung Kim,* Soo Gyeong Cho,† Chang Kon Kim, Mi-Ri Kim, and Hai Whang Lee

*Department of Chemistry, Inha University, Incheon 402-751, Korea. *E-mail: kckyung@inha.ac.kr*
†Agency for Defense Development, P.O. Box 35-5, Yuseong, Daejeon 305-600, Korea
Received November 27, 2012, Accepted January 8, 2013

Prediction of physicochemical properties of organic molecules is an important process in chemistry and chemical engineering. The MSEP approach developed in our lab calculates the molecular surface electrostatic potential (ESP) on van der Waals (vdW) surfaces of molecules. This approach includes geometry optimization and frequency calculation using hybrid density functional theory, B3LYP, at the 6-31G(d) basis set to find minima on the potential energy surface, and is known to give satisfactory QSPR results for various properties of organic molecules. However, this MSEP method is not applicable to screen large database because geometry optimization and frequency calculation require considerable computing time. To develop a fast but yet reliable approach, we have re-examined our previous work on organic molecules using two semi-empirical methods, AM1 and PM3. This new approach can be an efficient protocol in designing new molecules with improved properties.

Key Words : Organic molecules, MSEP method, Semi-empirical methods, Prediction of physicochemical properties, Root-mean-square deviation

Introduction

Before designing a new compound with improved physical or mechanical property, a usual practice is to predict its properties before synthesis. If the virtual candidate meets the requirement, next job is to start synthetic work. If not, the researcher has to go back to evaluate more compounds. This kind of search and prediction method is a tedious job for bench chemists. One of the best methods doing this task is to utilize Quantitative Structure Property Relationship (QSPR) method.[1] In QSPR method, molecular properties are predicted by using some analytical solutions, which are expressed as a function of various molecular descriptors derived from molecular structures. There are many different types of molecular descriptors:[2] topological, geometrical, electrostatic, and quantum mechanical ones. The simple descriptors are easily calculated from the molecular formula and connectivity. However, these simple descriptors cannot distinguish isomers or different spatial orientation of atoms in a molecule. In this sense, it is essential to include three-dimensional (3D) structures of molecules. To obtain proper 3D descriptors, Politzer and coworkers developed an excellent method,[3] which has been applied successfully to predict various physicochemical properties of organic and high energetic molecules. This method, denoted as Density method hereafter, calculated electrostatic potential (ESP, in kcal/mol) on the envelope of electron density of 0.001 e/bohr^3.[4,5] ESP is given rigorously by Eq. (1).

$$V(r) = \sum_A \frac{Z_A}{|R_A - r|} - \int \frac{\rho(r')dr'}{|r' - r|} \quad (1)$$

where, $V(r)$, Z_A, and $\rho(r')$ refer to ESP values at distance r, charge on nucleus A, and the electron density of the molecule, respectively. Using these ESP values, various statistical descriptors were developed as shown in Eqs. (2)-(4).[6]

$$\sigma_{tot}^2 = \sigma_+^2 + \sigma_-^2 = \frac{1}{m}\sum_{i=1}^{m}[V^+(r_i) - \overline{V}_s^+]^2 + \frac{1}{n}\sum_{j=1}^{n}[V^-(r_j) - \overline{V}_s^-]^2 \quad (2)$$

$$\nu = \frac{\sigma_+^2 \sigma_-^2}{\sigma_{tot}^2} \quad (3)$$

$$\Pi = \frac{1}{n}\sum_{i=1}^{n}|V(r_i) - \overline{V}_s| \quad (4)$$

where, σ_{tot}^2 (in kcal^2/mol^2), ν, and Π (in kcal/mol) are total variance of $V(r)$, a balance parameter between positive (σ_+^2) and negative (σ_-^2) ESP values, and average deviation of $V(r)$, respectively. $V^+(r_i)$ and $V^-(r_j)$ are the positive and negative ESP values on the surface, and $\overline{V}_s^+$, $\overline{V}_s^-$, and $\overline{V}_s$ are the positive, negative and total average values, respectively. Kim and coworkers simplified the Density method by calculating ESP values on van der Waals (vdW) molecular surface of molecules.[7] This method, denoted MSEP method, is conceptually simple than the Density method, because vdW surface is well documented and easy to describe using the vdW radii of atoms. The MSEP method gave exactly the same results as the Density method and applied to explain solid densities[8] and impact sensitivities[9] of high energetic molecules. In MSEP method, the first step is to optimize 3D structures of molecules using quantum chemical programs, such as Gaussian packages. We have optimized all the molecular structures using the hybrid density functional theory,

B3LYP[10] with the 6-31G(d) basis set because this combination of method and basis set, B3LYP/6-31G(d), could give reasonable structures compatible to the correlated MP2 method.[11]

Application of the above methodology to a database containing more than a few thousand molecules is a formidable task because the process of getting minimized structures is a very slow step. To overcome this problem, we have to choose some simpler but yet reliable method for structural optimization and frequency calculations. Molecular mechanics (MM) could be the best method due to its speed and simplicity.[12] MM method, however, poses a problem. In MM, molecular properties related to the electron distribution could not be obtained because this method does not consider electrons in the formalism. Moreover, MM method poses another serious problem in dealing with conjugated functional group such as nitro group, which is most abundant in high energy density molecules (HEDM). The next alternative method will be semi-empirical approach, which substitutes some important integrals for experimental data in solving Schrödinger equation and is known to be fast and somewhat reliable in geometry prediction if the molecules under consideration are similar to those employed in parameter developing process.[13] Therefore, we chose to use two well-known semi-empirical methods, AM1 and PM3, to examine their performance in the prediction of physicochemical properties of organic molecules.

Calculations

The molecules considered in this work are the same as those in earlier works. To compare the speed, the same initial structures were used in this work. As is well-known, the most time consuming processes in quantum chemical calculations is to optimize geometry and sometimes to perform frequency calculation. In this work, the molecular structures were optimized by using AM1 and PM3 hamiltonians in Gaussian 03 package.[14] Frequency calculations were also done to examine if the structures corresponded to minima in the potential energy surface (PES). Some molecules with a labile group tend to give small negative frequencies in initial optimization, but they were removed by re-optimizing the structures after adjusting the atomic coordinates corresponding to the imaginary frequencies using the GaussView 3.0 graphic program.[15] The computation of ESP values on the vdW surface is quite easy and straightforward even at the B3LYP/6-31G(d) method because this calculation corresponds to a single point calculation. Therefore, we chose to use the B3LYP/6-31G(d) method to calculate ESP values on the AM1 or PM3 optimized molecular surface. The molecular descriptors were subjected to the multiple linear regressions described earlier.[7] The semi-empirical optimized structures may deviate from the B3LYP-optimized structures. To examine the difference in geometries, we calculated the root-mean-square deviations (RMSD) between DFT-optimized and semi-empirical-optimized structures.

Results

The physicochemical properties considered in this work are normal boiling points (T_{bp}),[6] heats of fusion (ΔH_{fus}),[6] heats of sublimation (ΔH_{sub}),[16] heats of vaporization (ΔH_{vap}),[17] liquid density (ρ_{liq}),[17] and crystal density (ρ_{cry}).[17] The results of multiple correlations are summarized in Table 1 and the predicted values and the molecular descriptors used in the correlations are summarized in Table S1 (supporting information). The plots of calculated vs experimental values using AM1 and PM3 hamiltonians are shown in Figure 1. Normal boiling points (in K) of 100 organic molecules are examined by using the dual-parameter equation shown in Eq. (5).[6]

$$T_{bp} = \alpha(AREA) + \beta\sqrt{\nu\sigma_{tot}^2} + \gamma \quad (5)$$

where *AREA* (in Å^2) is vdW surface area and α, β, and γ are coefficients of multiple linear equation that can be the sensitivity of each independent variable. Inspection of Table 1 shows that the dual-parameter equation for normal boiling points is well represented by Eq. (5). The correlation coefficients (r) of 0.929 (AM1) and 0.934 (PM3) are similar to that of the original MSEP method (r = 0.936).[8] This is quite surprising because the pro of semi-empirical methods lies in its speed but the con lies its poorer performance regarding molecular structures and properties than *ab initio* or density functional method. To see the structural variation on going from DFT to semi-empirical methods, average and maximum RMSD values are calculated with respect to DFT structures and the results are summarized in Table 2.

Inspection of Table 2 shows that the average RMSD value is 0.0596 and 0.0588 Å for AM1 and PM3, respectively. The maximum RMS values of 0.479 (AM1) and 0.490 Å (PM3) indicates that the semi-empirical optimized structures are not much different from the corresponding DFT structures.

Table 1. Summary of multiple linear regressions of some physicochemical properties

Coeff.	Method	T_{bp}	ΔH_{vap}	ΔH_{sub}	ΔH_{fus}	ρ_{liq}	ρ_{cry}
α^a	AM1	2.543	5.230	0.000461	0.113	1.161	1.410
	PM3	2.528	5.227	0.000462	0.111	1.144	1.419
	MSEP	2.496	5.335	0.000448	0.113	1.172	1.436
β^a	AM1	22.64	2.456	0.131	1.643	0.0100	0.0333
	PM3	22.86	2.384	0.150	1.575	0.0103	0.0390
	MSEP	23.15	2.146	1.690	1.667	0.00960	0.0344
γ^a	AM1	-28.86	-32.52	1.781	-6.890	0.0208	0.0419
	PM3	-23.06	-31.83	1.636	-6.431	0.0281	0.0315
	MSEP	-23.80	-32.35	-2.060	-6.824	0.0190	0.0229
r^b	AM1	0.929	0.925	0.951	0.933	0.986	0.987
	PM3	0.934	0.933	0.950	0.935	0.986	0.981
	MSEP	0.936	0.947	0.951	0.928	0.984	0.988

[a]Coefficient of multiple correlation in Eqs. (5)-(10). [b]Linear regression coefficient.

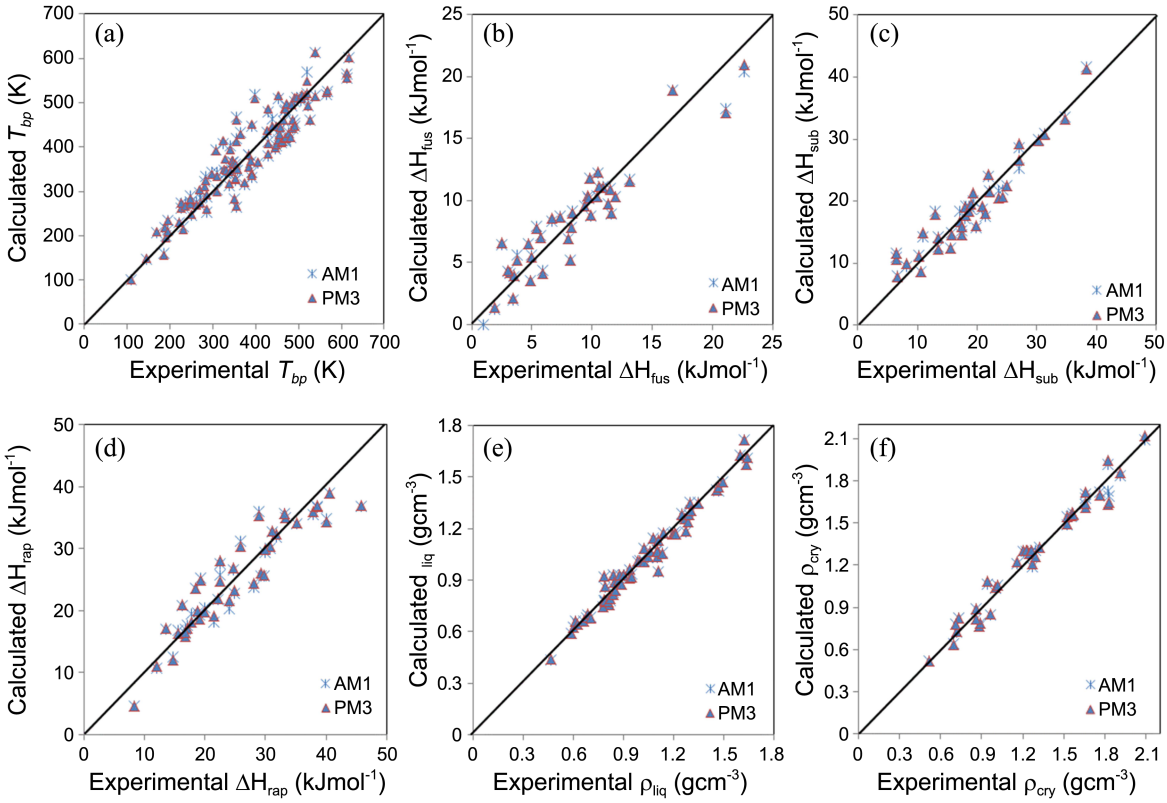


Figure 1. Plots of calculated *vs* experimental values computed using semi-empirical methods. (a) Normal boiling points, (b) Heats of fusion, (c) Heats of sublimation, (d) Heats of vaporization, (e) liquid density, and (f) crystal density.

Table 2. Average and maximum RMSD values (in Å) of semi-empirical optimized structures relative to the DFT optimized structures

RMSD	Method	T_{bp}	ΔH_{vap}	ΔH_{sub}	ΔH_{fus}	ρ_{liq}	ρ_{cry}
Average	AM1	0.0596	0.0453	0.0983	0.0472	0.0634	0.0676
	PM3	0.0588	0.0310	0.105	0.0520	0.0567	0.0697
Maximum	AM1	0.479	0.479	0.565	0.355	0.479	0.469
	PM3	0.490	0.0630	0.492	0.370	0.370	0.490

Table 1 shows that such a small structural variation has almost no effect in deriving a useful QSPR equation for boiling points of organic compounds.

Next, we examined various heats (in kJ mol^{-1}) related to phase changes - ΔH_{fus}, ΔH_{sub}, and ΔH_{vap}. The dual-parameter equations for these properties are shown in Eqs. (6)-(8).[6,16,17] The QSPR results are also shown in Table 1 and the calculated values and the relevant molecular descriptors are summarized in Tables S2-S4 (supporting information).

$$\Delta H_{fus} = \alpha(AREA) + \beta(\nu\pi) + \gamma \quad (6)$$

$$\Delta H_{sub} = \alpha(AREA)^2 + \beta(\nu\sigma_{tot}^2) + \gamma \quad (7)$$

$$\Delta H_{vap} = \alpha(\sqrt{AREA}) + \beta(\sqrt{\nu\sigma_{tot}^2}) + \gamma \quad (8)$$

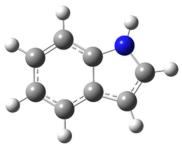
Total number of molecules considered for heat of fusion, heat of sublimation, and heat of vaporization are 37, 34, and 40, respectively. The correlation coefficients for these properties are almost as good as the corresponding values for MSEP method. It seems surprising that there is not much effect by using computationally cheaper semi-empirical methods, which is also supported by smaller RMSD values shown in Table 2 (average RMSD < 0.2 Å and maximum RMSD < 0.6 Å).

Finally, we examined liquid and crystal densities (in g cm^{-3}) of organic molecules. All the properties studied above are related to the phase changes of organic molecules but the density is an intrinsic property of a molecule in each phase. In this sense, these densities might be more sensitive to molecular structures. The QSPR equations for liquid and solid densities are shown in Eqs. (9)-(10).[17] The QSPR results are also shown in Table 1 and the predicted values and the relevant molecular descriptors are summarized in Tables S5-S6 (supporting information).

$$\rho_{liq} = \alpha\left(\frac{MW}{AREA}\right) + \beta(\pi) + \gamma \quad (9)$$

$$\rho_{cry} = \alpha\left(\frac{MW}{AREA}\right) + \beta\left(\frac{\sigma_{tot}^2}{AREA}\right) + \gamma \quad (10)$$

Table 3. Computing time[a] required to perform calculations of indole



Method	Optimization	Frequency	RMSD[b]
B3LYP/6-31G(d)	1:32:33.8	1:03:49.4	0.00
AM1	0:00:05.2	0:00:13.9	0.0179
PM3	0:00:11.5	0:00:13.1	0.0656

[a]The calculations were performed using Intel Pentium 4 Dual CPU 3.40 GHz Linux PC and time format is in hh:mm:ss. [b]RMSD values in Å.

where *MW* is molecular weight of a molecule. Total number of molecules for liquid and solid densities is 61 and 36, respectively. Table 1 shows that the correlation coefficients for both properties are almost similar to the corresponding values obtained by MSEP method. This result comes from the fact that the molecular structures in the data set are similar to those optimized using the B3LYP/6-31G(d) method as confirmed by the smaller average and maximum RMSD deviations for both semi-empirical methods.

At this point, it would be interesting to see the effectiveness of our new approach. As an example, indole was chosen as a test molecule to compare the computing time required for structural optimization and frequency calculation. The result is shown in Table 3.

Table 3 shows that the computing time for AM1 and PM3 optimization and frequency calculations is almost negligible compared to the corresponding value for the B3LYP/6-31G(d) method. However, the RMSD values for AM1 and PM3 hamiltonians are 0.0179 and 0.0656 Å, respectively, which is quite insignificant in considering structural changes. This shows that the semi-empirical methods employed in this work are sufficient to derive meaningful QSPR relationships for physicochemical properties of organic molecules.

Conclusions

In this work, we re-examined QSPR of absolute properties, liquid and solid densities, and properties related to phase changes, boiling point, heats of fusion, sublimation, and vaporization using semi-empirical methods. All six properties were satisfactorily predicted using the AM1 and PM3 hamiltonians. Such good results are originated from the fact that the organic molecules considered in this work are ordinary molecules that can be optimized with reasonable accuracy using semi-empirical methods. This work suggests that semi-empirical method is an alternative and reliable method in dealing with large database in chemical engineering.

Acknowledgments. This work was supported by Defense Acquisition Program Administration (DAPA) and Agency for Defense Development (ADD) and Inha University.

References

1. Jurs, P. C. In *Encyclopedia of Computational Chemistry*; Vol. 4, Schleyer, P. v. R., Ed.; John Wiley & Sons: 1998; p 2320.
2. Karelson, M. In *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.
3. Politzer, P.; Murray, J. S. In *Reviews in Computational Chemistry*; Vol. 2, Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishing: New York, 1991; p 273.
4. Bader, R. F. W.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. *J. Am. Chem. Soc.* **1987**, *109*, 7968.
5. Bader, R. F. W.; Henneker, W. H.; Cade, P. E. *J. Chem. Phys.* **1967**, *46*, 3341.
6. Politzer, P.; Murray, J. S. In *Quantitative Treatments of Solute/Solvent Interactions*; Elsevier Amsterdam: 1994; p 243.
7. Kim, C. K.; Lee, K. A; Hyun, K. H.; Park, H. J.; Kwack, I. Y.; Kim, C. K.; Lee, H. W.; Lee, B.-S. *J. Comput. Chem.* **2004**, *25*, 2073.
8. Kim, C. K.; Cho, S. G.; Kim, C. K.; Park, H.-Y.; Zhang, H.; Lee, H. W. *J. Comput. Chem.* **2008**, *29*, 1818.
9. Kim, C. K.; Cho, S. G.; Li, J.; Kim, C. K.; Lee, H. W. *Bull. Korean Chem. Soc.* **2011**, *32*, 4341.
10. Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *Chem. Phys. Lett.* **1997**, *270*, 419.
11. Johnson, B. G.; Gill, P. M. W.; Pople, J. A. *J. Chem. Phys.* **1993**, *98*, 5612.
12. Kappé, A. K.; Casewit, C. J. In *Molecular Mechanics Across Chemistry*; University Science Books California, 1997.
13. Sadlej, J. In *Semi-Empirical Methods of Quantum Chemistry*; Ellis Horwood, Ltd., Chichester, 1985.
14. *Gaussian 03*, *Revision B.05*, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian, Inc.: Wallingford, CT, 2003.
15. Frish, A., Nielsen, A. B., Holder, A. J. *GaussView* 3.0, Gaussian, Inc.: Pittsburgh, PA, 2003.
16. Politzer, P.; Murray, J. S.; Grice, M. E.; Desalvo, M.; Miller, E. *Mol. Phys.* **1997**, *91*, 923.
17. Murray, J. S.; Brinck, T.; Politzer, P. *Chem. Phys.* **1996**, *204*, 289.