Available online at www.sciencedirect.com
ScienceDirect
Tetrahedron
Tetrahedron 64 (2008) 1759–1764
www.elsevier.com/locate/tet

Predicting experimental yields as an index to rank synthesis routes: application for Diels–Alder reactions

Kenzi Hori [a,*], Makoto Sakamoto [a], Toru Yamaguchi [a], Michinori Sumimoto [a], Katsuhiko Okano [b], Hidetoshi Yamamoto [a]

[a] *Department of Materials Science and Engineering, Graduate School of Science and Engineering, Yamaguchi University, 2-16-1 Tokiwadai, Ube 755-8611, Japan*
[b] *Process Technology Research Laboratories, Daiichi Sankyo Co., Ltd, Kitakasai, Edogawa, Tokyo 134-8630, Japan*

Received 18 October 2007; received in revised form 30 November 2007; accepted 30 November 2007
Available online 8 December 2007

Abstract

It is possible to create novel synthetic routes for compounds using synthesis route design systems (SRDS). We have been investigating an *in silico* screening protocol, which makes it possible to reduce the number of SRDS experiments in developing new synthesis routes. However, there still remains the problem of how to rank synthesis routes for experiments. The experimental yield is considered to be one of the most important factors in determining which synthesis route is better. The present study describes an attempt toward predicting the trends of experimental yields for organic synthesis by fusing computational chemistry and chemoinformatics. We examined whether the prediction of experimental yields for Diels–Alder reactions is feasible using activation energies obtained from Density Functional Theory (DFT) calculations together with the experimental conditions. A partial least squares analysis using these values gave correlation equations for the experimental yields. If it is possible to construct similar correlation equations for other reactions, then SRDS synthetic routes could be ranked on the basis of their predicted yields, and an order can be determined before beginning the experiments.

1. Introduction

Synthesis route design systems (SRDS) such as LHASA,[1] EROS[2] and AIPHOS[3] have been developed. They have made it possible to create new synthetic routes for many compounds. The KOSP (Knowledgebase-Oriented Synthesis Planning) program,[4] one of the AIPHOS family of programs, is now commercially available and has been used practically to create synthesis routes for target compounds. However, there are no guarantees that the routes created are suitable for actual application or production.

It is likely that theoretical methods such as molecular orbital (MO) and Density Functional Theory (DFT) calculations will be useful in screening synthesis routes, i.e., the *in silico* screening of these routes.[5] This screening makes it possible to reduce the number of experiments required by SRDS. We have been succeeding in shortening the time periods required for screening synthesis routes using the transition state database (TSDB).[5c,6]

Even though theoretical methods can correctly judge whether synthesis routes can produce target compounds, there still remains one problem: which synthesis route should be examined first? There are no standard solutions for this problem. For example, when discussing synthesis routes for a chiral compound, the enantiometric selectivity serves as the conclusive factor. A synthesis route using commercially available intermediates may be preferred to those using compounds, which need to be synthesized. Harmful or dangerous compounds should not be involved in the synthesis route. The presence of side reactions should also be considered when ranking the routes.

The experimental yield should be one of the most important factors that determine the superiority of various synthesis routes. The prediction of this outcome in organic synthesis has not been attempted until now because the yields are the

* Corresponding author. Tel.: +81 836 85 9238; fax: +81 836 85 9201.
E-mail address: kenji@yamaguchi-u.ac.jp (K. Hori).

0040-4020/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tet.2007.11.108

final result of the experiments. If it is possible to predict the yields of individual synthesis routes,[7] then the experiments can be conducted based on the expected yields. We have already shown that it is possible to predict the yield trends for S_NAr reactions by using the energies from the MO calculations, other energies such as activation energies, and the heats of the reactions.[8] The energetic relationship is directly correlated to the reaction mechanism. In order to predict the outcome of the reactions, partial least squares (PLS) analysis was performed on data sets based on isolation methods. However, we did not understand why the yields correlated so well with the calculated energies.

In this study, multivariate analysis was applied to predict the trends of the experimental yields from Diels–Alder reactions using the calculated activation energies. For this purpose, PLS analysis was used to obtain a relationship between the experimental yields and experimental conditions such as reaction temperature, reaction time as well as the activation energies obtained from the DFT calculations at the B3LYP/6-31G(d) level of theory.[9]

2. Results and discussion

2.1. Method to predict experimental yields using PLS analysis

The Diels–Alder reaction is a typical secondary reaction, and its rate constant k has a relationship with the initial concentration of the reactants as shown in Eq. 1,[10]

$$kt = \frac{1}{b-a}\ln\frac{a(b-x)}{b(a-x)} (a \neq b), \tag{1}$$

where t, and a, b are the reaction time and the initial concentrations of diene and dienophile, respectively. The concentration of the product is represented as x. Eq. 2 shows the experimental yield Y_{obsd},

$$Y_{obsd} = 100 \times x_F/a, \tag{2}$$

where x_F is the final concentration of the product. The yields from equipment such as HPLC and GC are similar to the theoretical value.

The kinetic constant k in Eq. 1 is represented by the frequency factor A, the gas constant and temperature (R and T), and an observed activation energy E_a (obsd),

$$k = A\exp\left(\frac{-E_a}{RT}\right). \tag{3}$$

We cannot predict an experimental yield by combining Eqs. 1 and 3 since we do not have E_a (obsd) values for new synthesis routes from SRDS before performing the experiments. Although a quantum chemical method can calculate the activation energies, E_a (calcd), it is almost impossible to use the energies to evaluate the rate constants of reactions. However, it may be possible to assume that the E_a (obsd)s for a series of reactions are proportional to those from theoretical calculations as follows,

$$E_a(\text{obsd}) \cong mE_a(\text{calcd}). \tag{4}$$

If Eq. 4 is applicable, we can obtain Eq. 5 by combining Eqs. 1, 3 and 4 as follows,

$$\frac{1}{b-a}\ln\frac{a(b-x)}{b(a-x)} \cong tA\exp\left(\frac{-mE_a(\text{calcd})}{RT}\right). \tag{5}$$

The logarithm of both sides are taken, and then:

$$z_{obsd} = \ln\left(\frac{1}{b-a}\ln\frac{a(b-x)}{b(a-x)}\right) \cong \ln t + \ln A - \frac{mE_a(\text{calcd})}{RT}. \tag{6}$$

It is probable that the frequency factor A does not change very much over a series of reactions such as the Diels–Alder reaction in the present study, and the factor m of Eq. 6 can be determined by a multivariate analysis using not only the experimental conditions, but also the calculated activation energies.

The solvent effect must be considered in performing organic synthesis to maximize their experimental yields. Therefore, we have to consider how the solvent affects the reaction mechanism. It is appropriate to consider logarithms of the dielectric constants of the solvents instead of the actual values themselves, since the free energies of reactions are considered.[11]

According to the arguments mentioned above, it is reasonable to consider Eqs. 7 and 8, and thus a multivariate analysis was used to correlate z_{obsd} with the logarithms of the dielectric constants, reaction times t_F, reaction temperature T and the calculated activation energy E_a (calcd) as explanatory variables as follows,

$$z_{calcd} = \alpha\ln\varepsilon + \beta\ln t_F - \gamma\frac{E_a(\text{calcd})}{RT} + \delta \cong \ln t + \ln A - \frac{mE_a(\text{calcd})}{RT}, \tag{7}$$

and then

$$z_{obsd} = \ln\left(\frac{1}{b-a}\ln\frac{a(b-x)}{b(a-x)}\right) \cong \alpha\ln\varepsilon + \beta\ln t_F - \gamma\frac{E_a(\text{calcd})}{RT} + \delta = z_{calcd}. \tag{8}$$

Once we obtain a correlation equation between z_{calcd} and z_{obsd}, the final concentration of the product $z_{F\ (calcd)}$ is calculated using Eqs. 9a and 9b by giving the values necessary for the reaction,

$$x_{F(calcd)} = (b - pa)/(1 - p), \tag{9a}$$

$$p = \frac{b}{a}\exp((b-a)\exp(z_{calcd})). \tag{9b}$$

The expected yield Y_{expect} is therefore expressed as follows,

$$Y_{expect} = 100x_{F(calcd)}/a. \tag{10}$$

2.2. Method of calculation

Geometric optimizations and their vibration analyses were performed for all of the transition states, dienes, and dienophiles at the B3LYP/6-31G(d) level of theory using the Gaussian03 program.[12] All transition states obtained were confirmed to have only one imaginary frequency.

The dienes and dienophiles used for the DFT calculations are shown in Figure 1. In the present study, the products are expressed as combinations of numbers for dienes, and alphabetical letters for dienophiles. For example, **1b** represents the product of the reaction of 1,2,3,4,5-pentachlorocyclopenta-1,3-diene (**1**) and methyl acrylate (**b**).

The activation energies were defined as the energy difference between the total energy of a TS and the sum of the energies for the corresponding diene and dienophile. We have observed that the activation energies based on this definition are very consistent with experimental values at the present level of theory.[13] Although it is possible to consider both *endo* and *exo* products in the Diels–Alder reaction, the former is generally predominant. For example, the *exo* addition product of **1j** was not observed, and thus that reaction was ignored. Figure 2 displays the two types of transition state geometries for **1j** giving *endo* products.

In the *endo-anti* TS, the chlorine atom on the sp^3 carbon points toward the outside, whereas in the *endo-syn* TS, the atom points toward the inside. The C–C distances in the *endo-anti* and *endo-syn* TS were calculated to be ca. 2.23 Å. These lengths represent the general distances for TSs in the Diels–Alder reaction.[14] The *anti*/*syn* ratio was observed to be 90.8:9.2.[15] In the case where two types of products were observed, we apply 'weighted' activation energies[16] for later PLS analyses. Similar weighted activation energies were also calculated for reactions yielding both *endo* and *exo* products.

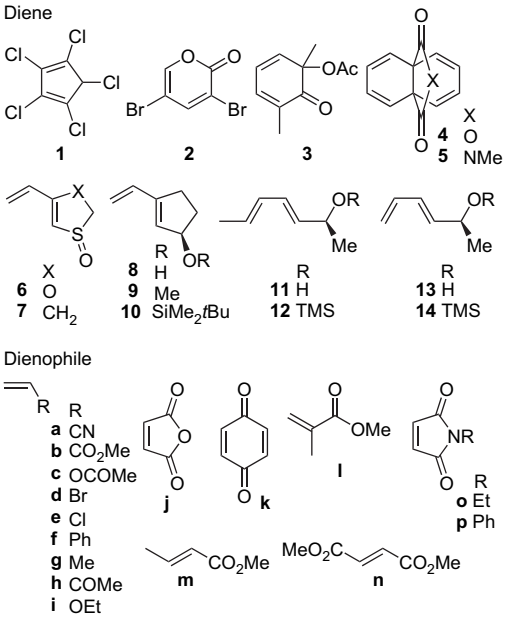

Figure 1. Dienes and dienophiles adopted for the PLS analysis.

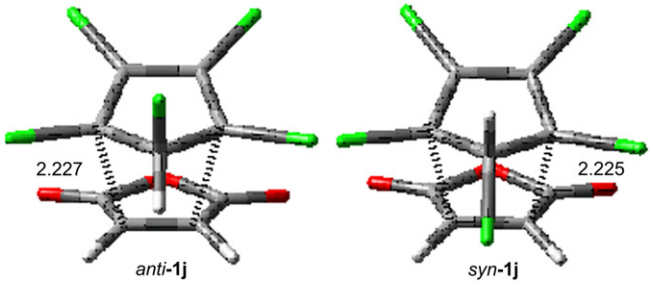

Figure 2. TS structures for *endo-anti* and *endo-syn* products of **1j** (lengths in Å unit).

The Chemish program (Ver.4.12) developed by Funatsu and Arakawa was used for the PLS and GA-PLS (Genetic Algorithm-Based PLS) analyses.[17] This program can perform the Cross-validation technique. The values of the empirical parameters for the GA-PLS computation are as follows: the population (N_p) is 50, the probability of initial variable selection (P_i) is 0.3, the probability of crossover (P_r) is 0.3, the probability of mutation (P_m) is 0.01, and the number of generations (N_g) is 500. These values were determined to be optimal after several GA-PLS computations while changing the values of the empirical parameters. Table 1 lists all of the values used for the PLS analysis.

2.3. Multivariate analysis

Although multivariate analysis was performed using all calculated values, we obtained no clear correlations between z_{obsd} and z_{calcd} as shown in Figure 3a. In other words, there seems to be no correlation between the experimental yields and the theoretical activation barriers. A detailed analysis indicated that the data related to **4** and **5**, as well as to **1a**, **1k**, **2i**, **2l** and **9p** (the circled points) deviated greatly from the correlation. The modified data set excluding these points produced a good correlation in the GA-PLS analysis. Principal component analysis made it clear that the three descriptors discussed above are important for modeling. The PLS analysis was performed again with the modified data set, and three correlation equations were obtained. Figure 3b shows a z_{calcd}–z_{obsd} plot for the correlation equation Eq. 11, with the best-captured variance (R^2=0.956) and the cross-validated R^2 (Q^2=0.940).

$$z_{calcd} = -4.0831\ln\varepsilon + 0.206\ln t + 0.073(-E_a/RT) + 5.407. \quad (11)$$

According to the values of R^2 and Q^2, this correlation equation is considered to be reliable. In fact, almost all of the points of the z_{calcd}–z_{obsd} plot are concentrated on the diagonal line, and thus the obtained correlation equation was appropriate for expressing the relationship between z_{obsd} and the explanatory variables.

In order to confirm that Eq. 10 produces a Y_{calcd} similar to the experimental value, a TS giving an *endo* product was sought for the reaction between **2** and *N*-ethylmaleimide (**o**).[15g] The activation energy was calculated to be 17.9 kcal mol^{-1}. The

Table 1
The calculated E_a values and the actual reaction conditions used for PLS analysis and GA-PLS analysis

Diene	Dienophile	E_a (calcd)	Yield (%)	Temp (°C)	Time (h)	*a* (mol/L)	*b* (mol/L)	ε
1	**a**	20.3	62	100	6	0.79	0.8	2.274
1	**b**	19.6	60	66	6	0.26	0.32	2.274
1	**c**	21.1	73	120	8	0.26	0.32	2.274
1	**d**	20.8	90	100	18	6.5	25	6.081
1	**e**	22.5	97	100	26	6.5	26.2	6.081
1	**f**	19.3	73	101	0.83	0.79	0.8	2.274
1	**g**	20.8	90	100	6	7.8	25	60.81
1	**j**	21.3	73	105	3	0.26	0.32	2.274
1	**k**	22.4	99	112	2.5	0.26	0.32	2.274
2	**a**	20.4	90	100	12	0.1	0.3	2.379
2	**i**	14.5	75	50	48	0.1	0.3	8.93
2	**l**	17.4	84	100	24	0.1	0.3	8.93
2	**m**	22.8	82	100	96	0.1	0.3	2.379
2	**n**	20.7	80	100	72	0.1	0.3	2.379
2	**b**	17.3	84	100	5	0.1	0.3	2.379
2	**h**	15.4	84	100	5	0.1	0.3	2.379
3	**j**	14.4	53	80.1	26	0.302	0.303	2.284
4	**k**	20.0	68	80.1	16	0.05	0.07	2.284
4	**t**	17.6	85	80.1	16	0.052	0.095	2.284
5	**k**	19.3	88	80.1	16	0.07	0.072	2.284
6	**p**	17.9	89	70	168	1.042	1.042	2.379
7	**p**	16.1	72	80	48	0.81	0.81	2.379
8	**p**	8.3	78.4	23	36	0.909	0.909	2.379
9	**p**	12.3	73	23	24	0.175	1.009	2.379
10	**p**	13.8	84	23	36	2.24	2.24	2.379
11	**p**	8.4	78.3	25	120	0.333	0.333	2.284
12	**p**	12.5	74.5	25	240	0.333	0.333	2.284
13	**j**	8.1	82.5	23	72	0.33	0.33	2.284
14	**j**	9.2	82.5	23	120	0.33	0.33	2.284

experiment for this reaction at 100 °C in toluene for 12 h gave a 76:24 ratio mixture of *endo* and *exo* products at a 92% yield. Eq. 11 gave z_{calcd}=2.052 by applying the calculated E_a as well as the reaction conditions. From this z_{calcd} value and Eqs. 9a and 9b, Y_{calcd} was calculated to be 84.9%, which differs by only 7.3% from the experimental value. It is believed that this difference is small enough to estimate the crude yield of the reaction, and to determine the order for the synthesis experiments.

2.4. Correlation equation

It was necessary to determine why the z_{obsd} for several compounds had to be excluded from the PLS analysis shown in Figure 3. Diene **2** has two types of carbon atoms participating in this reaction. Although the one carbon is greatly activated by both the carbonyl group and the bromine atom, the other has no such substituents. Dienophile **i** has an ethoxyl group. As these substituents change the electronic properties of the carbons, the two C–C distances in **2i** (TS) become unequal. Specifically, one of the C–C bond distances between the activated carbons was calculated to be 2.934 Å, which was longer by 1.041 Å than the other (1.893 Å), as shown in Figure 4. In the geometry from the IRC at 5.01 amu$^{1/2}$ Bohr, these C–C lengths were calculated to be 1.558 and 2.137 Å, respectively. The former C–C bond has almost completely formed, while the latter is still in the process of making a new bond. The IRC calculations show the deviation from the concerted mechanism of the Diels–Alder reaction, although the mechanism does not change completely from a continuous to a stepwise reaction. The features in the TS geometry of **1a** and **2i** fall into this category. Another effect

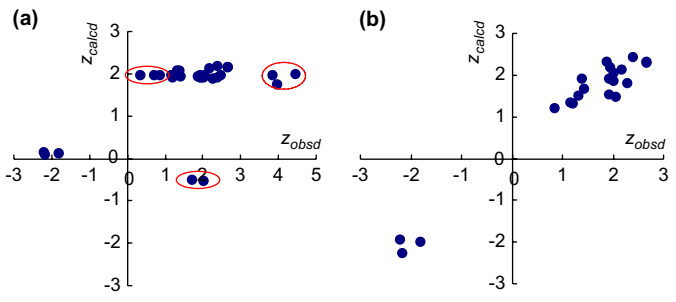

Figure 3. z_{calcd}–z_{obsd} plots based on the PLS analysis using (a) all the calculated data and (b) those of the modified data set.

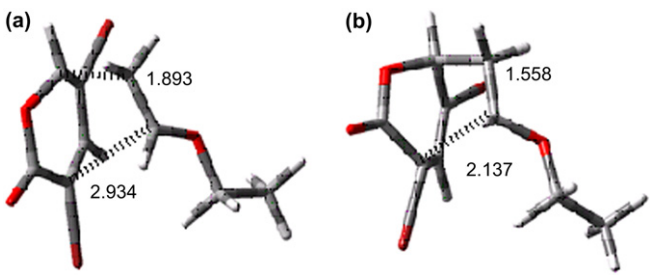

Figure 4. Geometry for **2i** on the IRC at (a) TS (0.0 amu$^{1/2}$ Bohr) and (b) 5.01 amu$^{1/2}$ Bohr.

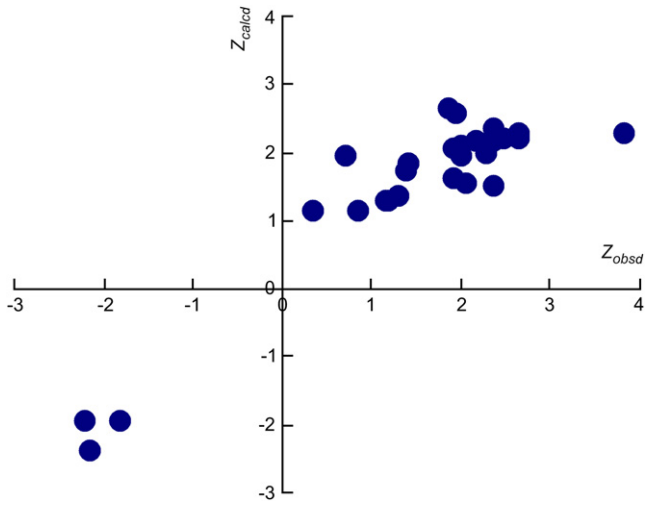

Figure 5. z_{calcd}–z_{obsd} plot based on the PLS analysis using the modified concentrations.

is considered in 1,1-disubstituted dienophile **l**, because there should be a large steric hindrance in **2l** (TS), in which one C–C length (2.773 Å) is longer by 0.841 Å than the other (1.932 Å). Therefore, these data should be excluded from the PLS analysis.

The reactants **4** and **5** contain two diene fragments in their structure. Both fragments can participate in a Diels–Alder reaction with the dienophiles. This means that the concentration of the dienes should be doubled in Eq. 1. Since the dienophile **k** has two double bonds, twice the concentration of the dienophile should be used in the PLS analysis. Compound **5k** has two diene parts in **5** as well as two ethylene parts in **k**. Another PLS analysis was performed using the z_{obsd} calculated by doubling their initial concentrations, to produce Eq. 12,

$$z_{calcd} = -4.328\ln\varepsilon + 0.377\ln t_F + 0.065\left(-\frac{E_a(\text{calcd})}{RT}\right) + 5.057, \quad (12)$$

where R^2 (0.856) and Q^2 (0.826) indicated that the reliability of Eq. 12 was not as good as that of Eq. 11, as shown in Figure 5.

According to the R^2 and Q^2 values, the correlation equation is acceptable. This is because we did not try to estimate actual yields but rather used rough experimental yields. Thus, Eq. 12 is more suitable for ranking the synthesis routes than Eq. 11.

3. Conclusions

In the present study, we tried to predict trends in experimental yields observed in Diels–Alder reactions using the activation energies from the DFT calculations, and reaction conditions such as the reaction temperature, reaction time, and dielectric constants of the solvents. The PLS analysis gave a clear correlation between the experimental yields and the described data set. Moreover, the correlation equation approximately predicted the yield of **2o**. These results showed that rough yields from Diels–Alder reactions could be predicted using Eq. 10, which was obtained by combining the second order kinetic equation and the Arrhenius equation with the PLS analysis. A similar procedure will likely be successful for other reactions such as ene, Wittig, and aldol reactions. Once it becomes possible to obtain correlation equations to predict the yields of different synthesis reactions, one can compare the yields from different synthesis routes using theoretical calculations. The synthetic routes for SRDS could be ranked on the basis of their predicted yields, and the order of experiments can be determined before any synthesis experiments are started.

Acknowledgements

This study was supported in part by a Grant-In-Aid from the Ministry of Education, Sports and Culture of Japan.

References and notes

1. Corey, J. E.; Cheng, X.-M. *The Logic of Chemical Synthesis*; John Wiley and Sons: New York, NY, 1989.
2. Gasteiger, J. *Chim. Ind. (Milan)* **1996**, *64*, 714–721.
3. Funatsu, K.; Sasaki, S. *Computer Chemistry Series 2, ''AIPHOS''*; Kyoritsu Shuppan: Tokyo, 1994.
4. (a) Satoh, K.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 316–325; (b) *AIPHOS/KOSP V1.0*; Fujitsu: Tokyo, 2003.
5. (a) Hori, K.; Sadatomi, H.; Okano, K.; Sumimoto, M.; Miyamoto, A.; Hayashi, S.; Yamamoto, H. *J. Comput.-Aided Chem.* **2007**, *8*, 57–66; (b) Hori, K.; Okano, K.; Yoshimura, K.; Nishida, A.; Yamamoto, H. *J. Comput.-Aided Chem.* **2005**, *6*, 30–36; (c) Hori, K.; Yamaguchi, T.; Okano, K. *J. Comput.-Aided Chem.* **2004**, *5*, 26–34.
6. (a) Hori, K.; Nagoshi, Y.; Yamazaki, S. *J. Comput.-Aided Chem.* **2000**, *1*, 89–97; (b) The TSDB is now accessible at the site, http://www.tsdb.jp.
7. (a) Not only as an index to ranking synthesis routes, prediction of the experimental yields is a powerful tool for optimizing organic reactions. (b) Roberts, B. A.; Strauss, C. R. *Molecules* **2004**, *9*, 459–465.
8. Okano, K.; Satoh, K.; Takahasi, H.; Hori, K. *J. Comput.-Aided Chem.* **2005**, *6*, 57–66.
9. (a) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724–728; (b) Dunning, T. H., Jr.; Hay, P. J. *Modern Theoretical Chemistry*; Schaefer, H. F., III, Ed.; Plenum: New York, NY, 1976; Vol. 3, pp 1–28; (c) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270–283.
10. When the initial concentrations of the reactants are the same, it is necessary to use a different equation: $kt = ((1/a - x) - (1/a))$.
11. Inamoto, N. *Hammett rule*; Maruzen: Tokyo, 1983; pp 97–99.
12. 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 03, Revision C.02*; Gaussian: Wallingford, CT, 2004.
13. Hori, K. *Petrotech* **2007**, *30*, 413–419.

14. Manoharan, M.; De Proft, F.; Greerlings, P. *J. Org. Chem.* **2000**, *65*, 7971–7976.
15. (a) Williamson, K. L.; Hsu, Y.-F. L. *J. Am. Chem. Soc.* **1970**, *92*, 7385–7389; (b) Auksi, H.; Yates, P. *Can. J. Chem.* **1981**, *59*, 2510–2517; (c) Williamson, K. L.; Hsu, Y.-F. L.; Lacko, R.; Youn, C. H. *J. Am. Chem. Soc.* **1969**, *91*, 6129–6138; (d) Krow, G. R.; Carey, J. T.; Zacharias, D. E.; Knaus, E. E. *J. Org. Chem.* **1982**, *47*, 1989–1993; (e) Reitz, A. B.; Jordan, A. D., Jr.; Maryanhoff, B. E. *J. Org. Chem.* **1987**, *52*, 4800–4802; (f) Macaulay, J. B.; Fallis, A. G. *J. Am. Chem. Soc.* **1990**, *112*, 1136–1144; (g) Cho, C.-G.; Kim, Y.-W.; Lim, Y.-K.; Park, J.-S.; Lee, H.; Koo, S. *J. Org. Chem.* **2002**, *67*, 290–293.
16. The weighted activation energies were obtained by applying the product ratios to each of the activation energies, i.e., E_a (**1j**)=0.918E_a (*endo-anti*)-+0.092E_a (*endo-syn*).
17. Funatsu, K.; Arakawa, M. *Chemish (Ver. 4.12)*; ChemInfoNavi: Iwakuni, 2004.