Fast Evaluation of Potential Synthesis Routes Using Transition State Database (TSDB)

Kenji Hori\textsuperscript{a,*}, Aki Hasegawa\textsuperscript{b}, Noriaki Okimoto\textsuperscript{b}, Suzuko Yamazaki\textsuperscript{c}

\textsuperscript{a} Division of Materials Science, Graduate School of Science and Technology for Innovation, Yamaguchi University, Tokiwadai, Ube 755-8611, Japan.
\textsuperscript{b} Laboratory for Computational Molecular Design, RIKEN Center for Biosystems Dynamics Research (BDR), 6-2-4 Furuedai, Suita, Osaka, 565-0874, Japan.
\textsuperscript{c} Division of Earth Science, Biology, and Chemistry, Graduate School of Science and Technology for Innovation, Yamaguchi University, Yamaguchi, 753-8512, Japan

(Received October 15, 2019; Accepted November 21, 2019)

Information of transition states of similar reactions is the key to locating those of unknown reactions. In order to utilize this feature, we are constructing a database, called QMRDB, which gathers results of quantum mechanical calculations for elementary reactions as well as those for related molecules. Another database (TSDB) stores information of name reactions in organic synthesis. Retrieval results from these databases are used for analyzing reaction mechanisms which have not been experimentally examined. We developed a cloud system managing both the two databases and theoretical calculations. The present paper describes the summary of the TSDB cloud system and how to use it to perform \textit{in silico} screenings for synthesizing drug candidates.

Key Words: Transition state database, fast evaluation of potential reactions, theoretical calculations, drug candidate

Introduction

In developing synthesis routes, organic chemists create several ones which are likely to produce the target. We can use synthesis route design systems (SRDS) such as AIPHOS to create synthesis routes on the basis of chemoinformatics \cite{1,2,3} without deep knowledge of organic synthesis. In both cases, hard experimental works including tries and errors are required to complete target syntheses. It has been considered that theoretical calculations are useless for developing synthesis routes.

We pointed out that the computational chemistry is applicable to analyzing synthesis routes \cite{4,5} which were created by organic chemists or SRDS, and have never challenged for synthesizing new compounds. We applied this method to several targets \cite{6,7,8}, in which the most complicated target is I with five plausible routes shown in Scheme 1\cite{9}.

It was calculated that Route A does not produce the
target since there is a side reaction with the lower barrier. TSSs of Routes D and E were optimized and ΔG‡ (the free energy of activation) values were calculated to be 57.0 and 47.5 kcal/mol, respectively. These routes should not be adopted since their precursors were much more complicated than the target itself.

Although the barrier of Route B with BF₃ as an additive is not so low (49.1 kcal/mol), the precursor 2 is easy to obtain so that this route is worth trying.

It was confirmed that Route A produced the side product as expected. Although reactions of Route B in benzene, toluene and xylene did not proceed, that in xylene with BF₃•OEt₂ at 120 °C produced the target with the yield of 21%. It means that in silico screening worked.

We called the expectation on the basis of theoretical calculations “in silico” screening as shown in Figure 1. In order to complete the screening, we have to confirm existence of TS structures of main and side reactions, and to calculate ΔG‡ and ΔG (the free energy of reaction) values as well as solvent effects for reactions. After these considerations, we can determine their effectivities as synthesis routes. These processes exclude useless synthesis routes and make it possible to rank the routes for future experimental works. However, in silico screenings have to be finished before patience of organic chemists are exhausted.

### Construction of TSDB Cloud System

#### QMRDB and TSDB

It is common knowledge that locating TS of reactions is difficult very much so that it takes long CPU time. In order to perform fast evaluation of potential synthesis routes, information of similar reactions is extremely useful for analyzing a new reaction. Therefore, we intended to construct a database with information of theoretical calculations related to name reactions.

We first constructed a database storing optimized structures with related keywords, their SMILES strings, chemical equations of the jpeg format, reaction names with “/TS/” in the keyword fields. The relations among reactants, products and TS are...
The meanings of information of (A) for the elementary reaction are as follows:
(1) **Rn_ID** : "-1" means that information relates to an elementary reaction.
(2) **ID_R1, ID_P1**: ID numbers of reactants and products. "0" separates them. Molecules with negative ID numbers are used for calculating ΔG² and ΔG values of the reaction.
(3) "-88888888" expresses the end of data for reactants and products.

Name reactions in organic synthesis usually consist of several elementary reactions. Their relation is expressed such as (B) shown above. TS1 and TS2 express ID numbers for TS data and "-99999999" means the end of data. A database for the name reactions are called "transition state database (TSDB)" which stores the relation between elementary reactions in QMRDB.

The line below (B) shows that the reaction consists of two elementary reactions; the first reaction proceeds to form an intermediate and the second one is its internal reaction to produce acetonitrile and Cl₂P(OH)(=O). Numbers such as 3194 and 4416 between "-88888888" and "-99999999" are ID numbers specifying the elementary reactions. From the information, a TSDB window for the dehydration reaction of acetoamide with Cl₂P=O is created as shown in Figure 2. This window also reveals the name of reaction, overall reaction equation, two ΔG² values (36.8 and 26.3 kcal/mol) for each elementary reaction.

### TSDB Cloud System

It is necessary to construct a system which makes it possible efficiently to analyze reaction mechanisms using TSDB/QMRDB. The TSDB cloud system is constructed to combine the two databases with related three programs as shown in Figure 3.

The first is the cStructure program searching reactions which are similar to a target one. We can access the two databases via internet. The PostgreSQL program searches keywords. It is possible for the OpenBabel program [10] to perform structure retrievals. The program calculates Tanimoto coefficients by using a SMILES string of reactant or product, and both of them as queries for similar reaction searches in TSDB/QMRDB. The first 10-20 mechanisms are displayed in Web browser on the basis of the magnitude of the coefficients.

The JSmol program [11] displays molecular structures of retrieval results in Web browser. TS coordinates can be downloaded and used for constructing initial structures for new TS calculations as will be discussed later.

The second is the sStructure program which adds substituents to accurate positions of reference TSs and makes inputs for the Gaussian program. The program sends Gaussian Jobs to and accepts results from the cloud server which manages theoretical calculations.

A single command is enough to put SMILES string, MOL file [12], optimized coordinates and log file with vibration frequency calculation of an optimized structure into QMRDB. The rMol program in the cloud server manages this process. The cStructure program is also used for storing information of elementary and name reactions in TSDB/QMRDB.

At present, QMRDB system includes 2020 TS structures for 260 types elementary reactions together with 10,864 optimized coordinates of molecules. By using information in QMRDB, 425 reactions of 286 different types of name reactions are stored in TSDB.

### Substitution Method

It has to be emphasized that retrieval information from TSDB/QMRDB is not that for a target reaction but those for similar reactions useful for *in silico* screenings. Therefore, we always have new calculations for corresponding targets. A TS structure of a similar reaction extracted from the database is used for constructing a good initial structure for a succeeding TS calculation.

The sStructure program can handle TS motifs to construct initial structures for new calculations shown in Scheme 2.

We can easily substitute a fragment in the reference TS with a complicated substituent in the following order;
1. Add two methyl groups to the position [D].
2. Select three atoms in the reference TS motif and
those of the substituent, for example, \([A]-[C]\) and \([a]-[c]\), respectively.

3. Merge the substituent with those of the TS motif by overlapping the selected fragments to make an initial structure.

4. Perform Optimization from the initial structure under a constraint condition such as the fixed distance between \([B]\) and \([D]\).

5. Extract the resultant structure for a succeeding TS optimization.

6. Locate the real TS for the reaction by using the previously extracted coordinates.

The calculation using the initial structure thus formed usually takes short time to locate the real TS one. In order to conduct the substitution method for \textit{in silico} screenings, we developed a program which merges a substituent with a TS motif and implemented in the iStructure program.

In this case, \(\Delta G^\ddagger\) and \(\Delta G\) values for the reaction were calculated to be 10.8 and 1.3 kcal/mol. The barrier is low enough for the reaction to proceed although the product is unstable a little.

It has to be pointed out that this procedure locates a TS structure which may not be the most stable one in all the possible conformations. We need a conformation analysis which finds the TS structure with the smallest \(\Delta G^\ddagger\) value for the target reaction. We are now investigating the method for this purpose.

**Application of \textit{In Silico} Screening**

**Target Molecules**

Molecular docking calculations are useful to select candidate compounds with high affinities for target proteins. We have done such calculations for a target protein, protein phosphatase methylesterase 1 (PME-1) [13] in the CREST project of Funatsu as a principal investigator[14]. In the project, a library with huge number of drug candidate compounds (a very large-scale virtual library, VLSVL) was created by applying name reactions to drug-like molecules derived from compound database ZINC15[15]. A part of VLSVL molecules was extracted to apply molecular docking calculations to select candidate compounds which bind to PME-1.

Figure 4 shows twelve candidate compounds with relatively good docking affinities. Candidate compound were created using specified reactions such as aldol reaction, olefin metathesis and Strecker reaction. We chose three molecules for \textit{in silico} screening with aid of the TSDB cloud system since they are considered to be less toxic than the others. The reaction mechanisms used for VLSVL were first examined to confirm whether or not TSs for the reactions exist. The 6-31G(D) level of theory was adopted to locate all the TS structures by use of the Gaussian 09 program [16]. It was easy to construct initial structures of TS optimizations using the iStructure program shown above.

The aldol reaction is used to create 6 and 7 in VLSVL and their synthesis routes are expressed in Eqs. (1) and (2),
The key of the reactions is whether or not enolate ions such as 9 and 10 form so that they are supposed to exist for TS calculations.

ΔG‡ values for the aldol reactions were calculated to be 13.3 and 8.8 kcal/mol for 6 and 7, respectively. These values are enough small for the reactions to proceed. Both of the reactions are exothermic since ΔG values are -3.0 and -5.9 kcal/mol.

Strecker reaction for this molecule was calculated to have a high barrier, as high as 54.0 kcal/mol. We cannot use this reaction to synthesize this candidate.

On the basis of the in silico screening for the targets, we conducted the experiments for 7. Although we have done several efforts, no reactions proceeded. It does not mean formation of the Li enolate ion 10.

Alternative Reactions

We have to create alternative synthesis routes in such case that synthesis routes adopted for the VLSVL construction was experimentally confirmed not to proceed. In the present case, two alternative routes were considered for further investigations.

The one uses McMurry coupling [17] shown in Eq. (4) A coupling of two ketones is expected to proceed in reaction mixtures. The NO₂ group of 14 is planned to be reduced to the NH₂ one after its synthesis. It was calculated that there exist TSs for the reaction paths to form 14 after the formation of 13.

It is possible to consider three combinations of ketones in the reaction mixture. They are 11+11, 12+12 and 11+12 producing 13. The relative stability of intermediates was examined and their ΔG values were calculated to be 44.1, 55.7, and 47.8 kcal/mol, at the B3LYP/LANL2DZ level of theory, respectively. As 13 is less stable than that of 11+11, the intermediate will not be formed even though 11 and 12 are mixed in solution. We have not adopted experiments of this synthesis route.

The other is Mukaiyama aldol reaction [18] shown in Eq. (5). As the reaction usually uses Lewis acid to activate carbonyl compounds, one BF₃ molecule was included in analyzing the reaction mechanism. It was calculated that the two steps were required to form 17 by the reaction of 15 and 16. The ΔG‡ and ΔG values for the step were calculated to be 28.8 and -6.4 kcal/mol at the B3LYP/6-31G(D) level of theory, respectively. 17 changes easily its form to 18 after treatment of the reaction mixture and its dehydration leads to the formation of the precursor 14 of 7. Therefore, this reaction is worth executing experiments.

It was confirmed that three different types of conformers of Mukaiyama reagents including 16 formed. According to MS spectra of the product after adding 12 to
the reaction mixture, the final one was determined to be not the target but the dimer 19.

Further investigations are required to clarify what makes this difference between theoretical calculations and experimental results.

Concluding Remarks

Chemoinformatic tools easily design new functional molecules as well as their synthesis routes. We have described that in silico screenings using the TSDB cloud system is useful to confirm whether or not the created synthesis routes are applicable to synthesizing targets. It is, however, very difficult to develop their real synthesis routes in experiments.

Unfortunately, we could not obtain the target compound although theoretical considerations suggested the possibility of their synthesis. It is necessary to make continuous efforts to prove the potential of in silico screenings for developments of synthesis routes for organic compounds.

Acknowledgement

The authors acknowledge the support of the Core Research for Evolutionary Science and Technology (CREST) project “Development of a knowledge-generating platform driven by big data in drug discovery through production processes” of the Japan Science and Technology Agency (JST). A part of this work was supported by a Grant of the project for Formation of the Core Research Center from Yamaguchi University.

References and Notes

[1] K. Funatsu, S. Sasaki, Computer Chemistry Series 2, Kyoritsu Shuppan: Tokyo, Japan, 1994.
[2] J. Gasteiger, Chim. Ind.,(Milan), 64, 714 (1982).
[3] Ellen R. Laird and William L. Jorgensen, J. Org. Chem., 55, 9-27(1990).
[4] H. Yamamoto T. Yamaguchi K. Yoshimura, M. Sumimoto, K. Hori, J. Synth. Org. Chem. Jpn., 70, 722 (2012).
[5] K. Hori, T. Sadatomi, K. Okano, A. Miyamoto, H. Kuroda, M. Sumimoto, H. Yamamoto, Molecule, 15, 8289 (2010).
[6] K. Hori, H. Sadatomi, K. Okano, M. Sumimoto, A. Miyamoto, S. Hayashi, H. Yamamoto, J. Comput. Aided Chem., 8, 57-66 (2007).
[7] K. Hori, K. Okano, K. Yoshimura, A. Nishida, H. Yamamoto, J. Comp. Aided Chem., 6, 30-36 (2005).
[8] K. Hori, T. Yamaguchi, K. Okano, J. Comp. Aided Chem., 5, 26-34 (2004).
[9] K. Hori, M. Sumimoto, T. Murafuji, AIP Conference Proceedings, 1702, 090019 (2015).
[10] James L. Melville, J. D. Hirst., J. Chem. Inf. Model., 47, 2626-634(2007).
[11] http://jmol.sourceforge.net/
[12] A. Dalby, J. G. Nourse, W. D. Houenshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J.Lauf, J. Chem. Info. Model. 32, 244(1992).
[13] E. Ogris, X. Dui, K. C. Nelson, E. K. Mak, X. X. Yu, W. S. Lane, D. C. Pallas, J. Biol. Chem., 274, 14382–14391(1999).
[14] https://www.jst.go.jp/kisoken/crest/en/project/44/e44_01.html
[15] T. Sterling, J. J. Irwin. J. Chem. Inf. Model. 55 (2015.,
[16] Gaussian 09, Revision D.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria,M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Liyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian, Inc., Wallingford CT, 2013.
[17] J. E. McMurry, M. P. Fleming, J. Am. Chem. Soc. 96, 4708(1974).
[18] T. Mukaiyama, K. Narasaka, K. Banno, Chem. Lett., 323(1974).