ORGANIC
LETTERS

2005
Vol. 7, No. 16
3541−3544

ROBIA: A Reaction Prediction Program

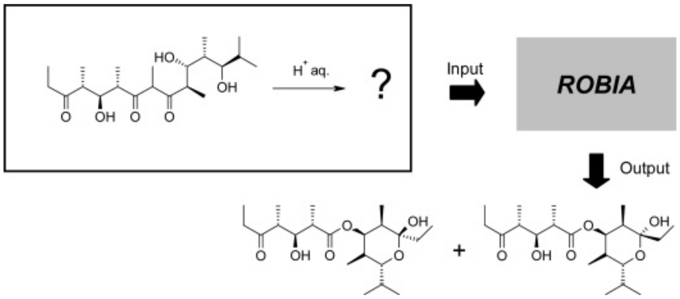
Ingrid M. Socorro,[†] Keith Taylor,[‡] and Jonathan M. Goodman*,[†]

University Chemical Laboratory, Unilever Centre for Molecular Science Informatics, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K., and Elsevier MDL, 14600 Catalina Street, San Leandro, California 94577

jmg11@cam.ac.uk

Received May 31, 2005

ABSTRACT



A reaction prediction program, ROBIA, has been developed. This interactive computer program predicts the products of organic reactions from the starting materials and the reaction conditions, on the basis of the selected transformations within its database. This mechanistic approach generates a large number of products, from which the most important are selected using filters and molecular modeling calculations. The procedure has been applied to the possible biosynthesis pathway of dolabriferol.

The prediction of the products of organic reactions is a very challenging task, as many competing processes must be analyzed. We have developed a program, ROBIA (Reaction Outcomes By Informatics Analysis), which generates possible reaction pathways and automatically assesses the most favorable route.

Since the development of the first programs[1] for computer-aided organic synthesis a few decades ago, many programs have been developed to assist chemists in planning synthesis,[1−4] predicting reaction outcomes,[5−9] drug design, and predicting metabolic routes.[10−11] Our work has been focused on the development of computational tools for the prediction and analysis of chemical reactivity. As a result, a new interactive computer program for reaction prediction, ROBIA, has been developed with the purpose of helping chemists to anticipate, analyze, and understand their results. The system is able to predict the outcome of organic reactions given the starting materials and conditions using selected organic transformations (Table 1). The program achieves reaction prediction by combining general knowledge of organic chemistry with molecular modeling.

The collection of a set of organic chemistry rules, along with two programming languages (Java[12] and MDL Cheshire[13]) and the appropriate algorithm, leads to a rule-based program in which knowledge of reactions has been coded. The program applies this series of rules to determine the reactivity

[†] University of Cambridge.

[‡] Elsevier MDL.

(1) (a) Corey, E. J. *Pure Appl. Chem.* **1967**, *14*, 19. (b) Johnson, A. P.; Marshal, C.; Judson, P. N. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 411. (c) Johnson, A. P.; Marshal, C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 418. (d) Johnson, A. P.; Marshal, C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 426.

(2) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*.

(3) Funatsu, K.; Sasaki, S. *Tetrahedron Comput. Methodol.* **1988**, *1*, 27.

(4) Satoh, K.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 316.

(5) Jorgensen, W. L.; Laird, E. R. *Pure Appl. Chem.* **1990**, *62*, 1921.

(6) Ihlenfeldt, W. D.; Gasteiger, J. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613.

(7) Höllering, R.; Gasteiger, J.; Steinhauer, L.; Schulz, K.; Herwig, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482.

(8) Satoh, H.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 173.

(9) Sello, G. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 713.

(10) Borodina, Y.; Sadym, A.; Filimonov, D.; Blinova, V.; Dmitriev, A.; Poroikov, V. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1636.

(11) Klopman, G.; Dimayuga, M.; Talafous, J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1320.

(12) Java Version 1.4.1.; http://java.sun.com/

(13) Cheshire Studio Version 3.0.0.54.; Elsevier MDL, 14600 Catalina Street, San Leandro, CA 94577.

10.1021/ol0512738 CCC: $30.25
Published on Web 07/12/2005

Table 1. Contents of the Transformation Block

type of transformation
nucleophilic substitution
enolate formation reaction
aldol reaction
retro-aldol reaction
dehydration of aldol product
claisen condensation
alkylation of enolates
Michael addition
acetal formation−decomposition
hemiacetal formation−decomposition
Diels−Alder reaction

of the molecules, making decisions on primary aspects of organic reactivity, such as the location of reactive sites and which bonds are to be broken or made. By this method a mechanistic approach is taken to reaction prediction.

Molecular and quantum mechanics calculations can make relevant contributions in the process of structure selection. Thus, molecular modeling has been integrated with the rule-based program for this purpose. The use of shell scripts allows the whole calculation process to be automated. These scripts are responsible for launching the molecular modeling programs used (Macromodel[14] and Jaguar[15]), calling Maestro[16] scripts, and running other programs we have developed to analyze the results of the calculations. The Maestro scripts prepare the input structures and set all of the parameters to perform energy minimizations and conformational searches. In this way, the program is able to make use of molecular modeling at any time in an automated way.

Thus, the program goes from the reactants to the products through a sequence of intermediate steps in which all possible intermediate structures are formed and then filtered and selected according to the coded rules and the reaction conditions (Figure 1). The program first makes use of the *reactive sites block*, which determines possible reactive sites of the input molecule. Then, it uses the *chemical transformations block* that contains a library of chemical reactions from which one is selected (Table 1). It is our aim to extend this list to more reactions. The intermediates formed are then filtered by the *filters block*. It contains rules to filter structures that are unlikely to be formed and structures that have already been formed in the same step. If there are any chiral centers with uncertain configuration or new chiral centers formed during the reaction, the program generates and considers all possible stereocenters.

Finally, molecular and/or quantum mechanics calculations are performed on all of the structures. This way, structures with lower energy are selected. ROBIA has been coded using the Java programming language and MDL Cheshire chemical scripting language. The user interface is MDL ISIS/Draw, where the structures of the reactants are entered and the products are displayed. The user can also visualize any intermediate formed during the process here.

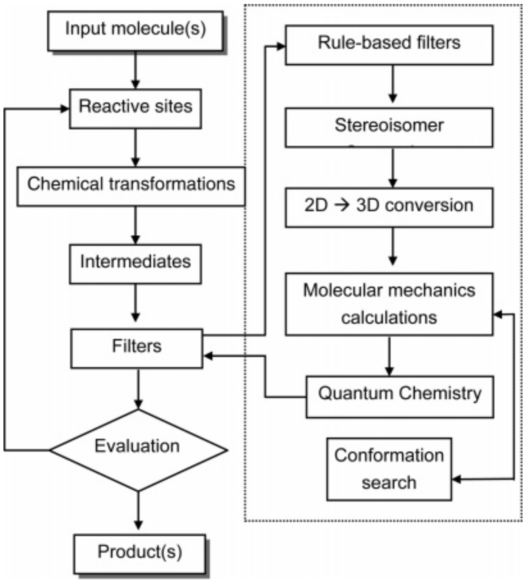
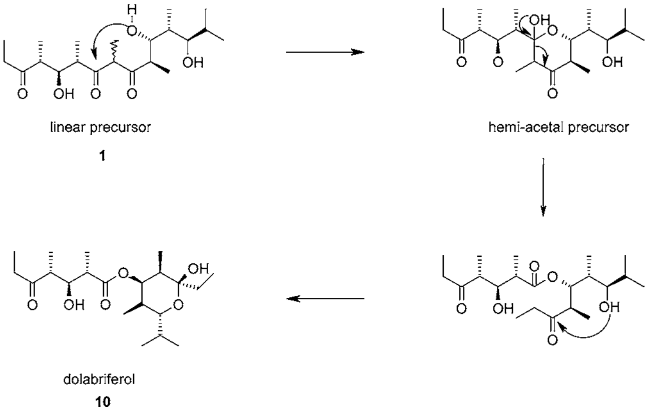


Figure 1. Algorithm.

An application of ROBIA is shown in the example outlining the capabilities of the program. Dolabriferol[17] is a complex natural product of uncertain biosynthetic origin. A pathway from the linear precursor (Scheme 1) has been suggested. However, an enormous number of competing reactions are possible if this linear precursor **1** is allowed to form acetals, do aldol reactions, and so on. To analyze the suggestion, it is necessary to consider all of the possible schemes and decide if the one leading to dolabriferol (**10**) is more likely than all of the rest. This would be an extremely

Scheme 1. Proposed Rearrangement To Form Dolabriferol



[14] Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem*. **1990**, *11*, 440−467.

[15] Jaguar version 4.2; Schrodinger, Inc., Portland, Oregon, 2000.

[16] Maestro Version 5.0.019; Schrodinger Inc, Portland, Oregon, 2000.

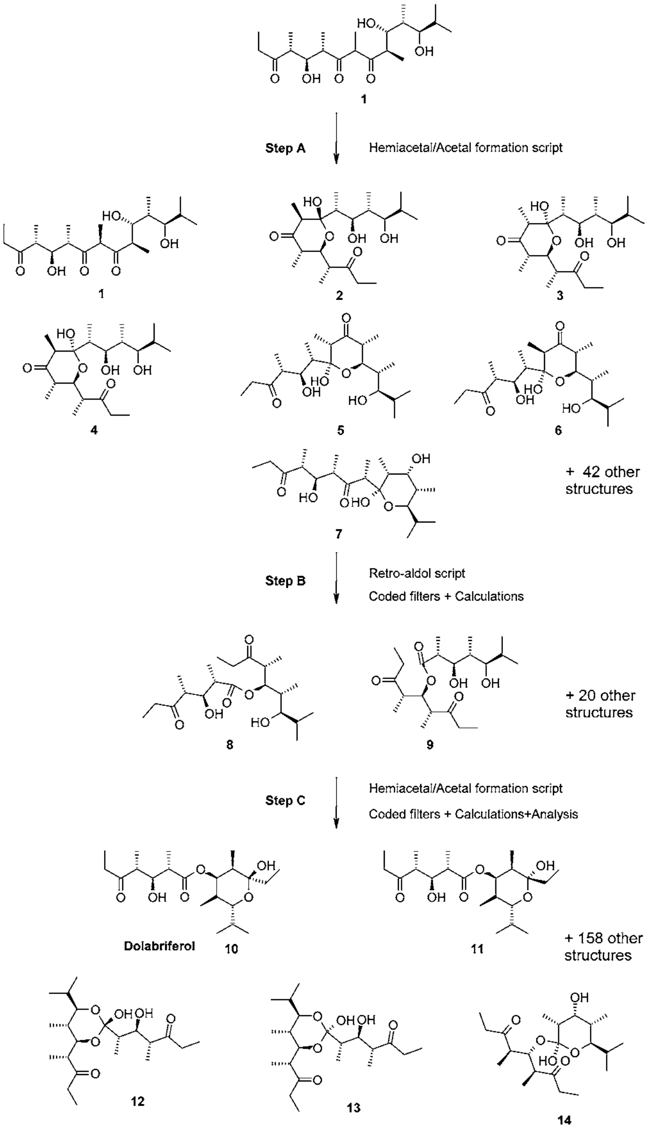
[17] Ciavatta, M. L.; Gavagnin M.; Puliti R.; Cimino G. *Tethrahedron* **1996**, *52*, 12831.

time-consuming task to do by hand, and mistakes would inevitably occur. ROBIA is able to automate the process.

We have applied the program to investigate this possible pathway under thermodynamic conditions. In this example we compare the reaction scheme generated by the program when the linear precursor **1** (under aqueous acidic conditions) is used as the input molecule to the proposed rearrangement to dolabriferol (Scheme 1).

Scheme 2 shows a selection of the molecules generated by the program in this reaction pathway. The sequence of steps the program performs from the input molecule **1** to the products is shown in the algorithm in Figure 1. Thus, after analysis of **1** the program generates all possible hemiacetals and acetals (step A in Scheme 2) according to its rules (scripts contained in the *chemical transformations block*). The reactant remains after this step to allow for the possibility step A has no effect. Four-membered ring structures are removed because of ring strain compared to all other generated structures by the *filters block*. Then, all possible stereoisomers are formed from these structures by assigning a configuration to nonspecified chiral centers while maintaining the configuration of the rest. Next, ROBIA transforms these 2D generated structures into 3D structures ready for the calculations. The structures obtained are then optimized by doing an energy minimization on them in MacroModel.[14] This is followed by a Monte Carlo[18] conformational search to find the most stable conformation for each. After analysis of the results obtained from these calculations, 23 structures are selected, **2**−**7** being among them.

Scheme 2. Some of the Structures Predicted and Generated by the Program


1
Step A
Hemiacetal/Acetal formation script
1
2
3
4
5
6
7
+ 42 other structures
Step B
Retro-aldol script
Coded filters + Calculations
8
9
+ 20 other structures
Step C
Hemiacetal/Acetal formation script
Coded filters + Calculations+Analysis
Dolabriferol
10
11
+ 158 other structures
12
13
14

The program then continues applying its rules on these selected structures, discarding the rest. After application of the retro-aldol transformation script (step B), a second set of intermediate structures is formed. The previous sequence of filtering, generating all stereoisomers, and transforming the structures into 3D is repeated for this new set of molecules. They are then optimized and a conformational search is done on them. After analysis of the results all the structures are selected as possible intermediates leading to a set of potential products formed after step C. In this last step all possible hemiacetals and acetals are formed. As before, the same sequence of steps is applied on them, from structure filtering to doing conformational searches on them. Finally, single point ab initio calculations on all the products are performed using the 3-21G basis set and a continuum water model in Jaguar.[15]

This level of theory was chosen because it is fast and gives a much better comparison between competing functional groups than molecular mechanics. Higher levels of theory are easily accessible through this process but are much more time-consuming. These calculations are also done on the rest of the structures generated by the program that were not filtered by the rule filters or after analysis of the results of the conformational searches. The results of these calculations show that structures **7** and **10**−**14** (Table 2) have the lowest energy values. Energy profiles leading to these structures are shown in Figure 2, which represents the energies of all of the structures from **1** leading to **7** and **10**−**14** but does not show transition state energies. Structures **10**−**14** are

Table 2. Single Point ab initio Calculation Results Using RHF/3-21G/Water Basis Set of the Lowest Energy Structures Generated by ROBIA

structure	*E* (kJ/mol)
7	12.3
10	14.3
11	0.0
12	14.3
13	−9.4
14	−6.9

[18] Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379.

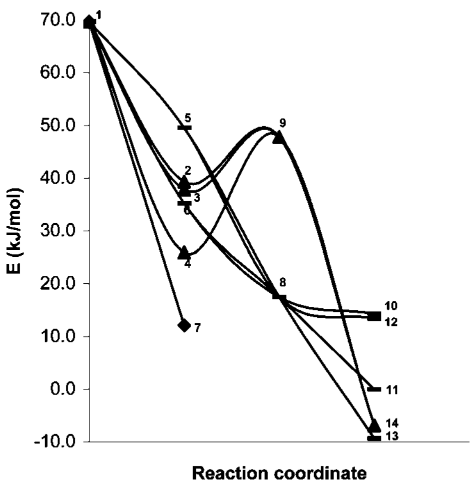


Figure 2. Energy (RHF/3-21G/water) pathways leading to products **7** and **10**−**14** from linear precursor **1**, through intermediates **2**−**6** (step A) and **8** and **9** (step B).

among the products generated. They are dolabriferol **10** and its stereoisomer **11** and three hemiortho esters **12**, **13**, and **14**. RHF/3-21G/water calculation gives a lower energy for the hemiortho ester than for the ester. This is inconsistent both with experimental observations and higher levels of theory (B3LYP/6-31G**), and so **12**, **13**, and **14** are discarded. We note, however, that **12** and **13** would form from **8**, which goes on to form dolabriferol **10**. Structure **14** forms from **9**, which is a much higher energy pathway than the competing route to dolabriferol. Since **8** and **9** share all the same functional groups, the RHF/3-21G/water calculation is likely to give an acceptable estimate of their relative energy. From the remaining structures **7**, **10**, and **11**, the structure selected as final product is the lowest energy one, **11**. It corresponds to dolabriferol's anomer at the hemiacetal center. However, it would rapidly equilibrate in water with its stereoisomer dolabriferol.

Thus, the proposed rearrangement to dolabriferol assumes that the molecule can pick out one pathway amongst a huge number of possibilities. ROBIA is able to generate this matrix of possibilities, evaluate them all, and come to the conclusion that this is a feasible pathway, as it will be preferred to all of the competing routes. In this particular example ROBIA generated 234 structures from which a mixture of dolabriferol **10** and its stereoisomer **11** was selected as final output.

As shown, ROBIA uses a mechanistic approach to reaction prediction, generating possible intermediate structures from the reactants to the products. Thus, the program could be able to anticipate results making possible the prediction and analysis of unknown reactions as well as the investigation of biosynthetic and metabolic pathways. The time required for the example shown was a few seconds for structure predictions and 1−2 h per structure for the molecular modeling calculations on a desktop PC, which can easily be distributed onto a cluster. Further uses of the program could include the checking of reaction databases, as a teaching tool in undergraduate lectures, and as an assistant in synthesis design to test reactions in the forward direction and check for possible side reactions.

Acknowledgment. We thank Elsevier MDL for financial support.

Supporting Information Available: All structures generated by ROBIA for the example shown and a tree diagram of possible reaction pathways starting from the linear precursor. Energies for RHF/3-21G/water and Monte Carlo searches. This material is available free of charge via the Internet at http://pubs.acs.org.

OL0512738