J Comput Aided Mol Des (2007) 21:351–357
DOI 10.1007/s10822-007-9120-4

ORIGINAL ARTICLE

Computational assessment of synthetic procedures

Jonathan M. Goodman · Ingrid M. Socorro

Received: 31 July 2006 / Accepted: 8 May 2007 / Published online: 22 June 2007

Abstract Synthetic chemistry is hard because some reasonable looking molecules cannot be made, because there are errors in the chemical literature, because it is easy to miss reaction possibilities and because even the shape of molecules is very difficult to determine. We propose an approach to the computational analysis of reactions that tries to circumvent these difficulties, by restricting the analysis to simple rules for reactivity that can generate a large number of competing pathways. This huge ensemble is filtered using computational methods to pick out the most likely pathways, and to suggest possible products.

Keywords Computer-aided synthesis · Conformation analysis · Chemical informatics · Data analysis

Challenges for synthetic chemistry

There have been many extraordinary achievements in synthetic chemistry. Highly complex molecules have been created and made on an industrial scale [1]. General asymmetric catalysts have been designed [2]. It may be tempting to conclude that it is always possible to use recipes in the chemical literature to make very complex systems, and that there are people available who will be able to make any molecules. In practice, organic synthesis is very hard, and computational assistance should be useful.

In this article we highlight four of the problems that must be overcome to create an effective computational procedure for synthesis design. First, it must be determined if it is possible to make the molecule at all. Some simple molecules cannot be made. Second, the literature precedent for the synthesis must be carefully analysed. Unfortunately, the chemical literature is not 100% reliable. Third, it must be possible to predict the products of new reactions. This is sometimes extremely difficult. Finally, it must be possible to determine the three-dimensional shape of the molecules, and this is not always straightforward.

Molecules that cannot be made

Are there molecules which look reasonable on paper, which have all their valencies satisfied, which do not include unusual elements, and yet which cannot be made? Some collections of atoms exist in the vacuum of space, or for a split second at very low temperatures. No organic chemist would attempt a synthesis of a molecule such as CH, as the carbon atom with just one covalent bond is clearly very reactive. Such a species could not be studied in solution at room temperature, and we use this criterion to decide what can and cannot be made. However, there may be molecules that appear reasonable, at first sight, which, nevertheless, cannot be made. If there are molecules that can never be synthesized, it is reasonable to ask what is the simplest inaccessible molecule. With this example, it should be possible to be aware of some of the features that will make synthesis impossible.

We recently searched for impossible molecules [3], focusing on branched hydrocarbons. The sequence illustrated in Fig. 1 shows molecules with increasing levels of branching, and increasing steric clashes between the branches. At some point, the steric clashes will be sufficient to break carbon-carbon bonds, and so some of these

J. M. Goodman (✉) · I. M. Socorro
Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
e-mail: J.M.Goodman@ch.cam.ac.uk

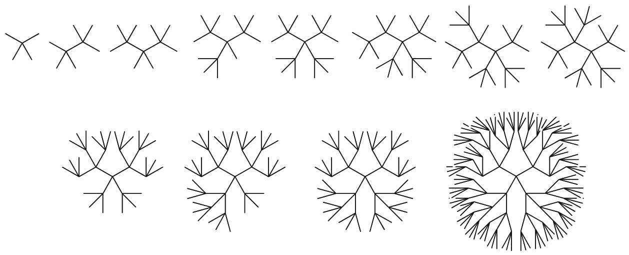


Fig. 1 Alkanes with increasing branching: where does crowding become so extreme that the molecules cannot be made?

saturated, acyclic hydrocarbons will be impossible to make.

By calculating the strain energy of this series using force fields and density functional theory, and comparing these data with experimental data in the literature, we conclude that the simplest molecule that cannot be made will be $C_{16}H_{34}$ as illustrated in Fig. 2. We mean by this that it will not be possible to make the molecule and store it in a bottle in the normal way for organic chemicals. It may be possible to make the molecule and store it at very low temperatures. If this is the case, adding an extra methyl group to make $C_{17}H_{36}$ will further destabilize the system. In the eighteen months since this study was completed, nobody has been able to synthesize either of these molecules.

Chemical space is not quite as extensive as is estimated by counting possible isomers, unless allowance for this type of structure is made. $C_{16}H_{34}$ has a molecular mass of only 226, and so is less than half the maximum size of molecules that obey Lipinski's rules for oral availability [4].

Literature that cannot be trusted

Synthetic chemistry depends on the literature, as procedures are recorded precisely, and repeated. The standard method of recording analytical data about new compounds is well established, and is common between all major journals, with minor variants. How much confidence should we have in this standard procedure? An experimental data checker [5] has recently been reported which provides an opportunity to explore and quantify this question.

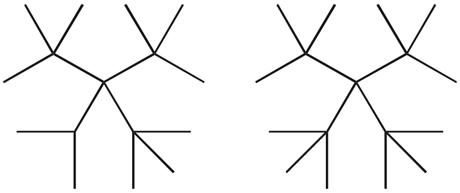


Fig. 2 Two impossible alkanes ($C_{16}H_{34}$ and $C_{17}H_{36}$)

The data checker reads papers and reports which are cut and pasted from a word processor, locates and parses the experimental data, and cross-correlates it to check that it is self-consistent. It also uses a small database that provides information about reasonable values for different types of experimental data. For example, the data checker tries to find the number of carbon atoms in a molecule using the mass spectral data, if it is reported, and then checks that the carbon NMR does not have more signals than there are carbon atoms. If there are too many signals, then a warning message appears. It is possible that this is correct, but it is worthwhile checking this unusual result.

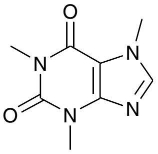
The data checker's initial test showed that organic chemistry papers are about 97% accurate. Some of the 3% of issues that are highlighted for checking may well turn out to have been recorded correctly. However, there are errors that the data checker would not detect, if the data are self-consistent but wrong. This might happen if an author has accidentally swapped spectra within a series of molecules, or made a typographical error with a small but significant effect. It is possible that this figure is an overestimate. 97% correct is a high value, but it means that there is a lot of erroneous data in the chemical literature.

We have extended the program to consider journal citations in papers [6]. A survey of highly cited papers within several issues of leading journals revealed that papers are often cited with variants on the correct names, volumes and authors. For example, a paper by Parr on density functional theory [7] is cited frequently, but not always cited correctly. The number of pages, the authors' initials, the volume of the journal, the name of the journal, and the year of publication were all correct in more than half of the citations. Some of these errors are corrected when the papers are listed in commercial databases, in part because erroneous information, such as the page number of the end of the paper, may be omitted when the information is transferred to a database.

Literature citation is much simpler than recording analytical data about new molecules, and yet there is still a significant error rate. Some experimental quantities, such as solubility, have a large variation in the recorded literature. For example, the solubility of caffeine appears to have varied substantially over the last century, Table 1. There are many examples of misleading physical measurements that have propagated through the scientific literature.

Structural determination is a very demanding problem, and there are examples of structures that have been misdetermined and then corrected. For example, the structure of glabrescol, Fig. 3, was originally reported as being *meso* [8] but was later shown to be C_2 symmetric [9]. Later calculations [10] showed that the *meso* structure was less favourable for ion-binding than the C_2 structure.

Table 1 Solubility of caffeine



Solubility	Data source
2.132 g l^{-1}	*Gazzetta Chimica Italiana* 1926, *56*, 896–901
896.2 g l^{-1}	*J. of Pharm. Sciences* 1985, *74*, 132–135
22 g l^{-1}	http://departments.oxy.edu/tops/Caffeine/CAFFEINE-T.pdf *''We're trying to determine the precise solubility. If any one knows, please let us know''*
Moderate	http://ptcl.chem.ox.ac.uk/MSDS/CA/caffeine.html
Slight	Wikipedia: http://en.wikipedia.org/wiki/Caffeine

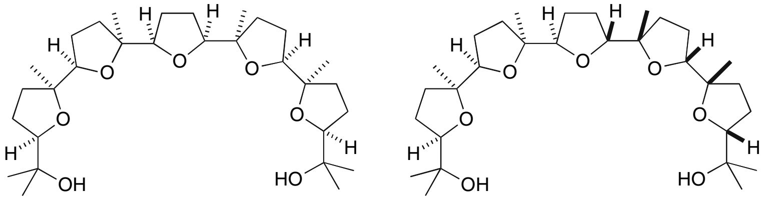


Fig. 3 Glabrescol: the originally proposed *meso* structure (*left*) and the correct C_2 structure (*right*)

The original report of the structure of amphidinoketide I [11] was unable to assign the stereochemistry of the molecule. The stereochemistry was later determined by total synthesis of all the diastereoisomers of the structure [12]. These syntheses demonstrated that the four diastereoisomers were very similar to each other. They could be distinguished by ^{13}C NMR, but only just. Had the molecule been slightly different, perhaps with a 1,5 stereochemical relationship in place of one of the 1,4 relationships, then the diastereoisomers may have been indistinguishable by usual techniques. This could have lead to the situation where low-molecular weight molecules had been isolated and characterised, but their structures could not be confirmed, even by total synthesis (Fig. 4).

Reactivity that is hard to predict

Figure 5 shows a fairly simple unsaturated ester, and straight-forward reaction conditions (catalytic acid). Is this enough information for a skilled chemist to be able to predict what will happen? In 1964 Fleming and Woodward reported the reaction in Fig. 5 [13] and the product was not identified, as it was not the main subject of the study.

A re-examination of the process forty years later [14] was able to prove the structure of the product, by X-ray crystallography, and provide a computational explanation for its formation (Fig. 6).

This process is undoubtedly a difficult one to predict. Other processes may be straightforward to predict, but may be easy to miss in a molecule with many other features. For example, the equilibrium illustrated in Fig. 7 may look surprising at first, but can be readily explained as an acetal rearrangement. It is not so clear, however, which of the three structures should be thermodynamically preferred. An experimental and computational study [15] was able to resolve this question.

The same substructure is the core of the natural product zaragozic acid, which showed interesting biological activity [16]. Would the potential rearrangement always be noticed? The natural product appears to strongly favour just one of the acetal isomers, but this is not the case for the unsubstituted analogue (Fig. 7), for which both the isomers with chair six membered rings are significantly populated. A computer program which could spot and highlight this possible rearrangement would be useful, even if the precision of the calculation of the likely ratio were not very high, as the suggestion of the alternative structures would be enough to ensure that an experimentalist checks for the possibility. The spectra of the isomers are extremely similar, and it would be easy to mistake one for another if all of the spectra were not available (Fig. 8)

Conformations that are difficult to analyse

Conformation analysis is critical, and it is easy to overlook low energy conformations. An example of this is the

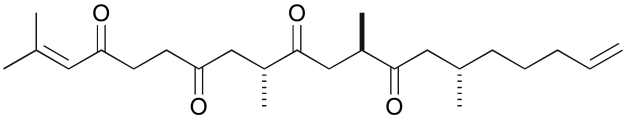


Fig. 4 Amphidinoketide I

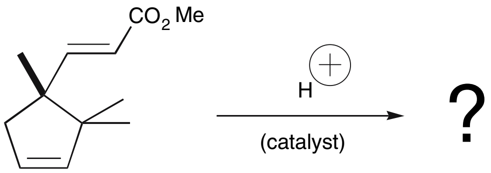


Fig. 5 What is the product?

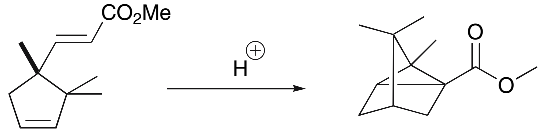


Fig. 6 The rearrangement product identified by X-ray crystallography and investigated using computational methods [14]

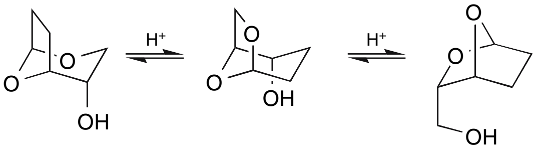


Fig. 7 An acetal rearrangement

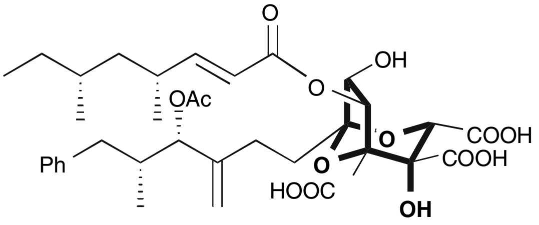


Fig. 8 Zaragozic acid

preferred conformation of unbranched alkanes. It is easy to assume that the extended conformation should be one of the most thermodynamically accessible forms, as there are no unfavorable interactions in such a structure. However, for longer alkanes in vacuo or in polar solvents, the extended structure can be hundreds of kilocalories per mol higher in energy than the global minima [17]. Further, molecular dynamics and systematic searching are unlikely to find the low energy structures in a reasonable time. Fortunately, genetic algorithms [18] seem to work much more effectively. In such cases, it is easy to mistake a conformation that looks reasonable for one which is significantly populated. The random generation of a few reasonable structures is unlikely to give good results.

This is important for flexible natural products, such as amphidinoketide I (Fig. 4) and PM toxin A (Fig. 9) [19]. The lowest energy conformation we have found for PM toxin A is not extended but has two turns, creating three chains which are roughly parallel, and which allow the alcohols to form intramolecular hydrogen bonds as well as making possible attractive van der Waals interactions between the chains.

Computational strategy

How can these problems be circumvented? Our approach is based on three ideas:

(i) Extract simple rules from the literature
(ii) Trust computational chemistry
(iii) Produce suggestions rather than definitive answers

Using idea (i) we restrict our study of the literature to reactions that are unquestionably important, and have been reported many times by independent studies. This rule is very restrictive, and means we omit a lot of interesting and important chemistry. However, even with a small number of rules, we can generate complex processes with conclusions that are not obvious.

Different approaches have been used in the development of programs for reaction prediction [20–27]. These programs, which have chemical knowledge based on a library of known chemical reactions or on a mathematical theory, usually achieve their results by making use of heuristics based on empirical or formal rules, and experimental or estimated physical data. ROBIA's approach to reaction prediction uses a hybrid system that combines rule-based techniques along with molecular mechanics and quantum chemistry. This approach allows the program to generate all possible reaction pathways, on the basis of the selected transformations within its database, and to evaluate and select the most feasible ones.

This is not the approach that synthetic chemists take. Experts usually are familiar with a very large number of examples. Reasoning based on a few simple rules should, therefore, be complementary to expert reasoning, and may sometimes come up with different suggestions.

For example, enolate formation, Fig. 10, is an important process, which can occur if a particular arrangement of atoms is present in a structure. Acetone can form only one enolate, but more complex molecules can form several different enolates. Applying the acetone template to a more complicated example, as illustrated, can generate all the possible enolates.

Trusting computational chemistry, idea (ii), is not natural for experimental chemists, and computational chemistry is certainly not infallible. However, once again,

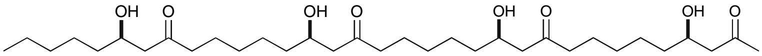


Fig. 9 PM toxin A

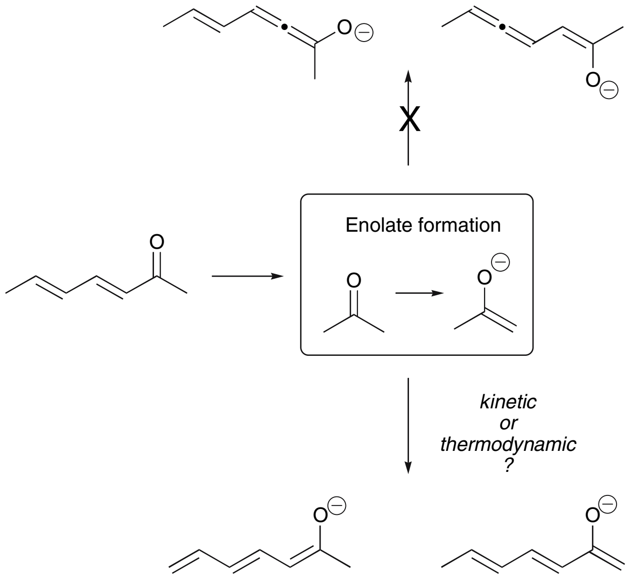


Fig. 10 ROBIA's knowledge of enolate formation

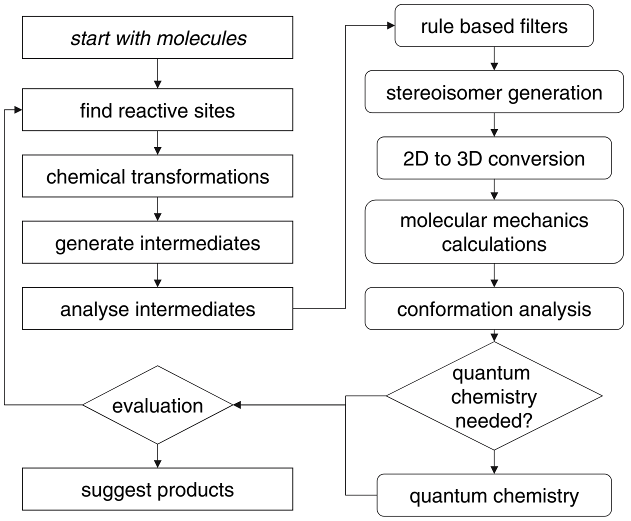


Fig. 11 The ROBIA process

it is different to the human approach to reaction analysis, and so may produce answers that are different and useful. Automation of computational procedures is straightforward, so it is possible to analyse a very large number of competing pathways without direct human intervention.

For the example in Fig. 10, the energies of the competing enolates can readily be calculated by computational methods. The allene-type products are much higher in energy, and could be omitted by a filter that looks for allenes, and so save the expense of undertaking a calculation. The choice of a filter or a calculation would depend on the reaction of interest.

We do not aim to produce a single answer, but rather a list of suggestions, ordered by probability-idea (iii). This is a pragmatic conclusion, based on the approximations within the procedure.

Putting these rules together, we have written a program called ROBIA [28] [29]. This is written in the Java programming language [30] and also uses MDL's Cheshire language [31]. Molecular modelling is handled through the Maestro interface to MacroModel and Jaguar [32]. In principle, there are many molecular modelling programs which should be able to perform this part of the analysis.

An outline of the ROBIA process is given in Fig. 11. Molecules are entered as MOL files [33]. These structures are analysed for their reactive sites, and a list of possible transformations is generated. So far, this covers only a small area of chemistry, but it should be possible to extend this to include more reactions. The transformations are used to generate new molecules.

At this stage of the process, there may be a very large number of new molecules. It would not be helpful for ROBIA to present a list of thousands of molecules as the answer to a synthetic question. The list is reduced first by applying filters. If appropriate, the user can decide to omit certain high-energy structural features, such as allenes or four membered rings, if it is unlikely that these will form under the reaction conditions. If the formation of such structures is possible, the filters can be omitted, and all the structures can be passed to the computational chemistry section of the program.

The two-dimensional structures are converted into three-dimensional structures, generating all the possible diastereoisomers in cases where the reaction forms a new chiral centre, or else the stereochemistry of the starting materials is unknown. These structures are subjected to an exhausting conformation search to ensure that all low energy conformations have been found. This is done using molecular mechanics. The low energy conformations may then be re-optimised using higher levels of theory, if molecular mechanics is unlikely to be able to compare the structures effectively.

The new structures, filtered and ordered by energy, are then passed back to the main ROBIA process. The analysis may stop at this point, or the new structures may be re-submitted to the reaction conditions, to generate even more structures through further reactions.

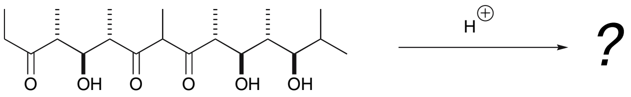


Fig. 12 What is the product of this reaction?

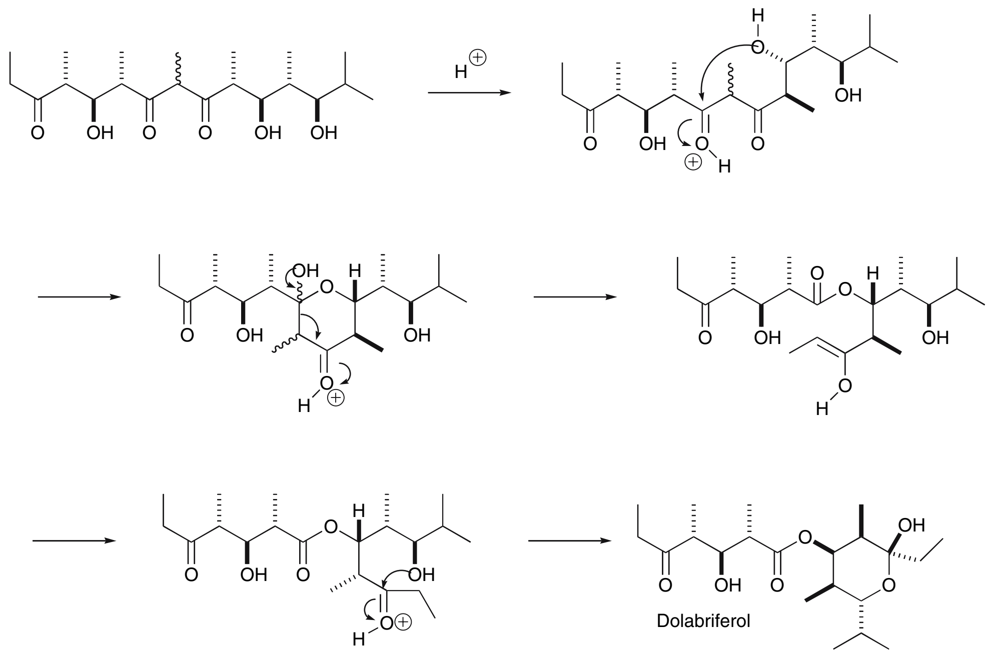


Fig. 13 A plausible sequence to form dolabriferol

Results

The procedure has been applied to the complicated polyketide illustrated in Figure 12. It is thought that this may be a precursor to dolabriferol [34]. The pathway, however, is not obvious.

ROBIA takes the input structure, and allows it to undergo acetal and hemi-acetal formation, enolisations and retro-aldol reactions, generating a very large number of new structures and new stereoisomers. All of these were assessed using computational chemistry, and the lowest energy ones were picked out.

At the first step, there are many different acetals and hemiacetals that can be formed. Some of these form structures that are suitable for retro-aldol reactions. After this, there is another choice of many different acetals and hemi-acetals. It is possible to go through all of the possibilities by hand, but it would be very hard to be sure that none of the hundreds of possibilities had been omitted, and also very hard to decide which of many similar structures was energetically preferred. The computational approach provides confidence that all of the structures have been generated, and all of their energies have been calculated in a reproducible way. A set of 234 different possible intermediates was generated, for each of which an energy was calculated. The most likely pathways were chosen by moving between the lowest energy structures which were available at each stage of the reaction pathway. A low energy pathway is illustrated in Fig. 13. Hundreds of competing structures are omitted from the diagram.

Conclusions

Four of the challenges facing synthetic chemists have been investigated. We have shown that there are some simple molecules that cannot be made, and that some mistakes in the chemical literature can be found by computational procedures. A computational process, ROBIA, has been developed to aid in the analysis of synthetic procedures, by suggesting likely products. This provides an alternative to the processes usually used by synthetic chemists, and includes detailed conformational analyses that require computational methods. Because it does not analyse synthetic procedures in the same way that chemists do, it may produce suggestions which are not obvious from a human analysis, but which seem reasonable once they have been suggested.

Acknowledgments We thank Elsevier MDL and Unilever for financial support.

References

1. Paterson I, Anderson EA (2005) Science 310:451
2. Yoon TP, Jacobsen EN (2003) Science 299:1691
3. de Silva KM, Goodman JM (2005) J Chem Inf Model 45:81
4. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Del Rev 23:3
5. Adams SE, Goodman JM, Kidd RJ, McNaught AD, Murray-Rust P, Norton FR, Townsend JA, Waudby CA (2004) Org Biomol Chem 2:3067
6. Russell B, Goodman JM University of Cambridge, manuscript in preparation
7. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785

8. Harding WW, Lewis PA, Jacobs HM, McLean S, Reynolds W, Tay L, Yang J-P (1995) Tetrahedron Lett 36:8137
9. (a) Xiong ZM, Corey EJ, (2000) J Am Chem Soc 122:4831. (b) Xiong ZM, Corey, EJ (2000) J Am Chem Soc 122:9328. (c) Morimoto Y, Iwai T, Kinoshita T (2000) J Am Chem Soc 122:7124
10. Bellenie BR, Goodman JM (2001) Tetrahedron Lett 42:7477
11. Bauer I, Maranda L, Young KA, Shimizu Y (1995) Tetrahedron Lett 36:991
12. Walsh LM, Goodman JM (2003) Chem Comm 20:2616
13. Fleming I, Woodward RB (1973) J Chem Soc, Perkin Trans I, 1653
14. Davies JE, Fleming I, Goodman JM (2003) Org Biomol Chem 1:3570
15. Dominey AP, Goodman JM (1999) Org Lett 1:473
16. Nadin A, Nicolaou KC (1996) Angew Chem, Int Ed Engl 35:1622
17. Goodman JM (1997) J Chem Inf Comput Sci 37:876
18. Nair N, Goodman JM (1998) J Chem Inf Comput Sci 38:317
19. Hayakawa H, Ohmori M, Takamichi K, Matsuda F, Miyashita M (1997) Chem Commun 1219
20. Jorgensen WL, Laird ER (1990) Pure Appl Chem 62:1921
21. Höllering R, Gasteiger J, Steinhauer L, Schulz K, Herwig A. (2000) J Chem Inf ComputSci 40:482
22. Satoh H, Funatsu K (1996) J Chem Inf Comput Sci 36:173
23. Sello G (1992) J Chem Inf Comput Sci 32:713
24. Agarwal KK, Larsen DL, Gelernter HL (1978) Comput Chem 2:75
25. Ugi I, Bauer J, Blomberger C, Brandt J, Dietz A, Fontain E, Gruber B, vScholley-Pfab A, Senff A, Stein NJ (1994) Chem Inf Comput Sci 34:3
26. Hendrickson JB, Parks CA (1992) J Chem Inf Comput Sci 32:209
27. Zefirov NS, Baskin II, Palyulin VA (1994) J Chem Inf Comput Sci 34:994
28. Socorro IM, Taylor K, Goodman JM (2005) Org Lett 7:3541
29. Socorro IM, Goodman JM (2006) J Chem Inf Model 46:606
30. Java Version 1.4.1. http://java.sun.com/ (accessed Sep 2005)
31. Cheshire Studio, version 3.0.0.54.; Elsevier MDL: San Leandro, CA
32. Maestro, version 5.0.019; Schrodinger Inc.: Portland, Oregon, 2000
33. MDL CTfile Formats http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp (accessed March 2007)
34. Ciavatta ML, Gavagnin M, Puliti R, Cimino G (1996) Tetrahedron 52:12831