Advanced Review

Automatic reaction mapping and reaction center detection

William Lingran Chen,[1]* David Z. Chen[2] and Keith T. Taylor[1]

A reaction center is the part of a chemical reaction that undergoes changes, the heart of the chemical reaction. The reaction atom–atom mapping indicates which reactant atom becomes which product atom during the reaction. Automatic reaction mapping and reaction center detection are of great importance in many applications, such as developing chemical and biochemical reaction databases and studying reaction mechanisms. Traditional reaction mapping algorithms are either based on extended-connectivity or maximum common substructure (MCS) algorithms. With the development of several biochemical reaction databases (such as KEGG database) and increasing interest in studying metabolic pathways in recent years, several novel reaction mapping algorithms have been developed to serve the new needs. Most of the new algorithms are optimization based, designed to find optimal mappings with the minimum number of broken and formed bonds. Some algorithms also incorporate the chemical knowledge into the searching process in the form of bond weights. Some new algorithms showed better accuracy and performance than the MCS-based method. © 2013 John Wiley & Sons, Ltd.

How to cite this article:
WIREs Comput Mol Sci 2013. doi: 10.1002/wcms.1140

INTRODUCTION

A chemical reaction is a process that transforms one set of chemical substances to another. A reaction mechanism describes in detail exactly what takes place at each stage (elementary reaction) of an overall chemical reaction (transformation). Traditionally, isotope-labeling experiments are used to study the mechanism. Some atoms that are expected to be involved in bond changes in a reactant structure are substituted with the corresponding isotopes. The positions of these isotopes in the product structures are then identified using certain techniques such as NMR spectroscopy.[1] This establishes the atom–atom mapping (AAM) relationship between reactant and product structures. The AAM information is then used to determine the changed part of the reaction—the reaction center. More specifically, which bonds in the reactant are broken, which bonds in the product are formed, and which bonds' orders are changed during the reaction. The knowledge of the AAM and reaction center of each elementary reaction constitutes the foundation for establishing the entire mechanism of a reaction.

It is interesting to note that new technologies are continued to be explored to study reaction mechanisms. For example recently, electrospray ionization mass spectrometry has been successfully applied to corroborate the mechanism of several organic reaction proposals.[2]

In silico reaction mapping offers an alternate approach for identifying optimal AAMs and reaction centers automatically, which can then be used to study the reaction mechanisms. This approach derives the AAM and reaction center data directly from the reactant and product structures of a chemical reaction via graph matching or other searching algorithms, and thus is much faster and cheaper than the experimental approaches.

It should be pointed out that the AAM and reaction center of an overall reaction may differ from those of its elementary reactions. This is especially true for some multiple-stage reactions where each stage is an elementary reaction. Furthermore, some

*Correspondence to: WilliamLingran.Chen@Accelrys.com
[1]Accelrys, Inc., San Ramon, CA, USA
[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

DOI: 10.1002/wcms.1140

 © 2013 John Wiley & Sons, Ltd.

reaction equations (such as those in some traditional reaction databases) may represent multiple-step reactions where each step is a complete transformation of its own. The reaction mapping algorithms discussed in this article consider only the overall structural changes of an input reaction.[3] Keeping this point in mind is of importance when interpreting the AAM and reaction center data from the mapping algorithms, especially when using them to study the underlying reaction mechanism.

The information on the AAM and reaction center of chemical reactions has many important applications. It can be used for reaction classification[4] and reaction database development. The AAM and reaction center information also make it more specific and flexible for the reaction substructure search (RSS).[5]

In recent years, the AAM and reaction center information has played an increasingly important role in biochemistry and systems biology (see Box 1).[6] Biochemical reactions are catalyzed by enzymes. In systems biology, one of the major research areas is the metabolic modeling of a cell. One of its focuses is on the deep understanding of the mechanisms of a particular organism at the molecule level. A metabolic network reconstruction breaks down metabolic pathways (see Box 2) into their respective reactions and enzymes and analyzes them within the perspective of the entire network. The AAM can be used to help trace single atoms in the network[7,8] and to deduce the metabolic pathways that are followed by a relevant molecule (such as metabolite or drug).[9] Furthermore, the AAM information can also be used to determine the conservation ratios of atoms in metabolic reactions. The AAMs and reaction centers of biochemical reactions can also be used to reveal the reaction mechanisms, which, in turn, can be used to identify and analyze metabolic pathways[10] and to classify biochemical reactions and enzymes in terms of the mechanisms.[11,12]

Before going further to discuss the subject, it is necessary to first introduce some definitions that will be used in this review. This is because different terms have been used in the literature for the reaction center. For example, the reaction center may be defined as the bonds that are changed (broken, formed, or order changed).[13] The reaction center may also be called reacting center.[14,15] A major disadvantage of this purely bond-based definition is that an isolated reacting atom cannot be included as part of a reaction center. An example of which is an epimerization reaction where a stereoisomer is transformed into its chiral counterpart, but no other transformation occurs, and thus there are no bonds to mark as reaction centers.

Here, we define the reaction center as the atoms and bonds that are directly involved in the bond and electron rearrangement of a reaction.[16] Atoms and bonds in the reaction center are called reacting atoms and reacting bonds, respectively. There are three different types of reacting bonds: the reactant bond that is broken, the product bond that is formed, and the bond that involves bond order change during the reaction. According to the above definition, the isolated reacting atom is part of the reaction center, and the atom whose stereoconfiguration is inverted in the epimerization reaction is the reaction center. It should also be noted that the reaction center that includes both reacting atoms and reacting bonds is also called reaction site in some literature.[13,17] In this article, we use reaction center and reaction site interchangeably.

Another important term is reaction mapping, which establishes the relationship between the reactant and product structures. Reaction mapping is also called AAM. The AAM establishes the one-to-one relationship between reactant and product atoms. That is, the AAM indicates which reactant atom becomes which product atom during the reaction. Strictly speaking, these two terms are not exactly the same. The reaction mapping includes two types of mappings: the AAM and the bond–bond mapping (BBM) between reactant and product structures. As we will see later, some reaction mapping algorithms were designed to first find the BBM and then derive the AAM from the former. For simplicity, the reaction mapping program will be called reaction mapper.

Establishing the AAM between reactant and product structures and detecting reaction centers for a general reaction are two closely related problems. For a simple, balanced organic reaction, the establishment of AAM between reactant and product and the detection of its reaction center are intuitive for human beings and straightforward for a computer program. However, owing to the complexity of the chemical reactions themselves and also owing to the style chemists use to draw reaction schemes in their research papers, many organic reactions in reaction databases are unbalanced. The situation can become more complicated when multiple reactants lead to multiple, complex products. Therefore, automatic assignment of AAMs and detection of reaction centers still remains one of the most challenging tasks in cheminformatics. Many efforts have been made to develop heuristics to handle different types of special cases. An efficient reaction mapping program that can handle over 85% of reactions of a large reaction database with millions of reactions[18] is already considered a good tool. Therefore, although reaction

 © 2013 John Wiley & Sons, Ltd.

mappers have been widely used to establish AAMs and detect reaction centers for reaction database production, manual verification by human experts for certain complex reactions is still required to ensure the quality of the reaction databases.

Driven by the increasingly interest in studying biological systems at the molecule level, in particular the development of biochemical reaction databases, such as the KEGG LIGAND database,[19] several novel AAM algorithms have recently been introduced. In this article, we will review the development of the major methodologies for automatic reaction mapping and reaction center detection.

In this article, the reaction center is highlighted in red unless otherwise explicitly stated. Reacting bonds may also be highlighted using hash marks: (one crossing line indicates the bond order change, and two crossing lines the bond broken or formed. The AAM is indicated using integer numbers in two ways: 1, 2, 3 or .1., .2., .3. (each number carries a prefix of dot and suffix of dot).

BOX 1 SYSTEMS BIOLOGY

Systems biology is an emerging interdisciplinary research field that focuses on complex interactions within biological systems. These systems may be entire species, organisms, or groups of cells, or groups of molecules. It uses a more holistic perspective method to study the behavior of groups of interacting biological components functioning as a system. Some of the major systematic measurement technologies used in systems biology includes genomics, proteomics, bioinformatics, mathematical, and computational models. Many systems biology studies involve metabolic networks or cell signaling networks. As a sidenote, the word 'systems' in the term systems biology is plural.

BOX 2 METABOLIC PATHWAY

Metabolic pathways are series of chemical reactions occurring within a cell. It involves the step-by-step modification of an initial molecule called a substrate to form another product called a metabolite. Metabolites can be intermediates or end products. In each pathway, a principal chemical is modified by a series of chemical reactions catalyzed by enzymes. There are many distinct pathways that coexist within a cell. These pathways together form the so-called metabolic network.

A molecule called a substrate enters a metabolic pathway depending on the needs of the cell and the availability of the substrate. An increase in concentration of anabolic and catabolic intermediates and/or end products may influence the metabolic rate for that particular pathway.

EARLY WORK

In 1938, Weygand[20] first proposed a systematic procedure to classify reactions based on the bonds formed or broken in the course of the reaction. This method was further developed by Theilheimer[21] and used as the foundation for indexing the famous series: *Synthetic Methods of Organic Chemistry*. The idea of automatic detection of the reaction center was first suggested by Vleduts.[22]

FRAGMENT-ASSEMBLY-BASED METHODS

Lynch and coworkers[23,24] developed an automatic method for the detection of the overall structural changes of organic reactions. This is achieved by breaking the reacting molecules down into sets of fragments, eliminating the fragments that remain unchanged, and finally assembling the reaction site from the remaining features. The major shortcoming of this approach is that it is generally impossible to determine the exact location of the reaction sites within their parent structures because of the ambiguities introduced during the fragmentation process.

COMMON SUBSTRUCTURE-BASED METHODS

A molecular structure can be conveniently described as a graph where a vertex represents an atom, and an edge represents a bond. Therefore, graph algorithms can be applied to molecular structures. We will use structure and graph, (sub)structure matching and (sub)graph matching interchangeably.

For two given molecular structures, there may be zero to multiple substructures that are common, the largest of which is called the maximum common substructure (MCS).[25] Take structures A and B in Figure 1 as an example. There are many possible substructures that are common to these two structures, three of which are shown in Figure 1 and also highlighted in bold in structures A and B. The substructure 3 is the largest substructure that is common to structures A and B and thus it represents the MCS of these two structures. It should be mentioned that the maximum common substructure is also called maximal common substructure in some literature.[25]

 © 2013 John Wiley & Sons, Ltd.

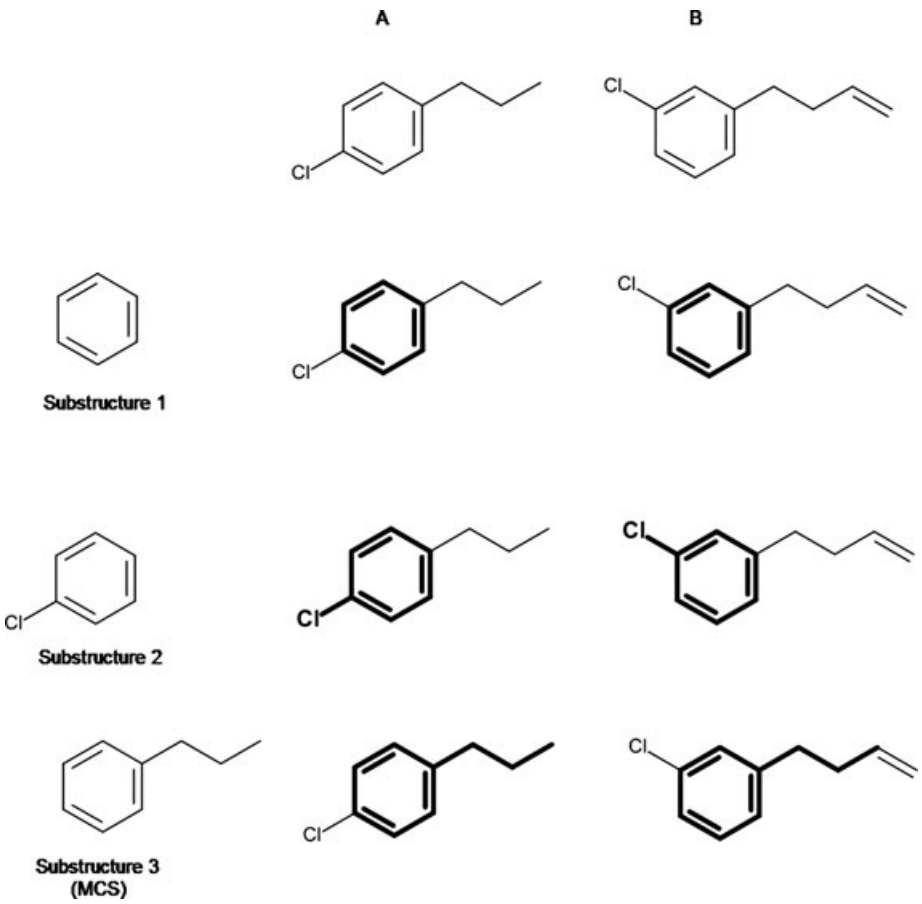

FIGURE 1 | The concept of MCS. The substructures contained in the original structures are highlighted in bold.

The MCS problem is to determine all possible MCSs of two given structures. If the detected MCS is isomorphic to both of the given structures, such an MCS problem belongs to the structure isomorphism problem. On the other hand, if the MCS is isomorphic to the smaller of the two structures, this kind of the MCS problem is the substructure isomorphism problem. Therefore, the structure and substructure isomorphism problems are only two special cases of the more general MCS problem.

It has been proven that both the substructure isomorphism problem and the MCS problem belong to a class of difficult problems called NP-complete[26] problems. NP-complete problems have no known efficient algorithms to find the exact solutions. Although solutions to NP-complete problems may be verified in polynomial time, there are no known algorithms that can find the exact solutions in polynomial time, and the only known methods to find exact solutions require exploring all possible solutions. As a result, to improve performance, MCS algorithms employ sophisticated heuristics to narrow the search space.

For both the structure and substructure isomorphism problems, there are some known conditions that can be used to guide the search process. For example, for both the structure and substructure isomorphism problems, the number of neighbors attached to each atom as well as the number of atoms and bonds of the smaller (query) structure can be used to guide the search. Furthermore, there exist efficient heuristic solutions for both the structure and substructure isomorphism problems, such as using molecular hash-codes or sets of graph invariants for the case of the structure isomorphism problem, or the use of screens and fingerprints for the case of the substructure isomorphism problems. For the structure isomorphism problem, an efficient nonheuristic solution is to use canonical structure representations. For the general MCS problem, none of these conditions and techniques can be used, presenting the main difficulty of the general MCS problem. More detailed discussion on the MCS problem and its algorithms can be found in recent review articles.[25,27]

The MCS has many applications, such as structure–activity relationship. Many traditional

 © 2013 John Wiley & Sons, Ltd.

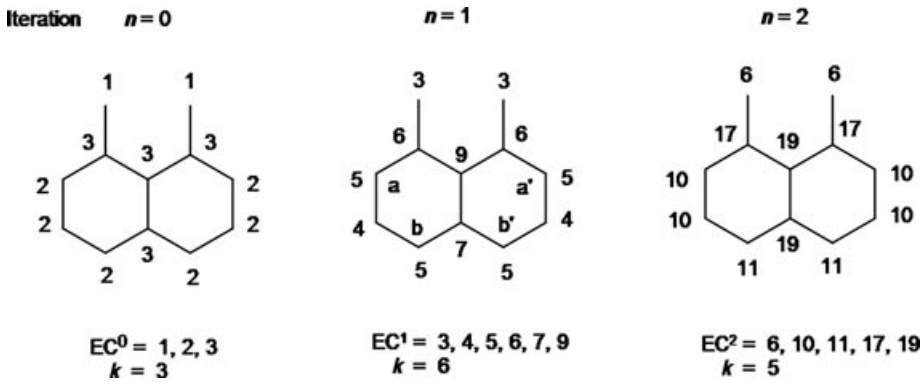

FIGURE 2 | Example of the calculation of EC values in the Morgan algorithm.

reaction mappers are based on the MCS algorithm. In this section, we will first discuss three types of reaction mapping algorithms that are based on finding the large common substructures between reactants and products.

Extended-Connectivity-Based Methods

In this section, we will discuss a special class of reaction mapping algorithms that are based on the Morgan algorithm[28] for calculating extended connectivity (EC). Those EC values are then used to find large, but not necessarily the maximum, common substructures between reactants and products.

Lynch–Willett's EC-Based Method

In 1977, Lynch and Willett[17] reported an efficient method for the automatic detection of reaction centers based on the EC of the Morgan algorithm.[28] The procedure consists of identifying one or more large substructures common to both sides of the reaction.

The Morgan algorithm can be employed to detect equivalent atoms within a single molecular structure based on the concept of EC. The EC values are calculated using the following procedure:

1. Assign to each atom i an initial EC value (EC_i^0) equal to the number of nonhydrogen atoms attached to that atom.
2. Calculate the number (k) of different EC values that have been assigned.
3. Establish an iterative process to calculate a new EC value (EC_i^n) for each atom i:
   The nth-order (via the nth iteration) EC value (EC_i^n) of atom i is calculated by summing the $(n-1)$th EC values EC_i^{n-1} of all adjacent atoms of atom i: $EC_i^n = \sum EC_i^{n-1}$.
4. Calculate the number (k') of different values in the set of new EC values.
5. If $k' > k$,
   a. Assign the new EC values to the corresponding atoms.
   b. Set k equal to k'.
   c. Go to step 4 to repeat the summation process.
6. Else if $k' \leq k$, terminate the process.
7. Induce a partial ordering among the atoms using the last set of EC values assigned to the atoms.

The example in Figure 2 illustrates the application of this technique for introducing a partial ordering among the atoms of a molecular structure. In this example, the iterative process will be terminated after two iterations ($n = 2$), and the EC values assigned after the first iteration ($EC^1{}_i$) will be used to introduce the partial ordering.[28]

From Figure 2, it can be seen that the final EC values (EC_i^1) reveal the intramolecular equivalences. For example, the two methyl groups have the same EC value of 3, the atoms a and a' have the same EC value of 5, and so on. The nth-order EC value of the center atom i, $EC^n{}_i$, represents a circular substructure of radius n bonds. For briefness, this kind of substructure will be referred as the EC-based substructure. As such, the EC value may be regarded as a 'hash' of the corresponding substructure.

It should also be noted from Figure 2 that EC values may fail to distinguish some atoms that have different environments. In the above example, EC values cannot distinguish atoms a and a' from b and b', respectively. These four atoms have the same EC value of 5. To address this problem, several variations of the EC procedures were proposed.[29,30] Additional properties (such as atom type and the surrounding

 © 2013 John Wiley & Sons, Ltd.

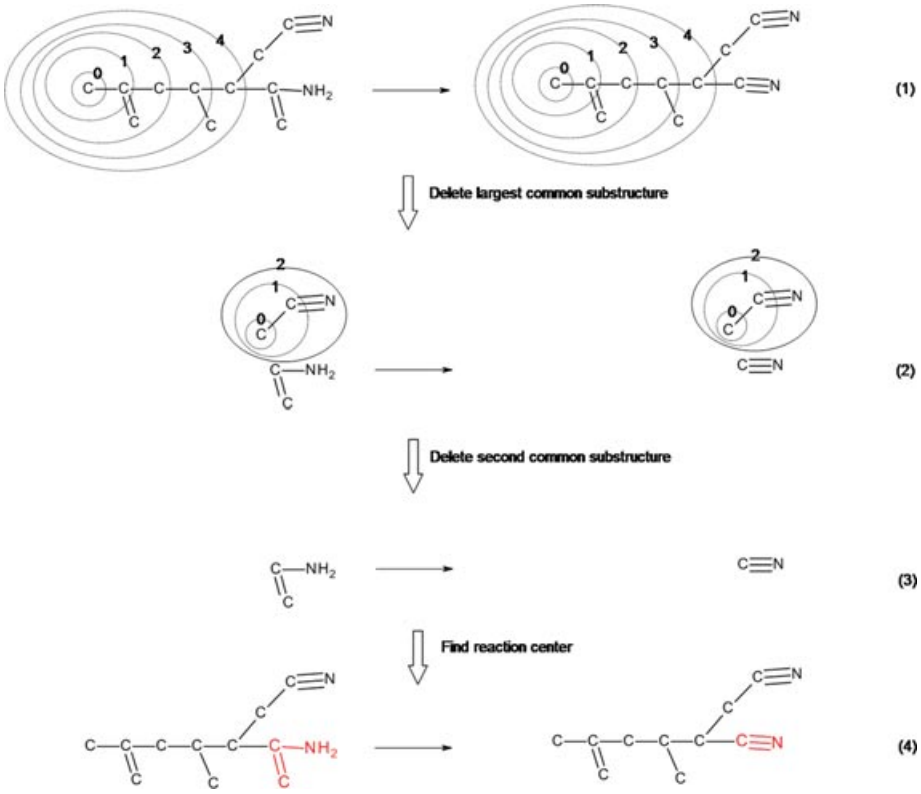

FIGURE 3 | The EC-MCS-based procedure to find reaction center (marked in red). (Adapted from Ref 17. Copyright 1978, American Chemical Society.)

bond pattern) may also be included in the procedure to increase its discriminatory power.

The standard Morgan algorithm described above is modified to detect the intermolecular equivalences. The initial EC value of each atom i, EC_i^0, is an integer derived from the atom type and the bond pattern of the atom. The higher order EC values are obtained using the following equation: $EC_i^n = 2EC_i^{n-1} + \sum_i EC^{n-1}$, where the summation is over all adjacent atoms of atom i.

Because the nth-iteration EC value of the center atom i, may be considered to represent a circular substructure of radius n bonds, if $EC_{r_i}^n = EC_{p_j}^n$, the two corresponding substructures that are centered at the reactant and product atoms r_i and p_j may be considered to be identical.

Lynch and Willett's procedure for matching the substructures of two reacting molecules and identify the reaction center is as follows:

1. Calculate the higher iteration EC values for all reactant and product atoms until there are no remaining pairs of atoms for which $EC_{r_i}^n = EC_{p_j}^n$.
2. Mark the pair(s) of atoms for which $EC_{r_i}^{n-1} = EC_{p_j}^{n-1}$, that is, those reactant–product atom pairs that are at the center of identical circular substructures with a radius of $(n-1)$ bonds. This kind of MCS is referred to as the EC-based maximum common substructures (EC-MCS) of reactant and product.
3. Delete all atoms contained in the EC-MCS from the reactant and product.
4. Repeat the above process until all substructures that are common to both reactant and product are eliminated. The remaining atoms and bonds constitute the reaction center.

Take reaction 1 (Figure 3) as an example, a large substructure common to reactant and product is obtained after four iterations; the substructures have a radius of four bonds. Deleting these substructures from both the reactant and the product leads to the reaction diagram 2 (Figure 3). Repeating the above process leads to the detection and elimination of the C—C≡N substructure with a radius of two bonds from Eq. (2). Finally, Eq. (3) is obtained

 © 2013 John Wiley & Sons, Ltd.

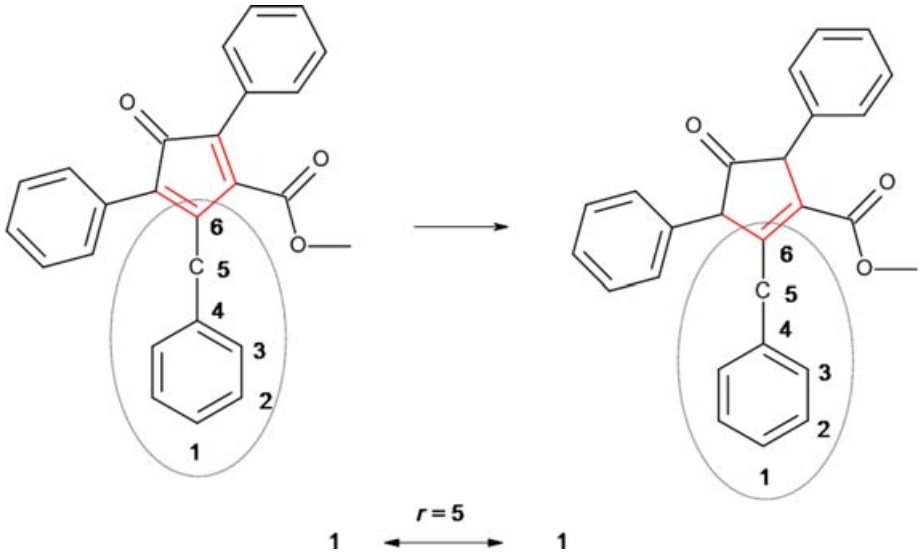

FIGURE 4 | The bonds attached to the outermost atoms (r6, p6) are differently oriented in the two structures. These two atoms are reacting atoms and thus should not be deleted. The matched substructures are marked using ellipses. The reaction center is highlighted in red. (Adapted from Ref 17. Copyright 1978, American Chemical Society.)

(Figure 3). The reaction centers detected for reaction 1 are marked in red in Eq. (4) (Figure 3).

The actual implementation of the above algorithm has three additional features:

1. Allow for multiple equivalences (such as three equivalent fluorine atoms in $-CF_3$).
2. Define a minimal match radius of two bonds, because if a radius of one bond is used, incorrect mappings will increase significantly.
3. For a match radius r, delete only the atoms within $(r - 1)$ bonds.

This last condition is used to avoid incorrectly deleting atoms that are actually part of the reaction center (see Figure 4 for an example). A consequence of this restriction is that the number of atoms that will be deleted is slightly reduced.

It should be pointed out that even with the above additional restrictions, the substructures detected may be smaller than the MCS or, in a small number of cases, nonisomorphic substructures identified as equivalent due to the limitation of the EC procedure, as noted previously. Therefore, it is not always possible to specifically identify the bonds changed in the reaction. For example, reactions that lead to ambiguous mappings cannot be processed (see Figure 5 for example). Also, reactions where the reactant and product are too small, and thus no pairs of atoms have a matching radius that is greater than 1, cannot be handled (Figure 6).

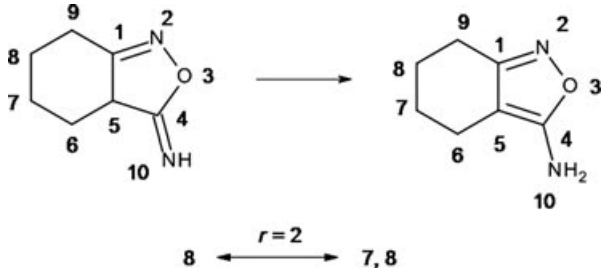

FIGURE 5 | Reactions that lead to ambiguous mappings cannot be processed. (Adapted from Ref 17. Copyright 1978, American Chemical Society.)

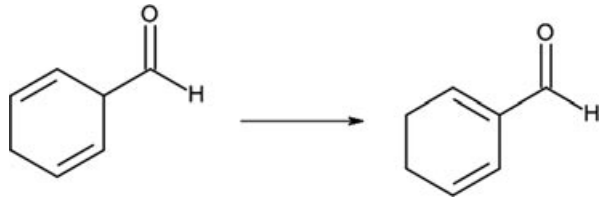

FIGURE 6 | This reaction cannot be handled because the reactant and product are too small, and thus no pairs of atoms have a matching radius that is greater than 1. (Adapted from Ref 17. Copyright 1978, American Chemical Society.)

The functional-group shift is a main cause of the failures of the Lynch–Willett algorithm because it is difficult for the matching algorithm to detect this kind of change. An example of this kind of reaction is shown in Figure 7. It should be noted that this reaction is also an example that contradicts the

 © 2013 John Wiley & Sons, Ltd.

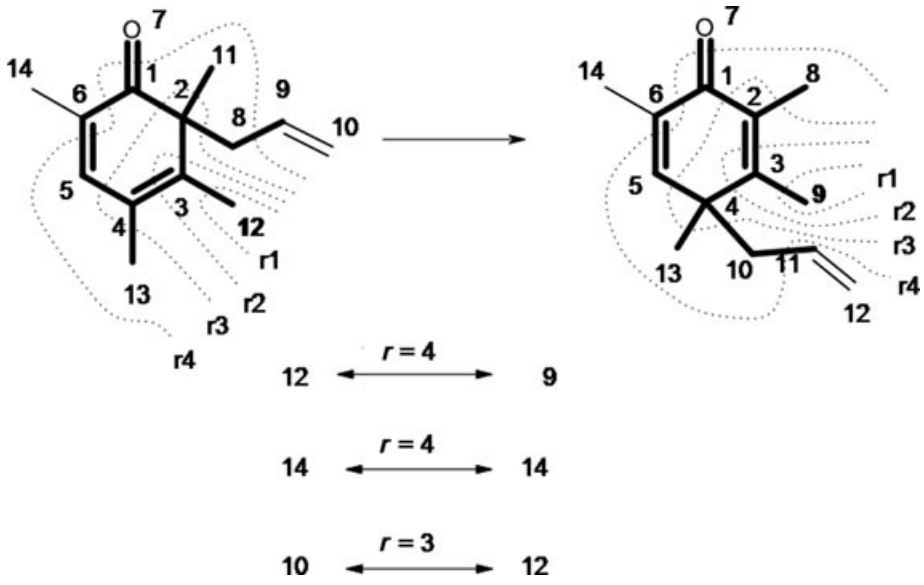

FIGURE 7 | This reaction is an example that contradicts the assumption that equal EC values correspond to identical substructures. The substructures obtained after 4th iteration are highlighted in bold. (Adapted from Ref 17. Copyright 1978, American Chemical Society.)

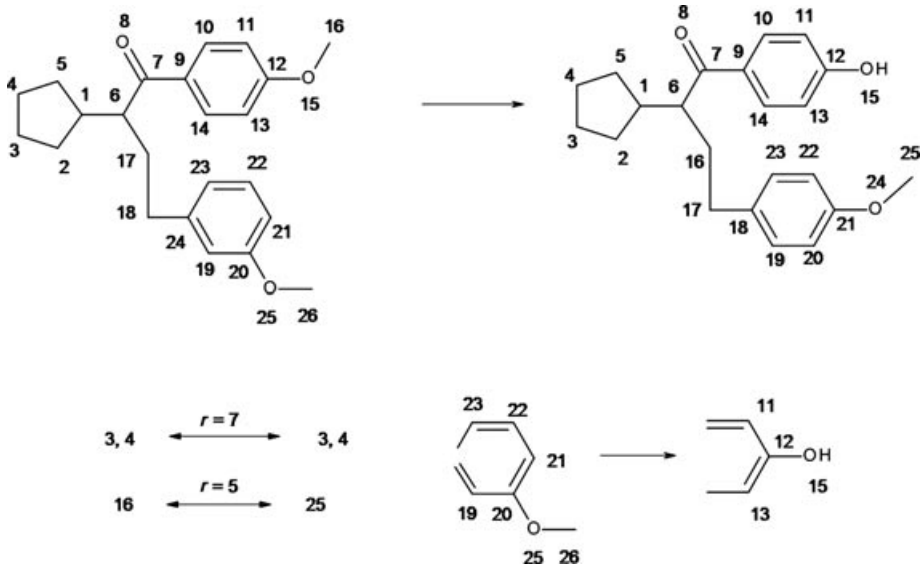

FIGURE 8 | In this reaction, the invalid equivalence between reactant atom 16 and product atom 25 is obtained. (Adapted from Ref 17. Copyright 1978, American Chemical Society.)

assumption that equal EC values correspond to identical substructures. The reactant atom 12 and the product atom 9 have the same fourth-order EC value, but the two corresponding substructures that center at those two atoms are certainly not identical.

If an atom involved in the change is matched with a nonreacting atom, incorrect mappings will be obtained. For example, in the reaction shown in Figure 8, the invalid equivalence between reactant atom 16 and product atom 25 was obtained.

However, this limitation is offset by the superior performance of this method for processing larger number of reactions. An implementation of this algorithm produced analyses for 92.6% of a sample file of 340 one-reactant and one-product reactions.[17]

It is interesting to note that Lynch and Willett's above method was initially undertaken to provide an alternative means for obtaining the guiding information for reaction center detection method based on the MCS algorithm, as suggested by Vleduts (see the *Vleduts' MCS-Based Algorithm* section below).[31] The

 © 2013 John Wiley & Sons, Ltd.

Vleduts algorithm and the Lynch and Willett procedure described here together laid an important foundation for the further development of the common substructure-based methods for automatic detection of reaction centers.

Accelrys EC-Based Method

In the late 1980s, four major reaction databases were developed for REACCS[32] with a total of about 90,000 reactions. During that time, many companies had also built their own private reaction databases for use with REACCS. The reactions in these databases usually contained reaction center information but lacked AAM relationships between reactant and product atoms. It would have been prohibitively expensive to manually assign AAMs for all the reactions in the existing databases. Furthermore, there are many nonstoichiometric and other complicated reactions in REACCS databases with the following characteristics:

- Missing pieces from either side.
- The presence of alternative products.
- Multiply used reagents.
- Deceptively simple transformations due to symmetry or similarity.

All of the above factors make it a challenging task to automatically assign the reaction center and AAM. In 1988, a program called Automatic Reacting Center Perception, later renamed as the Automapper, was developed at MDL (now Accelrys) for automatic reaction center perception and AAM assignment.[33] The Automapper meets the following requirements:

- Has high reliability over large databases of realistic, complicated reactions.
- Handles unbalanced and other ill-behaved reactions.
- Allows human intervention when necessary.

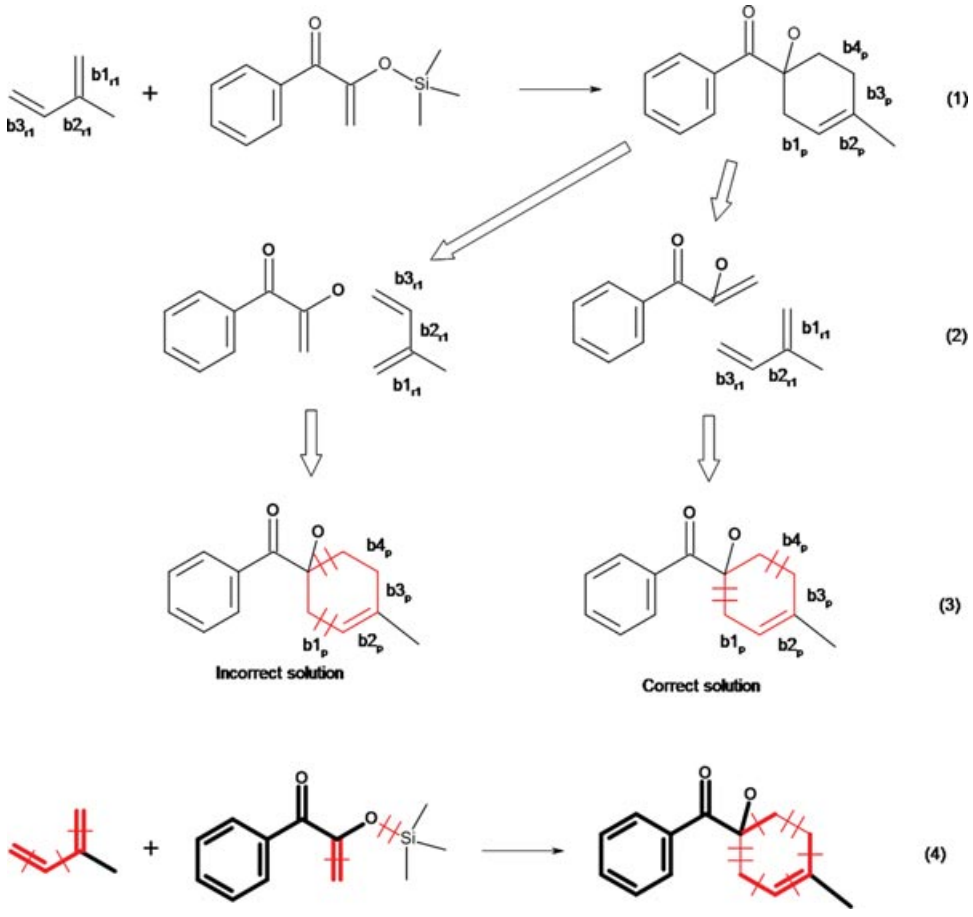

FIGURE 9 | (1) The Diels–Alder reaction. (2) Two sets of MCSs. (3) Two possible solutions based on the MCSs: the incorrect solution (left) and correct solution (right). (4) The output solution from the Automapper program. In both (3) and (4), the reacting bonds are highlighted in red and also with hash marks. The MCSs in (4) are highlighted in bold. (Adapted from Ref 35. Copyright 1988, Springer.)

© 2013 John Wiley & Sons, Ltd.

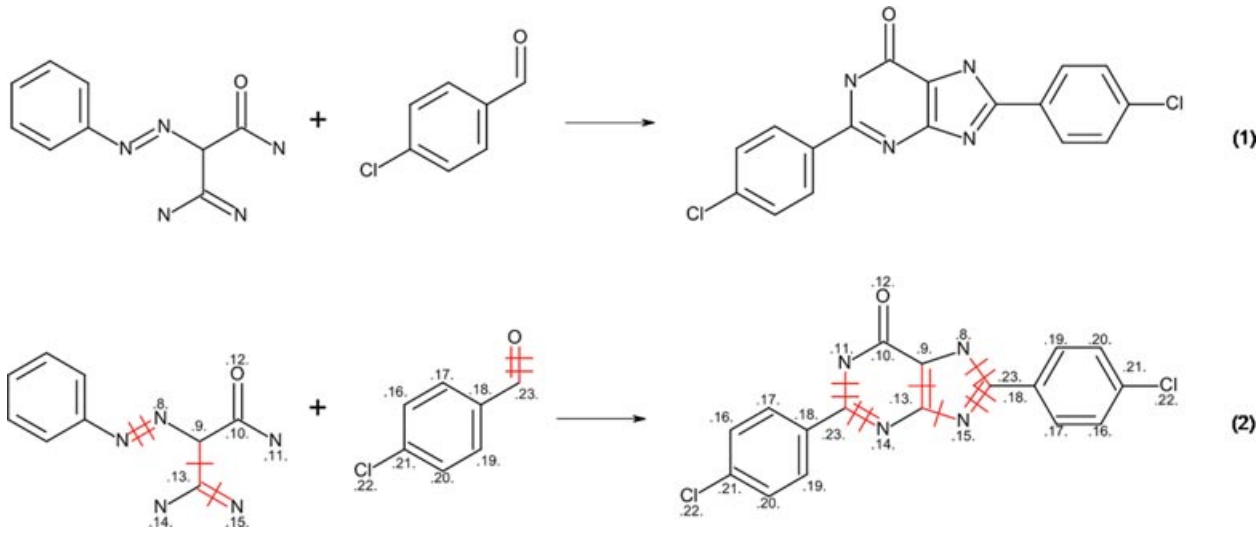

FIGURE 10 | Analysis of an unbalanced reaction. (1) Original reaction. (2) The reaction with the AAM marked and the reaction bonds highlighted in red and also with hash marks. (Adapted from Ref 35. Copyright 1988, Springer.)

Although the basic procedure of the Automapper program is similar to the classical approach,[17,31] it employs several techniques to handle complicated reactions. It works by iteratively detecting many common substructures, not simply the largest ones, and evaluating each possible solution using some heuristics to select the best one. Take the Diels–Alder reaction[34] in Figure 9 as an example. From this figure, it can be seen that there are two sets of MCSs between reactant and the product structures. In the first set of the MCS (Figure 9 (2), left), the double bond $b1_{r1}$ of the first reactant matches the double bond $b2_{p}$ of the product, whereas in the second set of the MCS (Figure 9 (2), right), both double bonds of the first reactant match the single bonds of the product. Therefore, from the matching point of view, the first set of MCSs is better than the second set. However, using the first MCS set leads to the incorrect result for this reaction [Figure 9 (3), left]. The Automapper chooses the correct solution as the best one using an extensively tuned evaluation function [Figure 9 (3), right]. The output solution from the Automapper program is shown in Figure 9 (4).

To solve nonstoichiometric reactions, Automapper takes into account the existence of alternate substructures (alternate products) and many-to-one mappings of identical fragments. Consider the reaction[35] shown in Figure 10 (1) as an example. At the first glance, this reaction seems well balanced. However, it actually is not. Both phenyl rings in the product contain a chlorine atom at the *para* position, and thus must come from the second reactant.

FIGURE 11 | Automapper found the correct AAM and reaction center for this simple, unbalanced reaction. Note: The reacting bonds are marked in red and also with hash marks.

The phenyl ring on the first reactant, on the other hand, is lost completely. The Automapper recognizes this complexity and assigns AAM and reaction centers accordingly [Figure 10 (2)]. The absence of AAM numbers on the phenyl fragment of the first reactant and on the oxygen atom of the second reactant indicates that they are lost during the course of the transformation.

Figure 11 shows a simple, unbalanced reaction for which Accelrys' Automapper found the correct AAM and reaction center. The MCS-based method failed to handle this reaction (see Figure 17).

Stereochemistry plays an important role in many reactions. An important feature of the Automapper is that it can handle stereochemistry. In addition to reaction centers and mapping numbers, atoms in mapped reactions have two properties that specify changes in stereoconfiguration in a reaction:

- .ret. Stereogenic center retains its stereoconfiguration during the reaction.
- .inv. Stereogenic center inverts stereoconfiguration during the reaction.

 © 2013 John Wiley & Sons, Ltd.

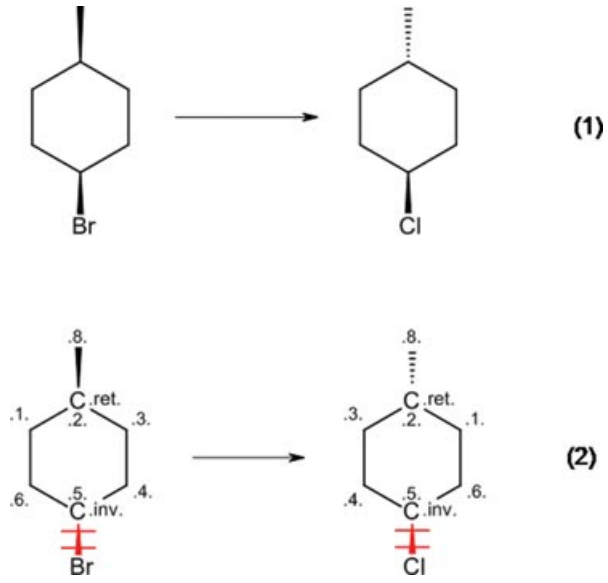

FIGURE 12 | An example showing that whether the stereoconfiguration of a stereogenic center is changed cannot be determined by simply mapping an UP bond to an UP bond or a DOWN bond to a DOWN bond. (1) Original reaction. (2) Automapper results: atom .2.'s stereoconfiguration is retained, whereas atom .5.'s reversed. Reacting bonds are highlighted in red.

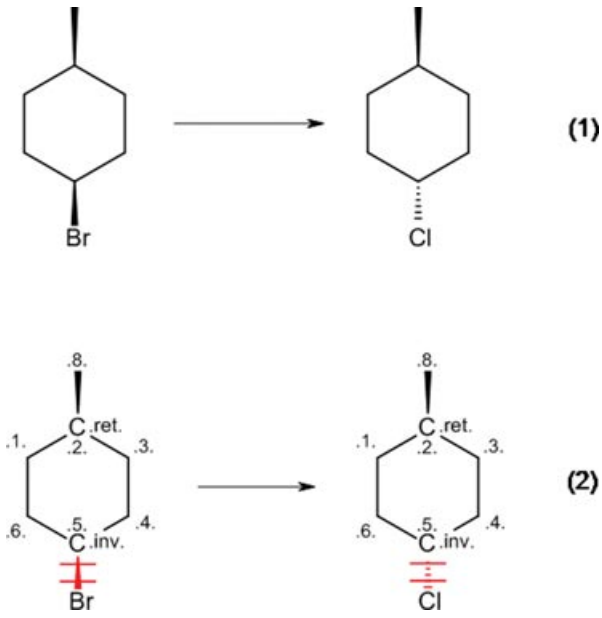

FIGURE 13 | The reaction (1) in this figure is identical to Figure 12 (1) except that the product structure was flipped over. The Automapper generated the same result [shown in (2)] as that in Figure 12 (2). The reacting bonds are highlighted in red.

It should be noted that whether the stereoconfiguration of a stereogenic center is changed cannot be determined by simply mapping an up bond to an up bond or a down bond to a down bond. Consider the reaction shown in Figure 12 (1). In this reaction, the two C—C stereo bonds in this reaction were drawn differently—one in the reactant is an up bond, whereas the other one in the product is a down bond. On the other hand, both the C—Br bond in the reactant and the C—Cl bond in the product were drawn as up bonds. At the first glance, the stereoconfiguration of atom C.2. should be inverted, whereas that of C.5. should be retained. However, the Automapper detects and uses the actual parity of each stereogenic center in the reacting center. It produces results that are actually opposite to the initial visual interpretation, as shown in Figure 12 (2). Flipping over the product structure of Figure 12 (1) leads to the same reaction, as shown in Figure 13 (1). The Automapper mapping result for this reaction is shown in Figure 12 (2). Comparing Figure 12 (2) with Figure 13 (2), it can be seen that although the stereo bonds in the two product structures were drawn differently, the Automapper generated the same correct results: the stereoconfiguration of atom C.2. is retained, whereas that of atom C.5. is inverted.

When the Automapper procedure was published in 1988, it was the only automatic reaction center detection method that could handle nonstoichiometric reactions. The program has since been used in Accelrys' many core cheminformatics products (such as REACCS, ISIS/Host, ISIS/Base, Accelrys Direct, Accelrys Draw, and Accelrys Cheshire). It has also been used as a tool by reaction database builders such as FIZ CHEMIE for developing ChemInform Reaction Library[36] and for researchers to detect reaction centers and assign AAMs to a huge number of reactions and RSS queries. After many years of enhancement, Accelrys' EC-based Automapper is probably one of the most mature, most function-rich, and fastest reaction mappers available on the market.

MCS-Based Methods

In this section, we will introduce the reaction-mapping algorithms that are based on the MCS algorithm.

Vleduts' MCS-Based Algorithm

A procedure to apply the MCS algorithm to the automatic detection of reaction centers was first proposed by Vleduts[31] in 1977. His algorithm involves the identification of the maximum substructures common to the both sides of the reaction. The reaction center can then be detected by identifying the bonds that are not included in this MCS but have one or both of their atoms included in it.

 © 2013 John Wiley & Sons, Ltd.

However, in contrast to the problem of structure isomorphism, MCS isomorphism was little studied due to the greater complexity of the problem and the limitations of computational power at that time. Therefore, Vleduts predicted that such an algorithm, implemented upon the computers of that time, would probably be limited to handling only molecular structures that do not exceed 10–15 atoms. To tackle this problem, he described a procedure for reducing the number of iterative mappings that must be performed by using certain guiding information of reactant–product atom equivalences. The latter can be conveniently obtained by comparison of the Wiswesser Linear Notation[37,38] symbol strings of the reacting molecules. This suggestion has led to the development of other faster but approximate approaches for the detection of reaction centers (see '*EC-Based Methods*').

McGregor–Willett's MCS-Based Method

As discussed in the previous section, the Lynch–Willett's EC-based method is fast for reaction center detection, but there are many cases where this method fails. To address some of the problems of the Lynch–Willett method, McGregor and Willett[39] developed an efficient MCS-based approach for the detection of the reaction center. Their method is a two-step procedure. First, the Lynch–Willett method is used to identify the preliminary reaction site for a given reaction. Second, the reaction site obtained in the first step is used as the starting point to identify the MCS via an MCS search algorithm.

The main advantage of this two-step procedure is that because the preliminary reaction sites are much smaller than the entire reacting molecules, the efficiency of the MCS search and reaction center detection is considerably increased.

The MCS algorithm used in this work was developed by McGregor[40] in 1982. Backtracking[41] is a refinement of the brute force method. The backtracking algorithm is used to find all (or some) solutions to some computational problems. The algorithm incrementally builds candidates to the solutions, and abandons each partial candidate *b* ('backtracks') as soon as it determines that *b* cannot possibly be completed to become a valid solution. To guide the backtracking search in such a way that good solutions will be found earlier, the search tree is ordered dynamically.

In the McGregor's algorithm, a maximum common subgraph (MCS) of two given graphs is defined to be the common subgraph that contains the largest possible number of arcs (edges). In addition, to detect reaction centers, a weaker definition of a substructure is used. First, an MCS may consist of multiple disconnected fragments, so that it is possible for two atoms of a structure to be included in a substructure even if the bond that connects them is not included.

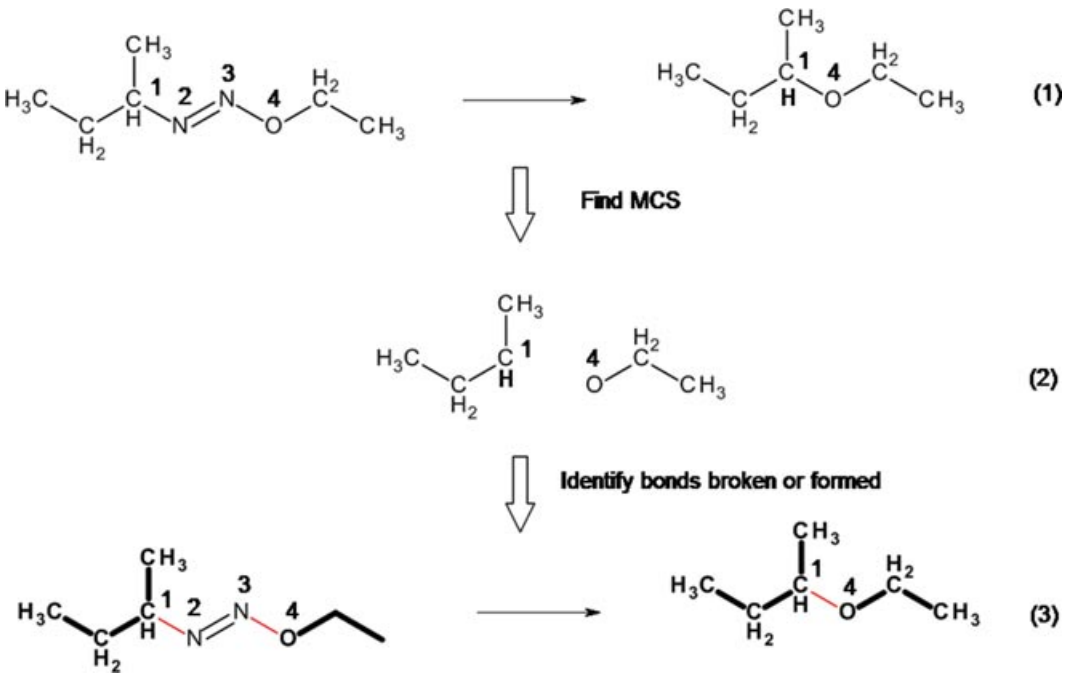

FIGURE 14 | (1) Reaction. (2) MCS. (3) The reaction with reacting bonds highlighted in red. The MCSs in (3) are highlighted in bold. (Adapted from Ref 39. Copyright 1981, American Chemical Society.)

 © 2013 John Wiley & Sons, Ltd.

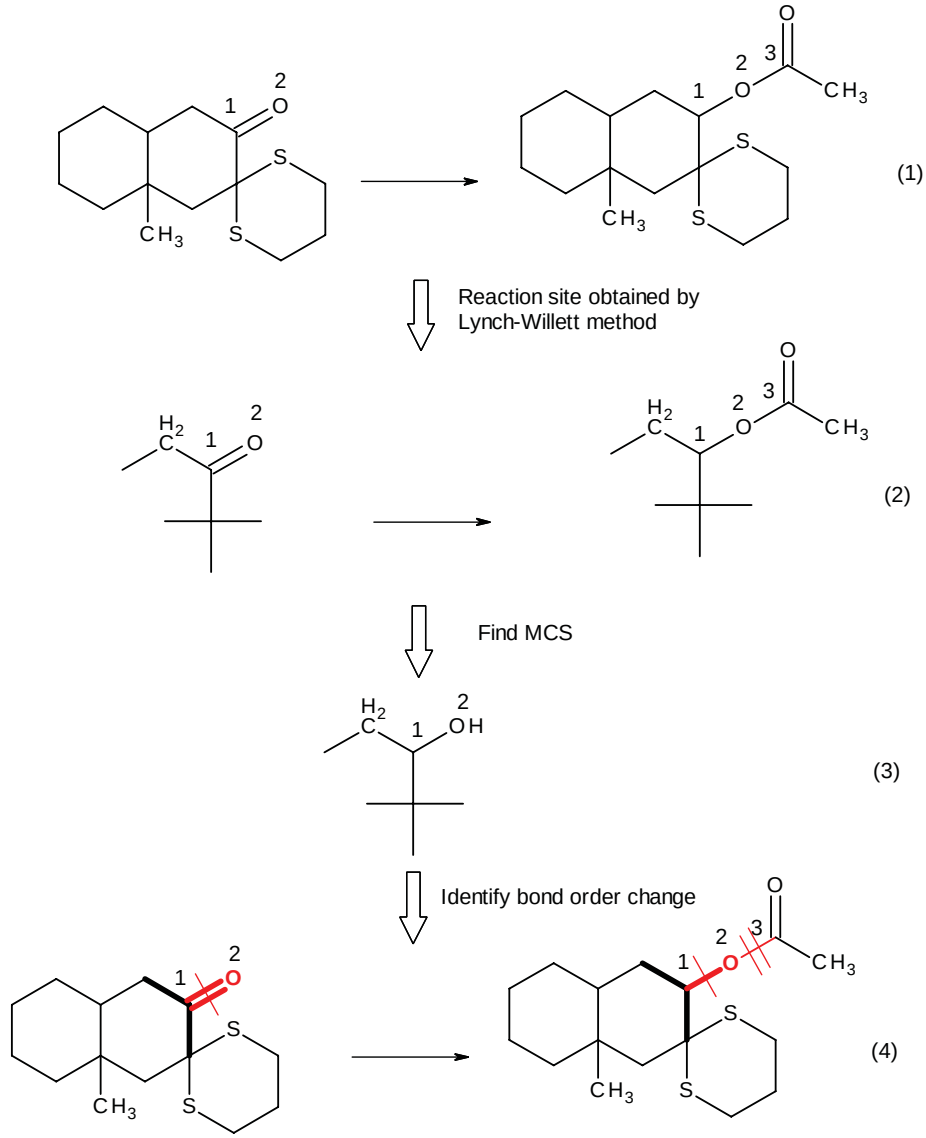

FIGURE 15 | (1) Reaction. (2) The reaction sites obtained using the Lynch–Willett method. (3) MCS. (4) The reaction with reaction centers highlighted in red and also with the hash marks, and the MCSs are highlighted in bold. Note: In reaction site (2) and the MCS (3), two unnumbered carbon atoms have 'free' valences of one and three, respectively. (Adapted from Ref 39. Copyright 1981, American Chemical Society).

Second, two bonds with different bond orders are allowed to match to each other. The first condition is introduced to facilitate the detection of bonds that are broken or formed during the course of the reaction. For example, the MCS of the reactant and product structures of reaction 1 in Figure 14 is shown in Figure 14 (2). Although atoms 1 and 4 are included in the MCS, the bond that connects these two atoms in the product is excluded. With such an MCS available, the bonds that are formed in the product can be identified by comparing the MCS with the product structure. Similarly, bonds that are broken in the reactant can be identified by comparing the MCS with the reactant structure. The bonds in the reaction center of the reaction Figure 14 (1) are highlighted in red in Figure 14 (3).

The second condition is used to help identifying bond order changes. Recall that in the Lynch–Willett method, the substructure that is common to a reactant–product pair must not contain any reacting atom and reacting bond, and thus, detecting bond order changes using that method is difficult. With the above second condition, an MCS may contain reaction centers that involve bond order changes. In such a case, the MCS is used to establish the direct AAM relationship between reactant reacting atoms and the product reacting atoms. This makes it straightforward to identify the bond order changes.

© 2013 John Wiley & Sons, Ltd.

We use the reaction 1 shown in Figure 15 as an example to illustrate the entire McGregor–Willett procedure. First, applying the Lynch–Willett method to this reaction leads to the reaction site shown in Figure 15 (2). Second, applying the McGregor MCS search algorithm to this reaction site leads to the MCS [see Figure 15 (3)]. Third, comparing the MCS with the reactant's and product's reaction sites in Figure 15 (2) indicates that atoms 1 and 2 in the product must come from the atoms 1 and 2 of the reactant, respectively, and thus, the double bond between the atoms 1 and 2 of the reactant must have been changed into the single bond between the atoms 1 and 2 in the product. Because atom 3 of the product is not included in the MCS, the bond between the atoms 2 and 3 of the product must be newly formed during the course of the reaction; finally, the full reaction with the reaction center highlighted is shown in Figure 15 (4).

It should be stressed that similar to the Lynch–Willett method, the McGregor–Willett approach is also designed to detect the overall structure changes of a reaction, which may not directly reflect the true reaction mechanism. Take the hydrolysis of an ester as an example (Figure 16). In this case, the application of the MCS obtained will lead to the conclusion that the alkyl-oxygen bond (between atoms 2 and 3) has been broken irrespective of the actual mechanism of the reaction.

It should also be noted that the McGregor–Willett method may fail for some unbalanced reactions. For example, the reaction Figure 17 (1) involves the hydroxylamine on the carbonyl group. However, because the MCS algorithm is unaware of the reagent that is missing in the reaction equation, the product oxygen atom 3 is presumed to be the same as atom 3 in the reactant. And thus, the bond changes in the reaction are incorrectly analyzed as shown Figure 17 (3). The correct reaction center is already shown in Figure 11. It should be noted that the MCS shown in Figure 17 (2) is similar to that in Figure 14 (2). That is, both the MCSs consist of multiple disconnected fragments. Furthermore, in Figure 14, the atoms 1 and 4 are included in the MCS, the bond that connects these two atoms in the product is excluded in the MCS. Similarly, in Figure 17, the atoms 2 and 3 are included in the MCS, the bond that connects these two atoms in the reactant is excluded in the MCS. The example shown in Figure 17 indicates that care must be taken in using an MCS that consists of multiple disconnected fragments.

To demonstrate the efficiency of their two-step method, McGregor and Willett applied it to a set of 140 reactions in which both the reactant and the

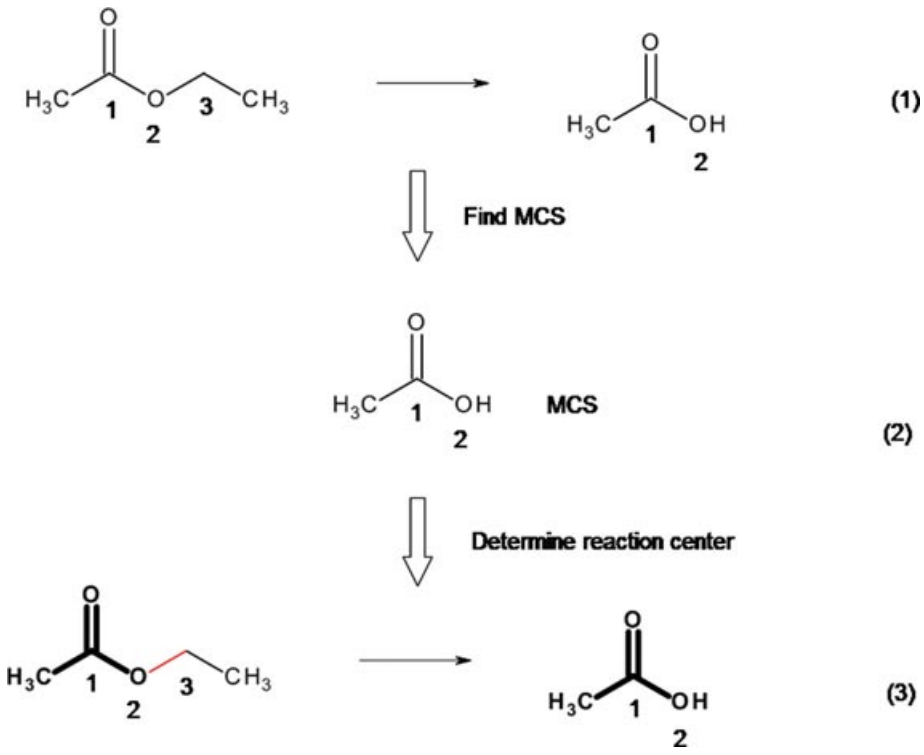

FIGURE 16 | (1) Reaction. (2) MCS. (3) The reaction with reacting bond highlighted in red and the MCSs are in bold.

 © 2013 John Wiley & Sons, Ltd.

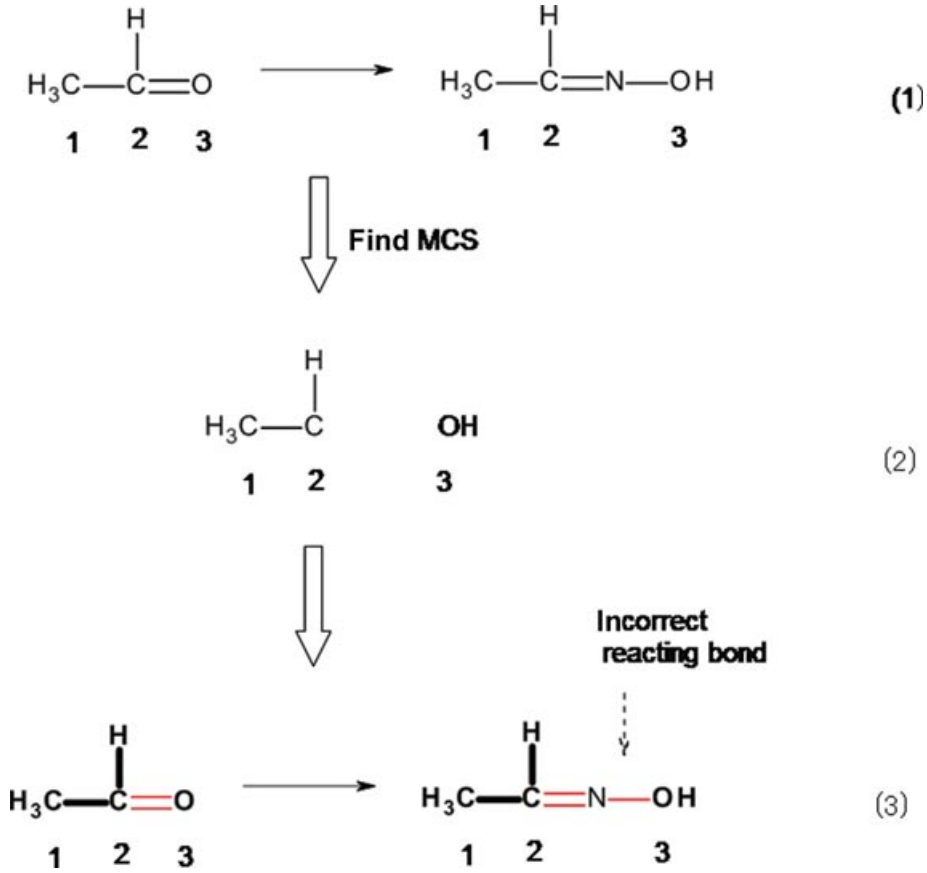

FIGURE 17 | (1) Reaction. (2) MCS. (3) The reaction with the reacting bonds highlighted in red, and the MCSs marked in bold. (Adapted from Ref 39. Copyright 1981, American Chemical Society.)

product structures each consists of fewer than 25 atoms or bonds. It is interesting to note that for the identification of the reaction centers, using the reaction sites as input to their procedure is over 200 times faster than using the complete molecule structures as input, a significant improvement of the performance. Finally, as the authors pointed out in their paper, this two-step procedure is efficient but approximate. Some of the limitations are certainly inherited from the first step—the EC-based method.

Funatsu et al. MCS-Based Method

In 1988, Funatsu et al.[13] described an MCS-based algorithm for the recognition of the reaction centers. Unlike Lynch and coworkers whose interest in development of reaction center detection methods was to use the obtained reaction centers as indexing terms for the retrieval of reaction information, Funatsu et al. recognized the importance of reaction center identification for reaction prediction and synthesis design.

Similar to the McGregor–Willett method, the MCS algorithm used in the Funatsu et al. method also ignores the bond orders. They call such an MCS as the maximal common skeleton structure. However, there are slight differences of the criteria for choosing the MCS between the two methods. In the McGregor–Willett method, the MCS is the common substructure that contains the largest number of bonds, whereas in the Funatsu et al. approach, the MCS is the one that contains the largest number of atoms. Furthermore, if there are multiple candidates, Funatsu et al. chose the one that contains the larger number of bonds and has the smallest number of differences in the bond orders in the final MCS. On the other hand, McGregor and Willett[39] use a more sophisticated, weighted approach to choose the final MCS. This ensures that a mapping involving a correspondence between two multiple order bonds does not take priority over some other mapping that preserves more of the structure when handling those reactions that involve bond-order changes.

The main difference between the above two methods is in that the McGregor–Willett method

 © 2013 John Wiley & Sons, Ltd.

uses the reaction sites obtained by the Lynch–Willett method as the input for the MCS algorithm, whereas the Funatsu et al. approach uses the pair of complete reactant and product structures as the input for searching MCS.

To improve the performance of finding the MCS, Funatsu et al. use a Morgan algorithm-based procedure similar to the one described by Lynch and Willett[17] to obtain the guiding information of reactant–product atom equivalences. Specifically, instead of directly using the EC-based largest common substructure of the reactant and product as an approximate MCS as Lynch and Willett did, Funatsu et al. use the former to determine the starting atom pair for their MCS procedure to speed up the process of identifying the MCS.

The Funatsu et al. procedure for detecting reaction sites for reactions with a single reactant and a single product is as follows:

1. Input the connectivity matrices of the reactant (R) and product (P) and replace them with the corresponding adjacency matrices.
2. Calculate EC values for both reactant and product. Obtain all the EC-MCSs (they call EC-MCS as the pie-maximal common skeletal structure) and collect a set of center atoms of EC-MCSs.
3. Find all possible MCSs between the reactant and the product using the pairs of center atoms of EC-MCSs obtained in step 2 as the starting points of the MCS search algorithm.
4. Choose the largest MCS among all the MCSs obtained in step 3 as the final solution and keep a copy of it. Update the adjacency matrices by eliminating the MCS atoms from the corresponding matrices of reactant and product, respectively. Repeat steps 2 to 4 until no more atom correspondence between the reactant and the product exists.
5. Determine the reaction center based on the correspondence of atoms associated with the MCS as well as comparing the connectivity matrices of the reactant and product.

The procedure used by Funatsu et al. for calculating EC values in step 2 is slightly different from that used by Lynch and Willett. Funatsu et al. EC values for both reactant and product are calculated using the following method:

1. EC_i^0 = atomic number × 10 + the number of adjacent bonds of atom i.

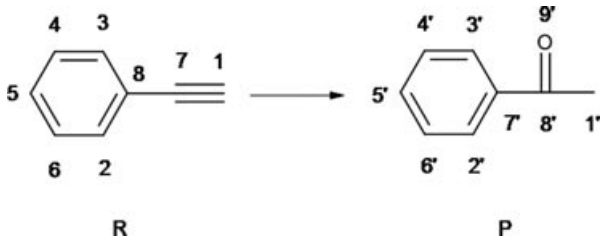

FIGURE 18 | The structures for reactant R and product P. (Adapted from Ref 13. Copyright 1988, Elsevier.)

2. $EC_i^n = 4EC_i^{n-1} + \sum EC_i^{n-1}$, where the summation is over all adjacent atoms of atom i.
3. Repeat step 2 until $EC_{ri}^n \neq EC_{pj}^n$ for all reactant–product atom pairs.

Consider the reaction shown in Figure 18 as an example. After the 4th iteration, there are no more pairs of reactant–product atoms that have the same EC values. For the first four iterations (0, 1, 2, 3), atoms 5 and 7 of the reactant and the atom 5′ of the product have the same EC values (62, 372, 2232, 13,394). However, atom 7 is not in an aromatic ring, whereas both atoms 5 and 5′ are members of aromatic rings. Therefore, the atoms (5,5′) are the only reactant–product atom pair that has the same EC values (62, 372, 2232, 13,394) and matchable properties, and the corresponding substructure with a radius of 3 bonds and centered at atoms 5 and 5′, respectively, is common to the reactant and product. Therefore, the pair of atoms (5,5′) is chosen as the starting point to carry out the backtracking search for MCS.

It should be noted that if the six-membered rings in both reactant and product are not aromatic, then another atom pair (7,5′) must also be chosen as the starting point for the MCS search. This is another example that shows that the EC procedure used by Funatsu et al. would fail to distinguish atoms 5 and 7 in such a simple structure.

The MCS algorithm that Funatsu et al. use in step 3 is also a backtracking[41] based search algorithm. However, unlike McGregor's MCS algorithm that dynamically orders the search tree at each match step, in the Funatsu et al. method, the search correspondence matrix is preordered, that is, the pair of atoms chosen as the starting points of reactant and product structures are ranked 1, and the atoms at α- and β-positions are separately ranked 2 and 3, and so on.

The above procedure for detecting reaction site for the reaction with a single reactant and a single product can be extended to deal with a reaction that

 © 2013 John Wiley & Sons, Ltd.

consists of multiple reactants and a single product as shown below:

1. Input the connectivity matrices of the reactants (R_1, R_2) and product (P) and replace them by the adjacency matrices.
2. Calculate EC values for reactants (R_1, R_2) and product (P). Obtain all the EC-MCSs and collect a set of center atoms of EC-MCSs.
3. Find all the MCSs by comparing R_1 and P, and R_2 and P, respectively, using the pairs of center atoms of EC-MCSs obtained in step 2 as the starting points of the MCS search algorithm.
4. Choose the largest MCS among all the MCSs obtained in step 3 as the final solution and keep a copy of it. Update the adjacency matrices by eliminating the MCS from the corresponding matrices of reactant and product, respectively. Repeat steps 2 to 4 until no more atom correspondence between the reactant and the product.
5. Determine the reaction center based on the correspondence of atoms associated with the MCS as well as comparing the connectivity matrices between reactants R_1 and R_2 and the product P.

Take the Diels–Alder reaction in Figure 19 (1) as an example. The larger MCS of this reaction (i.e., the skeletal structure of maleic anhydride) is obtained by comparing R_2 with P. This MCS is eliminated from the adjacency matrix. Then MCS is searched again and a skeleton of butadiene is detected. By comparing this new MCS and the connectivity matrices of R_1, R_2, and P, the reaction sites are detected [Figure 19 (2)].

Reactions that consist of multiple reactants and multiple products ($R_1 + R_2 \rightarrow P_1 + P_2$) can be handled using the above procedure. This can be achieved by splitting the reaction into simpler forms, each of which contains only one product ($R_1 + R_2 \rightarrow P_1$ and $R_1 + R_2 \rightarrow P_2$).

Similar to the McGregor–Willett method, the Funatsu et al. approach also has the limitation with regard to where the oxygen atom comes from in some reactions. As an example, consider the first step of the Pechmann condensation [Figure 20 (1)]. This step of the reaction involves the formation of one product from two reactants. The MCS algorithm identified two MCSs: the skeletons of phenol and β-ketoester [Figure 20 (2)]. Both MCSs contain seven atoms, but the former has one more bond than the latter. Therefore, the phenol MCS is chosen as the solution. This leads to the reaction site as shown in [Figure 20 (3)]. However, there is a problem with regard to where the oxygen atom 1 of the product comes from. In Figure 20 (3), this oxygen atom supposedly comes from the first reactant. If the substituent at the β-position of the second reactant is larger than the methyl group (say, an ethyl group), then the β-ketoester MCS will be larger than the phenol MCS and thus the former will be chosen as the solution. In this case, the oxygen atom in the corresponding

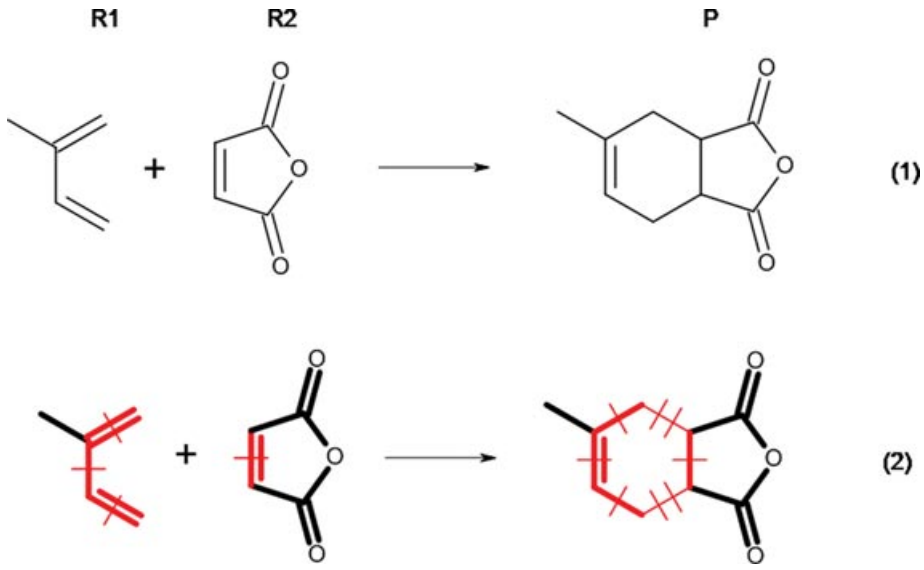

FIGURE 19 | (1) The Diels–Alder reaction that consists of two reactants and one product. (2) The reaction sites recognized are marked in red, and the MCSs are highlighted in bold. (Adapted from Ref 13. Copyright 1988, Elsevier.)

© 2013 John Wiley & Sons, Ltd.

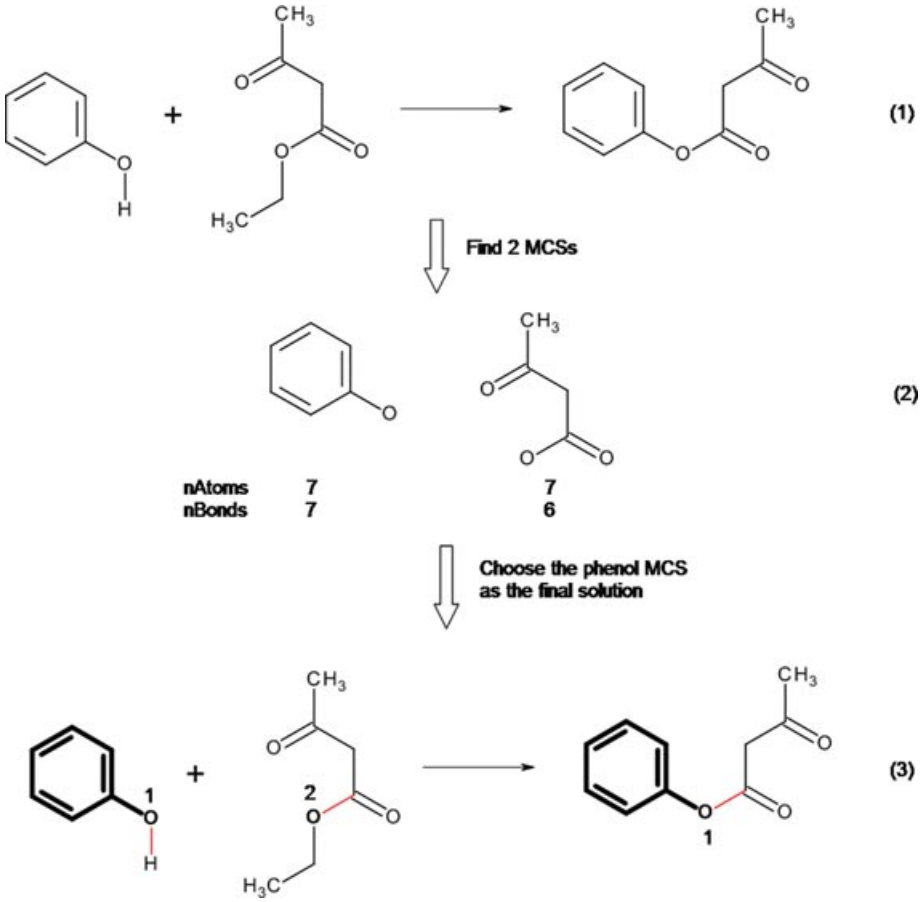

FIGURE 20 | (1) The first step of the Pechmann reaction. (2) Two MCSs. (3) The reaction with reacting bonds highlighted in red, and the phenol MCS highlighted in bold. (Adapted from Ref 13. Copyright 1988, Elsevier.)

product will be treated as coming from the second reactant. This kind of relative-MCS-size-dependent result is certainly not limited to only the ester formation reaction. Several other reactions also cause similar problems.

In some cases, there are multiple same-size (same numbers of atoms and bonds) MCSs between a reactant and a product that possess the same characteristics. In such a case, there will be multiple possible AAMs and reaction centers. Consider the Favorskii rearrangement (Figure 21).[42] Funatsu et al. program outputs eight possible candidate reaction sites for this reaction. Two of the most interesting ones are shown in Figure 21. From this figure, it can be seen that in the first reactant, the bond may be broken between atoms 1 and 2 [Figure 21 (1)] or between atoms 2 and 3 [Figure 21 (2)]. A mechanistic study of this reaction using isotopic labeling technique reveals that this reaction proceeds via an intermediate that contains the cyclopropanone substructure [Figure 21 (3)].[43] This experimental result indicates that there is an equal probability of bond breaking between atoms 1 and 2 and atoms 2 and 3. Therefore, two possible solutions proposed by the computer program are consistent with the experimental results.

In the above discussion, it is assumed that the EC-MCS can be obtained and thus the corresponding pairs of center atoms of EC-MCS can be used as the starting points of the MCS algorithm. However, in some cases, the EC-MCS cannot be obtained. Funatsu et al. use a special procedure to deal with one of the special cases where the reactant and product have the identical skeletal structure, and the structure has high symmetry. For such structures it is often difficult to obtain the sets of correspondent nodes; furthermore, the MCS algorithm has to examine many combinations to obtain the correspondence between atoms. In such a case, Funatsu et al. use the Morgan algorithm to canonicalize the two structures. The atom orders thus obtained are used to determine the correspondences between the reactant and product atoms.

 © 2013 John Wiley & Sons, Ltd.

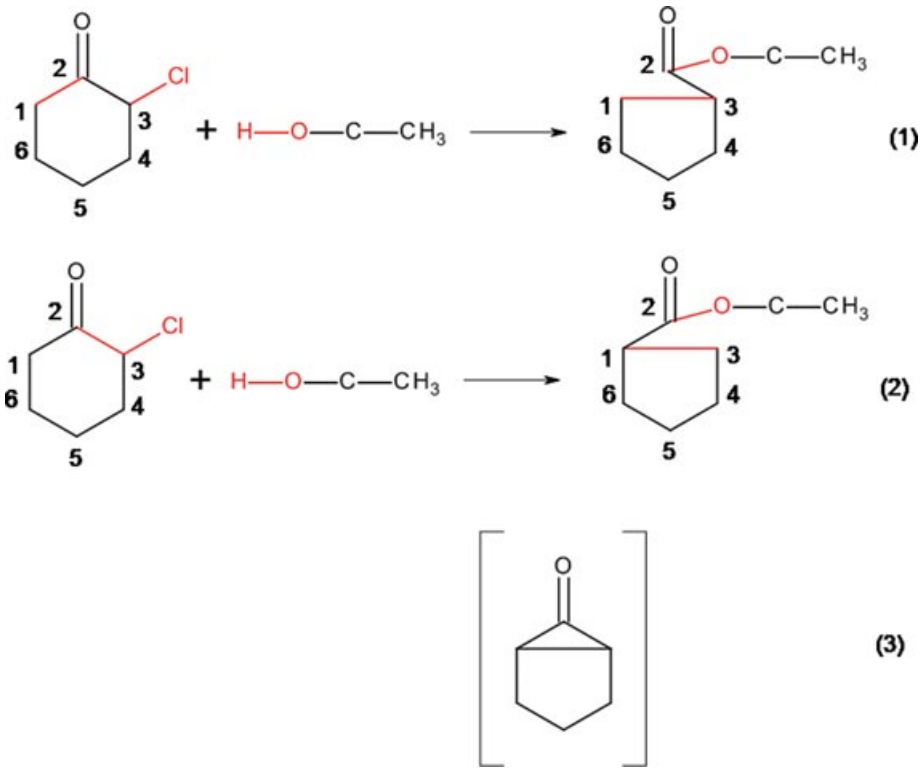

FIGURE 21 | (1) and (2) are two possible reaction sites (highlighted in red) of the Favorskii rearrangement. (3) The reaction intermediate that contains the cyclopropanone substructure. (Adapted from Ref 13. Copyright 1988, Elsevier.)

Maximum Common Edge Substructure-Based Method

Recently Körner and Apostolakis[44] introduced a new method for reaction mapping that also employs an MCS algorithm. However, their approach is more sophisticated than the previous ones, and their use of the MCS algorithm is quite different from other MCS-based methods. The Körner–Apostolakis method is based on three assumptions: (1) low temperature assumption (the valid reaction mechanism converts the reactants to the products with the lowest activation energy); (2) single transition state assumption (the reaction involves only a single transition state); and (3) additivity assumption (the activation energy for the transition state (called Imaginary Transition State Energy, ITSE) is the sum of the activation energies of reacting bonds. Their optimization goal is to minimize the ITSE. Therefore, they called their method the ITSE-based method.

The problem of minimizing the ITSE is solved via finding the MCS of weighted edge graphs, which are derived from the reactant and product structures (graphs). Before going into more detail, let's first briefly discuss the relationship between a graph and its corresponding edge graph. In graph theory, the edge graph $E(G)$ of undirected graph G is another graph $E(G)$ that represents the adjacencies between edges of G. The edge graph is also called line graph.[45] One of the most important edge graph theorems is that with one exceptional case the structure of G can be recovered completely from its edge graph.[46] In the context here, the nodes of an edge graph represent the bonds in the original molecular structure. In 2002, Raymond et al.[47] reported a rigorous algorithm for perceiving the maximum common edge subgraphs (MCESs) and applied it to the calculation of graph similarity.

Körner and Apostolakis extended the algorithm of Raymond et al. to deal with the weighted MCEG. Similar to the previous MCS-based algorithms, they also use different weights for matching bonds with different multiplicity. However, Körner and Apostolakis go one step further by introducing weights for other bonds too depending upon the types of atoms that form the bond. For example, CC σ-bonds have a weight of 1.5, C—N-amine, C—O-ester, and C—S-thioester bonds have a weight of 0.48, and all other bonds have a weight of 1. In addition, the weight of the bond is increased by 0.02 for each additionally mapped π-bond. The weights directly correspond to the cost of not matching the bonds that induce the weight in the first place. The edge graph

 © 2013 John Wiley & Sons, Ltd.

matching identifies the BBM and the reacting bonds. Those nodes that are not matched represent the reacting bonds (broken or formed bonds) of the original molecular structures.

It should be noted that the AAM is derived from the BBM in the second step, and there are some complications to obtain the AAM from the edge graph matching results. For example, when a single atom is transferred from a reactant to a product, it does not retain any of its bonds, and thus that atom is not included in the MCES. Besides, single mapped bonds between atoms of the same element type lead to two possible mappings for those atoms. Their method for solving those problems is as follows. First, all bonds that were not mapped in the MCES are removed from the original reactant and product structures. For a balanced reaction, the remaining structures at the two sides of the reaction must be isomorphic. They then use the MCS algorithm to directly match the remaining atoms based on the atom symbols.

Unlike several new reaction mapping algorithms that can deal with only the balanced reactions (see the following sections), Körner and Apostolakis[44] proposed some techniques to deal with unbalanced reactions. (1) In the atom mapping step, all atoms that cannot be mapped are simply removed. (2) Adding the missing atoms to the corresponding side of the reaction to make it balanced. (3) Assuming that the structural fragments that only appear at one side of the reaction do not change.

Like the Funatsu et al. method,[13] the Körner–Apostolakis algorithm can also find multiple solutions for a given reaction that can be useful for study alternative reaction mechanisms. All the hydrogen atoms are ignored during the matching process. The number of reacting hydrogen atoms is derived from the sum of the valence change of heavy atoms in the obtained mappings. When two or more mappings are obtained, they are sorted based on the number of reacting hydrogen atoms.

Adding different weights to the bonds connecting to different types of atoms allows the graph matching process to be guided by the chemical knowledge. This improves the accuracy of the reaction mapping results. To test the robustness and accuracy of their method, Apostolakis et al. applied their algorithm to the KEGG database (~6700 reactions).[48] They also validated their method against the manual mappings found in the BioPath database (1500 biochemical reactions).[49] The results show that in 98% of cases, the automatically generated reaction mappings are consistent with the manually annotated mappings. Although, as the authors pointed out, the agreement of the two independent methods for a particular reaction mapping is no proof of its correctness; the high agreement between these two independent approaches is, however, quite impressive. To assess the improvement obtained by using bond weights, they performed both weighted and nonweighted reaction mapping with the same algorithm for the BioPath database. As expected, the weighted method had fewer incorrect cases than the nonweighted one (14 vs. 52).

Let us consider an interesting example given in Figure 22. The mechanism of the isochorismate synthase reaction (RXN00053) stored in BioPath is shown in Figure 22(1). In this case, the carboxy group is transferred across the conjugated system through a 1,3-shift, and the reaction involves only one bond broken in the reactant and one bond formed in the product. It looks quite simple. The Körner and Apostolakis' algorithm, however, found a different reaction mapping for RXN00053, as shown in Figure 22(2). In this mechanism, although there is also only one bond broken in the reactant and one bond formed in the product, the reaction has an additional four bonds changing their bond order in the reactant and product structures—the hydroxyl group undergoes a 1,5-shift across the conjugated system and causes a cascading shifting of single and double bonds of the conjugated ring system. This predicted result is much more complicated than the one in the BioPath database.

So, which mechanism is correct? To answer this question, they reviewed the literature. They found that up to 2003 the enzyme isochorismate synthase was assigned to the class 'EC 5.4.99 Transferring Other Groups' of the enzyme classification system. The reaction mapping in the BioPath database is consistent with this class. However, in 2003 the isochorismate synthase was reassigned to the category 'EC 5.4.4 Transferring hydroxy groups'.[50] The new class is consistent with the reaction mechanism implied by the ITSE mapping.[51] It should be noticed that the products (1) and (2) in Figure 22 are identical constitutionally, but different stereochemically. Because the ITSE algorithm does not take into account of stereochemistry, it found the same solution—the bond broken, formed, bond order changes—for both reactions (1) and (2) based on the bond weights. The reaction RXN00053 in the BioPath was corrected based on the ITSE result.

It is interesting to note that the authors attribute their result to the effects of the bond weights: a C—O bond is more easily broken than a C—C bond because the corresponding weights for matching are 1.0 and 1.5, respectively.[52] Although this explanation is

 © 2013 John Wiley & Sons, Ltd.

reasonable, it still gives one the impression that the ITSE algorithm performed some magic to find those four double/single bond shifts.

In fact, the same reaction mapping result as shown in Figure 22(2) can be obtained by using the MCS-based algorithm. To explain this, let us first highlight with the bold bonds the largest common substructure, which does not include the bond broken or formed between the reactant and the product for both reactions (see Figure 23). From Figure 23 it can be seen that the common substructures in (1) and (2) contain 13 and 15 heavy atoms, respectively. Therefore, the second common substructure is the MCS for RXN00053. The C—O bond in the reactant must be broken, and the C—O bond in the product must be formed. Furthermore, based on the AAM and BBM from the MCS algorithm, it is quite intuitive to see that the bond orders of four bonds in the ring were changed during the reaction process [Figure 23 (2)].

It should be pointed out that, here, we do not intend to play down the importance of the bond weights. In fact, the bond weights were used to allow the MCS algorithm to match bonds with different multiplicities, as discussed previously.[13,39] The point we want to make here is that for this specific case, the concept of MCS or MCES plays a critical role.

A major shortcoming of the Körner–Apostolakis method is that it does not take into account stereochemistry. This can lead to incorrect reaction mapping results.[44] Furthermore, as the authors pointed out, their two-stage approach makes it complicated to solve this problem.[52] Besides, although the Raymond et al. MCES algorithm[47] is fast, it seems that this two-stage approach may not be as efficient as the MCS-based one-step methods because the former needs to perform the graph matching twice for the same reactant–product pair, whereas the latter only once.

Another notable issue is that their method lacks a good way to deal with the reactions that contains several small molecules with symmetry, and the performance for dealing with such reactions can be quite poor. For example, as the authors pointed out, the mapping of the reaction shown in Figure 24 (RXN00048) could not be automatically completed even after 24 h of CPU time.[52]

A major advantage of the MCS-based approach for reaction mapping is that the idea is simple and straightforward. It can also be easily employed to handle bond order changes. However, this approach has some drawbacks. A notable shortcoming is that the MCS problem is NP-hard.[53] It can be time extensive for handling large, complicated structure pairs. Several heuristics can be employed to significantly improve the performance with the sacrifice of some accuracy.[25] It should be pointed out that even when the MCSs between reaction and product structure pairs are found, the MCS-based methods still cannot guarantee that the chemically correct reaction mappings will be found.

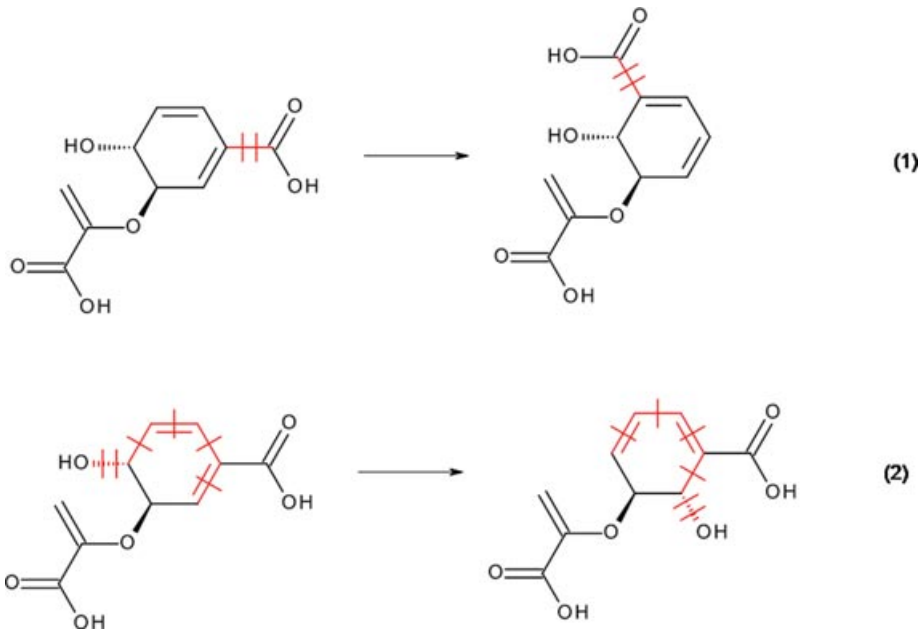

FIGURE 22 | Isochorismate synthase (RXN00053). (a) The reaction centers marked in the original BioPath database. (b) The solution of the ITSE method. Note: The reacting bonds are highlighted in red. (Adapted from Ref 52. Copyright 2008, American Chemical Society.)

© 2013 John Wiley & Sons, Ltd.

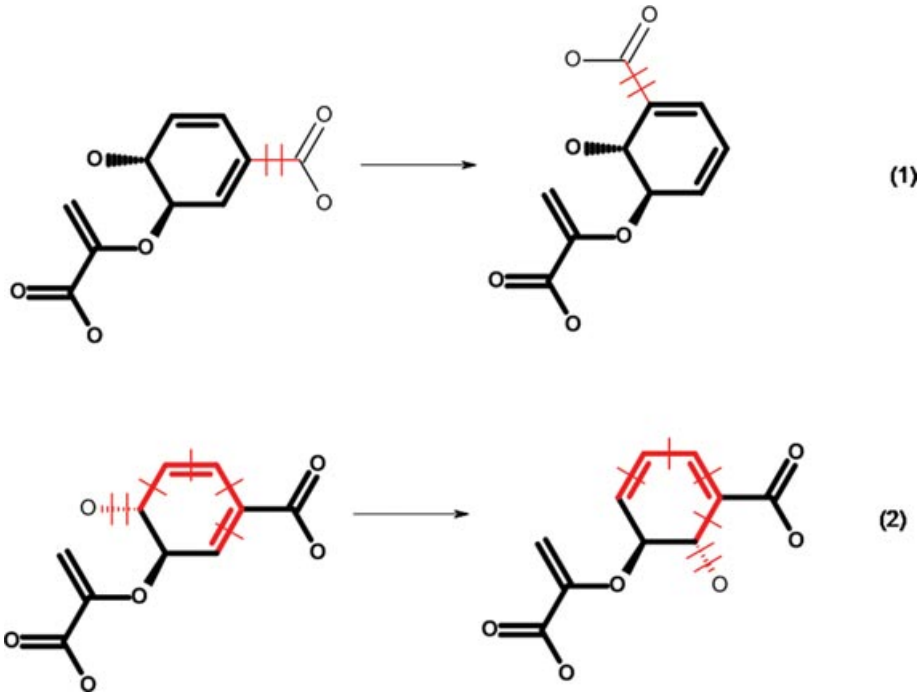

FIGURE 23 | Two reactions (1) and (2) are identical with the corresponding reactions in Figure 22 except that the largest common substructure between the reactant and the product that does not include reacting bonds were highlighted with the bold bonds. The reacting bonds are highlighted in red and also marked with the hash marks.

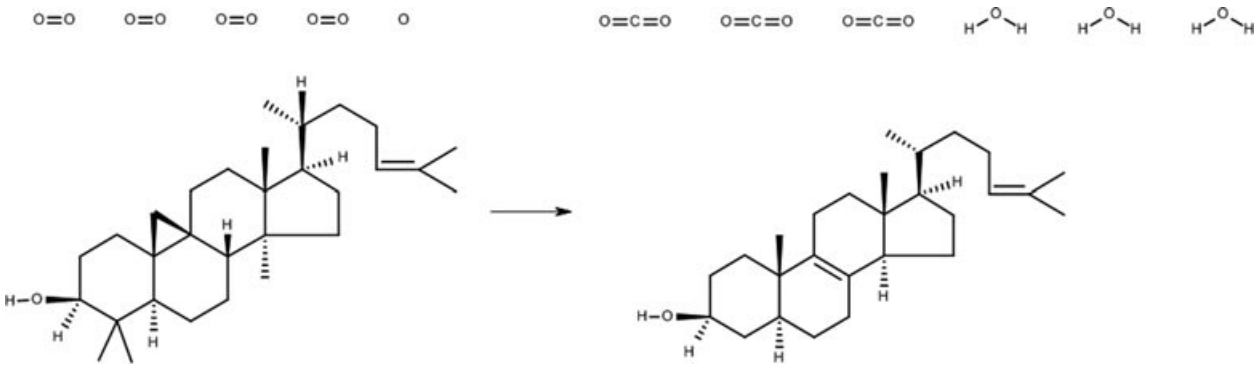

FIGURE 24 | The reaction RXN00048 of BioPath.[52] Note 1: Hydrogen atoms with no stereo bond connections are not shown here. Note 2: In BioPath, all reactants are grouped into one structure representation, and so are the products. (Adapted from Ref 52. Copyright 2008, American Chemical Society.)

OPTIMIZATION-BASED METHODS

Principle of Minimal Chemical Distance

In 1980, Jochum et al.[54] proposed an empirical formulation called the principle of minimal chemical distance, that most chemical reactions follow the shortest path for transforming the reactant structure to the product structure. In recent years, with the development of nontraditional chemical reaction databases, especially biological pathway databases (such as KEGG LIGAND databases[19]), interest has renewed in the development of more accurate methods for automatic reaction mapping and reaction center detection. Most of the new methods are optimization based. Their optimization goal is to find the AAM solution with the minimum number of bonds broken and/or formed. To a certain extent, Körner and Apostolakis' reaction mapping method described in the previous section is also optimization-based, but from the implementation point of view, it is closer to MCS-based methodologies.

 © 2013 John Wiley & Sons, Ltd.

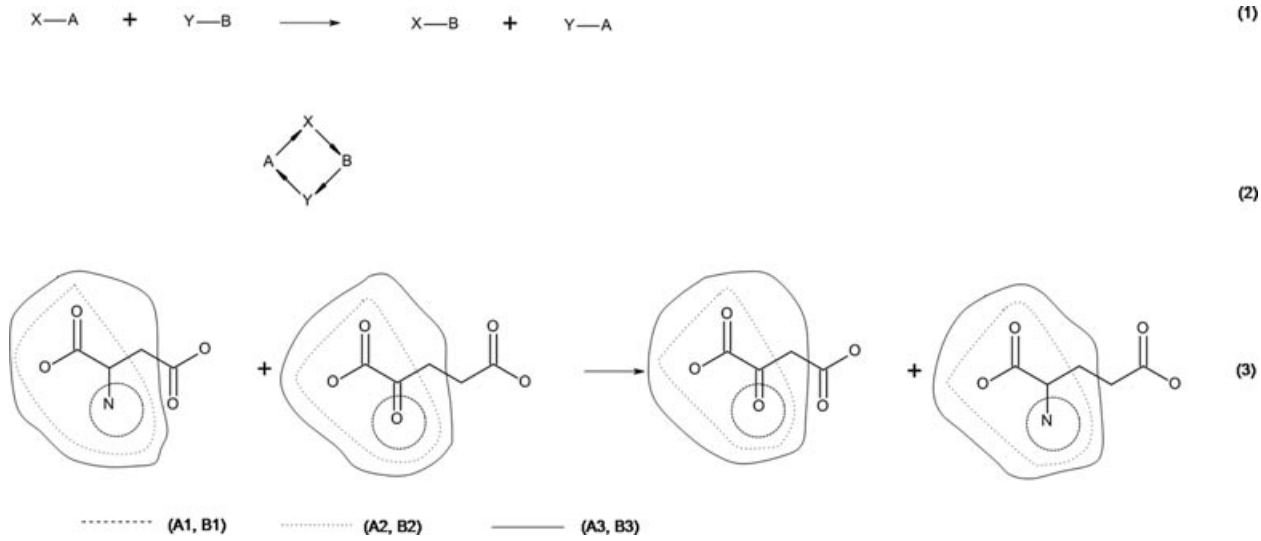

FIGURE 25 | (1) A special type of chemical reactions with two reactants and two products. All reaction structures are acyclic and can be split into two parts by cutting one bond ('–' corresponds to a cut). (2) A directed cycle of length 4 that is built from reaction (1). Chemical cuts X–A, Y–B, X–B, Y–A separately correspond to directed edges (*A*, *X*), (*B*, *Y*), (*X*, *B*), and (*Y*, *A*). (3) Example of a reaction instant that has the form (1). This reaction is catalyzed by a transaminase. There are three possible pairs for (*A*, *B*), where (A_1, B_1) is the most plausible. (Adapted from Ref 55. Copyright 2004, Mary Ann Liebert, Inc.)

Graph Isomorphism-Based Methods

The main principle of the traditional common substructure-based reaction-mapping algorithms discussed above is first to identify the unchanged part of the molecular structures, and then derive the changed sections of the structures—the reaction centers. Can one develop a reaction mapping algorithm that is based on the reversed order of the above two steps? The answer is yes. This is the basic idea of the graph isomorphism-based reaction mapping algorithms.

As discussed previously, the graph isomorphism is an equivalence relation on graphs. There are two new reaction mapping algorithms that are based on the determination of graph isomorphism.

Akutsu Algorithm

In 2004, Akutsu[55] reported a novel method for extracting mapping rules from enzymatic reactions. Instead of finding the MCS between a reactant and product pair, the Akutsu algorithm works by cutting reactant and product structures into smaller fragments. This process is called partitioning. Then, unique names are generated for all fragments. Those names are used to determine whether each reactant fragment is isomorphic to a product fragment. If all reactant fragments are isomorphic to their corresponding product fragments, the procedure ends. The bonds that are cut at the reactant site are the broken bonds in the reactants; the bonds that are cut at the product site are the formed bonds in the products. The optimization goal of the procedure is to find the minimum number of bonds broken and formed in the reaction. Therefore, the partition starts with cutting one bond for reactants and products. After all bonds are tried and no solution has been found, all combinations of a two-bond cut are performed. This process is repeated until an optimized solution is found.

Akutsu describes both theoretical and practical algorithms for dealing with a special type of reactions that take the form of Figure 25 (1), where A, B, X, and Y are trees. An example reaction of this type is given in Figure 25 (3).

It should be noted that only one bond needs to be cut for each structure of the above reaction. The practical algorithm (Algorithm 3 in his original paper) is shown below.

1. For all partitions (X_i, A_i) of reactant 1, compute Morgan names of X_i and A_i. For all partitions (Y_j, B_j) of reactant 2, compute Morgan names of Y_j and B_j.For all partitions (X_k, B_k) of product 1, compute Morgan names of X_k and B_k. For all partitions (Y_h, A_h) of product 2, compute Morgan names of Y_h and A_h.
2. For each A_i, examine whether there exists A_h that has the same Morgan name. If so, create an object A_i and use its Morgan name as its label.

 © 2013 John Wiley & Sons, Ltd.

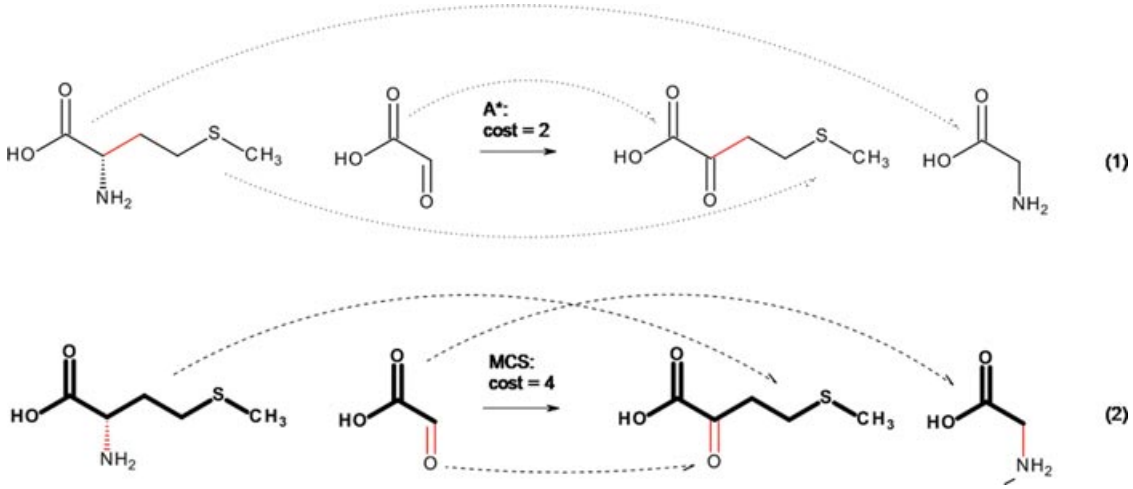

FIGURE 26 | R00652 methionine: glyoxylate aminotransferase reaction. (1) The atom mapping found by the A* algorithm requires only one bond broken and one bond formed; but this mapping is incorrect. (2) The atom mapping found by the MCS algorithm requires four operations: two bonds broken and two new bond formations; this is the correct reaction mechanism. The MCSs in (2) are highlighted in bold. The reacting bonds in both (1) and (2) are marked in red. (Adapted from Ref 55. Copyright 2004, Mary Ann Liebert, Inc.)

3. Handle B_j using the same procedure as in step 2.
4. For all pairs of objects (A_i, B_j), let $M[A_i, B_j] = 0$.
5. For all pairs (X_i, X_k) that have the same Morgan name, let $M[A_i, B_k] = 1$.
6. For all pairs (Y_j, Y_h) that have the same Morgan name, let $M[A_h, B_j] = 2$, if $M[A_h, B_j] = 1$.
7. Output (A_i, B_j) such that $M[A_i, B_j] = 2$.

The above algorithm works in $O(n^2)$ time. It should be noted that this algorithm cannot guarantee the chemical correctness of the results.

Take the reactant shown in Figure 25 (3) as an example. There are three possible pairs for (A, B), and the pair (A_1, B_1) is the most plausible. If hydrogen atoms and bond types are ignored and the removed edges are not taken into account, the procedure (called Algorithm 3c) in Figure 26 can find all three pairs.

The above algorithm was tested using enzymatic reaction data in the KEGG/LIGAND database (release 20.0).[56] This database contains 5238 enzymatic reactions, but only 2346 reactions belong to the reaction type shown in Figure 25 (a) and were actually used for testing. Among them 1912 reactions were successfully handled by Algorithm 3c with a success rate of 81.5%.

As Akutsu pointed out in the *Discussion* section of his paper, there are many challenges to overcome when processing practical reactions. First, the time complexity increases as the number of bonds to be cut at the same time (C) increases. In some cases, C can be 3 or more and thus, it can take $O(n^5)$ time or more. Second, no stereochemical information is considered in the current algorithms. Third, the algorithm cannot deal with reactions that involve ring modification. Fourth, the algorithm cannot handle unbalanced reactions that are missing small molecules; Akutsu pointed out that some preprocessing method should be developed in order to identify the omitted small compounds. Finally, if there are multiple mapping rules consistent with a given reaction, they should be scored based on chemical knowledge.

Overall, Akutsu's algorithms were designed for handling specific types of reactions, and are quite fast. However, they lack generality.

Crabtree–Mehta Algorithm

To extend Akutsu's algorithm to more general reactions, Crabtree and Mehta[57] reformulated the reaction mapping problem and generalized Akutsu's ideas by introducing a new concept of identity chemical reaction and related two theorems. They defined the identity chemical reaction as the reaction in which there is a one-to-one mapping between reactant graphs and product graphs such that each reactant graph is isomorphic with the corresponding product graph. Their first theorem (Theorem 4.2) is about using a unique graph name to determine whether a reaction is an identity chemical reaction or not. It states:

 © 2013 John Wiley & Sons, Ltd.

> 'The reaction is an identity reaction if and only if all reactant names match product names provided that the canonical labeling algorithm generates unique names for distinct molecules'.

Their second theorem (Theorem 5.1) is about how to perform the reaction mapping. This theorem reads:

> 'Any mapping of a valid chemical reaction is equivalent to cutting a set of bonds in the reactants and products such that the resulting equation is an identity chemical reaction'.

The above-mentioned definition and theorems together lay the foundation for their five reaction mapping algorithms: ExhaustiveBondSearch, FewestBondsFirst, FBF-Symbolized, ConstructiveCountVector, and CutSuccessiveLargest.

As its name indicates, the ExhaustiveBondSearch algorithm exhaustively searches the solution space for mappings that break the fewest number of bonds. It searches for patterns of broken bonds in the direction of increasing bond number. First, all combinations of breaking a single bond are considered until all structures can be matched or all combinations have been exhausted. Then, breaking two bonds are considered and so on. The pseudocode for ExhaustiveBondSearch is given below:

1. Create a bit pattern containing one bit for each bond in the reaction.
2. Initialize the bit pattern to all zeroes.
3. While there is at least one bit in the pattern set to zero:
    a. Create a new equation by breaking each bond represented by a 1 in the bit pattern.
    b. If the equation can be completely matched, then save it as a possible solution.
    c. Increment the bit pattern by one.

The four other algorithms are variations of the ExhaustiveBondSearch algorithm by introducing additional optimization strategies. For example, the FewestBondsFirst algorithm first searches all bit patterns containing a single 1, then followed by bit patterns containing two 1's, and so on, until a match is found. The FBFSymbolized algorithm is identical with FewestBondsFirst algorithm except that to eliminate some of the bit patterns, the former also employs the bond symbol (e.g., CO is the symbol of the bond between a carbon atom and an oxygen atom). Modifying the FBFSymbolized algorithm by adding a count vector to track the number of bonds by bond symbol on each side of the equation leads to a more efficient ConstructiveCountVector algorithm.

Unlike the above four algorithms that are designed to find an optimal solution, the fifth algorithm (CutSuccessiveLargest) is a greedy heuristic algorithm. It attempts to map reactants to products by successively cutting the largest structure on either side of the reaction until a valid mapping is found. It is the fastest one among all five algorithms but cannot guarantee to obtain the optimal solution.

Their Java-based implementation[58] offers three naming technologies: the Morgan algorithm,[28] the Nauty algorithm,[59] and the Faulon algorithm.[60]

Crabtree and Mehta compared the algorithms using the KEGG/LIGAND database (Version 20). In this test, the Nauty naming algorithm was used. Their FBF-symbolized, ConstructiveCountVector, and CutSuccessiveLargest were able to handle all the reactions with success rates of 99%, 99%, and 84%, respectively. In contrast, Akutsu's algorithm could only handle 45% of the reactions with a success rate of 82%. However, Crabtree and Mehta's algorithms are generally slower than the Akutsu's algorithm. In the above tests, FBF-symbolized, ConstructiveCountVector, and CutSuccessiveLargest took 774,471, 537,015, and 1400 seconds, respectively, whereas the Akutsu's algorithm took only 67 seconds.

It should be noted that the Crabtree–Mehta algorithms ignore double and triple bonds. They mentioned that their algorithms can be extended to map multigraphs that are obtained by representing double and triple bonds using two and three edges, respectively. Because multiple order bonds are common in molecular structures, the multigraphs would be quite complicated. Thus, mapping reactions with multiple multigraphs could be very expensive.

A*-Based Algorithm

Recently, Heinone et al.[61] reported a new reaction AAM algorithm based on Crabtree–Mehta's optimization philosophy. However, unlike the two-graph isomorphism-based algorithms discussed previously, Heinone et al. algorithm does not perform bond cut. Instead, it employs an A* search algorithm[62,63] to directly match reactant structures to product structures. The A* search algorithm[64] is widely used in path finding and graph traversal. It uses a best first search and finds a least-cost path from a given initial node to one goal node.

 © 2013 John Wiley & Sons, Ltd.

The objective function of their algorithm is to find an atom mapping that minimizes the graph edge edit distance. Given a pair of graphs G1, G2, the edge edit distance is defined as the minimum number of edge edit operations that is required to transform G1 to G2. In the reaction mapping term, the edge edit distance is the minimum number of broken, formed, and order changed bonds that are required to transform the reactant structure to the product structure.

The three most important components of the Heinone et al. algorithm are as follows:

- An A* type total path cost estimate to guide the search in the space of partial atom mappings.
- An extension operator for partial mappings that maintains the path cost estimates in constant time per edge.
- Pruning of A* search space by computing upper bounds on the optimal cost via fast greedy search.

The reactant atoms are numbered consecutively using breadth-first search. The search starts from an extreme atom of the largest reactant. It then iteratively processes the rest of the reactants in the order of their size. Although the algorithm does not impose the constraints on the reaction type or size, it does require that the reaction must be balanced.

Heinone et al. compared their A* algorithm with a greedy search, bipartite graph matching, and the MCS approach. The following results are worth noting. They implemented all algorithms in Java and computed with 4GB of memory and Intel Xeon X5355 CPU running at 2.66 GHz. They found that of the 6015 valid KEGG reactions, their A* algorithm managed to compute 5802 reactions, and their implementation of the MCS algorithm used by Hattori et al.[65] computed 5934 reactions. The two methods are similar in computational resource demands: both took less than one hour per reaction.[61]

With regard to the accuracy, Heinone et al. pointed out that the MCS fails especially on reactions with high minimum edit distance. These reactions are often large and have complex reaction mechanism. But no specific examples were given in the paper. However, they do cite a reaction instance for which the A* and the MCS algorithms found different atom mappings: the mapping with the minimum edit distance (cost = 2) found by their algorithm is incorrect, whereas the mapping with higher edit distance (cost = 4) found by the MCS algorithm is biochemically correct (see Figure 26).[66]

The equivalent AAMs can be eliminated within a mapping algorithm or be done afterward to classify the resulting mappings into equivalent classes. Heinone et al. implemented a VF2-based isomorphism algorithm[67,68] which can answer whether two atom mappings are isomorphic. This technique is directly integrated into the mapping algorithm itself.

Integer Linear Optimization-Based Methods

There are two integer linear optimization-based reaction mapping algorithms reported recently. The first lays the principle; the second extends it to handle bond weights.

First et al. Algorithm

First et al.[69] recently described an integer linear optimization-based method for mapping reaction and identifying multiple reaction mechanisms.

Linear optimization, also called linear programming (LP), is a mathematical method for determining a way to achieve the best outcome in a given mathematical model for some list of requirements represented as linear relationships.[70] It is a technique for optimizing a linear objective function, subject to linear equality and linear inequality constraints. A LP algorithm finds a point in the polyhedron where this function has the smallest (or largest) value if such a point exists.

If the unknown variables are all required to be integers, then the problem is called an integer programming (IP) or integer linear programming problem. However, in contrast to LP, in which even the worst case can be solved efficiently, IP problems are in many practical situations (those with bounded variables) NP-hard. If only some of the unknown variables are required to be integers, then the problem is called a mixed integer linear programming (MILP) problem. These are generally also NP-hard.[71] However, there are some important subclasses of IP and MILP problems that are efficiently solvable, most notably problems where the constraint matrix is totally unimodular and the right-hand sides of the constraints are integers. There exist several advanced algorithms for solving integer linear programs,[72,73] such as the cutting-plane method, branch and bound, and so on.

First et al. express the reaction mapping problem as an MILP model to identify a reaction mapping that minimizes the number of bonds broken and formed. The objective function consists of four summation terms. The first summation term is over reactant bonds with each term equal to one if the

 © 2013 John Wiley & Sons, Ltd.

bond breaks. The second is over product bonds with each term equal to one if the bond forms. The third is over tetrahedral atoms with each term equal to one if the stereochemistry changes. The fourth is over stereochemical double bonds with each term equal to one if the stereochemistry changes. The value of the objective function can be interpreted as the total number of bonds that break and form in the chemical reaction mechanism implied by the mapping.

They developed eight constraints for their MILP model. Constraints 1 and 2 require that each atom in the reactants/products maps to exactly one atom in the products/reactants, respectively; that is, there must be a one-to-one mapping between the atoms in the reactants and products. Constraint 3 states that only atoms of the same type can map to one another. Constraints 4 and 5 define a variable, which takes the value of one only if a reactant bond maps to a product bond. Constraints 6, 7, and 8 detect changes in stereochemistry during the reaction. Each solution of the model corresponds to a one-to-one mapping between reactant and product atoms. The model can be solved using standard MILP techniques, such as branch and bound.

The First et al. algorithm can find multiple optimal mappings. The equivalent mappings are resulted from the symmetries of reaction components. To address this problem, they employed a technique to break these symmetries of small molecules and terminal atoms by specifying an ordering for each pair of equivalent atoms. One atom in the pair is to map to an atom with a higher index than the other. The basic idea was also extended to deal with symmetries of symmetric rings.

First et al. implemented their method as a Web tool called DREAM[74]. It is freely available to the scientific community. DREAM accepts balanced chemical reactions in a variety of formats, including the Accelrys rxnfile.[14] If the input is provided by the Interactive Editor or in SMILES format,[75] the input will first be converted into an rxnfile. DREAM offers several options. It can produce either a single optimal mapping or multiple optimal mappings, each of which corresponds to a distinct reaction mechanism. The mapping solution can be obtained by minimizing the number of bonds that break or form in the reaction mechanism or by minimizing the number of bond order changes during the reaction. The result from DREAM is reaction mapping information that is stored in the AAM number column of the V2000 rxnfile format. If DREAM finds multiple mappings, it will generate a separate output rxnfile for each corresponding mapping. DREAM sends the results to the user by email. The result can also be viewed via its reaction viewer.

DREAM's output contains only AAM information; it does not offer any reaction center information. However, after the establishment of AAM relationships between reactant(s) and product(s), it is not difficult to derive the reaction center information. It should also be noticed that First et al. algorithm can only be applied to fully balanced reactions. For example, submitting the reaction shown in Figure 3 (1) to DREAM, produced the warning: 'There was a problem parsing your reaction file. Please check that your reaction is balanced.' After adding a methane molecule to the product side to balance the reaction, DREAM delivered a mapping result. As another example, we tested DREAM with reaction EC 2.6.1.28 (see Figure 27). It took more than 24 h to receive the results. It is impossible to judge the performance of DREAM based on the time required to receive the mapping result by email because DREAM may have to handle many requests. DREAM found six AAMs for this reaction. Two of them are shown in Figure 27. The AAM (1) is incorrect, whereas the AAM (2) is chemically correct. For this specific example, the chemically correct mapping can easily be obtained using the MCS-based algorithm.

The above example shows that a mathematically optimal mapping may not be the chemically correct solution. In the next section, we will show how to improve the accuracy of the MILP-based reaction mapping algorithm.

Latendresse et al. Algorithm

Most recently, Latendresse et al.[76] reported a new reaction mapping algorithm that is also based on the MILPs. This approach is essentially the same as that of First et al.[69] The main difference between the two methods is that the First et al. approach does not use bond weights, whereas Latendresse et al. method employs bond weights.

In principle, Latendresse et al. bond weights (they call them 'bond propensity') are similar to those of Körner and Apostolakis'.[44] For example, Körner and Apostolakis assign a weight of 1.5 to the C—C σ-bond, 0.48 to the C—N(amine), C—O(ester), and C—S(thioester) bonds.[44] The C—C bond weight is over three times larger than that for C—N, C—O, and C—S. For comparison, Latendresse et al. assign a bond propensity value of 400 to the C—C single bond, 56 to the C—N single bond, and 48 to the single C—O and C—S bonds. If the above-mentioned propensity values are all divided by 100, we get 4, 0.56, and 0.48. These weight values are now very close to the corresponding bond weight values of Körner and Apostolakis',

 © 2013 John Wiley & Sons, Ltd.

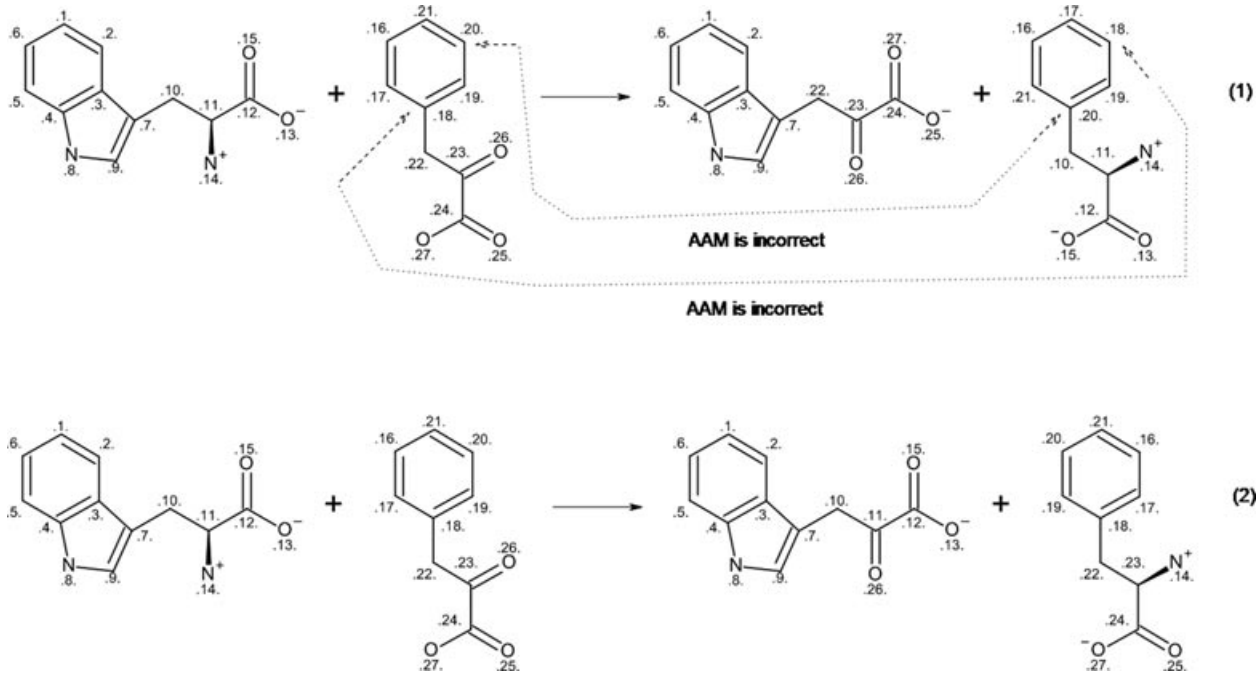

FIGURE 27 | DREAM found six AAMs for reaction EC 2.6.1.28. Two of them are shown in this Figure. (1) An incorrect AAM. (2) A chemically correct AAM. (Adapted from Ref 76. Copyright 2012, American Chemical Society.)

except the C—C bond weight of 4 that is almost three times higher than that of Körner and Apostolakis'. As expected, in both weight systems, the smaller a bond weight value is, the easier the bond is broken or made.

A main difference between the two systems is that the Körner and Apostolakis' bond weights are more general and can be applied to both chemical and biochemical reactions, whereas the bond weights used by Latendresse et al. were designed specifically for handling biochemical reactions. For example, Latendresse et al. do not assign any bond propensity value to the P—H bond, indicating that this bond does not exist in biochemical compounds.[76] Another difference that should be noticed is that Körner and Apostolakis apply the bond weights to the MCES matching process, whereas Latendresse et al. use the bond propensity values to calculate the parameters in the objective function of the MILP model.

Latendresse et al. algorithm can find multiple optimal mappings. Their reaction mappers use a post-processing step to eliminate the equivalent mappings. However, the use of the bond weights may allow the elimination of some chemically incorrect mappings. For example, as mentioned previously, First et al. DREAM found six mappings for reaction EC 2.6.1.28 (see Figure 27). Latendresse et al. algorithm found only the correct mapping.[76] They handle stereoconfiguration based only on the reaction depiction, and this may lead to incorrect mappings.[76]

To reduce the size of the MILP formulation and consequently to increase the speed in searching optimal mappings, they use an approximate approach to find similar ring structures between reactant and product. No detailed ring perception is involved. This step is done by a program that generates the MILP formulation. That program also determines whether this technique can be applied.

Latendresse et al. applied their approach on 7501 reactions of the MetaCyc database; it solved 87% of the models in less than 10 seconds. They reported an error rate of 0.9% by comparing their automatically identified AAMs to 2446 AAMs of the manually curated KEGG RPAIR database. They pointed out that their computational approach is the fastest and most accurate published to date.[76]

It should be noted that several recently published reaction mapping algorithms were not cited in their paper. To get the first-hand experience of their approach, we asked them to test several reactions discussed previously. Latendresse et al. reaction mapper did produce correct mappings for those tricky reactions. For example, their algorithm produced the correct reaction mapping for R00652 in 0.03 second for which Heinone et al. algorithm failed[61] (see Figure 26).[77] Latendresse et al. algorithm also produced the correct mapping for RXN00053 of Figure 23 in 0.05 second (see Figure 28).[78] For the Diels–Alder reaction shown in Figure 19,

 © 2013 John Wiley & Sons, Ltd.

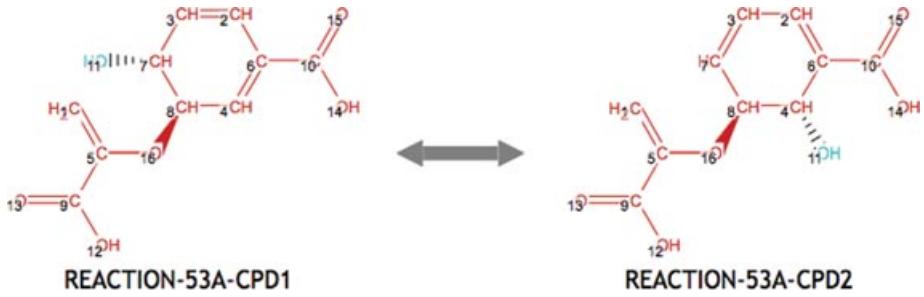

FIGURE 28 | Latendresse et al. reaction mapper produced correct reaction mapping for R00053 (see Figure 23). Latendresse et al. program did not mark the bond order changes. Note: their program marked the broken and formed bonds in black. (Reproduced with permission from Ref 76.)

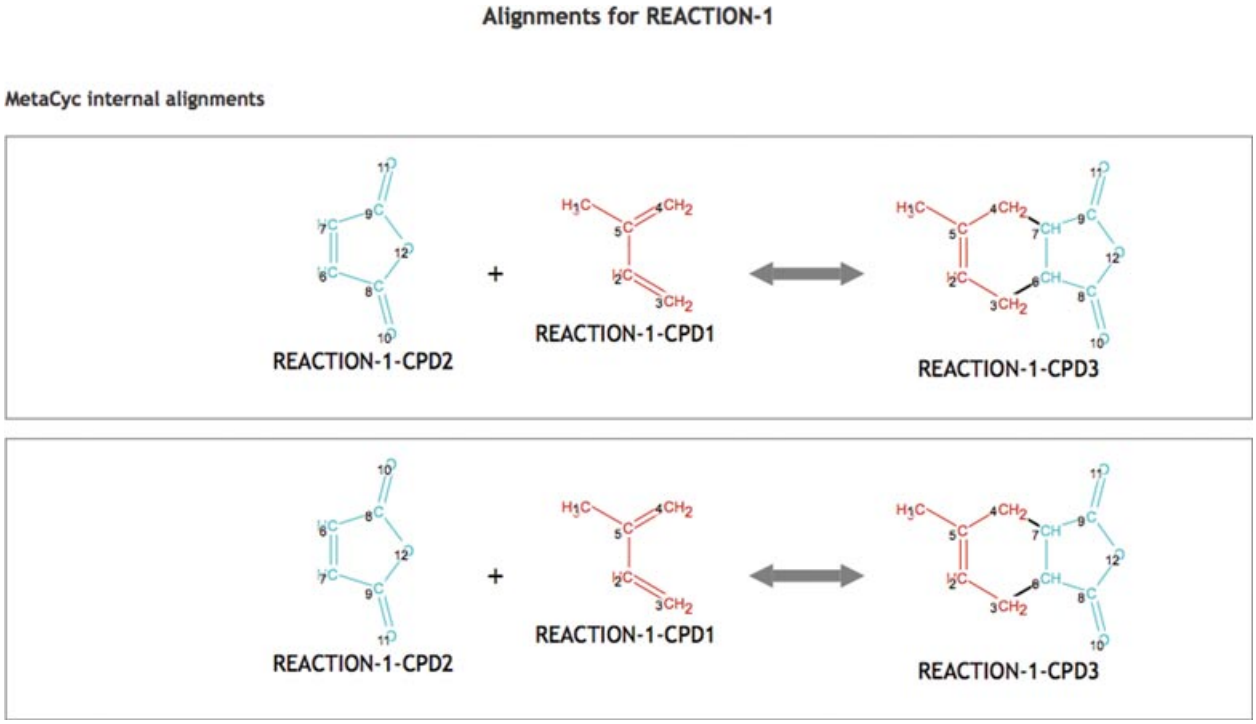

FIGURE 29 | Latendresse et al. reaction mapper produced two reaction mappings for the Diels–Alder reaction shown in Figure 19.[78] Latendresse et al. program did not mark the bond order changes. Their program marked the broken and formed bonds in black. (Reproduced with permission from Ref 76.)

Latendresse et al. algorithm produced two mappings in 0.24 second (see Figure 29).[78]

A nice feature of the MILP-based methods is that an MILP formulation is a general description of an optimization problem that can be solved by multiple MILP solvers.[79,80] As mentioned previously, the Körner–Apostolakis algorithm failed to complete mapping for RXN00048 (see Figure 24) after 24 h of CPU time.[52] Latendresse et al. applied the computation to the atom mapping on reaction RXN00048. Version 2.1.0 of SCIP took over 3 h, but version 3.0.0 is much faster and was able to solve it in 13 seconds. The IBM CPLEX solved the same problem in only 0.69 second.[81] As an interesting comparison, Accelrys' MCS-based reaction mapper (available in Pipeline Pilot[82]) and Automapper[33] took 0.85 and 0.033 seconds, respectively, to find a solution for RXN00048 on a Dell Precision M6500 laptop with Intel Core ™ i7 CPU Q720 @1.60 GHz 1.60 GHz, 8 GB RAM, 64-bit Windows 7 Ultimate Edition.

It should be noted that like many other optimization-based reaction mapping algorithms, Latendresse et al. algorithm can only deal with fully balanced reactions.

© 2013 John Wiley & Sons, Ltd.

CONCLUSIONS

From building and searching reaction databases to studying biochemical reaction mechanisms, identifying reaction AAM and reaction centers is a fundamental task in many applications. Reaction mapping in general is NP-hard. Owing to the complexity of reaction process and the limitation of chemical reaction representation (unbalanced reactions, overall reactions that involve multiple steps), reaction mapping still remains a challenging problem in cheminformatics.

The traditional reaction mappers usually employ either the EC- or MCS-based algorithms mainly pioneered by Vleduts,[22] Lynch and Willett.[17] They were mainly designed for building and searching large chemical reaction databases and thus required to be fast and have the capability to handle stereochemistry and unbalanced reactions. These mappers usually produce only one mapping per reaction.

Recent research work in this area was mainly driven by the need to study and handle reaction mechanism in biochemical reactions. The new focus has been placed more on accuracy and finding multiple optimal solutions. The major new developments include: (1) improving the existing MCS-based reaction mapping algorithms, (2) converting the molecular graph into an edge graph so that the BBM can be

TABLE 1 | Summary of Reaction Mapping Methods

Reaction mapping methods	Fragment-assembly-based methods	Lynch et al. algorithm	Early research project
	Common substructure-based methods	Extended-connectivity (EC)-based methods	Lynch–Willett's EC-based algorithm: classic reaction-mapping algorithm
			Accelrys' EC-based algorithm: fast, comprehensive approach, support unbalanced, complex reactions, one of the most widely used tool in reaction retrieval system
		Maximum common substructure (MCS)-based methods	Vleduts' MCS-based algorithm: classic reaction-mapping algorithm
			McGregor–Willett's MCS-based algorithm: apply MCS to EC-based reaction site to get good performance
			Funatsu et al. MCS-based method: use EC to choose starting atoms pairs for MCS search
		Maximum common edge substructure (MCES)-based method	Körner and Apostolakis' algorithm: first MCES-based reaction-mapping algorithm; employs bond weight to guide the MCES search
	Optimization-based methods	Graph isomorphism-based methods	Akutsu algorithm: first graph isomorphism-based reaction-mapping algorithm; designed to deal with a special class of reactions.
			Crabtree–Mehta algorithm: generalization of Akutsu's algorithm with much large application scope
		A*-based algorithm	Heinone et al. algorithm: first A*-based reaction-mapping algorithm with heuristics to prune search space
		Integer Linear Optimization (ILO)-Based Methods	First et al. algorithm: first ILO-based reaction-mapping algorithm that rigorously treats stereochemistry
			Latendresse et al. algorithm: combination of First et al. algorithm with bond weight concept

© 2013 John Wiley & Sons, Ltd.

directly obtained via the MCS algorithm, (3) introducing novel mapping algorithms.

The reaction mapping methods reviewed in this article is summarized in Table 1.

The underlying algorithms in these new mapping methods vary significantly. Some employ graph isomorphism algorithms, some use A* search algorithms, and some reply on the MILP technique. However, most of them share one common feature: they are optimization based, and they share the same optimization goal: finding optimal mappings with the minimal number of bonds broken and formed. Some new algorithms have shown better accuracy than the existing ones. However, it should be kept in mind that such mathematically optimal mappings may not always deliver a chemically correct solution.

Another feature, or more precisely speaking, limitation, that is common to most newly reported reaction mapping algorithms is that they are designed only to handle fully balanced reactions. This makes the mapping problems simpler and increases the mapping accuracy.

One promising strategy for further improving the mapping accuracy is to incorporate chemical knowledge into the searching process. The current method is to encode chemical knowledge as bond weights. The weight values are usually manually created based on chemist's knowledge and tuned after some tests. Only a limited number of types of bonds are assigned weights. In the future, machine learning technology may be employed to help produce more general, and more accurate bond weight values.

With regard to performance, the EC-based algorithm may still provide the fastest mapping approach owing to the simplicity of the Morgan algorithm. Among the novel reaction mapping algorithms, Latendresse et al. MILP-based algorithm may be the fastest one. However, it should be pointed out that objective comparison of performance of different algorithms is not an easy task because the performance depends on several factors: the algorithm itself, the implementation detail, and the hardware running the test. Even for the same algorithm, finding only one mapping will be much faster than finding all possible optimal mappings. In the latter case, early elimination of equivalent mappings that result from the symmetry will lead to significantly improved performance.

Equivalent AAMs can be eliminated within a mapping algorithm or be done in a postprocessing step to classify the resulting mappings into equivalent classes. This allows to output only unique optimal mappings. However, this approach cannot help improve the performance. A technique that can truly eliminate the generation of equivalent AAMs is to break the symmetry of small molecules and terminal groups so that the symmetric atoms will be treated as asymmetric ones. The basic idea can also be extended to handling symmetric rings. Chemical knowledge encoded as bond weights may also eliminate some mathematically optimal but chemically incorrect AAMs.

The most important requirements for next-generation reaction mapping algorithms include:

1. high performance,
2. high accuracy,
3. capability to handle unbalanced reactions,
4. capability to find multiple optimal mappings,
5. capability to handle stereochemistry.

The idea of a hybrid method as explored by McGregor and Willett[39] may be borrowed to build a next-generation reaction mapping algorithm with both high performance and high accuracy.

Finally, with more and more new reaction mapping algorithms being proposed recently, it seems that it is a time to establish a unified, open-source test benchmark for objectively comparing and evaluating the accuracy and performance of different reaction mapping algorithms.

REFERENCES

1. Wiechert W. ^{13}C metabolic flux analysis. *Metab Eng* 2001, 3:19–206.
2. Santos LS, Knaack L, Metzger JO. Investigation of chemical reactions in solution using API-MS. *Int J Mass Spectrom* 2005, 246:84–104.
3. Hendrickson JB. A general protocol for systematic synthesis design. *Top Curr Chem* 1976, 62:49–172.
4. Hendrickson JB, Chen L. Reaction Classification. In: Rague Schleyer PV, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF, Schreiner PR, eds. *The Encyclopedia of Computational Chemistry*. Vol. 4. Chichester, UK: John Wiley & Sons; 1998, pp. 2381–2402.
5. Chen L, Nourse JG, Christie BD, Leland BA, Grier DL. Over 20 years of reaction access systems from MDL: a novel reaction substructure search algorithm. *J Chem Inf Comput Sci* 2002, 42:1296–1310.
6. Available at: http://en.wikipedia.org/wiki/Systems_biology (Accessed September 8, 2012).
7. Rantanen, A, Rousu, J, Jouhten, P, Zamboni, N, Maaheimo, H, Ukkonen, E. An analytic and

 © 2013 John Wiley & Sons, Ltd.

systematic framework for estimating metabolic flux ratios from ^{13}C tracer experiments. *BMC Bioinform* 2008, 9:266–285.

8. Rousu J, Rantanen A, Maaheimo H, Pitkanen E, Saarela K, Ukkonen E. A method for estimating metabolic fluxes from incomplete isotopomer information. *Lect Notes Comput Sci* 2006, 2602:88–103.
9. Blum T, Kohlbacher O. Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J Comput Biol* 2008, 15:565–576.
10. Arita M. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA* 2004, 101:1543–1547.
11. Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J Am Chem Soc* 2004, 126:16487–16498.
12. Leber M, Egelhofer V, Schomburg I, Schomburg D. Automatic assignment of reaction operators to enzymatic reactions. *Bioinformatics* 2009, 25:3135–3142.
13. Funatsu K, Endo T, Kotera N, Sasaki SI. Automatic recognition of reaction site in organic chemical reactions. *Tetrahedron Comput Methodol* 1988, 1(1):53–69.
14. Available at: http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php (Accessed August 27, 2012).
15. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J Chem Inf Comput Sci* 1992, 32:244–255.
16. Chen L, Gasteiger J, Rose JR. Automatic extraction of chemical knowledge from organic reaction data: addition of carbon-hydrogen bonds to carbon-carbon double bonds. *J Org Chem* 1995, 60:8002–8014.
17. Lynch MF, Willett P. The automatic detection of chemical reaction sites. *J Chem Inf Comput Sci* 1978, 18:154–159.
18. Chen WL. Chemoinformatics: past, present, and future. *J Chem Inf Model* 2006, 46:2230–2255.
19. Avaialble at: http://www.genome.jp/kegg/ligand.html (Accessed on July 15, 2012).
20. Weygand C. *Organische-Chemische Experimentierkunst*. Vol. 1–3. Barth, Leipzig: Thieme; 1938.
21. Theilheimer W. *Synthetic Methods of Organic Chemistry*. Vol. 1. Basel, Switzerland: Karger; 1946.
22. Vleduts GE. Concerning one system of classification and codification of organic reactions. *Inf Storage Retr* 1963, 1:117–146.
23. Harrison JM, Lynch MF. Computer analysis of chemical reactions for storage and retrieval. *J Chem Soc C* 1970, 15:2082–2087.
24. Lynch MF, Willett P. The production of machine-readable descriptions of chemical reactions using Wiswesser line notations. *J Chem Inf Comput Sci* 1978, 18:149–154.
25. Chen L. Substructure and maximal common substructure searching. In: Bultinck P, Winter HD, Langenaeker W, Tollenaere JP, eds. *Computational Medicinal Chemistry and Drug Discovery*. New York: Marcel Dekker, Inc.; 2004, pp. 483–513.
26. Available at: http://en.wikipedia.org/wiki/NP-complete (Accessed August 27, 2012).
27. Ehrlich HC, Rarey M. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *WIREs Comput Mol Sci* 2011, 1:68–79.
28. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem Doc* 1965, 5:107–112.
29. Figueras J. Morgan revisited. *J Chem Inf Comput Sci* 1993, 33:717–718, and the references therein.
30. Chen L, Robien W. MCSS: a new algorithm for perception of maximal common substructures and its application to NMR spectral studies. 1. The algorithm. *J Chem Inf Comput Sci* 1992, 32:501–506.
31. Vleduts GE (British Library Research and Development Department, London, UK). Development of a Combined WLN/CTR Multilevel Approach to the Algorithmic Analyses of Chemical Reactions in View of Their Automatic Indexing. 1977. Report No. 5399.
32. Wipke WT, Dill JD, Peacock S, Hounshell D. *Search and Retrieval Using an Automated Molecular Access System*. Presented at the 182nd National Meeting of the American Chemical Society. New York; 1981.
33. Moock TE, Nourse JG, Grier D, Hounshell WD. The implementation of AAM and related reaction features in the reaction access system (REACCS). In: Warr WA, ed. *Chemical Structures*. Berlin, Germany: Springer-Verlag; 1988, pp. 303–313.
34. Sasaki T, Ishibashi Y, Ohno M. Catalysed cycloaddition reactions of α-silyloxy-α,β-unsaturated ketone and aldehyde. *Tetrahedron Lett* 1982, 23:1693–1696.
35. Yoneda F, Koga R, Higuchi M. A one-step synthesis of purine derivatives by the reaction of phenylazomalonamidamidine with aryl aldehydes. *Chem Lett* 1982, 3:365–368.
36. Avalaible at: http://www.fiz-chemie.de (Accessed August 27, 2012).
37. Smith EG. *The Wiswesser Line-Formula Chemical Notation*. New York: McGraw-Hill; 1968.
38. Wiswesser WJ. Historic development of chemical notations. *J Chem Inf Comput Sci* 1985, 25:258–263.
39. McGregor JJ, Willett P. Use of a maximal common subgraph algorithm in the automatic identification of the ostensible bond changes occurring in chemical reactions. *J Chem Inf Comput Sci* 1981, 21:137–140.

 © 2013 John Wiley & Sons, Ltd.

40. McGregor JJ. Backtrack search algorithms and the maximal common subgraph problem. *Softw Prac Exp* 1982, 12:23–34.
41. Available at: http://en.wikipedia.org/wiki/Backtracking (Accessed August 27, 2012).
42. Available at: http://en.wikipedia.org/wiki/Favorskii_rearrangement (Accessed August 27, 2012).
43. Loftfield RB. The alkaline rearrangment of alpha-haloketones. II. The mechanism of the Favorskii reaction. *J Am Chem Soc* 1951, 73:4707–4714.
44. Körner R, Apostolakis J. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J Chem Inf Model* 2008, 48:1181–1189.
45. Available at: http://en.wikipedia.org/wiki/Line_graph (Accessed August 27, 2012).
46. Whitney H. Congruent graphs and the connectivity of graphs. *Am J Math* 1932, 54:150–168.
47. Raymond J, Gardiner E, Willett P. RASCAL: calculation of graph similarity using maximum common edge subgraphs. *Comput J* 2002, 45:631–644.
48. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004, 32:D277–D280.
49. Reitz M, Sacher O, Tarkhov A, Trümbach D, Gasteiger J. Enabling the exploration of biochemical pathways. *Org Biomol Chem* 2004, 2:3226–3237.
50. Available at: http://www.chem.qmul.ac.uk/iubmb/enzyme/EC5/4/99/6.html (Accessed August 20, 2012).
51. Knaggs AR. The biosynthesis of shikimate metabolites. *Nat Prod Rep* 2001, 18:334–355.
52. Apostolakis J, Sacher O, Körner R, Gasteiger J. Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *J Chem Inf Model* 2008, 48:1190–1198.
53. Garey MR, Johnson DS. *Computers and Intractability—A Guide to the Theory of NP-completeness*. New York: Freeman; 1999.
54. Jochum C, Gasteiger J, Ugi I. The principle of minimal chemical distance (PMCD). *Angew Chem Int Ed Engl* 1980, 19:495–505.
55. Akutsu T. Efficient extraction of mapping rules of atoms from enzymatic reaction data. *J Comoput Biol* 2004, 11:449–462.
56. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res* 2002, 30:402–404.
57. Crabtree JD, Mehta DP. Automated reaction mapping. *ACM J Exp Algorithmics* 2009, 13:1–14.
58. Crabtree JD, Mehta DP, Kouri TM. An open-source java platform for automated reaction mapping. *J Chem Inf Model* 2010, 50:1751–1756.
59. McKay B. Practical graph isomorphism. *Congr Numer* 1981, 30:45–87.
60. Faulon JL, Collins MJ, Carr RD. The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *J Chem Inf Model* 2004, 44:427–436.
61. Heinonen M, Lappalainen S, Mielikainen T, Rousu J. Computing atom mappings for biochemical reactions without subgraph isomorphism. *J Comput Biol* 2011, 18:43–58.
62. Hart PE, Nilsson NJ, Raphael B. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans Syst Sci Cybernet* 1968, 4:100–107.
63. Dechter R, Pearl J. Generalized best-first search strategies and the optimality of A*. *J ACM* 1985, 32:505–536.
64. Available at: http://en.wikipedia.org/wiki/A*_algorithm (Accessed September 3, 2012).
65. Hattori M, Okuno Y, Goto S, Kanehisa M. Heuristics for chemical compound matching. *Genome Inform* 2003, 14:1–153.
66. Glover JR, Chapple CCS, Rothwell S, Tober I, Ellis BE. Allylglucosinolate biosynthesis in *Brassica carinata*. *Phytochemistry* 1988, 27:1345–1348.
67. Cordella LP, Foggia P, Sansone C, Vento M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans Pattern Anal Mach Intell* 2004, 26:1367–1372.
68. Cordella LP, Foggia P, Sansone C, Vento M. Performance evaluation of the VF graph matching algorithm. *Proc ICIAP* 1999, 99, 1172.
69. First EL, Gounaris CE, Floudas CA. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *J Chem Inf Model* 2012, 52:84–92.
70. Available at: http://en.wikipedia.org/wiki/Linear_programming (Accessed August 27, 2012).
71. Available at: http://en.wikipedia.org/wiki/NP-hard (Accessed August 27, 2012).
72. Padberg M. *Linear Optimization and Extensions*. 2nd Ed. Berlin, Germany: Springer-Verlag; 1999.
73. Beasley JE, ed. *Advances in Linear and Integer Programming*. Oxford, UK: Oxford Science; 1996.
74. Available at: http://selene.princeton.edu/dream (Accessed August 27, 2012).
75. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988, 28:31–36.
76. Latendresse M, Malerich JP, Travers M, Karp PD. Accurate atom-mapping computation for biochemical reactions. *J Chem Inf Model* 2012, 52:2970–2982.
77. Latendresse M, Personal Communication with William L. Chen, September 6, 2012.

 © 2013 John Wiley & Sons, Ltd.

78. Latendresse M, Personal Communication with William L. Chen, August 31, 2012.
79. Achterberg T. SCIP: solving constraint integer programs. *Math Program Comput* 2009, 1:1–41, SCIP Web Site.
80. IBM ILOG CPLEX. CPLEX: high-performance software for mathematical programming and optimization; 2012. See http://www.ilog.com/products/cplex/.
81. Latendresse M, Personal Communication with William L. Chen, September 10, 2012.
82. Available at: http://accelrys.com/products/pipeline-pilot/ (Accessed September 12, 2012.

FURTHER READING

Chen Lingran. Reaction classification and knowledge acquisition. In: Gasteiger J., ed. *Handbook of Chemoinformatics*. Vol. 1. Weinheim, Germany: Wiley-VCH; 2003, pp 348–388.

Peter Willett, ed. *Modern Approaches to Chemical Reaction Searching*. Proceedings of a Conferebnce organized by the Chemical Structure Association at the University of York, England, 8–11 July 1985. Aldershot, England: Gower Publishing Company Limited; 1986. The proceedings summarize the early development of chemical reaction indexing and searching.

 © 2013 John Wiley & Sons, Ltd.