Language-oriented rule-based reaction network generation and analysis: Description of RING

Srinivas Rangarajan, Aditya Bhan*, Prodromos Daoutidis**

Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

ABSTRACT

The input and output formats, and the structure of Rule Input Network Generator (RING), a computational tool for generation and analysis of complex reaction networks, are described with reference to the underlying algorithms from Cheminformatics and graph theory. RING consists of three modules: (a) a compiler that translates inputs written as a program in an English-like reaction language into internal representations and instructions, (b) a network generator that constructs an exhaustive reaction network from reaction rules and initial reactants specified, and (c) a post-processing module that can extract pathways, mechanisms, or lumps from the network based on user-specified instructions. RING can be used, in a rule-based manner, for constructing a large and complex reaction network from a set of elementary/overall reaction rules, and for elucidating transformations occurring in these networks through identifying pathways and mechanisms to specified products. RING is available open under GNU Lesser GPL.

1. Introduction

Complex reaction networks are at the core of numerous chemical and biochemical systems (Broadbelt & Pfaendtner, 2005); examples include pyrolysis (Froment, Van de Steene, Van Damme, Narayanan, & Goossens, 1976; Mohan, Pittman, & Steele, 2006) and acid-catalyzed (Corma, Huber, Sauvanaud, & O'Connor, 2008; Yaluris, Madon, & Dumesic, 1997) conversion of hydrocarbons or oxygenates, and biochemical networks such as glycolysis in *Saccharomyces cerevisiae* (Rizzi, Baltes, Theobald, & Reuss, 1997) and genome-scale models of *Escherichia coli* (Reed, Vo, Schilling, & Palsson, 2003). Such networks are also extensively prevalent in industrial reaction systems, particularly in petroleum refining, as the crude oil feed is a mixture of many compounds and a multitude of chemical transformations can potentially take place (Ho, 2008; Quann & Jaffe, 1996). Despite the large size, the reaction networks can be constructed from a relatively smaller set of chemical transformation rules. Automated network generators have been developed to construct such networks from initial reactants and pre-specified reaction rules.

Kinetic modeling is a common application of automated network generation, wherein the appropriate differential algebraic system of equations that captures the dynamics of the system is formulated. The model is then solved with kinetic parameters estimated, predicted, or specified, to obtain product yield information (Cortright & Dumesic, 2001). Network generation in conjunction with kinetic modeling has been extensively applied for hydrocarbon (Broadbelt, Stark, & Klein, 1994a; Van Geem et al., 2006) and biochemical systems (Faeder et al., 2003). Complex reaction networks, however, have also been analyzed for: (a) deriving topological properties such as average path length of the network (Arita, 2003, 2004), (b) identifying synthetic/degradation pathways (Gonzalez-Lergier, Broadbelt, & Hatzimanikatis, 2005), and (c) deriving and testing plausible mechanisms and overall rate expressions (Fan, Bertok, & Friedler, 2002; Fishtik, Callaghan, & Datta, 2004; Lin, Fan, Shafie, Bertok, & Friedler, 2009). The use of additional thermodynamic data in conjunction with the reaction network has further enabled quantitative analysis of networks in terms of: (a) generating thermodynamically meaningful flux distributions in biochemical systems (Henry, Broadbelt, & Hatzimanikatis, 2007), (b) extracting functional information such as regulatory sites in biological systems (Kummel, Panke, & Heinemann, 2006), and (c) identifying thermodynamically feasible synthesis routes (Li et al., 2004) to form chemicals, or biological degradation pathways (Finley, Broadbelt, & Hatzimanikatis, 2009) to decompose molecules.

All these examples require a constructed network as the first step and graph theoretical analysis of the network in subsequent steps; hence, an automated network generator can be used to construct a complex network prior to network analysis. In this context, we have developed a new computational tool, Rule Input Network

* Corresponding author.
** Corresponding author. Tel.: +1 612 625 8818; fax: +1 612 626 7246.
E-mail addresses: abhan@umn.edu (A. Bhan), daout001@umn.edu (P. Daoutidis).

Generator, or RING (Rangarajan, Bhan, & Daoutidis, 2010), for network generation and analysis of complex thermochemical reaction systems. RING takes initial reactants, chemistry rules and network analysis instructions as inputs written in a domain specific English-like reaction language. RING then generates an exhaustive network of reactions and analyzes the network in a post-processing step to obtain: (a) lumps of isomeric chemical species in the network, (b) reaction pathways from reactants to products, and (c) mechanisms that describe the overall chemical transformation of the initial reactants to specified products. RING is available open source (RING, 2011) under GNU Lesser GPL.

In our previous work (Rangarajan et al., 2010), the reaction generation feature in RING was introduced briefly and its application to biomass conversion was illustrated through four relevant reaction systems as examples. We have subsequently developed a domain specific English-like extensible reaction language as a user interface and a post-processing module, as discussed above. In this article, the first of a two-part series, we discuss all features of RING with illustrative examples. Applications of RING in complex network analysis of catalytic systems are discussed in the second part.

2. Network generation and analysis: a review

2.1. Network generators

Rule-based automated reaction network generators are computational tools that take in a set of molecules as reactants, and iteratively apply a set of reaction rules, to construct a comprehensive list of possible reactions. Network generators have been developed and applied in different fields such as pyrolysis and oxidation, catalysis, and biological systems. Table 1 lists and describes several of them. All automated network generators have five common and essential features (Tomlin et al., 1997, chap. 4). First, an unambiguous representation is required for molecules and reactions. This is usually represented as character strings for input and output (Broadbelt et al., 1994a; Prickett & Mavrovouniotis, 1997a; Ratkiewicz & Truong, 2003). Second, an internal representation of molecules is required, such as molecular trees, adjacency matrices, or chemical graphs, thereby enabling quick structure manipulation. Adjacency matrices are the most common representation format owing to their simplicity. An adjacency matrix 'M' of a molecule is a square matrix containing connectivity and bond order information between every two atoms. Thus, $M(i,j)=0$ implies that the ith and jth atoms are not connected while a positive nonzero value would indicate the strength of the bond (1 is a single bond, 2 is a double bond, etc.). The diagonal values indicate the number of unpaired electrons in the atoms. Third, an internal representation of reaction rules that can be applied iteratively on the molecules is required. A common representation scheme is to employ a matrix for reaction rules. Proposed initially by Dugundji and Ugi (Dugundji & Ugi, 1973, also Ugi et al., 1993) and later used in other tools such as NETGEN (Broadbelt et al., 1994a), the reaction matrix describes the transformations for an adjacency matrix representation of molecules. The reaction matrix is structured similarly, with off-diagonal values representing bond changes and diagonal entries representing changes in unpaired electron count. Further, positive/negative values indicate increase/decrease in bond order, bond formation/cleavage, or increase/decrease in unpaired electrons of the atom. Baltanas and Froment (1985) used a Boolean matrix to represent molecules for generation of networks for modeling paraffin cracking and isomerization on bifunctional catalysts. The Boolean matrix is similar to the adjacency matrix; however, bonds of a higher order (e.g. double bonds) and information on charges (such as +1 for carbenium ions) are stored separately. This method, therefore, is similar to

that of Dugundji and Ugi (1973). Transformations in RDL (Prickett & Mavrovouniotis, 1997a) and RDL++ (Hsu et al., 2008), on the other hand, are input by the user as English-language-like statements describing changes in the charge/bonding of atoms participating in the reaction rule which get directly applied on the internal graph description of molecules. Fourth, all network generators have a generation scheme that iteratively applies the reaction rules to all input and generated molecules so that the resultant network is exhaustive. The scheme should ensure that all possible reactions of a given set of reactants are generated corresponding to that reaction rule. Faulon and Sault (2001) describe such a generation scheme as deterministic network generation.

Combinatorial explosion is an important practical problem that can significantly increase execution time and lead to a large proportion of unimportant or improbable reactions. The fifth essential feature of most network generation tools, therefore, is to employ a systematic procedure to curtail this effect. When kinetic parameters are available a priori, quantitative estimates of the magnitude of the reaction rates allows for the identification of 'important'/'unimportant' reactions and species that should be included in, or excluded from, the network. For example, the tool NETGEN adopts rate-based (Susnow et al., 1997) network pruning criteria. This requires generation and kinetic modeling in concurrence because the rates calculated on-the-fly are used to determine if a particular species will react further. In the absence of such kinetic information, either topological or experts-based constraints can be provided. For example, species rank-based criteria (Broadbelt et al., 1995) network pruning criteria prevents reactions that involve species of ranks greater than a specified value, while the tools RDL (Prickett & Mavrovouniotis, 1997a) and RDL++ (Hsu et al., 2008) allow for the specification of constraints that molecules should satisfy to undergo a particular transformation. Faulon and Sault (2001) propose stochastic (or sampling) network generation algorithms, in contrast to the deterministic scheme, for concurrent generation and reduction of networks. These algorithms scale in polynomial time but require on-the-fly estimation of rate constants which is achieved, in their case, through quantitative structure property relationships for free-radical chemistries.

2.2. Topological network analysis

The transformations occurring on a compound to form various products can be obtained by finding: (a) a sequence of reactions that describe how the reactant gets converted into an intermediate, which could further react, and ultimately form the final product, and (b) a set of reactions that when taken together leads to an overall reaction without the net generation/consumption of any unstable reactive intermediate. The former represents "a pathway" from the initial reactant to the final product while the latter is a "mechanism" with a balanced overall stoichiometry. A *direct* mechanism is the minimal set of reactions that has zero net consumption/formation of reactive intermediates (Millner, 1964); minimality ensures that no nonzero subset of this set is also a mechanism. Direct mechanisms, thus, are conceptually similar to reaction cycles. An *overall* (or complete) mechanism, further, is a set of direct mechanisms so that the reactants of the overall reaction are all initial reactants, and the desired target molecule is one of the products. Overall mechanisms, thus, are conceptually similar to a set of reaction cycles that operate simultaneously in a reaction system to convert the initial reactant to final products.

Identifying pathways between two species in a network has wide applications in biological network analysis, and algorithms with/without tracking of individual atoms have been proposed for both weighted and unweighted networks (Croes, Couche, Wodak, & van Helden, 2006; Heath, Bennett, & Kavraki, 2010). Given a list of reactions, algorithms have been proposed for identifying possible

Table 1
A list of reaction network generators, and a description of their essential features, and their areas of application.

Name	Description	Remarks	References
NETGEN	Network generator and model builder based on 'BE' & 'R' matrices. Uses adjacency matrix representation of molecules and reaction rules.	(i) Rate based and rank based pruning. (ii) Application in gas phase pyrolysis, nanoparticle synthesis, and biochemical reactions. (iii) Linked to MOPAC2009 (2008) and NIST S&P database for thermochemistry.	Li et al. (2004), Broadbelt et al. (1994a), Broadbelt, Stark, and Klein (1994b, 1995, 1996), Susnow, Dean, Green, Peczak, and Broadbelt (1997), Wong, Li, Swihart, and Broadbelt (2004), Pfaendtner and Broadbelt (2008)
EXGAS	Kinetic model builder using a tree data structure for internal molecule representation.	(i) Applied in gas phase combustion and oxidation. (ii) A library of free radical chemistry rules used in reaction network generation. (iii) Tree representation system based on Chinnick, Baulch, and Asyscough (1988).	Heyberger et al. (2001), Warth et al. (2000)
COMGEN	Network generator based on chemical graph theory. String representation of molecules, reactant pattern based on Blurock (1995), and topological indices for molecule identification.	(i) Hydrocarbon gas phase chemistry. (ii) Thermochemistry was calculated from a database.	Ratkiewicz and Truong (2003)
RMG	Kinetic models of free radical chemistry of hydrocarbons. Kinetics estimated from semi-empirical relations obtained from theoretical calculations.	(i) Applications in hydrocarbon pyrolysis. (ii) Accurate calculations of kinetics and thermodynamics, and formulation of kinetic models to predict product yields and conversion.	Song (2004), Van Geem et al. (2006), Green et al. (2004)
RDL	English-like language based description of reaction rules. Object-oriented framework using elements of graph theory.	(i) Reaction network generated from scratch depending upon reaction rules input, thus offering flexibility in describing the system. (ii) Constraints on rules to prevent combinatorial explosion.	Prickett and Mavrovouniotis (1997a, 1997b, 1997c)
RDL++	Extends RDL with additional features to enable description of solid-acid catalyzed reactions of hydrocarbons.	(i) Applied in microkinetic modeling of heterogeneous catalytic systems, and data analysis and knowledge extraction in high-throughput experimentation.	Hsu et al. (2008), Katare, Caruthers, Delgass, and Venkatasubramanian (2004), Bhan et al. (2005), Caruthers et al. (2003)
KING	An automated mechanism generator. Uses 'BE' and 'R' matrix for molecule and reaction representation.	(i) Applied in combustion chemistry. (ii) Reactions are determined combinatorially, as a linear combination of elementary steps.	Di Maio and Lignola (1992)
Bio-NETGEN	Rule-based generation of biological reaction network. Graph based representation of molecules with each node being a building block of the macromolecule of the biological system.	(i) Application in reaction network generation in biological systems and subsequent dynamic modeling.	Blinov, Yang, Faeder, and Hlavacek (2005), Blinov, Faeder, Yang, Goldstein, and Hlavacek (2005), Faeder, Blinov, Goldstein, and Hlavacek (2005)
BNICE	Computational framework for generating and analyzing biological reaction pathways.	(i) Reaction rules are obtained from the enzyme function information in KEGG database. (ii) Incorporation of group-contribution based thermochemistry estimation for flux analysis and pathways prediction.	Henry et al. (2007), Henry, Broadbelt, and Hatzimanikatis (2010), Li et al. (2004), Finley et al. (2009)
SynBioSS	Modeling and simulation tool for synthetic biological systems. Complete enumeration of sets of biomolecular reactions based on user input of molecular parts involved in gene expression and regulation.	(i) Multiscale simulation of the generated reaction network using stochastic algorithms.	Hill, Tomshine, Wedding, Sotirpolous, and Kaznessis (2008)
BioTRANS	Computational tool for predicting metabolism of chemicals in a mixture. Generated paths of different compounds are interconnected through common metabolites. ODE models solved to predict the time profiles of the each of the compounds.	(i) Application in modeling of biotransformations of VOCs that commonly pollute water.	Mayeno, Yang, and Reisfeld (2005)

direct mechanisms using either combinatorial (Mavrovouniotis & Stephanopoulos, 1992; Otarod & Happel, 1992) or graph theoretic methods (Fan et al., 2002; Lin et al., 2009).

A concept closely related to that of mechanisms is the linear-algebra-based flux analysis wherein all possible independent basis set of solution vectors that solve the stoichiometric steady-state mass balance equations are identified (Papin et al., 2004). Two commonly applied flux analysis techniques in systems biology are extreme pathways (Schilling, Letscher, & Palsson, 2000) and elementary flux modes (Schuster et al., 1994).

3. The structure of RING: an overview

RING builds on and extends the network generators listed in Table 1 in two major aspects. First, RING provides a generic scope
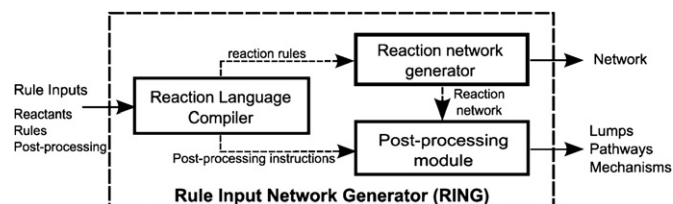

Fig. 1. The modular structure of RING. Dashed arrows show internal information flow, solid arrows refer to inputs and outputs.

by allowing for representing intermediates and describing reaction rules applicable to most organic chemical transformations. Second, RING allows for topological network analysis through identification of pathways and mechanisms in the generated network. Fig. 1

shows the input–output structure of RING as well as the three internal modules – a language compiler, a reaction network generator, and a post-processing module.

3.1. Reaction language compiler

The compiler acts as an interface between the user and network generation and post-processing components of RING by translating the inputs into relevant instructions used internally by the network generator. The inputs into RING are written as a program in an English-like reaction language and include the initial reactants, reaction rules, and set of post-processing instructions.

The utility of a language interface in network generation was demonstrated in RDL (Prickett & Mavrovouniotis, 1997a) and RDL++ (Hsu et al., 2008). The syntax of the language closely resembles chemistry parlance so that debugging a program written in the language essentially involves proofreading. RING's language interface has been developed using SILVER, an attribute grammar specification language (Van Wyk, Bodin, Gao, & Krishnan, 2010). In addition to providing an interface like in RDL++, the language, as discussed in detail in Rangarajan, Bhan, and Daoutidis (in press), can: (a) catch chemistry-specific inconsistencies as well as syntactical errors, (b) improve speed of execution through domain-specific optimization of input instructions, and (c) allow for independent extensions in terms of syntax and semantics to the original grammar.

3.2. Reaction network generator

The information on initial reactants and the translated reaction rules is used by the network generator to construct the reaction network by iterative application of the rules upon the initial reactants and the products generated thereof. The output from the network generator is a list of all possible species and reactions consistent with the rules specified. Several Cheminformatics algorithms to manipulate the chemical information of a molecule, extract topological information of molecules, and generate reactions have been adopted in RING. Further, standard Cheminformatics formats to input, represent, store, and retrieve the chemical information pertaining to molecules, such as elemental composition, valency, charge, bonding, and electronic information have also been used.

3.3. Post-processing module

Specific information regarding the constituents of/transformations within the network can be "mined" from the generated reaction network using the several post-processing options available in RING. For example, the generated reaction network may be large enough to preclude manual analysis to obtain specific information such as whether and how a particular molecule (or a class of molecules) is (are) formed. Such queries can be input in the form of instructions, or "rules", specifying the analysis that is sought.

3.3.1. Lumping
Complex reaction networks can be composed of several species that are structural isomers of each other. If these similar species are grouped together into lumps, the total network size can be reduced significantly. For example, Hsu et al. (2008) showed for acid-catalyzed aromatization of propane that the number of reactions without lumping structural isomers was more than a million while structural lumping resulted in only 605 distinct reactions. Analytical identification of lumps of molecules, instead of individual species, is a more feasible option when the number of species in a system is large (Quann & Jaffe, 1996) because of experimental difficulties in distinguishing between isomers. Furthermore,

lumping can lead to a reduced network that can be used to formulate a simplified lower-dimensional mathematical model that is easier to solve. Lumping strategies have been proposed in hydrocarbon processing wherein, a lump is represented either by a single likely structure and reactions are written on that basis (Quann & Jaffe, 1992), or as a mixture with an estimated internal distribution and an overall reaction computed on the basis of composition and kinetics information (Ranzi, Dente, Goldaniga, Bozzano, & Faravelli, 2001). In contrast to these methods, lumping methods for dimension reduction in kinetic models through mathematical transformations have also been proposed (Kuo & Wei, 1969; Li & Rabitz, 1989, a brief review in Ho, 2008).

Lumping instructions, as an input feature is not available in the tools listed in Table 1, though RDL++ (Hsu et al., 2008) can group together hydrocarbons in a post-processing step. In RING, however, a generic lumping scheme, that identifies functional equivalence between isomers, has been implemented. In general, two molecules can be lumped only when there is a one-to-one mapping between their atoms; mapping in this case exists when the two atoms belong to the same functional group. Such functional equivalence cannot be inferred from molecular formula and structure alone for non-hydrocarbon molecules such as oxygenates. We have, therefore, developed a new lumping algorithm that takes into consideration the environment of each atom in a molecule to decide the lump it belongs to. For example, ethanol (CH_3CH_2OH) and dimethyl ether (CH_3OCH_3) both have the same molecular formula; however, they have different functional groups. Ethanol has a C–OH group that the ether does not possess; hence, there is no one-to-one mapping between the oxygen atom of ethanol and the ether. Therefore, the two molecules are not lumped. On the other hand, 2-pentanol and 3-pentanol are both functionally equivalent (S2 in Supplementary material) are lumped together as secondary five-carbon alcohols. 1-Pentanol, being a primary alcohol, will naturally not be a constituent of this lump.

3.3.2. Pathway identification
RING can identify pathways to user-specified molecules from initial reactants that satisfy user-specified constraints on the nature of pathways (maximum length, presence/absence of reactions in the pathway, etc.). A reverse depth-first network traversal algorithm (Rangarajan et al., in press) is used to identify all possible pathways exhaustively.

3.3.3. Mechanism identification
In RING, all overall mechanisms from the initial reactants to user-specified products can be identified. Further, all direct mechanisms forming any specified molecule can also be found. The algorithms of Mavrovouniotis and Stephanopoulos (1992) and Fan et al. (2002) identify all possible direct mechanisms from a given list of reactions. Since mechanisms are only sought for specified products, these two algorithms cannot be directly used within RING. An adapted version of the pathways algorithm is instead used that seeks only the relevant direct or overall mechanisms.

4. RING: inputs and outputs

Inputs into RING are: (a) initial reactants, (b) global constraints, (c) reaction rules, and (d) post-processing instructions describing the desired analyses of the network. The outputs from RING are lists of: (a) species, (b) reactions, (c) lumps and their constituents, (d) pathways to specified products, (e) mechanisms to specified products, and (f) species or reactions extracted from the network from a molecule or reaction query. In this section, the inputs into and outputs from RING are discussed in the context of an example – dehydration of fructose to form HMF. The system is typically carried out in aqueous phase and is catalyzed by mineral acids (Antal, Mok,

Table 2
Sample input code for Fructose-to-HMF system.

Reaction rules	Post-processing
1. input reactant "O1C(CO)(O)C(O)C(O)C1(CO)"	27. lump all isomers{
2. input reactant "[H+]"	28. represent acyclic with farthest apart
3. //Defining global constraints	29. represent cyclic with farthest apart}
4. global constraints on Molecule f	30. find pathways to mol{
5. fragment a{	31. mol is "OCc1oc(C=O)cc1"
6. C+ labeled 1	32. } constraints {
7. any atom labeled 2 double bond to 1}	33. maximum length 11
8. ! Molecule contains a	34. contains <= 2 rule Hshift
9. Molecule.size < 15 && Molecule.charge < 2	35. eliminate similar pathways
10. fragment b {	36. } store in "HMFPathways.txt"
11. C labeled c1	37. find complete mechanisms to mol{
12. C labeled c2 double bond to c1	38. mol is "OCc1oc(C=O)cc1"} overall constraints{
13. X labeled x1 double bond to c2}	39. maximum length 12
14. ! Molecule contains b}	40. maximum cycles 4} cycles constraints {
15. //alcohol protonation rule	41. maximum length 4
16. rule alcoholprot{	42. contains <= 2 rule Hshift
17. neutral reactant r1{	43. eliminate similar mechanisms} store in "HMFMechanisms.txt"
18. C labeled c1 {! connected to > 1 O with any bond}	44. find mol{
19. nonringatom O labeled o1 single bond to c1}	45. mol is cyclic && mol.size > 5 && mol is neutral
20. positive reactant proton{	46. } store in "CyclicNeutralMolecules.txt"
21. H+ labeled h1}	47. find reactions{
22. constraints {r1.size < 15 && r1.size > 2}	48. rule is only alcoholprot
23. form bond (o1, h1)	49. reaction with 1 reactant mol{mol is cyclic}
24. modify atomtype (o1, O+)	50. } store in "CyclicAlcProtRxns.txt"
25. modify atomtype (h1, H)}	
26. //More reaction rules here	

& Richards, 1990). An illustrative program for this system is shown in Table 2.

4.1. Initial reactants

Initial reactants are represented as SMILES-like strings (Weininger, 1988). Lines 1 and 2 in Table 2 describe the reactants for the Fructose-to-HMF system (Fructofuranose and proton to represent the acid), written in SMILES. A SMILES parser has been developed for RING which recognizes the elements C, H, N, O, S, and P. On many occasions, inorganic atoms such as Al, Cu, Pt, etc. are involved in chemical transformations, for example, as active sites (catalysts, ligands) or as reagents (electrophiles, nucleophiles, etc.) in substitution reactions, and have to be represented and manipulated during reaction generation. To this end, a new feature, *composite atoms*, has been added to represent a miscellaneous group of atoms as a single atom for convenience in representation and manipulation. A composite atom, within RING, is a user-defined entity used to represent: (a) an atom other than the recognized elements or (b) a group of atoms considered together as a "unified atom". These atoms are represented as alphanumeric strings enclosed within curly braces, for example, Platinum atoms can be represented as '{Pt}'. The name given to a composite atom is arbitrary and does not have to be an element name. For example, two different types of active sites (AS) on a catalyst can be represented as {AS1} and {AS2}. These atoms have to be defined using "define composite atom" statement. Once defined, RING recognizes the composite atom during reaction rule specification and reaction
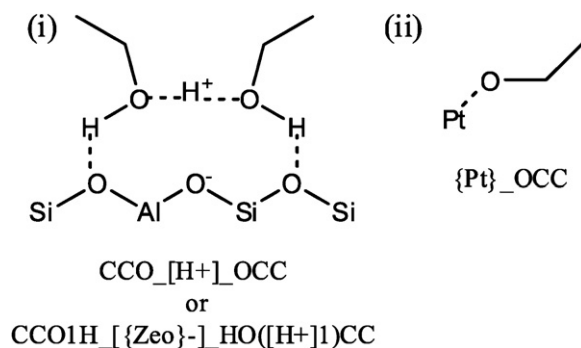

Fig. 2. SMILES strings with composite atoms and non-bonded interactions: (i) dimeric intermediate in DEE synthesis from ethanol on proton form zeolite catalysts (Chiang & Bhan, 2010); (ii) surface species formed upon dissociative adsorption of ethanol.

generation. Examples employing composite atoms are discussed in the second part of this series (Rangarajan et al., in press). A second feature that has been added into SMILES in RING is the definition of non-bonded interactions – partial and hydrogen bond. This interaction is represented using "_" and can be used to depict, for example, dimeric alcohol intermediates in zeolites (Chiang & Bhan, 2010) co-adsorbed on a zeolite, or adsorption of molecules on a metal surface as shown in Fig. 2. The use of SMILES, with these two additional features, allow for rigorous description of the intermediates involved in different types of chemistries.

4.2. Reaction rules

These describe the chemical transformations that take place in the reaction system (Rangarajan et al., 2010) and comprise of: (i) the set of atoms that participate in the reaction, called the reaction center or reactant patterns, (ii) the final electronic configuration of the atoms in the reaction center and the bonding between them, and (iii) constraints that govern the reaction center and the entire molecule. These constraints have to be satisfied by the reactant molecule prior to the application of the reaction rules. In addition, constraints can also be imposed on the products formed and are classified under molecular constraints.

Each reaction rule defines a particular chemical transformation which can be elementary or non-elementary, and be unimolecular or bimolecular. For a given reaction system, the reaction rules can be deduced from the literature and from knowledge of the operating conditions. For example, gas phase pyrolysis of hydrocarbons or oxygenates typically involves homolytic bond cleavage/formation, β-scission, hydride shifts, etc., while heterogeneous solid acid catalysis would involve adsorption/desorption of hydrocarbons, β-scission, alkylation, etc. RING allows for the specification of an additional reactant identical to one of the reactant(s). This feature has been implemented mainly to address elementary steps on surfaces where two adjacent identical surface species (or sites) are involved, such as desorption of surface hydrogen species on palladium to form molecular hydrogen (Conrad, Ertl, & Latta, 1974). This duplication allows for an exception whereby termolecular reactions can de described when the third reactant pattern is identical to one of the other two, such as in the case of dissociative adsorption of molecular hydrogen (the reverse of the desorption step) shown in Fig. 3. Lines 16–25 in Table 2 describe the reaction rule for protonation of a hydroxy group, shown in Fig. 4. The rule essentially describes the abstraction of a proton by the oxygen atom of an alcohol to form an oxonium ion.
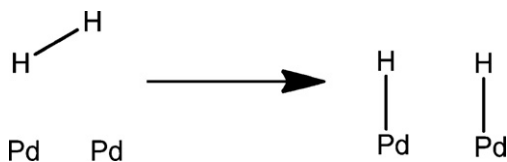
Fig. 3. Dissociative adsorption of molecular hydrogen on metal sites (Pd).
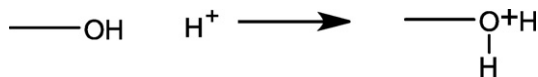

Fig. 4. Sample input reaction rule in RING – elementary step involving protonation of an alcohol.

4.2.1. Reactant patterns

Reaction centers (patterns) are written as shown in lines 17–21 in Table 2. The hydroxy pattern is described in lines 17–19 – a carbon atom, labeled 'c1', that is single bonded to an oxygen atom not in a ring, labeled 'o1'. Lines 20–21 describe a proton – a positive hydrogen atom labeled 'h1'. All atoms in a reactant pattern are labeled for unambiguous identification while describing transformations. Each reactant pattern describes one of the reactants. Rules can be unimolecular (single pattern), bimolecular (two patterns), ternary (consisting of two identical and a second distinct pattern), and intramolecular (special case of bimolecular rules). Therefore, bimolecular reactions have two reactant patterns which together constitute the reaction center. Ternary reaction rules can also be specified when two of the three patterns are identical. Bimolecular reaction rules can be set to also/only allow intramolecular reactions. In such a case, an intramolecular reaction will occur only if both the patterns exist in a single reactant.

4.2.2. Constraints specification

Constraints, as discussed in Section 2, are essential in preventing a combinatorial explosion in the size of the generated network. In RING, two types of constraints can be specified: molecular and atom environment constraints. The atom environment constraints, given along with reactant pattern specification, describe the spatial properties of the atoms in the reaction center. Specifically, they describe: (a) what the neighboring atoms and fragments are, and (b) topological characteristics of the atoms – ring atom, aromatic atom or allylic atom, largest size of the ring in which atom is present, etc. The molecular constraints are constraints on (a) charge, (b) size, (c) presence (or absence) of structural features, and (d) shape and presence (or absence) of topological characteristics. Atom constraints are specified as shown in line 18 of Table 2, wherein, the constraint restricts the carbon 'c1' to be connected to one or fewer oxygen atoms. This ensures that the carbon is not part of an ester or an acid functional group.

Molecular constraints are specified as shown in line 22 of Table 2. The carbonyl species has an upper and lower bound on the size (defined as the number of non-hydrogen atoms in it). Lines 17 and 20 also impose charge constraints on the reactants by using the prefixes "neutral" and "positive". Combined constraints involving both reactants and constraints on products can also be specified. This constraint specification scheme, inclusive of reactant and product molecular constraints and reaction center atom constraints, thus, provides the user with the capability of imposing restrictions to varying levels of severity.

4.2.3. Transformations

A reaction rule is complete when the final configuration of atoms and bonds of the reaction center is specified. Transformations in RING, similar to those in RDL (Prickett & Mavrovouniotis, 1997a) and RDL++ (Hsu et al., 2008) involve either changes in connectivity such as bond formation/cleavage and modification of the bond order, or changes in electronic configuration such as adding or removing charge, lone pairs, and unpaired electrons. The changes in electronic configuration are described by atom-type modifications – an atom of one atomtype changes to another atomtype (e.g., C forming C+). Lines 23–25 of Table 2 describe the transformations for the case of alcohol protonation: (a) formation of a single bond between oxygen atom, 'o1', and hydrogen atom, 'h1', (b) formation of an oxonium ion, and (c) loss of charge on the proton to form neutral hydrogen. The description of changes in the bond connectivity and electronic configurations, along with reactant pattern description, provides complete information on the initial and final state of each atom in the reaction center. This allows for: (a) catching incorrect rules that violate valency constraints of atoms directly by the language compiler prior to network generation, and (b) correctly identifying changes in topological characteristics of molecules such as in steps that break aromaticity. Such a description of transformations is independent of the type of chemistry and hence applicable to most organic chemical transformations – homogeneous/heterogeneous catalytic/non-catalytic chemistries.

The Fructose-to-HMF system has several additional rules: protonation of carbonyl and C=C groups (and their reverse steps), hydride shift, allyl shift, and dehydration of oxonium ion. S1 in Supporting material contains the corresponding reaction language code for these elementary steps. These rules, while only a subset of all acid-catalyzed elementary steps, form the likely steps for aqueous phase mineral acid catalyzed system at moderate temperatures (∼100 °C).

4.3. Global constraints

These are molecular constraints on the entire reaction system applicable to all molecules in the network. For example, global constraints can fix the maximum possible size of a molecule in a system. For the Fructose-to-HMF system, lines 4–14 describe molecular constraints applicable to all molecules at all times – the maximum size of a molecule is 15 atoms (line 9), the maximum allowed charge on a molecule is 1 (line 9), a positively charged atom can never be connected to any atom with a double bond (lines 5–8), and there are no consecutive double bonds (lines 10–14). It can be noted that the two fragments, 'a' and 'b', are defined in a manner similar to the reactant pattern definition, and refer to specific patterns in the molecule that should or should not be present in the molecule. Such fragment constraints can also be imposed in reaction rules.

4.4. Post-processing instructions

4.4.1. Lumping strategy

In RING, the user can specify: (a) if lumping based on functional equivalence is required, (b) if the representative molecule of the lump is the isomer that has branches farthest/closest apart, and (c) if further lumping of hydrocarbons, such as paraffins, olefins, naphthenes, and aromatics based on size is required. While line 27 specifies that lumping is required, lines 28–29 in Table 2 describe how acyclic and cyclic molecules satisfying the condition for lumping, are both represented by the constituent that has its branches farthest apart. For example, in the lump of five-carbon secondary alcohols (Section 3.3.1), the lump representative, according to line 28, would be 3-pentanol because its branch (the C–OH group) is farther apart from both the ends when compared to 2-pentanol, the other constituent.

4.4.2. Pathways identification

Pathway querying, in RING, consists of two steps – a description of the target molecule(s) and the nature of the pathways desired. Target molecules are described by molecular constraints, while

constraints on the nature of pathways – pathway constraints – are described subsequently to put a limit on path length, describe presence/absence of rules and molecules, and to describe constraints on the reactants and products in one or more reactions of the pathway. Lines 30–36 in Table 2 show a pathway query to HMF, "OCc1oc(C=O)cc1". The constraints on the pathway include a maximum length constraint of 11 reactions in the sequence, a limit on the number of occurrences of a reaction corresponding to the rule Hshift (hydride shift, see Supporting information), and a requirement that similar pathways[1] be eliminated.

4.4.3. Mechanism identification

Lines 37–43 in Table 2 describe an overall mechanism query to find mechanisms to HMF. The mechanisms can have several cycles; therefore, constraints can be imposed on the overall mechanism (lines 39–40) as well as on individual cycles (lines 41–43). The overall constraints restrict the total number of reactions and cycles in the complete mechanism. The individual cycle constraints restrict the number of reactions in each cycle, and limit the number of occurrences of a reaction corresponding to the rule Hshift. Further, the requirement "eliminate similar mechanisms" is imposed so that multiple cycles having the same overall reaction are not obtained.

4.4.4. Molecule and reaction queries

The generated network can be queried for specific components of the network – species and reactions. The molecule queries are given as shown in lines 44–46 in Table 2 wherein all molecules satisfying the specified molecular constraints – cyclic and neutral – are sought. Further, reaction queries seeking all reactions of the network satisfying specific types of reactions, reactants, or products, can also be input. Lines 47–50 in Table 2 show an example of reaction query for obtaining all alcohol protonation reactions involving a cyclic reactant.

4.5. Outputs from RING

4.5.1. Reaction network

The reaction network is a list of species and a list of reactions, both written in SMILES format. The format for reactions essentially involves concatenating SMILES strings of reactants and products of that reaction with appropriate delimiters ('.' demarcates two different reactants/products, while '»' separates reactants from products). A SMILES generator has been developed that constructs the individual strings for molecules from their internal representation and appends them together to get the reaction string. Fig. 5 shows a sample set of reactions constituting the network, in both SMILES and graphical form. Several chemistry packages parse SMILES strings to generate a graphical output of molecules and reactions; the reactions in Fig. 5 were generated using ChemDraw (CambridgeSoft, 2011).

4.5.2. Lumping results

Results from a lumping analysis are output as a list of lumps wherein each lump is represented by its representative molecule (identified based on user-specified criterion) and is followed by a list of species belonging to the lump. Fig. 6 shows, pictorially, one of the lumps identified by RING – a lump of Fructose derivatives containing three hydroxy groups, an oxonium ion, a methyl group, and a carbon–carbon double bond. The eight molecules shown in the figure are lumped into one lump represented by the lump representative shown. Once the molecules are lumped, each reaction can

C1(C(C(C(C(CO)O1)O)O)(CO)O.[H+]>>C1(C(C(C(C(CO)O1)O)O)(CO)[O+]
C1(C(C(C(C(CO)O1)O)O)(CO)[O+]>>C1(C(C([C+](CO)O1)O)O)CO.O
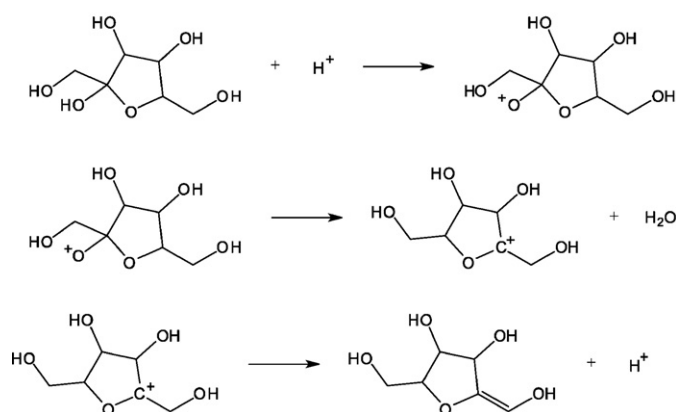C1(C(C([C+](CO)O1)O)O)CO>>C1(=CO)C(C(C(CO)O1)O)O.[H+]

Fig. 5. Sample reactions of the Fructose-to-HMF network generated by RING.

be rewritten in terms of lumps to get a lumped reaction. Since the number of lumps is lesser than the number of species in a network (about 550 species and 240 lumps), the number of lumped reactions is also fewer (about 650 lumped reactions from a network of about 1200 reactions). Lumping, thus, reduces the network size by a factor of 2. For some systems, such as propane aromatization, as shown in the accompanying paper, lumping is shown to lead to reduction in network size by several orders of magnitude (Rangarajan et al., in press).

4.5.3. Pathway identification results

Pathways are obtained as a list of reaction sequences wherein each sequence is, in itself, a list of reactions that traces the product formation. Fig. 7 depicts one of the pathways for the conversion of fructose to HMF as a sequence of reactions. The pathway involves three alcohol-protonation and dehydration steps, while one of the steps is essentially an enol–carbonyl tautomerism step. Antal et al. (1990) propose this sequence of steps as the dominant route for HMF synthesis from fructose in aqueous solutions catalyzed by mineral acids.

4.5.4. Mechanism identification results

Fig. 8 shows the overall (complete) mechanism that corresponds to the pathway described above. The mechanism has four catalytic cycles (Fig. 8) – three dehydration cycles and one isomerization cycle. The overall reaction, thus, is a balanced reaction with one molecule of HMF and three molecules of water formed from each fructose molecule. We can further query to obtain the constituent elementary step reactions of these individual cycles through identifying direct mechanisms.

5. Discussion

The input options and output features of RING allow for systematic rule-based elucidation of complex reaction networks through: (a) construction of the reaction network from reaction rules, and (b) analysis of the transformations in the network through postprocessing instructions (or rules). Such analysis is particularly relevant in biomass conversion as upgrading oxygenates to produce fuels and chemicals involves multiple thermochemical steps to remove oxygen atoms present in different functional groups (Corma, Iborra, & Velty, 2007; Huber, Chheda, Barrett, & Dumesic, 2005; Schmidt et al., 2007). RING currently does not take into consideration the kinetics of each of the steps and hence cannot apply features such as rate-based pruning (Susnow et al., 1997).

---

[1] Two pathways are considered to be dissimilar, or distinct, if one of them has at least one more instance of a reaction of any reaction type than the other.
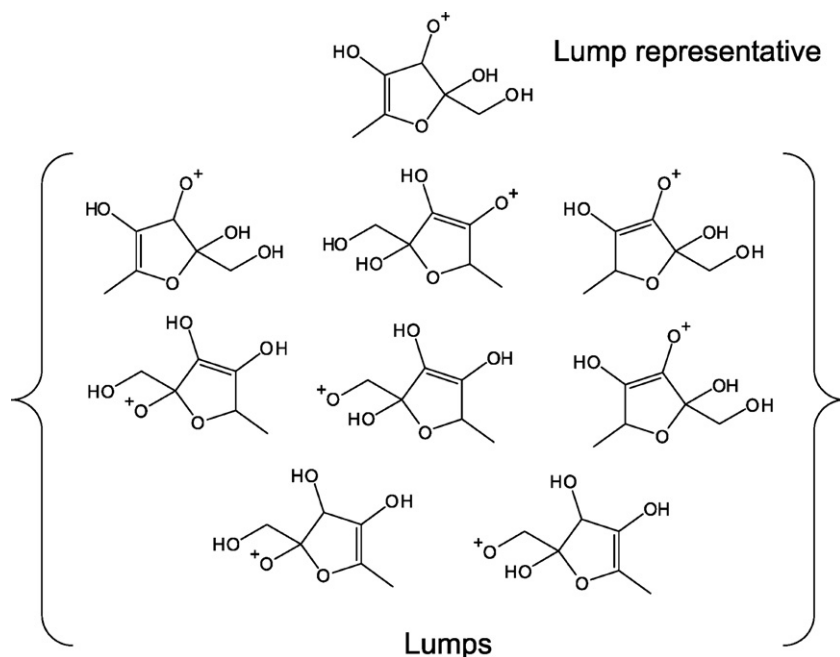
Fig. 6. Sample lump and its representative in the Fructose-to-HMF reaction network identified by RING.
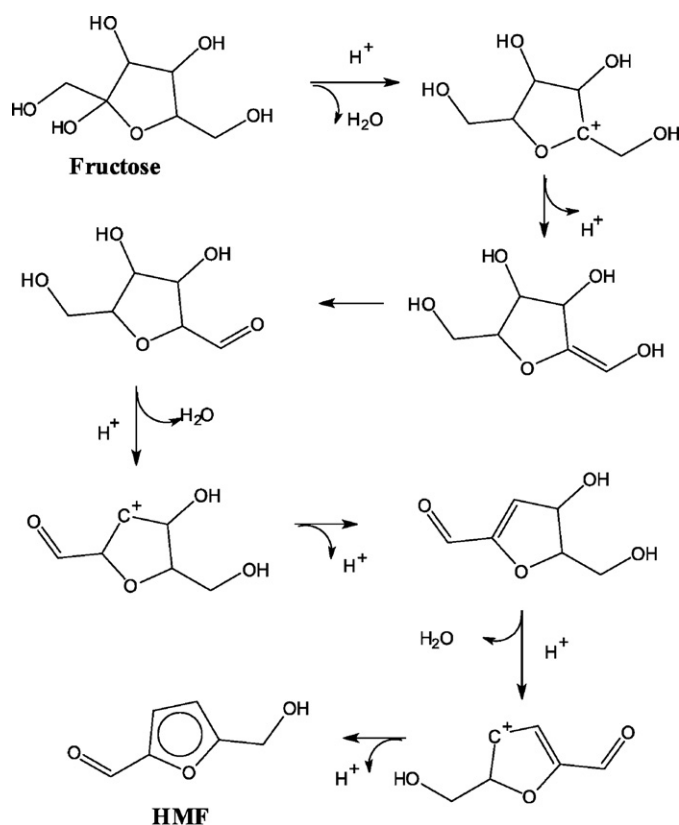

Fig. 7. One pathway from fructose to HMF in the Fructose-to-HMF reaction network identified by RING. The sequential representation of reaction is an adaptation of the reaction list that RING generates.

C1(C(C(C(C(CO)O1)O)O)(CO)O --> O + C1(=CO)C(C(C(CO)O1)O)O
C1(=CO)C(C(C(CO)O1)O)O --> C1(C(C(C(C=O)O1)O)O)CO
C1(C(C(C(C=O)O1)O)O)CO --> O + C1(=CC(C(O1)CO)O)C=O
C1(=CC(C(O1)CO)O)C=O --> O + c1(ccc(C=O)o1)CO

C1(C(C(C(C(CO)O1)O)O)(CO)O --> 3 O + c1(ccc(C=O)o1)CO

Fig. 8. Example of one complete mechanism having four catalytic cycles for HMF production from fructose, identified by RING.

The construction and analysis features of RING can be applied for analyzing complex reaction networks in three distinct classes of problems. We discuss each of these through illustrative examples in the second part of this series (Rangarajan et al., in press):
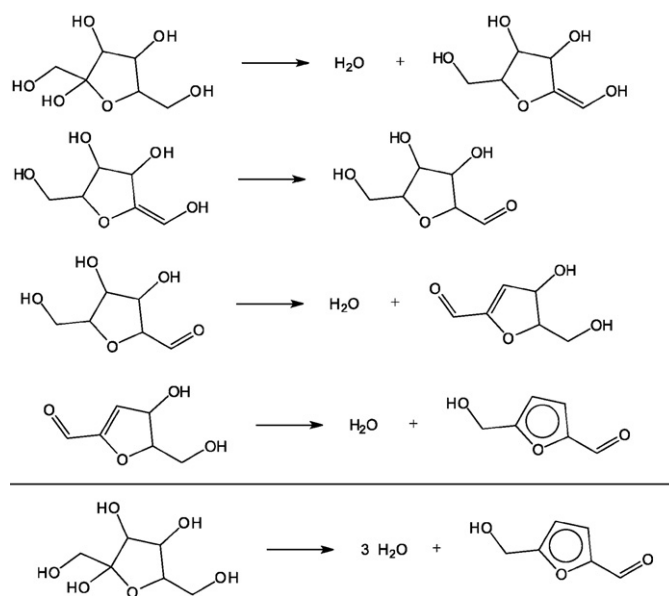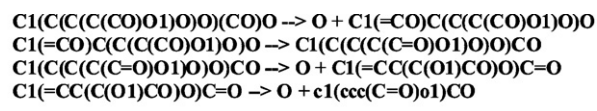
(a) If the elementary chemistry rules and their estimated kinetic parameters are available for a chemical system, RING can be used to identify a plausible set of dominant pathways that form specific products. One such case study is discussed in detail in Rangarajan et al. (in press), where we analyze the system of acid-catalyzed propane aromatization. In this example, we demonstrate three major features of RING – "encapsulation" of class of molecules using composite atom definitions, constraints-based extraction of pathways, and lumping of

isomers leading to orders of magnitude reduction in network size.

(b) Manual identification of all possible mechanisms exhaustively is cumbersome. Instead, RING can be used to generate and identify the mechanisms for the system. Specifically, if experimental data is available for a complex reaction network, RING can be used to test postulated elementary steps and propose a set of plausible pathways/mechanisms that describe how a product is formed in the system. In Rangarajan et al. (in press), an example involving the system of dehydration of glycerol to form acetone and a related system of acetone conversion to acetic acid is used for illustration. This example extensively uses the pathway identification feature of RING to query for specific pathways to products.

(c) RING can be used to identify all possible synthetic routes to a compound from prescribed starting chemicals using specific chemistries. These routes can be pruned further by identifying those that satisfy properties like highest atom-efficiency and constraints like the length of the pathway or the presence/absence of certain types of reactions. In Rangarajan et al. (in press), we identify possible atom-efficient synthetic routes to form large alcohols from small oxygenates. Here, we use the complete mechanism identification feature of RING because the calculation of atom-efficiency requires the knowledge of the stoichiometry of the overall reaction.

In summary, RING is a network generation and analysis tool developed using Cheminformatics and graph theoretical algorithms. It is composed of three components – a language compiler for a domain specific English-like reaction language, a reaction network generator that can construct an exhaustive network from specified reaction rules and initial reactants, and a post-processing module that enables (i) lumping of isomers to get a network of reduced size, (ii) finding pathways between initial reactants and specified products, and (iii) finding mechanisms for the formation of specified molecules in the network. RING can be used to construct reaction networks from elementary or overall reaction rules in a rule-based manner, and further, to elucidate the transformations occurring in complex chemical reaction networks through identifying reaction pathways, mechanisms, and lumps of isomers. The construction and analysis options in RING were described through an illustrative example involving acid catalyzed dehydration of fructose to HMF. Detailed case studies showcasing the network generation and analysis features of RING are discussed in the second part of this series (Rangarajan et al., in press). RING is distributed open source under GNU Lesser GPL version 2.1 license through a dedicated webpage (RING, 2011).

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compchemeng.2012.06.008.

References

Antal, M. J., Mok, W. S. L., & Richards, G. N. (1990). Mechanism of formation of 5-hydroxymethyl-2-furaldehyde from fructose and sucrose. *Carbohydrate Research*, 199(1), 91–109.

Arita, M. (2003). In silico atomic tracing by substrate–product relationships in *Escherichia coli* intermediary metabolism. *Genome Research*, 13, 2455–2466.

Arita, M. (2004). The metabolic world of *Escherichia coli* is not small. *Proceedings of the National Academy of Sciences*, 101(6), 1543–1547.

Baltanas, M. A., & Froment, G. F. (1985). Computer generation of reaction networks and calculation of product distribution in the hydroisomerization and hydrocracking of paraffins on Pt-containing bifunctional catalysts. *Computers & Chemical Engineering*, 9(1), 71–81.

Bhan, A., Hsu, S.-H., Blau, G., Caruthers, J. M., Venkatasubramanian, V., & Delgass, W. N. (2005). Microkinetic modeling of propane aromatization over HZSM-5. *Journal of Catalysis*, 235, 35–51.

Blinov, M. L., Yang, J., Faeder, J. R., & Hlavacek, W. S. (2005). Graph theory for rule-based modeling of biochemical networks. *Transactions on Computational Systems Biology VII: Lecture Notes in Computer Science*, 4230, 89–106.

Blinov, M. L., Faeder, J. R., Yang, J., Goldstein, B., & Hlavacek, W. S. (2005). 'On-the-fly' or 'generate-first' modeling? *Nature Biotechnology*, 23(11), 1344–1345.

Blurock, E. S. (1995). Reaction: System for modeling chemical reactions. *Journal of Chemical Information and Computer Science*, 35, 607–616.

Broadbelt, L. J., Stark, S. M., & Klein, M. T. (1994a). Computer-generated pyrolysis modeling – On the fly generation of species, reactions and rates. *Industrial & Engineering Chemistry Research*, 33(4), 790–799.

Broadbelt, L. J., Stark, S. M., & Klein, M. T. (1994b). Computer generated reaction networks: On-the-fly calculation of species properties using computational quantum chemistry. *Chemical Engineering Science*, 49(24 (Pt. 2)), 4991–5010.

Broadbelt, L. J., Stark, S. M., & Klein, M. T. (1995). Termination of computer-generated reaction mechanisms: Species rank-based convergence criterion. *Industrial & Engineering Chemistry Research*, 34(8), 2566–2573.

Broadbelt, L. J., Stark, S. M., & Klein, M. T. (1996). Computer generated reaction modelling: Decomposition and encoding algorithms for determining species uniqueness. *Computers & Chemical Engineering*, 20(2), 113–129.

Broadbelt, L. J., & Pfaendtner, J. (2005). Lexicography of kinetic modeling of complex reaction networks. *AIChE Journal*, 51(8), 2112–2121.

CambridgeSoft (2011). ChemDraw. http://www.cambridgesoft.com/software/ChemDraw/ Accessed October 2011.

Caruthers, J. M., Lauterbach, J., Thomson, K. T., Venkatasubramanian, V., Snively, C. M., Bhan, A., et al. (2003). Catalyst design: Knowledge extraction from high-throughput experimentation. *Journal of Catalysis*, 216, 98–109.

Chiang, H., & Bhan, A. (2010). Catalytic consequence of hydroxyl group location on the rate and mechanism of parallel dehydration reactions of ethanol over acidic zeolites. *Journal of Catalysis*, 271(2), 251–261.

Chinnick, S. J., Baulch, D. L., & Asyscough, P. B. (1988). An expert system for hydrocarbon pyrolysis reactions. *Chemometrics and Intelligent laboratory Systems*, 5, 39–52.

Conrad, H., Ertl, G., & Latta, E. E. (1974). Adsorption of hydrogen on palladium single crystal surfaces. *Surface Science*, 41, 435–446.

Corma, A., Iborra, S., & Velty, A. (2007). Chemical routes for the transformation of biomass into chemicals. *Chemical Reviews*, 107(6), 2411–2502.

Corma, A., Huber, G. W., Sauvanaud, L., & O'Connor, P. (2008). Biomass to chemicals: Catalytic conversion of glycerol/water mixtures into acrolein, reaction network. *Journal of Catalysis*, 257, 163–171.

Cortright, R. D., & Dumesic, J. A. (2001). Kinetics of heterogeneous catalytic reactions: Analysis of reaction schemes. *Advances in Catalysis*, 46, 161–264.

Croes, D., Couche, F., Wodak, S. J., & van Helden, J. (2006). Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356, 222–236.

Di Maio, F. P., & Lignola, P. G. (1992). KING, a kinetic network generator. *Chemical Engineering Science*, 47(9–11), 2713–2718.

Dugundji, J., & Ugi, I. (1973). An algebraic model of constitutional chemistry as a basis for chemical computer programs. *Topics in Current Chemistry*, 39, 19–64.

Faeder, J. R., Blinov, M. L., Goldstein, B., & Hlavacek, W. S. (2005). Rule-based modeling of biochemical networks. *Complexity*, 10, 22–41.

Faeder, J. R., Hlavacek, W. S., Reischl, I., Blinov, M. L., Metzger, H., Redondo, A., Wofsy, C., & Goldstein, B. (2003). Investigation of early events in Fc epsilon RI-mediated signaling using a detailed mathematical model. *Journal of Immunology*, 170(7), 3769–3781.

Fan, L. T., Bertok, B., & Friedler, F. (2002). A graph-theoretic method to identify candidate mechanisms for deriving the rate law of a catalytic reaction. *Computers & Chemical Engineering*, 26(3), 265–292.

Faulon, J.-L., & Sault, A. G. (2001). Stochastic generator of chemical structure 3. Reaction network generation. *Journal of Chemical Information & Computer Science*, 41, 894–908.

Finley, S. D., Broadbelt, L. J., & Hatzimanikatis, V. (2009). Computational framework for predictive biodegradation. *Biotechnology and Bioengineering*, *104*(6), 1086–1097.

Fishtik, I., Callaghan, C. A., & Datta, R. (2004). Reaction route graphs. I. Theory and algorithm. *Journal of Physical Chemistry B*, *108*(18), 5671–5682.

Froment, G. P., Van de Steene, B. O., Van Damme, P. S., Narayanan, S., & Goossens, A. G. (1976). Thermal cracking of ethane and ethane–propane mixtures. *Industrial & Engineering Chemistry Process Design and Development*, *15*(4), 495–504.

Gonzalez-Lergier, J., Broadbelt, L. J., & Hatzimanikatis, V. (2005). Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways. *Journal of American Chemical Society*, *127*(27), 9930–9938.

Green, W. H., Bhattacharjee, B., Oluwole, O., Song, J., Sumathi, R., Wijaya, C. D., et al. (2004). New methods for predictive chemical kinetics. *Preprint Papers – American Chemical Society, Division of Fuel Chemistry*, *49*(1), 323.

Heath, A. P., & Bennett, G. N., & Kavraki, L. E. (2010). Finding metabolic pathways using atom tracking. *Bioinformatics*, *26*, 1548–1555.

Henry, C. S., Broadbelt, L. J., & Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophysical Journal*, *92*(5), 1792–1805.

Henry, C. S., Broadbelt, L. J., & Hatzimanikatis, V. (2010). Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-Hydroxypropanoate. *Biotechnology and Bioengineering*, *106*(3), 462–473.

Heyberger, B., Battin-Leclerc, F., Warth, V., Fournet, R., Come, G. M., & Scacchi, G. (2001). Comprehensive mechanism for the gas-phase oxidation of propene. *Combustion and Flame*, *126*, 1780–1802.

Hill, A., Tomshine, J., Wedding, E., Sotirpolous, V., & Kaznessis, Y. (2008). SynBioSS: The synthetic biology modeling suite. *Bioinformatics*, *24*(51), 2551–2553.

Ho, T. C. (2008). Kinetic modeling of large-scale reaction systems. *Catalysis Reviews*, *50*(3), 287–378.

Hsu, S. H., Krishnamurthy, B., Rao, P., Zhao, C. H., Jagannathan, S., & Venkatasubramanian, V. (2008). A domain-specific compiler theory based framework for automated reaction network generation. *Computers & Chemical Engineering*, *32*(10), 2455–2470.

Huber, G. W., Chheda, J. N., Barrett, C. J., & Dumesic, J. A. (2005). Production of liquid alkanes by aqueous phase processing of biomass-derived carbohydrates. *Science*, *308*, 1446–1450.

Katare, S., Caruthers, J. M., Delgass, W. N., & Venkatasubramanian, V. (2004). An intelligent system for reaction kinetic modeling and catalyst design. *Industrial & Engineering Chemistry Research*, *43*(14), 3484–3512.

Kummel, A., Panke, S., & Heinemann, M. (2006). Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Molecular Systems Biology*, *2*, 2006.0034.

Kuo, J. C., & Wei, J. (1969). A lumping analysis in monomolecular reaction systems—Analysis of approximately lumpable system. *Industrial and Engineering Chemistry Fundamentals*, *8*, 124–133.

Li, C., Henry, C. S., Jankowski, M. D., Ionita, J. A., Hatzimanikatis, V., & Broadbelt, L. J. (2004). Computational discovery of biochemical routes to specialty chemicals. *Chemical Engineering Science*, *59*, 5051–5060.

Li, G., & Rabitz, H. (1989). A general analysis of exact lumping in chemical kinetics. *Chemical Engineering Science*, *39*, 1261.

Lin, Y.-C., Fan, L. T., Shafie, S., Bertok, B., & Friedler, F. (2009). Generation of light hydrocarbons through Fischer–Tropsch synthesis: Identification of potentially dominant catalytic pathways via the graph-theoretic method and energetic analysis. *Computers & Chemical Engineering*, *33*(6), 1182–1186.

Mavrovouniotis, M. L., & Stephanopolous, G. (1992). Synthesis of reaction mechanisms consisting of reversible and irreversible steps. 1. A synthesis approach in the context of simple examples. *Industrial & Engineering Chemistry Research*, *31*, 1625–1637.

Mayeno, A. N., Yang, R. S. H., & Reisfeld, B. (2005). Biochemical reaction network modeling: Predicting metabolism of organic chemical mixtures. *Environmental Science and Technology*, *39*, 5363–5371.

Millner, P. C. (1964). The possible mechanism of complex reactions involving consecutive steps. *Journal of the Electrochemical Society*, *111*, 228–232.

Mohan, D., Pittman, C. U., Jr., & Steele, P. H. (2006). Pyrolysis of wood/biomass for bio-oil: A critical review. *Energy & Fuels*, *20*(3), 848–889.

MOPAC2009 (2008). James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA. http://OpenMOPAC.net Accessed August 2010.

Otarod, M., & Happel, J. (1992). Studies of the structure of chemical mechanisms. *Chemical Engineering Science*, *47*(3), 587–592.

Papin, J. A., Stelling, J., Price, N. D., Klmat, S., Schuster, S., & Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, *22*(8), 400–405.

Pfaendtner, J., & Broadbelt, L. J. (2008). Mechanistic modeling of lubricant degradation. 2: The autoxidation of decane and octane. *Industrial & Engineering Chemistry Research*, *47*, 2897–2904.

Prickett, S. E., & Mavrovouniotis, M. L. (1997a). Construction of complex reaction systems. 1. Reaction description language. *Computers & Chemical Engineering*, *21*(11), 1219–1235.

Prickett, S. E., & Mavrovouniotis, M. L. (1997b). Construction of complex reaction systems. 2. Molecule manipulation and reaction application algorithms. *Computers & Chemical Engineering*, *21*(11), 1237–1254.

Prickett, S. E., & Mavrovouniotis, M. L. (1997c). Construction of complex reaction systems—3. An example: Alkylation of olefins. *Computers & Chemical Engineering*, *21*(12), 1325–1337.

Quann, R. J., & Jaffe, S. B. (1992). Structure-oriented lumping: Describing the chemistry of complex hydrocarbon mixtures. *Industrial and Engineering Chemistry Research*, *31*(11), 2483–2497.

Quann, R. J., & Jaffe, S. B. (1996). Building useful models of complex reaction systems in petroleum refining. *Chemical Engineering Science*, *51*(10), 1615–1631.

Rangarajan, S., Bhan, A., & Daoutidis, P. (2010). Rule-based generation of thermochemical routes to biomass conversion. *Industrial & Engineering Chemistry Research*, *49*(21), 10459–10470.

Rangarajan, S., Bhan, A., & Daoutidis, P. (in press). Language-oriented rule-based reaction network generation and analysis: Applications of RING. *Computers & Chemical Engineering*, http://dx.doi.org/10.1016/j.compchemeng.2012.06.003.

Ranzi, E., Dente, M., Goldaniga, A., Bozzano, G., & Faravelli, T. (2001). Lumping procedures in detailed kinetic modeling of gasification, pyrolysis, partial oxidation and combustion of hydrocarbon mixtures. *Progress in Energy and Combustion Science*, *27*, 99–139.

Ratkiewicz, A., & Truong, T. N. (2003). Application of chemical graph theory for automated mechanism generation. *Journal of Chemical Information and Modeling*, *43*, 36–44.

Reed, J. L., Vo, T. D., Schilling, C. H., & Palsson, B. O. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology*, *4*(9), R54.

RING (2011). http://gemini.cems.umn.edu/research/bhan/RING.html.

Rizzi, M., Baltes, M., Theobald, U., & Reuss, M. (1997). In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: II. Mathematical model. *Biotechnology and Bioengineering*, *55*, 592–608.

Schilling, C. H., Letscher, D., & Palsson, B. O. (2000). Theory for the systematic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, *203*(3), 229–248.

Schmidt, L. D., & Dauehauer, P. J. (2007). Hydride routes to biofuels. *Nature*, *447*, 914.

Schuster, S., & Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, *2*, 165–182.

Song, J. (2004). Massachusetts Institute of Technology, Ph.D. Dissertation.

Susnow, R. G., Dean, A. M., Green, W. H., Peczak, P., & Broadbelt, L. J. (1997). Rate-based construction of kinetic models for complex systems. *Journal of Physical Chemistry A*, *101*(20), 3731–3740.

Tomlin, A. S., Turanyi, T., & Pilling, M. J. (1997). M. J. Polling (Ed.), *Low-temperature combustion and auto-ignition* (p. 293). Amsterdam: Elsevier.

Ugi, I., Bauer, J., Bley, K., Alf, D., Dietz, A., Fontain, E., et al. (1993). Computer assisted solution of chemical problems—The historical development and present state of the art of a new discipline of chemistry. *Angewandte Chemie International Edition in English*, *32*, 201–227.

Van Geem, K. M., Reyniers, M.-F., Marin, G. B., Song, J., Green, W. H., & Matheu, D. M. (2006). Automatic reaction network generation using RMG for steam cracking of n-hexane. *AIChE Journal*, *52*(2), 718–730.

Van Wyk, E., Bodin, D., Gao, J., & Krishnan, L. (2010). Silver: An extensible attribute grammar system. *Science of Computer Programming*, *75*(1–2), 39–54.

Wong, H.-W., Li, X., Swihart, M. T., & Broadbelt, L. J. (2004). Detailed kinetic modeling of silicon nanoparticle formation chemistry via automated mechanism generation. *Journal of Physical Chemistry A*, *108*(46), 10122–10132.

Warth, V., Battin-Leclerc, F., Fournet, R., Glaude, P. A., Come, G. M., & Scacchi, G. (2000). Computer based generation of reaction mechanisms for gas-phase oxidation. *Computers and Chemistry*, *24*, 541–560.

Weininger, D. (1988). SMILES: A chemical language and information systems. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, *28*(1), 31–36.

Yaluris, G., Madon, R. J., & Dumesic, J. A. (1997). 2-Methylhexane cracking on Y zeolites: Catalytic cycles and reaction selectivity. *Journal of Catalysis*, *165*, 205–220.