Deductive Assistance for Elucidation of Chemical Reaction Pathways

Raúl E. Valdés-Pérez

January 1990

CMU-CS-90-104$_2$

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

A deductive framework is introduced to assist with elucidation of chemical reaction pathways. Given an overall, unbalanced, reaction scheme, a list of specific pathways is inferred that account for the transformation of the reagents into known products.

One use of the framework is to infer a list of simplest such pathways, i.e., those having fewest steps and species. This list can serve as an initial set of candidate hypotheses. For some reaction schemes, the list of simplest pathways contains only a few candidates.

The scope of our current implementation consists of all three-step pathways, having any number of species.

The author is supported by a Graduate Fellowship from the Avionics Laboratory at Wright-Patterson Air Force Base, awarded by Universal Energy Systems. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government or Universal Energy Systems.

1. Introduction

This report describes a deductive framework for partially automated elucidation of chemical-reaction pathways.[1] The framework has been implemented in a computer program that generates, for a specific reaction scheme, a list of pathways that account for the formation of known products from the reagents.

For example, the synthesis of the dibenzodiazocine $C_{17}H_{18}N_2$ with water by-product according to the (unbalanced) reaction scheme [2]:

$$C_7H_9N + CH_2O \rightsquigarrow C_{17}H_{18}N_2 + H_2O$$

gave rise to the following five, simplest pathways, each having six species and three steps:[2]

$$\begin{array}{rcl@{\qquad}rcl} C_7H_9N + CH_2O & \rightarrow & H_2O + C_8H_9N & A+B & \rightarrow & X+Y \\ CH_2O + C_8H_9N & \rightarrow & C_9H_{11}NO & B+Y & \rightarrow & Z \\ C_8H_9N + C_9H_{11}NO & \rightarrow & C_{17}H_{18}N_2 + H_2O & Y+Z & \rightarrow & T+X \end{array}$$

$$\begin{array}{rcl@{\qquad}rcl} C_7H_9N + CH_2O & \rightarrow & H_2O + C_8H_9N & A+B & \rightarrow & X+Y \\ CH_2O + C_8H_9N & \rightarrow & H_2O + C_9H_9N & B+Y & \rightarrow & X+Z \\ C_8H_9N + C_9H_9N & \rightarrow & C_{17}H_{18}N_2 & Y+Z & \rightarrow & T \end{array}$$

$$\begin{array}{rcl@{\qquad}rcl} C_7H_9N + CH_2O & \rightarrow & H_2O + C_8H_9N & A+B & \rightarrow & X+Y \\ 2C_8H_9N & \rightarrow & C_7H_9N + C_9H_9N & 2Y & \rightarrow & A+Z \\ C_8H_9N + C_9H_9N & \rightarrow & C_{17}H_{18}N_2 & Y+Z & \rightarrow & T \end{array}$$

$$\begin{array}{rcl@{\qquad}rcl} C_7H_9N + CH_2O & \rightarrow & H_2O + C_8H_9N & A+B & \rightarrow & X+Y \\ 2C_8H_9N & \rightarrow & C_{16}H_{18}N_2 & 2Y & \rightarrow & Z \\ CH_2O + C_{16}H_{18}N_2 & \rightarrow & C_{17}H_{18}N_2 + H_2O & B+Z & \rightarrow & T+X \end{array}$$

$$\begin{array}{rcl@{\qquad}rcl} C_7H_9N + CH_2O & \rightarrow & H_2O + C_8H_9N & A+B & \rightarrow & X+Y \\ 2C_8H_9N & \rightarrow & C_{16}H_{18}N_2 & 2Y & \rightarrow & Z \\ C_8H_9N + C_{16}H_{18}N_2 & \rightarrow & C_{17}H_{18}N_2 + C_7H_9N & Y+Z & \rightarrow & T+A \end{array}$$

For each pathway, a schematic version appears on the right, with A,B for the reagents, T for the dibenzodiazocine, and X for the water; the values of Y,Z vary according to the pathway.

---

[1] Others terms in use are 'network' and 'mechanism.' We use 'pathway' here to emphasize the goal of obtaining a sequence of steps from reagents to a target product. We use 'network' elsewhere [1] in the context of finding the complete set of reaction steps. We avoid the term 'mechanism,' because of its suggestion of a more detailed, physical account of how reactions proceed.

[2] All of these networks share the stoichiometry $2A + 3B \rightarrow 1T + 3X$, and the intermediate products Y,Z are wholly consumed at stoichiometric proportions of the reagents.

2. A Deductive Framework $\mathcal{D}$

Our framework relies on certain simplicity assumptions about chemical reactions, and on restriction of the scope to non-isomeric reaction steps. These assumptions and restrictions can be regarded as premises, which when conjoined with information about the molecular formulas of reagents and product(s), imply a list of pathways, as in the example of the last section.

Other premises are needed to obtain exactly the output of the first page by formal means, such as by a computer program. These auxiliary premises are discussed in Appendix A. To understand the framework's scope, only the premises of this section are needed.

The key premises defining the scope are the following:

- The overall (non-stoichiometric) reaction scheme is $A + B \rightsquigarrow T + (products)$.
- Neither reagent alone suffices to form the product T.
- Any reaction step has at most two reactants (termolecular steps are excluded).
- Any reaction step has at most two products.
- No reaction step involves on both sides a same molecular formula. This excludes isomerizations, e.g., $C \rightarrow D$, as well as catalytic steps, e.g., $C + X \rightarrow D + X$.[3]

To be explicit, the following reaction *steps* are included in the scope (each step can also be reversible):

$$\begin{array}{rcl} C & \rightarrow & E+F \\ C+D & \rightarrow & E \\ C+D & \rightarrow & E+F \end{array} \qquad \begin{array}{rcl} C & \rightarrow & 2E \\ C+D & \rightarrow & 2E \\ 2C & \rightarrow & E \\ 2C & \rightarrow & E+F \end{array}$$

Excluded are the steps $C \rightarrow E$, and $2C \rightarrow 2E$.

Other premises convenient to include are the total number S of species[4] contained in a pathway, and the number R of reaction steps.

These premises, together with the the molecular formulas of reagents and known products, deductively imply a finite list - possibly empty - of instantiated pathways.[5] As with any deduction, the truth of one of the pathways is necessary, if the premises hold. To obtain a list of simplest pathways accounting for a reaction scheme, we increment the premises S,R until a non-empty list results.

Summarizing, we assert that carrying out the following deduction $\mathcal{D}$ is both useful and practical:

$$\text{key premises} \wedge \text{molecular formulas} \wedge \text{S} \wedge \text{R} \Rightarrow \text{list-of-instantiated-pathways}$$

---

[3] Only a narrow meaning of 'catalytic,' referring to individual steps, is intended here; catalytic roles within the overall reaction pathway are allowed.

[4] In chemistry, a species refers to a molecule having distinct chemical properties. Molecules having identical molecular formulas, but differing chemically due possibly to distinct spatial configurations, i.e., isomers, are usually considered distinct species. Here, a species is distinguished *only* by its molecular formula; isomers are not considered.

[5] All pathways inferred must contain both reagents, as well as all the known products.

3. Use and Interpretation of $\mathcal{D}$

For now, let's pretend that carrying out the deduction (e.g., computing it) is not a problem. How is the list of instantiated pathways to be interpreted and used?

First, we stress the following points. The only information needed to carry out $\mathcal{D}$ on any reaction scheme is the molecular formulas of reagents and any known products (at least one); there is no need to supply a complete list of reaction products. No information about reaction mechanism, molecular connectivity, or molecular configuration is needed nor used. Also, $\mathcal{D}$ is general: it is not tied to any specific type of reaction chemistry.

If $\mathcal{D}$ infers an empty list, and if the original synthetic reaction scheme is known to occur, then one of the premises of $\mathcal{D}$ is false, and must be changed or dropped in order to find a pathway. Within this framework, the premises to change are the number of species S or steps R. For the reaction scheme of section 1, no pathways were found for values of S,R less than S=6, R=3.

Typically, the inferred instantiated pathways contain species not mentioned in the input reaction scheme. For the dibenzodiazocine reaction, the species C_8H_9N appears in each of the five pathways, but was not part of the input formulas. The basis for "conjecturing" these species via $\mathcal{D}$ is largely non-empirical; the species can be interpreted as intermediates (or by-products) that are convenient for the task of transforming reagents into known products under the constraint of simplicity.

Insight derived from viewing reaction pathways as inferences made by $\mathcal{D}$, and from carrying out $\mathcal{D}$ on several reaction schemes, has led to the following observations:

1. If only one product is known (i.e., the target T), then the list of simplest pathways contains only reactions on a path to T, e.g., no side reactions are present.
2. If more than one product is known, then a pathway in the simplest list may involve one of the products *off* a path to T, e.g., as a side product.
3. The inferred pathways involving R+1 steps are generally much fewer than the inferred pathways involving R steps, if both involve the same *number* S of species. If, moreover, they both involve the same *set* of S species, then the R+1 pathways *cannot* be more than the R pathways.
4. With other factors equal, reaction schemes involving more chemical elements generally result in a much shorter list of simplest pathways. This is because each element contributes a constraint that must be satisfied independently, so the problem is more constrained.
5. In general, identifying one more unknown product greatly reduces the total number of pathways simpler than, or as simple as, the "true," but presumably unknown, pathway. Our experience with $\mathcal{D}$ on several examples has shown that knowing one more product is a very powerful source of constraint.

3.1. Isomerization reaction steps

The scope of the current framework excludes isomeric and catalytic reaction steps. Such steps can be included at will in any of our pathways, without contributing, *at the abstract level of molecular*

formulas, to the transformation of reagents into products. Hence, any list of simplest pathways would not include such steps, even if they were within the scope of the framework. For example, the following pathway, shown schematically for illustrative purposes,

$$\begin{aligned} A + B &\rightarrow 2X \\ B + X &\rightarrow B + T \end{aligned}$$

would be rendered as

$$A + B \rightarrow 2T$$

at the abstract level of molecular formulas.

Of course, the true pathway of many reaction schemes does include isomeric or catalytic steps. Under an expanded scope of $\mathcal{D}$, the simplest pathways presumably would be ruled out, and more complex pathways including such steps considered until the correct one were found. Nevertheless, we have excluded such steps because pathways involving them tend to clutter tremendously the lists of non-simplest pathways.

Therefore, two recommendations are advanced for the practical use of the current framework. First, one can use it only on reaction schemes for which no isomeric and catalytic steps are expected. Second, one can work, as does the framework, at the abstract level of *species* $\equiv$ *molecular formulas*, thus ignoring isomers, etc. Working at that more abstract level serves to narrow the space of plausible pathways; the chemist can always descend to the conventional, mechanistic level as desired, e.g., by proposing a catalytic step as an addition to the abstract pathway.

4. An Implementation of $\mathcal{D}$

We are using a program implementation of $\mathcal{D}$ to propose hypotheses for a reaction-pathway elucidation program, called MeChem. Other component programs of MeChem are charged with discriminating among the pathways proposed, partly by interpreting experimental, kinetic data [1].

Our current, Common-Lisp implementation of $\mathcal{D}$ builds a priori a permanent catalogue of *schematic* reaction pathways, expressed in terms of (molecular) variables, i.e., A,B for the two reagents, T for the target product, and other symbols as needed to represent the required number of species.

The scope of the catalogue consists of all pathways involving at most three reaction steps.[6] The accompanying table shows the size of the different classes; the total number of undirected, schematic pathways stored is 15,494.[7]

---

[6] A three-step reaction that fulfills our key premises has at most eight species, so S=8, R=3 is the current boundary of the scope.

[7] An undirected pathway is one that abstracts the step directions, i.e., each step is a set of two sets of species. Considering further the possible assignments of step directions ($\rightarrow, \leftarrow, \leftrightarrow$) gives rise to a larger number of directed pathways.

Total undirected pathways

#steps	#species	pathways
1	3	2
1	4	1
2	3	5
2	4	60
2	5	72
2	6	16
3	3	3
3	4	659
3	5	4824
3	6	6773
3	7	2759
3	8	320

Matching a specific reaction scheme against the catalogue works as follows. The corresponding molecular variables within each stored pathway are bound to the molecular formulas of reagents and known products. Then, by algebraic manipulation, the values of the remaining variables are inferred, if there is enough constraint to do so. Any contradiction (e.g., due to contradictory steps, or to negative atomic coefficients inferred for the unknowns, etc.) causes the instantiated pathway to be rejected.

5. Conclusion

Our main contribution is to point out that a useful deduction can be formulated and tractably computed, in order to assist, by proposing simple, initial hypotheses, in the elucidation of chemical reaction pathways. The scope of our current implementation consists of pathways involving three or fewer reaction steps.

6. Acknowledgments

Profs. Jonathan Lindsey (Chemistry), Tom Mitchell, and Herbert Simon have regularly guided the larger network-elucidation effort. Prof. Bruce Buchanan (Univ. of Pittsburgh) suggested attention to ways of incorporating constraints (i.e., premises) earlier in the schematic-pathway enumeration algorithm, thus enabling the practical generation of the entire 3-reaction-step catalogue. François Lecouat was an early collaborator on the elucidation problem.

Prof. Gary Powers and Tamara Daugherty (Chemical Engineering) directed me to references on the matrix-algebraic treatment of chemical reactions. David Applegate assisted with the linear programming formulation of Appendix C2.

The author thanks Prof. Simon for his comments on a draft of this report.

References

[1] R. Valdes-Perez, "Learning retrodictive knowledge from scientific laws: the case of chemical kinetics," Technical Report CMU-CS-89-179, Carnegie Mellon University, 1989.

[2] T. H. Webb and C. S. Wilcox, "Improved synthesis of symmetrical and unsymmetrical 5,11-methano[b,f][1,5]dibenzodiazocines - readily available nanoscale structural units," *Journal of Organic Chemistry*, January 1990. To appear.

[3] E. Petersen, *Chemical Reaction Analysis.* Prentice-Hall, 1965.

[4] C. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity.* Prentice-Hall, 1982.

[5] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, 1986.

A Auxiliary premises

The following lists the auxiliary premises used to carry out the deduction $\mathcal{D}$:

- The elemental coefficients of any species are required to be non-negative integers, not all zero.
- No two steps within a pathway are identical.
- Any species contained in a pathway must be formable from the two reagents. For example, the schematic pathway

$$\begin{aligned} A + B &\rightarrow T + X \\ Y &\rightarrow 2X \end{aligned}$$

  is excluded because Y is not present, hence the step consuming it is spurious.
- For any element, the number of its atoms must balance across both sides of a reaction step.

In addition, our computer implementation uses for convenience the following two premises, regarding schematic pathways:

- For any pathway in the permanent catalogue, there must exist an assignment of legal, molecular formulas to the unknowns such that the steps are balanced (Appendix C2 describes how to carry out this test). A legal, molecular formula is one that satisfies the first condition in the previous list.
- No two pathways in the permanent catalogue are identical under any permutation of A.B, or any permutations of non-T products, or both.

These two premises are not strictly necessary, because unsatisfiable or redundant pathways would be intercepted at the stage when molecular variables are instantiated. However, these premises serve to reduce the catalogue size, as well as the computation time when instantiating variables.

B Another example

For the reaction scheme

$$C_8H_{18}O_3 + H_2O \rightsquigarrow C_4H_8O_2 + C_2H_6O$$

the list of simplest pathways is as follows:

$$\begin{aligned} C_8H_{18}O_3 &\rightarrow C_2H_6O + C_6H_{12}O_2 & \qquad A &\rightarrow X + Y \\ H_2O + C_6H_{12}O_2 &\rightarrow C_4H_8O_2 + C_2H_6O & \qquad B + Y &\rightarrow T + X \end{aligned}$$

$$\begin{aligned} C_8H_{18}O_3 + H_2O &\rightarrow C_2H_6O + C_6H_{14}O_3 & \qquad A + B &\rightarrow X + Y \\ C_6H_{14}O_3 &\rightarrow C_4H_8O_2 + C_2H_6O & \qquad Y &\rightarrow T + X \end{aligned}$$

$$\begin{aligned} C_8H_{18}O_3 + H_2O &\rightarrow C_4H_8O_2 + C_4H_{12}O_2 & \qquad A + B &\rightarrow T + Y \\ C_4H_{12}O_2 &\rightarrow 2C_2H_6O & \qquad Y &\rightarrow 2X \end{aligned}$$

Each pathway has five species and two reaction steps; the schematic versions are on the right.

C Satisfiable schematic pathways

This section discusses how to detect whether a schematic pathway such as

$$\begin{aligned} 2A &\rightarrow X+Y \\ A+B &\rightarrow Y \\ B+X &\rightarrow T+Y \end{aligned}$$

is a priori unsatisfiable, in the sense that there exists no assignment of legal molecular formulas to the variables such that all steps are balanced.

The pathway just shown is unsatisfiable, as can be seen by regarding and manipulating the steps as algebraic equations, or stoichiometric equations [3]:

$$\begin{aligned} 2A &= X+Y \\ A+B &= Y \\ B+X &= T+Y \end{aligned}$$

Doing so, one infers that $T+B=0$, so that either both T,B are zero, or one is the negation of the other. In either case, an illegal molecular formula is required. We have developed a procedure to carry out by computer program the detection of unsatisfiability.

C1. A preliminary result

We show the following:

Proposition 1 *A schematic pathway is satisfiable if and only if its stoichiometric (algebraic) equations are satisfied by some assignment of positive integers (zero excluded) to the variables.*

If the pathway is satisfiable, then there is a satisfactory positive-integer assignment. Consider a satisfactory assignment of molecular formulas to the pathway variables. For each variable, count the number of atoms in its assignment, e.g., count 3 atoms in H_2O. Assign these counts to the corresponding algebraic variable. Because elemental conservation also implies conservation of total atoms, the sum of counts on one side of an equation equals the sum on the other side. Therefore, there is a satisfactory positive-integer assignment to the algebraic variables.

If there is a satisfactory positive-integer assignment to the stoichiometric equations, then the pathway is satisfiable. For each positive-integer assignment (say, 3 to the algebraic 'A'), assign the molecular formula having that number of hydrogen atoms to the corresponding pathway variable (H_3 to the pathway 'A'). The resulting instantiated pathway is seen to balance (although the steps may, irrelevantly, not be empirically plausible).

Both senses of the implication are shown, hence the proposition is proven.

C2. The formal procedure

To find an assignment of positive integers to the stoichiometric equations, we first re-formulate the latter as a matrix equation

$$A_{R\times S}\, X_{S\times 1} = 0_{R\times 1} \tag{1}$$

where matrix A has the stoichiometric coefficients, vector X corresponds to the pathway variables, R, S are respectively the number of steps and species, and 0 is the zero vector. Next, positive-integer assignments are enforced by the constraints

$$x_i \geq 1,\ i = 1, \ldots, S \tag{2}$$

one for each entry x_i in $X_{S\times 1}$. Finding a solution to equation 1 and constraints 2 is the problem of checking the consistency of a linear program [4], which is computable almost instantly for the size of problems considered here. Any Simplex implementation of linear programming has a subroutine to check consistency (e.g., see [5]).

Finally, we note that if the above linear program has a solution, then it has also a solution in the rational numbers, because consistency checking could be carried out exactly while preserving rationality. Then, any rational solution is convertible to a positive-integer solution via multiplication by the least common multiple of the denominators.