$$P(x,y) = 2D''(x,y) - D'(x,y) = \frac{1}{2}xC(x^4,y^4) + \frac{1}{2}xE^2(x,y)C(x^2,y^2) + xyE'(x,y)E(x,y)C(x^2,y^2) \quad (13)$$

Care must be taken that $P(x,y)$ does not contain meso (R_+-R_- or R'_+-O-R'_-) forms.

Let $Q(x,y)$ be the generating function of achiral configurations of polyethers containing a labeled C-C bond or a C-O-C unit, in which meso forms are not included. Meso forms of polyethers containing one labeled C-C bond or a C-O-C unit can be achiral. Therefore

$$Q(x,y) = \sum_{i=2}^{\infty}\sum_{j=0}^{k} q_{ij}x^i y^j = \frac{1}{2}[E(x,y) - 1]^2 + \frac{1}{2}[E(x^2,y^2) - 1] + \frac{1}{2}y[E'(x,y) - 1]^2 + \frac{1}{2}y[E'(x^2,y^2) - 1] \quad (14)$$

Let $S(x,y)$ be the generating function of achiral configurations of R-R forms and of R′-O-R′ forms in which the meso forms are also not included. So

$$S(x,y) = [E(x^2,y^2) - 1] + y[E'(x^2,y^2) - 1] \quad (15)$$

Meso forms however are achiral and must be counted. Let $M(x,y)$ be the generating function of the meso forms of polyethers. Then

$$M(x,y) = \frac{1}{2}[C(x^2,y^2) - 1] - \frac{1}{2}[E(x^2,y^2) - 1] + \frac{1}{2}y[C'(x^2,y^2) - 1] - \frac{1}{2}y[E'(x^2,y^2) - 1] \quad (16)$$

and

$$G(x,y) = \sum_{i=1}^{\infty}\sum_{j=0}^{k} g_{ij}x^i y^j = P(x,y) - Q(x,y) + S(x,y) + M(x,y) = \frac{1}{2}xC(x^4,y^4) + \frac{1}{2}xE^2(x,y)C(x^2,y^2) + xyE'(x,y)E(x,y)C(x^2,y^2) - \frac{1}{2}[E(x,y) - 1]^2 + \frac{1}{2}[C(x^2,y^2) - 1] - \frac{1}{2}y[E'(x,y) - 1]^2 + \frac{1}{2}y[C'(x^2,y^2) - 1] \quad (17)$$

where g_{ij} is the number of achiral configurations of polyethers. Some results from this algorithm are given in Table III.

REFERENCES AND NOTES

(1) Slanina, Z. *Contemporary Theory of Chemical Isomerism*; D. Reidel Publishing Co.: Dordrecht, 1986.

(2) Fujita, S. *J. Math. Chem.* **1990**, *5*, 99–121.

(3) Fujita, S. *Theor. Chem. Acta* **1990**, *77*, 307; **1989**, *76*, 247.

(4) Muller, W. R. *J. Comput. Chem.* **1990**, *11*, 223.

(5) Wang, J.; Wang, Q. *Tetrahedron* **1991**, *47*, 2969.

(6) Balaban, A. T. *Chemical Applications of Graph Theory*; Academic Press: London, 1976; pp 25–62.

A Canonical Representation of Multistep Reactions

RAÚL E. VALDÉS-PÉREZ

School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Received May 28, 1991

A canonical representation of multistep reactions or reaction networks is introduced. This representation has been applied to the efficient generation of reaction pathways for computer-assisted elucidation. Other potential applications are to chemical information systems and to intelligible depiction of pathways in publications. An algorithm is presented that canonicalizes a multistep reaction from a specification of the starting materials.

INTRODUCTION

An ideal canonical representation of a multistep reaction or network would have the following consequences. First, it would facilitate information retrieval within chemical databases that store information on multistep reactions. Second, it would enhance the intelligibility of multistep reactions as they are depicted in publications. Third, it would find application in computer algorithms that generate reaction networks for theoretical investigations or for modeling a chemical reaction.

This paper proposes a canonical network representation that can serve the three purposes above. The canon was discovered in the context of the design of a program to generate pathway hypotheses for building models of a chemical reaction. In this application there was a critical need to avoid redundant generation of multiple pathways that are identical except for permutations among steps and among reactants and products within single steps.

A desirable feature of the canon is that a given reaction network can easily be rearranged to follow the canon without rewriting the network in another representation, e.g., as a matrix. A second feature is that canonicalizing a network is conveniently done by hand. Lastly, canonicalized networks convey well the notion of flow, since steps appear in the network "soon" after their reactants become available; this will become clearer below.

Related Work. Brandt and von Scholley[1] describe a canonical numbering of reaction matrices in which a matrix represents the changes in the covalent bonds of molecules that occur as a result of a single reaction step. Fujita[1] likewise describes a method to represent the molecular–structural changes induced by individual reactions. However, our search in the literature for canonical representations of multistep reactions has not turned up any precursors to this work.

A CANONICAL REPRESENTATION

The canon consists of three conditions that an ordered reaction network must fulfill; these are discussed in what follows. For convenience, we assume that a given network is to be represented on a page, one step per line, so that it makes sense to speak of one step being higher or lower than another. For now, we also assume that the starting materials of the reaction network are known; later in the paper we describe an algorithm to infer a set of starting materials from a given network. Finally, we assume that each reaction step is irreversible; the handling of reversible steps will be discussed separately later.

After each condition below there appear two schematic networks that represent identical sets of steps, but which are depicted differently. In each case, the network on the left violates the condition, whereas the network on the right satisfies it. Also in each case the starting materials are A and B.

Condition 1: **Reactants Must Be Available.** Every reactant of any step in the canonical representation must either be a starting material or be formed at a higher step.

$A + B \rightarrow X$	$A + B \rightarrow X$
$X + Z \rightarrow T + A$	$B + B \rightarrow Y + Z$
$B + B \rightarrow Y + Z$	$X + Z \rightarrow T + A$
$Y + Z \rightarrow T$	$Y + Z \rightarrow T$

Step 2 on the left violates the condition.

Condition 2: **Network Labels Must Be Ordered.** First some preliminary definitions. The steps in the canonical representation of the given network are numbered from top to bottom starting with 1 and ending with the total number of steps R. The *index* of a network species is defined as 0 if the species is a starting material, otherwise as n if the highest step forming the species is numbered n. The *label* of a step is defined as an ordered set of integers in *descending* order. A step's label is built by collecting the indices of each reactant in the step, followed by sorting in descending order.

Now the condition can be stated: the step labels must appear in *ascending* order, treating the step labels as if they were ordinary numbers.

$(0\,0)\ A + B \rightarrow X$	$(0\,0)\ A + B \rightarrow X$
$(0\,0)\ B + B \rightarrow Y + Z$	$(0\,0)\ B + B \rightarrow Y + Z$
$(2\,2)\ Y + Z \rightarrow T$	$(2\,1)\ X + Z \rightarrow T + A$
$(2\,1)\ X + Z \rightarrow T + A$	$(2\,2)\ Y + Z \rightarrow T$

Step 3 on the left violates the condition.

Since the number of reactants per step can vary, the label of any step having less than the maximum found in the network can be filled out with −1's at the right until reaching the maximum, as in the following example, again having starting materials A and B.

$(0\ \ {-1})\quad A \rightarrow X + Y$

$(1\ \ 0)\quad B + X \rightarrow T$

An informal interpretation of this condition is that each step should appear in the canonicalized network as soon as its reactants become available, while waiting its turn for steps that were already enabled and should appear higher than it.

Condition 3: **Species Must Be Ordered.** Different network species may share a same index, and different steps may share a same label, hence an ordering criterion is needed to remove the residual ambiguity. Any criterion may be used, for example, a lexicographic order on a canonical string representation of molecules would be a good choice because of its universality. On the other hand, if the canonical network representation is only being used internally within an application program, then a more ad hoc criterion such as an arbitrary list may be adequate (and was used in the author's Ph.D. Thesis[3]).

The following completes the canonical network representation. Given two steps having identical labels, the reactants of the two steps are internally sorted according to the chosen criterion. Then the two steps are compared, possibly determining which step should appear higher. If the steps have identical reactant species, then the products are used to determine priority in a manner analogous to the case of the reactants. This final test must decide priority, otherwise the two steps are completely identical, which should not occur.

The following example uses the criterion of alphabetic order.

$(0\,0)\ B + B \rightarrow Y + Z$	$(0\,0)\ A + B \rightarrow X$
$(0\,0)\ A + B \rightarrow X$	$(0\,0)\ B + B \rightarrow Y + Z$
$(2\,1)\ X + Z \rightarrow T + A$	$(2\,1)\ X + Z \rightarrow T + A$
$(2\,2)\ Y + Z \rightarrow T$	$(2\,2)\ Y + Z \rightarrow T$

Step 1 on the left violates the condition, since A,B precedes B,B according to alphabetic order.

Uniqueness. Any network depiction satisfying the above three conditions cannot be represented in any different way while still satisfying the conditions. This is easily seen by realizing that when listing the network steps from top to bottom, at each point there is a unique next step to add to the list; adding any other step would violate one of the conditions.

We emphasize that uniqueness so far has depended on the assumptions that all steps are irreversible and that the starting materials are known.

Handling Reversible Steps. If some network steps are reversible, then the above canon does not determine a unique representation. For example, the following two depictions of a same network both fulfill the three previous conditions (a lexicographic order is used for condition 3, and the starting materials this time are A, B, and C).

$A + B \rightarrow W + X$	$A + B \rightarrow W + X$
$B + C \rightarrow Z$	$B + C \rightarrow Z$
$W + X \rightleftharpoons Z$	$X + X \rightarrow$ (any)
$X + X \rightarrow$ (any)	$Z \rightleftharpoons W + X$

This example shows that adjustments to a network of reversible steps that involve merely the selection of which species will be leftmost will not guarantee a unique representation.

To guarantee a unique depiction of a network having reversible steps, we will introduce below a constructive algorithm that determines at each decision point a unique step to add to the canonical representation.

Applications. The three conditions above have been applied to the problem of efficiently generating pathway hypotheses from data on observed products of a chemical reaction.[3] The canonical representation is used to avoid redundant generation of multiple pathways that are identical except for permutations, e.g., of individual steps. In this application, pathways are built up incrementally by adding a step at a time, so that the canonicalization serves to prohibit adding new steps that would render the partial pathway uncanonical. The algorithm for pathway generation will be the subject of a future paper.

Another possible application of canonicalization concerns judging the similarity of reaction pathways. For example, the pathways underlying oxidation of the various hydrocarbons could be compared on the basis of a canonical representation in a search for abstract similarities.

AN ALGORITHM TO CANONICALIZE A MULTISTEP REACTION

Given a set of reaction steps, each of which is possibly reversible, and a set of reaction starting materials, the following algorithm determines a unique, canonical representation of the network of steps.

```
canonicalize(steps, starting_materials)
  initialize present_species to equal the starting_materials.
  initialize remaining_steps to equal the set of all given irreversible steps.
  expand each reversible step into two irreversibles & add to remaining_steps.
  until remaining_steps is empty do all of the following:
     collect each step in remaining_steps whose reactants are a subset
       of the present_species.
     select from that collection the unique step S having the smallest label.
     add to the network the step S (depicted reversibly, if its reversed
       version exists).
     delete S (and its reversed version, if applicable) from remaining_steps.
     append the products of S to present_species.
  return the network.
```

We will examine now the worst-case computational complexity of a straightforward implementation of this abstract algorithm, where R is the number of given steps. There are two iterations over the **remaining_steps**: in the **until** header and at the first statement below the **until**. Also, the number of **present_species** is proportional in the worst case to the number of steps added so far. Hence, checking whether a

Chart I

$CH_3CHO \rightarrow CH_3 + CHO$	$CH_3CHO \rightarrow CHO + CH_3$	(0 -1)
$CHO \rightarrow CO + H$	$CHO \rightarrow CO + H$	(1 -1)
$H + CH_3CHO \rightarrow H_2 + CH_3CO$	$CH_3 + CH_3CHO \rightarrow CH_3CO + CH_4$	(1 0)
$CH_3 + CH_3CHO \rightarrow CH_4 + CH_3CO$	$2CH_3 \rightarrow C_2H_6$	(1 1)
$CH_3CO \rightarrow CH_3 + CO$	$CH_3CHO + H \rightarrow CH_3CO + H_2$	(2 0)
$2CH_3 \rightarrow C_2H_6$	$CH_3CO \rightarrow CH_3 + CO$	(3 -1)

step's reactants are a subset of **present_species** involves $O(R)$ comparisons. Since there are three nested iterations of complexity $O(R)$ each, the overall worst-case complexity is $O(R^3)$.

A straightforward implementation seems sufficient in practice, since the worst case is unlikely to occur and the number of reaction steps is usually modest. If a more ambitious implementation were needed, an analogy with the problem in computer science of determining and updating the applicability of rules in a production system would seem a fruitful source of algorithms and data structures[4] (in this analogy, an irreversible step corresponds to a production, and the present species correspond to working memory).

Inferring the Starting Materials Automatically. In general, a set of reaction steps does not imply a unique set of minimal starting materials, even if all of the steps are irreversible. The following network illustrates this, since A,B or W,X or Y,Z could be starting materials:

$$A + B \rightarrow W + X$$

$$W + X \rightarrow Y + Z$$

$$Y + Z \rightarrow A + B$$

This result implies that one cannot always depict uniquely a set of reaction steps unless the set of starting materials is specified. Of course, in some cases a unique set of minimal starting materials can be inferred. For example, the following truncated version of the previous network implies the unique minimal starting materials A and B:

$$A + B \rightarrow W + X$$

$$W + X \rightarrow Y + Z$$

Although we make no use of it in this paper, we will describe an algorithm to determine the complete set of possible, minimal, starting materials of a given reaction network. The algorithm makes use of a predicate **feasible**, defined as follows. A set of species are *feasible* starting materials of a multistep reaction, if every step in the network can eventually occur using those starting materials. For example, A and B are not feasible starting materials of the network $A + B \rightarrow X$. $C + X \rightarrow T$ because the second step can never occur. However, the set A, B, and C is feasible for this network. Here is the algorithm:

```
find_all_possible_minimal_starting_materials(steps)
  set all_species to the complete set of species appearing in the steps.
  set power_set to the power set of all_species.
  initialize possible_starting_materials to the empty set.
  for i from 1 to length(all_species) do the following:
    collect every set in power_set that has length i, is feasible in the steps,
      and is not a superset of a set already in possible_starting_materials.
    append the collection to possible_starting_materials.
  return the possible_starting_materials.
```

This algorithm is clearly worst-case exponential in the number of network species, since it iterates over the power set of species.

EXAMPLE

We will illustrate the canonicalization algorithm on a multistep reaction taken from Corio,[5] who depicts the reaction as shown on the left in Chart I. The canonicalized version is shown on the right, where we have ordered the individual species lexicographically; at the extreme right are the canonical step labels of the network on the right. The reader can judge whether the canonicalized network conveys slightly better the sense of mass flow through the multistep reaction.

CONCLUSION

This paper has introduced three conditions that suffice to determine a canonical representation of a multistep reaction, given knowledge of the starting materials and assuming that all steps are irreversible. In case some of the steps are reversible, there is an algorithm that produces a unique network depiction, again given knowledge of the starting materials.

In general, one cannot infer a unique set of minimal starting materials for a given multistep reaction, so the requirement that the starting materials be specified cannot appear to be relaxed. An algorithm is presented that infers the set of all possible, minimal starting materials of a given set of steps.

We have applied the canonical network representation within an algorithm to generate pathway hypotheses for building models of a reaction; this algorithm will be the subject of a future paper. The canonical representation proposed here may find use in chemical information systems. Finally, the canon could serve to present multistep reactions more uniformly in publications, as well as more intelligibly, since the use of conditions 1 and 2 above helps to convey a good sense of flow.

ACKNOWLEDGMENT

This paper is based on the author's Ph.D. Thesis in Computer Science at Carnegie Mellon University. This research was supported by an Air Force Laboratory Graduate Fellowship, administered by the Southeastern Center for Electrical Engineering Education (SCEEE).

REFERENCES AND NOTES

(1) Brandt, J.; von Scholley, A. An Efficient Algorithm for the Computation of the Canonical Numbering of Reaction Matrices. *Comput. Chem.* **1983**, *7*, 51–59.

(2) Fujita, S. A Novel Approach to the Linear Coding of Individual Organic Reactions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 128–137.

(3) Valdes-Perez, R. E. Machine Discovery of Chemical Reaction Pathways. Ph.D. Thesis, Carnegie Mellon University, 1990 (School of Computer Science, CMU-CS-90-191).

(4) Forgy, C. L. Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. *Artif. Intell.* **1982**, *19*, 17–37.

(5) Corio, P. L. Theory of reaction mechanisms. In *Topics in Current Chemistry: Relationships and Mechanisms in the Periodic Table*; Springer-Verlag: Berlin, 1989; Vol. 150.