Exploring Chemistry with SMT

Rolf Fagerberg[1], Christoph Flamm[2], Daniel Merkle[1], Philipp Peters[1]

[1]Department of Mathematics and Computer Science, University of Southern Denmark
[2]Institute of Theoretical Chemistry, University of Vienna

Introduction

This is to our knowledge the first time **Satisfiability Modulo Theories** (**SMT**) is used for modeling chemistry. We present an SMT-based tool which allows the chemist to explore chemical spaces in a highly flexible way, permitting a diverse array of chemical search problems to be attacked. This includes solving traditional chemical synthesis planning problems, and investigating chemical reaction patterns.
The latter includes the fairly new **Inverse Reaction Mechanism Problem** (**IRMP**): In the IRMP, it is assumed that an underlying reaction mechanism is known. The question to ask is, whether it is possible to map a different set of chemical reactions (**rules**) to this mechanism. This opens new perspectives on synthesis planning, using established knowledge to find similar, but new, chemical patterns.

Modeling Chemistry

The Reaction Mechanism

A reaction mechanism is a combination of elementary reactions. It defines how many and which molecules react in each of these single reactions. Chemists usually denote the elementary reactions of the reaction mechanism as follows:

$$\mathbf{m(A + B \rightarrow C + D),}$$

A, B, C, D denote the molecules and **m** the multiplicity of the reaction in the mechanism (cmp. Fig. 6).
The *balance* **bal(v)** of molecule $\mathbf{v} \in \mathbf{V}$ in a reaction mechanism is defined as an integer number indicating its net production or consumption over the entire synthesis:

$$\mathbf{bal(v)} = \sum_{\mathbf{e} \in \mathbf{E}} \mathbf{m_e}(\mathbf{1_{e^+}(v) - 1_{e^-}(v)})$$

where $\mathbf{1}_\alpha$ is the multiplicity function on the multiset α. If $\mathbf{bal(v) < 0}$, **v** is a reactant of the overall synthesis, if $\mathbf{bal(v) > 0}$, **v** is an end product. If $\mathbf{bal(v) = 0}$, either molecule **v** does not take part in the synthesis, or is produced and consumed in equal amount during the synthesis.
The *overall reaction* of a reaction mechanism is defined by summing up the two sides of all reactions (including multiplicities) in the mechanism, cancelling out equal amounts of identical molecules appearing on both sides.

The Molecules

Each molecule is represented by a **vector of functional groups**. Position **i** in molecule **A**'s vector provides the number of occurrences of the functional group $\mathbf{x_i}$ in **A**. There is one global set of functional groups, which is determined by the user, based on the used chemistry. I.e., what functional groups are deemed relevant to model the considered molecules and reactions (cmp. Fig. 1). This vector representation neglects the spatial structure, i.e., only the minimal number of occurrences of a functional group is noted, not its position(s) in the molecule.

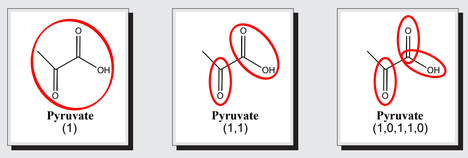

Figure 1: Possibilities of modeling molecules as vectors of functional groups

The Chemical Reactions

We here model an elementary reaction by its change of the number of occurrences of the functional groups of the reactants, i.e., the change of their vector representations. In addition, preconditions are specfied, which must hold for the reactants vector. However, in chemistry, it may also be necessary that a specific functional group appears in a reactant for the reaction to take place. We call such a modeling of a reaction a **rule**.

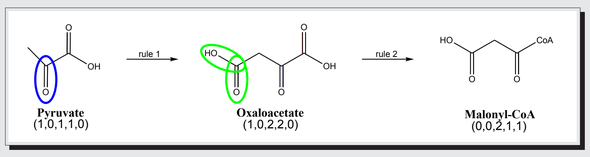

Figure 2: precond(pyr) = (1,0,0,0,0), change(pyr) = (0,0,1,1,0)

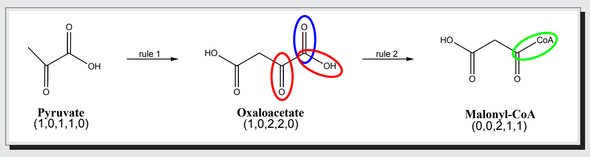

Figure 3: precond(oxa) = (1,0,1,1,0), change(oxa) = (-1,0,0,-1,1)

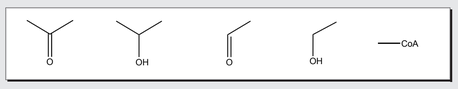

Figure 4: The functional groups in the vectors, in the order as used in Fig. 2 and 3

SMT-Approach

Definitions

```
;declaring molecules, functional groups and reactions
(declare-datatypes () ((MOL   A C D E)))
(declare-datatypes () ((SUB   a b c d e f)))
(declare-datatypes () ((REACT react1 react2 ... reactn)))

;declaring functions
(declare-fun NrOfGroups (MOL SUB) Int)
(declare-fun STOI (REACT MOL) Int)
(declare-fun ID (MOL MOL) Bool)
(declare-fun REACTANT (MOL) Bool)
(declare-fun PRODUCT (MOL) Bool)

;asserting if 2 molecules in mechanism are in the
;same equivalence class, then foreach functional group
;the number has to be identical
(assert (forall ((mol MOL)(mol2 MOL))
     (=>  (=(ID mol mol2) true)
     (forall ((sub SUB))
          (= (NrOrGroups mol sub) (NrOrGroups mol2 sub))
            ))))
```

Constraints to a rule

```
(assert (and
     ; stoichiometry constraints:
     ; (the stoi. coeff. of A needs to be the negative of
        D, etc.)
     (< (STOI react1 A) 0)
     (= (STOI react1 A) (STOI react1 C))
     (= (STOI react1 A) (- (STOI react1 D)))
     (= (STOI react1 D) (STOI react1 E))
     (or  (and
          ;preconditions
            (>= (NrOrGroups A a) 1)  (>= (NrOrGroups A b)
               1)
            (>= (NrOrGroups A d) 1)  (>= (NrOrGroups C e)
               1)
          ;changes made to A, which results in D
            (= (NrOrGroups D a) (- (NrOrGroups A a) 1))
            (= (NrOrGroups D b) (- (NrOrGroups A b) 1))
            (= (NrOrGroups D c) (+ (NrOrGroups A c) 1))
            (= (NrOrGroups D d) (- (NrOrGroups A d) 1))
            (= (NrOrGroups D e) (NrOrGroups A e))
            (= (NrOrGroups D f) (NrOrGroups A f))
          ;changes made to C, which results in E
            (= (NrOrGroups E a) (+ (NrOrGroups C a) 1))
            (= (NrOrGroups E b) (+ (NrOrGroups C b) 2))
            (= (NrOrGroups E c) (NrOrGroups C c))
            (= (NrOrGroups E d) (NrOrGroups C d))
            (= (NrOrGroups E e) (- (NrOrGroups C e) 1))
           (= (NrOrGroups E f) (NrOrGroups C f))
     ))))
```

Postprocessing

The solution output by the SMT-solver contains a rule mapping and a set of vector values, and is thus expressed in the vector representation of molecules. The spatial structure of molecules is neglected, implying that false positives can occur in the sense that some found solutions may not have corresponding real-world chemical reactions. Fig. 5 presents, how a vector of functional groups can represent several real world molecules.
An automated post-processing method allows us to filter our set of SMT-solutions, and retain only chemically viable solutions consisting of existing real-world chemical reactions. The method is based on the existence of large chemical databases of reactions, such as **KEGG** [Ogata et al. 1999].

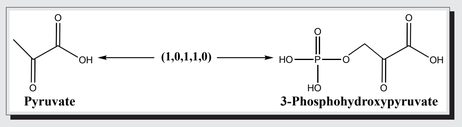

Figure 5: Vector representation to real molecules

Results: Reaction Mechanisms - Pentose Phosphate Pathway

In the **Pentose Phosphate Pathway** (PPP) are 5 molecules **Fructose-6-phosphate** molecules created out of 6 **Ribulose-5-phosphate** molecules.
Fig. 6 shows a set of 7 2-to-2 reactions on which a set of abstract rules were mapped to. The molecules within these rules consist of only one functional group (themselves), a change in a group is a change of the molecule. An overall reaction was provided. The multiplicities of reactions are not restricted and are part of the solution (Fig. 7).

$\mathbf{m_1}\times$	$\mathbf{A + B \rightarrow C + D}$
$\mathbf{m_2}\times$	$\mathbf{E + F \rightarrow G + H}$
$\mathbf{m_3}\times$	$\mathbf{I + J \rightarrow K + L}$
$\mathbf{m_4}\times$	$\mathbf{M + N \rightarrow O + P}$
$\mathbf{m_5}\times$	$\mathbf{Q + R \rightarrow S + T}$
$\mathbf{m_6}\times$	$\mathbf{U + V \rightarrow W + X}$
$\mathbf{m_7}\times$	$\mathbf{Y + Z \rightarrow Å + Ø}$
$\sum$:	$\mathbf{6 \cdot A + B \rightarrow 5 \cdot H + K}$

Figure 6: Abstract mechanism

$2\times$	$\mathbf{A \rightarrow C}$
$2\times$	$\mathbf{A + C \rightarrow D + E}$
$2\times$	$\mathbf{D + E \rightarrow G + H}$
$2\times$	$\mathbf{A + G \rightarrow D + H}$
$1\times$	$\mathbf{D \rightarrow F}$
$1\times$	$\mathbf{D + F \rightarrow I}$
$1\times$	$\mathbf{B + I \rightarrow H + J}$
$\sum$:	$\mathbf{6 \cdot A + B \rightarrow 5 \cdot H + J}$

Figure 7: Concrete mechanism

$2\times$	(1 2 0 1 0 1)	$\rightarrow$	(0 2 0 1 1 1)
$2\times$	(1 2 0 1 0 1) +(0 2 0 1 1 1)	$\rightarrow$	(0 1 1 0 0 1) + (1 4 0 1 0 1)
$2\times$	(0 1 1 0 0 1) + (1 4 0 1 0 1)	$\rightarrow$	(0 2 1 0 0 1) + (0 3 0 1 1 1)
$2\times$	(1 2 0 1 0 1) + (0 2 1 0 0 1)	$\rightarrow$	(0 1 1 0 0 1) + (0 3 0 1 1 1)
$1\times$	(0 1 1 0 0 1)	$\rightarrow$	(1 0 0 1 0 1)
$1\times$	(0 1 1 0 0 1) + (1 0 0 1 0 1)	$\rightarrow$	(0 3 0 0 1 2)
$1\times$	$\mathbf{B}$ + (0 3 0 0 1 2)	$\rightarrow$	(0 3 0 1 1 1) + $\mathbf{J}$

Figure 8: Vector based reaction mechanism

In a second step, the mechanism of Fig. 7 serves as underlying pattern. A mapping of rules, derived from real chemical reactions, to the mechanism provides a possibly new synthesis mechanism. This is referred to be the Inverse Reaction Mechanism Problem. The result will be highly dependent on the given set of rules; in this example we focus on finding the PPP. The transition from Fig. 7 to 8 describes the mapping of rules of sugar chemistry.

$\mathbf{A = C_5H_{11}O_8P}$	$\mathbf{B = H_2O}$	$\mathbf{C = C_5H_{11}O_8P}$
$\mathbf{D = C_3H_7O_6P}$	$\mathbf{E = C_7H_{15}O_{10}P}$	$\mathbf{F = C_3H_7O_6P}$
$\mathbf{G = C_4H_9O_7P}$	$\mathbf{H = C_6H_{13}O_9P}$	$\mathbf{I = C_6H_{14}O_{12}P_2}$
$\mathbf{J = H_3PO_4}$		

Figure 9: Sum formulas of molecules, fitting to the solution vectors of Fig. 8

Based on the SMT solution (Fig. 8), the post-processing step searches for known real-world mechanisms. Fig. 9 presents possible sum formulas for molecules and Fig. 10 gives a hypergraph representation of the Pentose Phosphate Pathway.

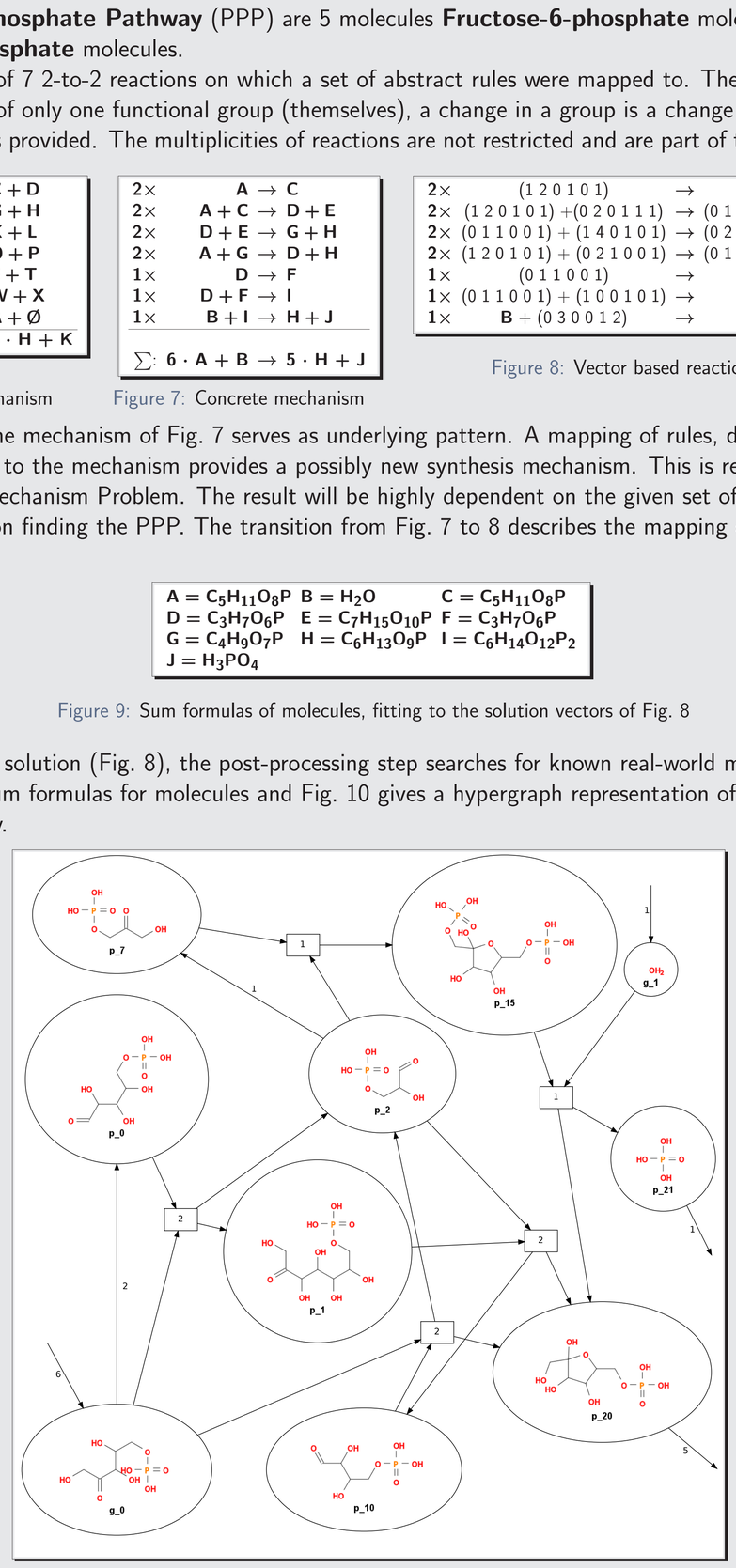

Figure 10: Hypergraph representation of Pentose Phosphate Pathway

Results: Reaction Pathways - Biosynthesis of 3-Hydroxypropanoate

The biosynthesis of **3-Hydroxypropanoate** (**3HP**) was investigated here as an instance of the Inverse Reaction Mechanism Problem. I.e., an abstract cascading mechanism ($\mathbf{A \rightarrow B \rightarrow C \rightarrow \ldots}$) and the chemistry (a set of 19 chemical rules with 10 functional groups) were given. The solution of this instance identifies possible pathways from **Pyruvate** to the desired product **3HP**. The vectors of the reactant and the product were predefined. The concrete reaction mechanisms provided by the SMT-Solver were post-processed using the KEGG database. 3 of the 27 found pathways from the post-processing are shown in Fig. 11. The functional groups marked dashed disappear in the subsequent reaction, the bold-marked functional groups are pre-conditional for the reaction to take place.

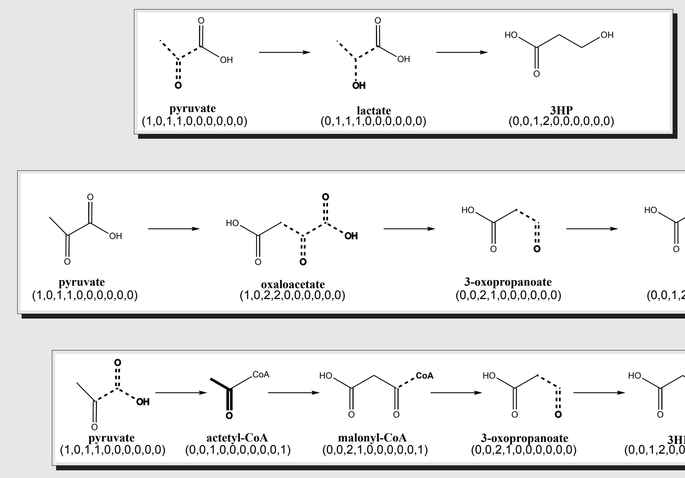

Figure 11: 2- to 4-step synthesis pathways to 3HP.

Additionally, by post-processing a concrete mechanism of length 2, a solution for the IRMP could be found that does not produce 3HP. A pathway was found that employs exactly the same reaction pattern as the synthesis of 3HP but is based on a different set of molecules. The alternative two-step-pathway synthezises 2-phospho-D-glycerate from 3-phosphohydroxypyruvate (cmp. Fig, 12).

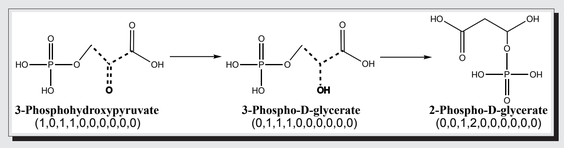

Figure 12: Alternative synthesis pathway

Selected References

- M. H. Todd. Computer-aided organic synthesis. *Chem Soc Rev*,34:247-266, 2005.
- E. J. Corey. General methods for the construction of complex molecules. *Pure Appl Chem*,14:19-38, 1967.
- H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27:29, 1999.