Atom Mapping with Constraint Programming

Martin Mann[1], Heinz Ekker[1], Peter F. Stadler[1−5], and Christoph Flamm[1]

[1]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria, [2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany, [3]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany, [4]Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany, and [5]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
{mmann,hekker,studla,xtof}@tbi.univie.ac.at

Abstract. The mass flow in a chemical reaction network is determined by the propagation of atoms from educt to product molecules within each of the constituent chemical reactions. The Atom Mapping Problem for a given chemical reaction is the computational task of determining the correspondences of the atoms between educt and product molecules. We propose here a Constraint Programming approach to identify atom mappings for "elementary" reactions. These feature a cyclic imaginary transition state (ITS) imposing an additional strong constraint on the bijection between educt and product atoms. The ongoing work presented here identifies only chemically feasible ITSs by integrating the cyclic structure of the chemical transformation into the search.

1 Introduction

For chemical reactions often only educt and product molecules are known. The underlying mechanism, i.e., the chemical bonds that are broken or newly formed to transform the educt molecule into the product, is unknown. Equivalently, it is unknown which atom in the educt corresponds to which atom in the product. Traditionally, such knowledge is gained by isotope labeling experiments, that is, by substituting certain atoms in an educt molecule with chemically identical but physically recognizable variants that are then identified in the product molecules by means of NMR or similar methods [25]. Such approaches produce a mapping between the atoms present in the educt and product molecules and thus identify the chemical bonds that have changed. Knowledge of the reaction mechanism enables for instance the analysis and identification of metabolic pathways [3] or the classification of reactions and enzymes in terms of the mechanisms [19, 20].

The *in silico* identification of correct atom mappings is computationally non-trivial and an extensively studied task. First approaches analyzed the adjacency information within educts and products [9] using branch-and-bound search following the Principle of Minimal Chemical Distance [17] or used topological indexing based on Morgan numbering [21]. More recent methods operate directly

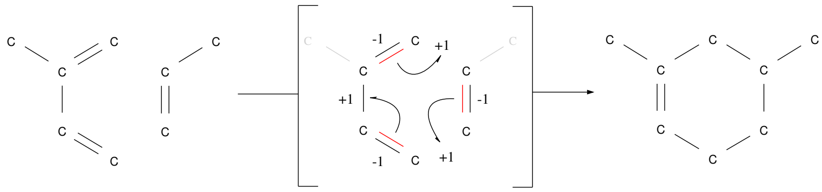

Fig. 1. Example of a Diels-Alder reaction. The ITS is an alternating cycle defined by the bonds that are broken (in red) and the bonds that are newly formed.

on graph representations of the molecules. For instance, searching for Maximum Common Edge Subgraphs (MCES) [8, 13, 14, 20, 23], an NP-hard problem, or the use of specialized energetic criteria [2, 18] allows for the identification of the static parts of the reaction and, subsequently, of the atom mapping. Another class of algorithms iteratively decomposes the molecules until only isomorphic sub-graphs remain [1, 4, 7] since it was shown by Akutsu that the MCES approaches fail for certain reactions [1].

Here, we propose a new approach to identify chemically feasible atom mappings given educt and product molecules as input. This approach makes explicit use of the observation that most reactions exhibit a cyclic transition state [16], i.e., the chemical bonds that are broken or formed are arranged in an alternating cycle. This class of mechanisms includes in particular all pericyclic reactions such as the Diels-Alder reaction, which is shown in Fig. 1 together with its transition state. We use this knowledge and focus on the identification of the cyclic *imaginary transition sub-graph* (ITS) because once identified the overall atom mapping is easily derived. For the identification of cyclic ITS candidates, constraint satisfaction problems are formulated for different cycle lengths. A fast graph matching approach is used successively to identify the overall atom mapping for each ITS solution. In the following, we will detail the problem, our constraint programming approach to identify the cyclic ITS, and how to extend an ITS candidate to a complete atom mapping for the chemical reaction.

2 Problem Definition

Given are two sets of molecules, the educts and products of a chemical reaction, each with n atoms. Both educts and products are represented by a single, not necessarily connected, undirected graph denoted $I = (V_I, E_I)$ for educts/input and $O = (V_O, E_O)$ for products/output. Each molecule corresponds to a connected component. Nodes in a molecule graph represent atoms labeled with the respective atom type $l(x)$. Following the principle of mass conservation it follows $|V_I| = |V_O|$. Edges encode covalent chemical bonds between atoms. More precisely, it is often convenient to use a multi-graph representation, in which each bonding electron pair is represented as an edge. Non-bonding electron pairs thus correspond to loops in the multi-graph. For the CSP formulation it will be more convenient, however, to use an ordinary graph representation and to label each edge $\{x, y\} \in E_I \cup E_O$ with its bond order: single, double or triple bonds are represented by a single edge with labels 1, 2, or 3, respectively. The matrix elements $\mathcal{I}_{x,y}$ denote the number of shared bond electron pairs for the edge between the atoms x and y in the educt graph I, i.e., in practice $\mathcal{I}_{x,y} \in \{0, 1, 2, 3\}$. $\mathcal{O}$ is defined accordingly. If necessary, non-bonding electron pairs can be represented by the diagonal entries $\mathcal{I}_{x,x}$ and $\mathcal{O}_{y,y}$. Thus, the matrices $\mathcal{I}$ and $\mathcal{O}$ encode the adjacency information of the educt and product graphs, respectively.

Consider a function $\alpha : V_I \to V_O$ mapping the nodes of I onto the nodes of O and a matrix $\mathcal{Q}$ rows and columns indexed by V_I. Then we denote by $\mathcal{Q} \circ \alpha$ the matrix with entries $\mathcal{Q}_{\alpha(x),\alpha(y)}$ with rows and columns indexed by V_O. Thus $\mathcal{R}^\alpha = \mathcal{O} - (\mathcal{I} \circ \alpha)$ is well defined.

Definition. An *atom mapping* is a bijective mapping $m : V_I \to V_O$ such that

1. $\forall_{x \in V_I} : l(x) = l(m(x))$ (preservation of atom types)
2. $\mathcal{R}^m \vec{1} = 0$ (preservation of bond electrons)

The *reaction matrix* $\mathcal{R}^m$ encodes the imaginary transition state (ITS) [11, 15]. This definition of m is a slightly more formal version of the Dugundji-Ugi theory [9]. Our notation emphasizes the central role of the (not necessarily unique) bijection m. Since we consider I and O as given fixed input, the atom mapping m uniquely determines $\mathcal{R}^m$. The pair $(m, \mathcal{R}^m)$, furthermore, completely defines the chemical reaction. It therefore makes sense to associate properties of the chemical reaction directly with the atom map m.

Equivalently, the ITS can be represented as a graph $R = (V_R, E_R)$ so that E_R consists of the edges in I that are removed in O and the edges in O that were not present in I as well as the atom nodes $x \in V_R$ with at least one adjacent edge. Each edge $\{x, y\} \in E_R$ is labeled by the changes in bond order $\mathcal{R}^m_{x,y} \neq 0$. See Fig. 1 for an example. We note that in a slightly more general setting we can regard $R = (V_R, E_R)$ as a multi-graph consisting of all electron pairs that are formed or removed.

It is important to note that the existence of an atom mapping m as defined above does not necessarily imply that $\mathcal{R}^m$ is a chemically plausible ITS.

We say that two edges $\{x, y\}, \{y, z\} \in E_R$ in R are alternating if $\mathcal{R}^m_{x,y} + \mathcal{R}^m_{y,z} = 0$. A simple cycle in R of size $k > 2$ is given by the node sequence $(v_1, v_2, \ldots, v_k, v_1)$ with $v_i \in V_R$, $\{v_i, v_{i+1}\} \in E_R$, and $\forall i < j \leq k : v_i \neq v_j$. Such a simple cycle is called alternating if all successive edges as well as the ring closure $\{v_2, v_1\}, \{v_1, v_k\}$ are alternating.

Definition. An atom map m is *homovalent* if $\mathcal{R}^m_{xx} = 0$ for all $x \in V_R$. A homovalent reaction is *elementary* if its ITS R is a simple alternating cycle. Thus $\mathcal{R}^m_{x,y} \in \{-1, 0, +1\}$ holds for all elementary homovalent reactions.

In the following we outline a novel algorithm for finding atom maps for elementary homovalent reactions that is guaranteed to retrieve all possible mappings given $\mathcal{I}$, $\mathcal{O}$, and the atom labels $l(x)$ for $x \in V_I \cup V_O$.

Of course, not all I,O pairs that are educts and products of chemical transformation admit an atom mapping m with a homovalent elementary ITS. This

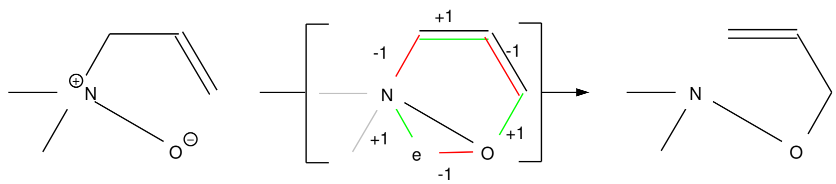

Fig. 2. The Meisenheimer rearrangement [22] transforms nitroxides to hydroxylamines. It does not admit a simple alternating cycle as ITS when molecules are represented as graphs whose vertices are atoms. An extended representation, in which the additional electron at the oxygen is treated a "pseudo-atom" can fix this issue. In such a representation an additional "charge separation" rule has to be introduced that allows an electron and a positive charge (here at the nitrogen in the product) to annihilate. This would disturb the bijectivity of m, however.

will in general be the case for multi-step reactions and for the so-called ambivalent reactions, in which the number of non-bonding electron pairs (and thus the oxidation number of atoms) changes in the course of a reaction. Fig. 2, for example shows an example of a reaction for which it is not possible to find a simple circular ITS using the encoding above. It appears to be possible to extend the formalism outlined above also to reactions with charged atoms and radicals. This is much less well understood, however, and will require a deeper theoretical analysis in the future.

3 Constraint Programming Approach

The central problem to find an elementary homovalent atom mapping is to identify the alternating cycle defining the ITS R given the adjacency information of the educts $\mathcal{I}$ and products $\mathcal{O}$. This can be done via solving the Constraint Satisfaction Problem (CSP) as presented below. Note, due to the alternating edge condition within the ITS, we have to consider rings with an even number of atoms only. In practice, the ITS of elementary homovalent reactions involves $|V_R| = 4$, 6, or 8 atoms.

A CSP for an ITS of size $k = |V_R|$ is given by the triple (X, D, C) defining the set of variables X, according domains D_i, and the set of constraints C to be fulfilled by any solution.

We construct an explicit encoding of the atom mapping using k variables representing the ring in I and another set for the mapped nodes in O, i.e., $X = \{X^I_1, \ldots, X^I_k\} \cup \{X^O_1, \ldots, X^O_k\}$ with domains $D^I_i = V_I$ and $D^O_i = V_O$.

To find a bijective mapping we have to ensure $\forall i \neq j : X^I_i \neq X^I_j$ and $\forall i \neq j : X^O_i \neq X^O_j$, i.e., a distinct assignment of all variables. To enforce atom label preservation we need arc consistency for $l(X^I_i) = l(X^O_i)$, i.e. we have to enforce $\forall e \in D^I_i : \exists p \in D^O_i : l(e) = l(p)$ as well as $\forall p \in D^O_i : \exists e \in D^I_i : l(p) = l(e)$. Analogously, homovalence is represented by $(\mathcal{I}_{X^I_i, X^I_i} - \mathcal{O}_{X^O_i, X^O_i}) = 0$. Due to the alternating ring condition, each atom can loose or gain at most one edge during a reaction. Thus, we can further constrain the variables with $|\,\mathrm{degree}(X_i^I) - \mathrm{degree}(X_i^O)| \leq 1$; where degree$(v)$ gives the out-degree of node v.

Finally, we have to encode the alternating cycle structure of the ITS in the mapping, i.e., for the sequence of bonds with indices 1-2-..-k-1. For all ring pair indices (i, j) we therefore require pairs with even index i to correspond the formation of a bond, i.e., we enforce $(\mathcal{O}_{X_i^O,X_j^O} - \mathcal{I}_{X_i^I,X_j^I}) = 1$, while all odd indices i are bond breaking $(\mathcal{O}_{X_i^O,X_j^O} - \mathcal{I}_{X_i^I,X_j^I}) = -1$ accordingly.

In order to avoid symmetric solutions, we introduce order constraints on the input variables: $(\forall i > 1 : X_1^I < X_i^I)$; where $X_i < X_j$ denotes $\exists(x, y) \in D_i \times D_j : x < y$ using e.g. an index order on the nodes. This ties the smallest cycle node to the first variable X_1^I and prevents the rotation-symmetric assignments of the input variables. Note, since we constrain the bond $(1, 2)$ to be a bond breaking $(\mathcal{O}_{X_1^O,X_2^O} - \mathcal{I}_{X_1^I,X_2^I} = -1)$, the direction of the cycle is fixed and all direction symmetries are excluded as well.

Although the CSP is defined above for domains of nodes $v \in V_I \cup V_O$ it can be easily reformulated using integer encodings of the atom nodes allowing the application of standard constraint solvers such as `Gecode` [12]. This enables the use of efficient propagators for most of the required constraints, such as the algorithm of Regin [24] for globally unique assignments. Only a few binary constraints, e.g. to ensure atom label preservation or the ring bonding, require a dedicated implementation, which poses no serious obstacles.

All solutions for this CSP are chemically valid ITS candidates. In order to check whether or not a true ITS is found we have to ensure that the remaining atoms, i.e., those that do not participate in the ITS, can be mapped without further bond formation or breaking. This is achieved using a standard graph matching approach as discussed in the following.

4 Overall Atom Mapping Computation

Given the CSP formulation from above, we can enumerate all valid ITS candidates for all possible ring sizes $k \in \{4, 6, 8\}$. For a CSP solution we denote with a_i^I and a_i^O the assigned values of the variables X_i^I and X_i^O, respectively. Once the ITS candidate is fixed, we can reduce the problem to a general graph isomorphism problem with a simple relabeling of the ITS edges. Thus, we derive two new adjacency matrices $\mathcal{I}'$ and $\mathcal{O}'$ from the original matrices $\mathcal{I}$ and $\mathcal{O}$, resp., as follows: For all ring pairs (i, j) within the ring sequence 1-2-..-k-1, we change the corresponding adjacency information to a unique label using $\mathcal{I}'_{a_i^I,a_j^I} = \mathcal{O}'_{a_i^O,a_j^O} \in \{f, b\}$ encoding if a bond between the mapped ITS nodes is formed (f) or broken (b). All other adjacency entries are kept the same as in $\mathcal{I}$ and $\mathcal{O}$, respectively.

Given these updated, "ITS encoding" adjacency matrices $\mathcal{I}'$ and $\mathcal{O}'$, the identification of the overall atom mapping m reduces to the graph isomorphism problem based on $\mathcal{I}'$ and $\mathcal{O}'$. Thus, all exact mappings of $\mathcal{I}'$ onto $\mathcal{O}'$ are valid atom mappings m of an elementary homovalent reaction, since the encoded ITS respects all constraints due to the CSP formulation. The graph matching can be done using fast and efficient algorithms as the VF2-algorithm [6], which is among the fastest available [5]. Since almost all molecular graphs are planar, even faster algorithms [10] might be applicable as well.

5 Discussion

We have presented here a novel constraint programming approach to identify atom mappings for elementary homovalent reactions. The incorporation of the cyclic ITS structure within the search ensures the chemical feasibility of the mapping that is not guaranteed by standard approaches that attempt to solve Maximum Common Edge Subgraph Problems [1].

The formulation of the CSP using only the atoms involved in the ITS results in a very small CSP that can be solved efficiently. Thus, it is well placed as a filter for ITS candidates for the subsequent, computationally more expensive graph matching approaches. While not described here, the CSP could be easily extended to find the entire atom mapping by introducing additional matching variables for all atoms participating in the reaction, all constrained to preserve atom label, node degree, and bond valence information. The solutions of such an extended CSP are the desired chemically feasible atom mappings m. This involves a much larger search space, however.

At present, we consider elementary homovalent reactions only, i.e., for reactions in which the transition state is an elementary cycle with an even number of atoms. The CSP formulation can be easily extended to odd ITS cycles ($k \in \{3, 5, 7\}$), but different ring layouts have to be considered. Furthermore, such reactions are not homovalent, i.e., at least one atom participating in the ITS is gaining or losing non-bonding electrons, which requires some moderate changes in the formulation of the constraints.

Constrain programming appears to be a very promising approach to solving atom mapping problems since it provides a very flexible framework to incorporate combinatorial constraints determined by the underlying rules of chemical transformations.

References

1. T. Akutsu. Efficient extraction of mapping rules of atoms from enzymatic reaction data. *J. Comp. Biol.*, 11:449–62, 2004.
2. J. Apostolakis, O. Sacher, R. Körner, and J. Gasteiger. Automatic determination of reaction mappings and reaction center information. 2. validation on a biochemical reaction database. *J. Chem. Inf. Mod.*, 48:1190–1198, 2008.
3. M. Arita. The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci. USA*, 106:1543–1547, 2004.
4. T. Blum and O Kohlbacher. Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *Journal of Computational Biology*, 15:565–576, 2008.

5. L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. Performance evaluation of the VF graph matching algorithm. In *Proceedings of the 10th International Conference on Image Analysis and Processing*, ICIAP '99, page 1172. IEEE Computer Society, 1999.
6. L.P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1367–72, 2004.
7. J. D. Crabtree and D. P. Mehta. Automated reaction mapping. *J. Exp. Algor.*, 13:1.15–1.29, 2009.
8. M. J. L. de Groot, R. J. P. van Berlo, W. A. van Winden, P. J. T. Verheijen, M. J. T. Reinders, and D. de Ridder. Metabolite and reaction inference based on enzyme specificities. *Bioinformatics*, 25(22):2975–83, 2009.
9. James Dugundji and Ivar Ugi. An algebraic model of constitutional chemistry as a basis for chemical computer programs. *Topics Cur. Chem.*, 39:19–64, 1973.
10. D. Eppstein. Subgraph isomorphism in planar graphs and related problems. In *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms*, SODA '95, pages 632–40. Society for Industrial and Applied Mathematics, 1995.
11. S Fujita. Description of organic reactions based on imaginary transition structures. 1. introduction of new concepts. *J. Chem. Inf. Comput. Sci.*, 26:205–212, 1986.
12. Gecode: Generic constraint development environment, 2007. Available as an open-source library from `www.gecode.org`.
13. M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Heuristics for chemical compound matching. *Genome Informatics*, 14:144–53, 2003.
14. M. Heinonen, S. Lappalainen, T. Mielikäinen, and J. Rousu. Computing atom mappings for biochemical reactions without subgraph isomorphism. *J. Comp. Biol.*, 18:43–58, 2011.
15. J B Hendrickson. Comprehensive system for classification and nomenclature of organic reactions. *J Chem Inf Comput Sci*, 37:852–860, 1997.
16. Rainer Herges. Organizing principle of complex reactions and theory of coarctate transition states. *Angewante Chemie Int Ed*, 33:255–276, 1994.
17. C. Jochum, J. Gasteiger, and I. Ugi. The principle of minimum chemical distance (PMCD). *Angew. Chem. Int. Ed.*, 19:495–505, 1980.
18. R. Körner and J. Apostolakis. Automatic determination of reaction mappings and reaction center information. 1. the imaginary transition state energy approach. *J. Chem. Inf. Mod.*, 48:1181–1189, 2008.
19. M. Kotera, Y. Okuno, M. Hattori, S. Goto, and M. Kanehisa. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, 126:16487–16498, 2004.
20. M. Leber, V. Egelhofer, I. Schomburg, and D. Schomburg. Automatic assignment of reaction operators to enzymatic reactions. *Bioinformatics*, 25:3135–3142, 2009.
21. M. Lynch and P. Willett. The automatic detection of chemical reaction sites. *Journal of Chemical Information and Computer Sciences*, 18:154–159, 1978.
22. Jakob Meisenheimer. Über eine eigenartige Umlagerung des Methyl-allyl-anilin-N-oxyds. *Chemische Berichte*, 52:1667–1677, 1919.
23. J. W. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Computer-Aided Mol. Design*, 16:521–33, 2002.
24. J.-C. Regin. A filtering algorithm for constraints of difference. In *Proceedings of the 12th National Conference of the American Association for Artificial Intelligence*, pages 362–367, 1994.
25. W. Wiechert. ^{13}C metabolic flux analysis. *Metabolic Engineering*, 3:195–206, 2001.