A Critical Evaluation of Automatic Atom Mapping Algorithms and Tools

Nuno Osório[1(✉)], Paulo Vilaça[1,2], and Miguel Rocha[1]

[1] Centre of Biological Engineering, University of Minho, Campus de Gualtar, Braga, Portugal
nuno.m.c.osorio@gmail.com

[2] SilicoLife-Computational Biology Solutions for the Life Sciences, Braga, Portugal

Abstract. The identification of the atoms which change their position in chemical reactions is an important knowledge within the field of Metabolic Engineering. This can lead to new advances at different levels from the reconstruction of metabolic networks to the classification of chemical reactions, through the identification of the atomic changes inside a reaction. The Atom Mapping approach was initially developed in the 1960s, but recently suffered important advances, being used in diverse biological and biotechnological studies. The main methodologies used for atom mapping are the Maximum Common Substructure and the Linear Optimization methods, which both require computational know-how and powerful resources to run the underlying tools.

In this work, we assessed a number of previously implemented atom mapping frameworks, and built a framework able of managing the different data inputs and outputs, as well as the mapping process provided by each of these third-party tools. We evaluated the admissibility of the calculated atom maps from different algorithms, also assessing if with different approaches we were capable of returning equivalent atom maps for the same chemical reaction.

Keywords: Metabolic engineering · Chemical reactions · Atom mapping algorithms · Open-source software · Maximum common structure

1 Introduction

Cell metabolism is composed of chemical reactions which are catalysed by enzymes responsible for transforming the nutrients uptaken by the cell into energy and cellular building blocks. When needed, the cell uses its anabolic pathways to produce essential macromolecules, from energy and cellular building blocks, maintaining its regular behaviour [1].

Glimpsing the cells as industrial factories, the raw materials prices persistent climbing, and the reduction of their reserves, take researchers to build models which help to understand and optimize cellular systems (such as genetically

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_31

altered microorganisms) to produce native and non-native high-value industrial compounds like biofuels, antibiotics or aminoacids [2,3]. These approaches, largely applied in industry, help Metabolic Engineering to solve problems like tracing metabolic pathways from a metabolite A to a metabolite B [4], analysing the conservation of metabolites in metabolic networks [5], calculating all possible paths inside a metabolic network, from the initial to the goal atom, classifying chemical reactions (e.g. assigning EC numbers to enzymes) [6] or identifying which atoms are preserved.

All these applications have a common approach, crucial to accomplishing their goals: in a chemical reaction, performing the matching of its reactants' and products' atoms. This correspondence, called Atom Mapping, allows a correct atom trace of the desired reaction, identifying what are the changes between the reactants and products. Atom Mapping assigns a different index (integer number) to each atom from the reactions' substrates and tries to map these atoms onto the products, thus assigning them the same index. With this information, it is possible to determine what are the changes performed by a reaction (catalysed by specific enzymes). In other words, the atom mapping procedure identifies which are the broken/formed bonds or which bond's change their order [7].

The atom mapping approach allows diverse uses and applications, for instance, in the reconstruction of metabolic networks, which represents the atom level of the pathways, it will improve understanding of the metabolic network [8]. Atom mapping can also be used to do consistency checking of pathways [4], to analyse the conservation ratios of atoms in a reaction [5] and to classify chemical reactions based on their chemical transformation [6]. Also, to optimise drug design, it is necessary to predict which atoms, from the candidate drug, change during the chemical reaction. It may also be used to deduce the relevant pathways of a certain metabolite or a particular drug [9].

With this work, we aim to study strategies to collect atom mappings from databases, by analysing reaction databases and build a framework to extract atom mapping information; analyse methods for automatic atom mapping of reactions, by automatically extract atom mappings from published atom mapping software (API's); and evaluate comparison metrics of atom mapping, namely, evaluate against atom mapping from databases and other atom mapping algorithms.

Here, the comparison of four algorithms within four different frameworks was performed to verify the differences between each other, in terms of valid and equivalent maps assignment.

2 Methods

2.1 Data

A biological database was chosen to build our set of reactions, namely MetaCyc, from where 11575 reactions were collected, in which more than 90% had an associated atom map. The set contains balanced, not balanced, incomplete and

elemental reactions, with the objective of obtaining the most complete sample possible.

2.2 Algorithms

The group of tools and algorithms selected to perform the atom mapping process will be briefly described. Note that these tools use a combination of different algorithms to obtain their results.

MetaCyc. The atom mappings collected from the MetaCyc database [10] were calculated using the Minimum Weighted Edit-Distance metric (MWED) [11]. It uses a Mixed-Integer Linear Programming (MILP) approach, that identifies which bonds have more tendency to react. MWED finds multiple optimal maps, but with the particularity of having less symmetric maps, due to the introduction of bond weights which represent the tendency of a bond to break. Within the reactions, bonds can be broken, formed or change their type (e.g. single to double). The cost of a transformation is calculated taking into account the weights assigned to the bonds involved in the bond breaking/forming/changing process. The sum of the costs of all the changes in the chemical reaction results in the weight-edit distance of the reaction. This process only handles fully balanced biochemical reactions (reactions with the same number of atoms on both sides).

AutoMapper. AutoMapper performs the atom mapping based on Maximum Common Structure (MCS) and MILP algorithms. It provides some options on the mapping style: *Complete*: where all atoms are mapped; *Changing*: as the name indicates, only maps the atoms that have their bonds modified; *Matching*: only maps the atoms which do not have any bond modified.

Reaction Decoder Tool. The Reaction Decoder Tool (RDT) [12] calculates the atom maps for balanced and unbalanced reactions using MCS and MILP algorithms. It uses the Chemistry Development Kit (CDK) [13], a cheminformatics framework which offers diverse functionalities in molecular informatics (e.g. input/output features for SMILES or RXN files, rendering chemical structures, modelling, building chemical graphs - isomorphism checker or MCS searchers, fingerprinting or Nuclear Magnetic Resonance prediction, etc.).

ICMap. ICMap maps and determines the reaction's centres based on MCS and MILP approaches. Some chemical rules are applied to help the MILP approach finding the best possible map (e.g. breaking/forming hetero-atoms bonds are preferable to carbon-carbon bonds). It has some restrictions on the mapping process: it has a limit on the number of molecules in the reaction (no more than 15 on each side), on the molecules' size (no more than 100 non-hydrogen atoms) and on single atom mapping (single atoms without non-hydrogen bonds e.g. Phosphor or Sulphur). The ICMap cannot map a reaction in which all chemical bonds were broken and remade.

2.3 AtomMapper Framework

To ensure that the four algorithms followed the same analysis pipeline, it was implemented a framework, called AtomMapper Framework (AMF). AMF is 100% developed in Java^{TM} and joins different algorithms of atom mapping into a single program. It allows users to map their chemical reactions with different approaches and verify if their atom maps are equivalent or not.

AMF is also implemented as an abstraction that provides generic functionalities, which can be specified with the addition of new code. It is an universal, reusable software environment, which facilitates the development of additional applications. AMF defines which functions the user should implement (interface classes) and releases users of thinking in low-level details. It is especially useful for users wanting to test their own tools and algorithms, once it is easy to add new methods following the existing interfaces.

Figure 1 illustrates the two main step of the atom mapping process. On A the reading process and on B the atom mapping.

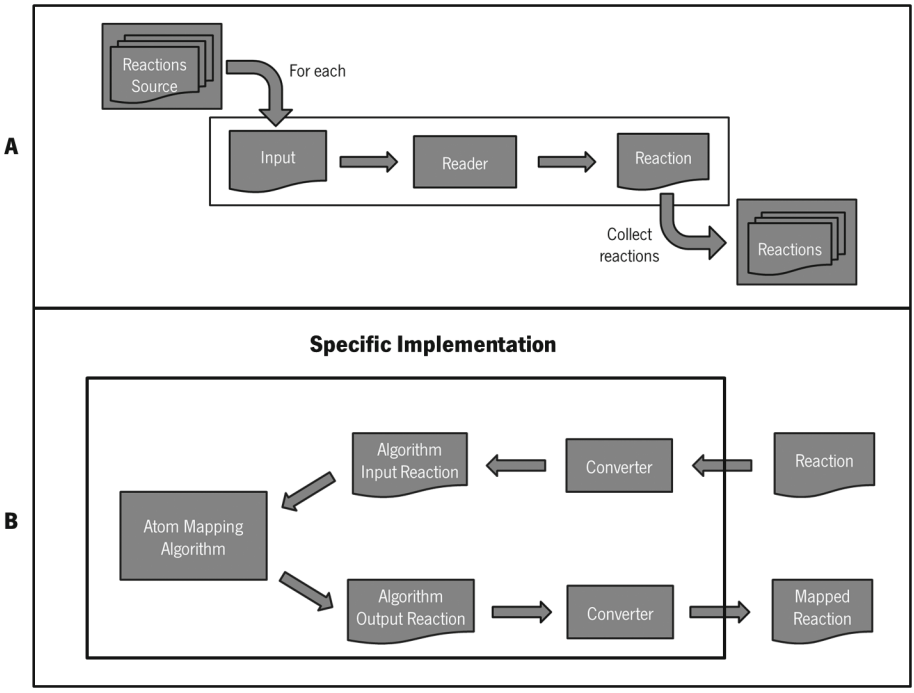

Fig. 1. Schematic representation of the AMF implementation philosophy. (A) It shows the reading of different types of input files to build a collection of reactions. (B) It represents the implementation needed to handle with each different algorithm input and output.

3 Results and Discussion

This section presents the results of the evaluation of different atom mapper algorithms. To do so, the Metacyc database was chosen as the reactions main set. It is constituted by 11575 different reactions, of which 10870 are already mapped, meaning that 705 reactions did not have a valid atom map on the Metacyc database.

It is important to differentiate a valid mapped reaction and an equivalent mapped reaction. A valid mapped reaction is a reaction where all atoms are assigned with a continuous numeration in both left and right sides, as well as both sides have the same elementary composition. An equivalent mapped reaction is a reaction for which different algorithms assigned the same atom linkage between left and right sides, i.e. all atoms in the right pair to the same atom in the left in both results, ignoring the individual numbers assigned to each atom (in one algorithm a right-left atom pair can have one label, while in the other algorithm the same pair has a different label, but they are the same pair).

The validation step will filter the reactions which have complete and plausible atom maps. This highlights the reactions for which their atom maps are comparable.

The first analysis of the mapping process was to consider the mappings provided by the four algorithms, checking the number of valid maps defined for each reaction. A total of 604 reactions were not mapped by any of the used atom mapping algorithms. This way, the number of admissible reactions decreased to 10971 valid reactions. Adding to this, the number of reactions with one or two valid maps was 1603, which is significantly lower when it is compared to the 9368 reactions with at least three valid maps. This indicates that over 80% of the reactions had three or four algorithms which were capable of assigning a valid map.

In terms of percentages, Metacyc presents 99.1%, AutoMapper 83.6%, RDT 99.8% and ICMap 40.9% of the whole set of reactions with at least one valid atom map. We can verify that the ICMap algorithm had the lowest percentage of valid maps, followed by AutoMapper algorithm, Metacyc database and RDT algorithm.

After analysing the behaviour of each individual algorithm, it was found that the MetaCyc and the RDT algorithms presented a similar number of reactions with valid maps assigned. The AutoMapper also presented a similar number, concerning the reactions with three and four valid atom maps, although, it did not have the same concordance with reactions containing one or two valid atom maps. About the ICMap, the numbers do not show very promising results, as its number of valid atom maps was less than half of the total reactions analysed and the concordance with the remaining algorithms was almost restricted to the reactions with four valid atom maps.

Figure 2A shows a Venn diagram with the intersection of the four sets of valid maps computed by each algorithm, assessing the reactions where pairs of algorithms are able to produce valid maps. Furthermore, the sum of all numbers of each oval form, gives the total number of valid reactions from each algorithm.

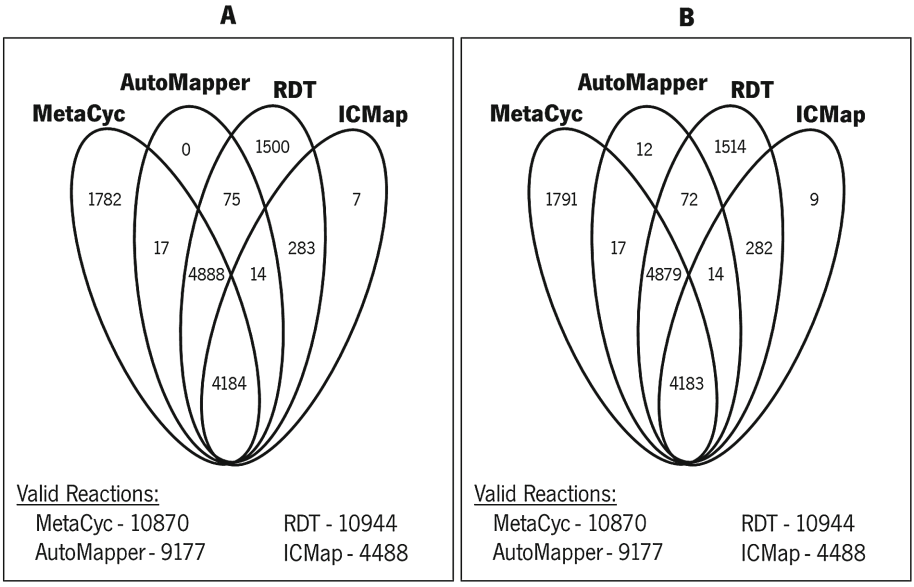

Fig. 2. Venn diagram showing the relations between the atom maps produced by the four algorithms: Metacyc, AutoMapper, RDT and ICMap sets, showing the intersection of reactions where each algorithm produced maps. (A) Counting of valid reactions, where intersections will show reactions where both algorithms produced valid maps; (B) Each intersection represents the number of reactions with equivalent atom maps assigned by the different algorithms. In both cases, if all numbers from an oval are added, it will represent the number of valid reactions on each set.

The four algorithms assign the same 4184 reactions as valid, corresponding to 38.1% of all reactions with at least one valid atom map (i.e. 10971 reactions). Nevertheless, if the ICMap algorithm is not considered in the analysis, the percentage of valid reactions raises from 38.1% to 82.7%, which represents 9072 reactions with three valid atom maps each. So, it may be admissible to say that the ICMap is pulling the number of common valid reactions down.

Having in mind that all analyses made so far do not imply that two valid maps, assigned to the same reaction, are equivalent, it is now time to check this. Considering all reactions from each set, and getting their atom maps, the comparison approach was performed to evaluate the atom maps equivalence.

Figure 2B shows the same representation from Fig. 2A, but now describing the comparison process. It represents the intersection of the four sets, and each intersection shows the number of reactions with equivalent maps between both algorithms. The intersection of the MetaCyc with the AutoMapper sets represents 9079 reactions with equivalent maps, which means 82.8%. The intersection of the AutoMapper with the RDT sets represents 9148 reactions, 83.4%, while 4479 reactions (40.8%) had equivalent atom maps calculated with the RDT and

the ICMap algorithms. Note that all percentages were calculated considering the 10971 reactions with at least one valid atom map.

When the intersection of more than two algorithms was analysed, the number of equivalent reactions tends reduce. The intersection of MetaCyc, AutoMapper and RDT represents 9062 reactions with three equivalent atom maps (82.6% of the valid reactions), still an interesting number. If it is now analysed the intersection of AutoMapper, RDT and ICMap, it joins 4197 reactions with three equivalent atom maps, with 38.3% of reactions. Finally, it was performed the intersection of all algorithms, and obtained 4183 reactions (38.1%), which were assigned with four equivalent atom maps for all analysed algorithms. Comparing the equivalence values with the ones from the validation, it is visible the high correlation between them. The ICMap was the algorithm with the lower percentage of valid atom maps. However, it was not significant in the comparison process, once it presented a similar percentage of equivalent atom maps.

Additionally, as referred before, 705 reactions did not have an atom map from the Metacyc database. Having into account that there are 604 reactions where none of the algorithms could provide a valid atom map, only 101 have the potential to have an atom map assigned by the remaining three algorithms. It was found that 14 reactions of those were assigned with four valid maps, all with four equivalent atom maps, which is a very interesting starting point to add new atom maps to the Metacyc database.

4 Conclusions

AMF enables the scientific community to explore the atom mapping process as well as, due its extensibility properties, be the base block to support additional implementation of atom mapping algorithms and comparison methods. It was shown that the studied algorithms had different behaviours: in the attribution of valid atom maps to this biological reactions set, they scaled from nearly 40% (ICMap) to almost 95% (RDT) of valid maps. However, despite this behaviour on the validation process, all algorithms, on the comparison step, had presented similar percentages of equivalent maps. Concerning the number of reactions which had four valid atom maps in the validation process, the majority had their atom maps considered equivalent, which proves the good precision of all tested algorithms. This may indicate that the atom mapping algorithms could assign different numbers to the atoms, but the matching of the left with the right reaction sides shows they are equivalent. The algorithms also had different techniques to assign the atom maps, which indicates that despite the theoretical differences, the result is somehow similar.

Acknowledgments. This study was partially supported by the Portuguese FCT under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by ERDF under the scope of Norte2020.

References

1. Heinonen, M., Lappalainen, S., Mielikäinen, T., Rousu, J.: Computing atom mappings for biochemical reactions without subgraph isomorphism. J. Comput. Biol. J. Comput. Mol. Cell Biol. **18**(1), 43–58 (2011)
2. Li, R., Townsend, C.A.: Rational strain improvement for enhanced clavulanic acid production by genetic engineering of the glycolytic pathway in Streptomyces clavuligerus. Metab. Eng. **8**, 240–252 (2006)
3. Rokem, J.S., Lantz, A.E., Nielsen, J.: Systems biology of antibiotic production by microorganisms. Nat. Prod. Rep. **24**, 1262–1287 (2007)
4. Arita, M.: Introduction to the ARM database: database on chemical transformations in metabolism for tracing pathways. In: Tomita, M., Nishioka, T. (eds.) Metabolomics, pp. 193–210. Springer, Tokyo (2005)
5. Hogiri, T., Furusawa, C., Shinfuku, Y., Ono, N., Shimizu, H.: Analysis of metabolic network based on conservation of molecular structure. Biosystems **95**, 175–178 (2009)
6. Yamanishi, Y., Hattori, M., Kotera, M., Goto, S., Kanehisa, M.: E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics **25**, i179–i186 (2009)
7. Fooshee, D., Andronico, A., Baldi, P.: ReactionMap: an efficient atom-mapping algorithm for chemical reactions. J. Chem. Inf. Model. **53**(11), 2812–2819 (2013)
8. Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., Palsson, B.O.: Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Natl. Acad. Sci. **104**, 1777–1782 (2007)
9. Blum, T., Kohlbacher, O.: Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. J. Comput. Biol. **15**, 565–576 (2008)
10. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., Karp, P.D.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. **42**, D459–471 (2014)
11. Latendresse, M., Malerich, J.P., Travers, M., Karp, P.D.: Accurate atom-mapping computation for biochemical reactions. J. Chem. Inf. Model. **52**(11), 2970–2982 (2012)
12. Rahman, S.A., Torrance, G., Baldacci, L., Cuesta, M.S., Fenninger, F., Gopal, N., Choudhary, S., May, J.W., Holliday, G.L., Steinbeck, C., Thornton, J.M.: Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics **32**, 2065–2066 (2016)
13. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E.: The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. J. Chem. Inf. Comput. Sci. **43**, 493–500 (2003)