The Minimum Information about a Tailoring Enzyme/Maturase data standard for capturing natural product biosynthesis - Supplementary Information

Mitja M. Zdouc[1]*, David Meijer[1], Friederike Biermann[1,2,3], Jonathan Holme[4], Aleksandra Korenskaia[5], Annette Lien[1], Nico L. L. Louwen[1], Jorge C. Navarro-Muñoz[1], Giang-Son Nguyen[4], Adriano Rutz[6], Anastasia Sveshnikova[7], Judith Szenei[8], Barbara Terlouw[9], Rosina Torres Ortega[1], Marc Feuermann[7], Alan J. Bridge[7], Justin J.J. van der Hooft[1,10], Tilmann Weber[8], Nadine Ziemert[5], Kai Blin[8] and Marnix H. Medema[1*]

1 Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

2 Institute of Molecular Bio Science, Goethe-University Frankfurt, Frankfurt am Main, Germany

3 LOEWE Center for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

4 Department of Biotechnology and Nanomedicine, SINTEF Industry, P.O. Box 4760 Torgard, N-7465, Trondheim, Norway

5 Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Germany

6 Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

7 SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, CH-1211, Geneva, Switzerland

8 The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark

9 Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, The Netherlands

10 Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa

*Corresponding authors: Mitja M. Zdouc (mitja.zdouc@wur.nl), Marnix H. Medema (marnix.medema@wur.nl).

Table of Contents

Supplementary methods

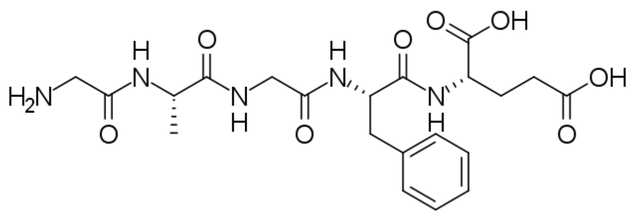
Protocol for the creation of MITE entries

This protocol describes the creation of MITE entries following the data standard schema as specified under:
https://github.com/mmzdouc/mite-preprint-reference/blob/main/schemas/mite/entry.json
Further, it assumes that the MITE entry is created for an enyzme that is already covered by a MIBiG entry.

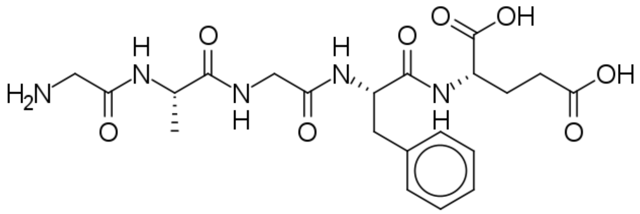
In MITE, the substrate specificity and enzymatic reaction are stored as a so-called **reaction SMARTS**, a line representation of the transformation of the substrate-product pair. The substrate (left) side of the reaction is used for substrate matching; the product (right) side of the reaction describes the instroduced changes. The most convenient way of creating such a reaction SMARTS is by drawing it in a chemistry drawing program. We recommend MarvinSketch by ChemAxon, which is free for individual, academic and non-commercial use and available for Windows, Mac, and Linux (https://download.chemaxon.com/marvin). The protocol was written assuming the use of MarvinSketch. Furthermore, MarvinSketch allows export of reaction **CXSMARTS**, which have expanded structure representation functions, explained in more detail in the respective section of the protocol. Of course, use of MarvinSketch is not mandatory and reaction SMARTS can be created in various ways (even written manually!). However, reaction SMARTS not created by MarvinSketch must be at least RDKit-compatible to be accepted by MITE.

1. From the publication, determine the substrate specificity, regioselectivity and the reaction that the enzyme performs. Make sure that the enzyme matches the one that is in the MIBiG entry. Either follow the tutorial steps below or watch the following video: https://youtu.be/WJDR_vQMY-s
    a. In your chemistry drawing program, start drawing the substrate (sub)structure that is going to be modified (in this protocol, all steps are shown using MarvinSketch). Also make sure to correctly depict the stereochemistry. Some enzymes are very specific with regard to their substrate, and large and specific

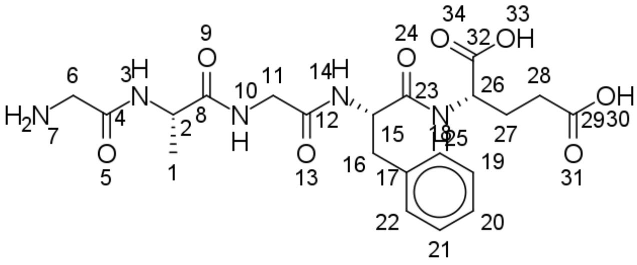
substrate/product structures need to be drawn. Other enzymes are very promiscuous and can therefore also work on a more generic substrate. This data can often be found in the paper, and it is essential to capture this information accurately. Below, a hypothetical peptide was drawn (GAXFE, where X indicates a non-specified amino acid, represented by a glycine due to its lack of residue).



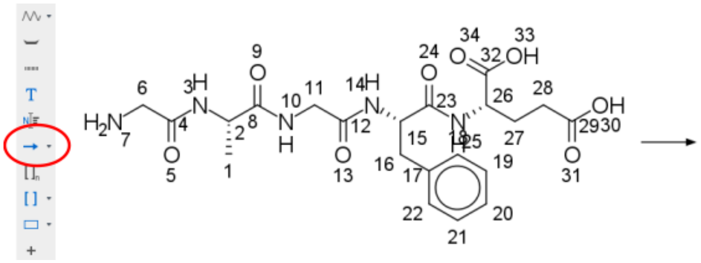
b. Next, the chemical structure must be turned from the Kekulé form into the aromatic form, else, the aromaticity information is not properly encoded. In MarvinSketch, select the structure, and in the menu, click **Structure** -> **Aromatic Form** -> **Convert to Aromatic Form**.



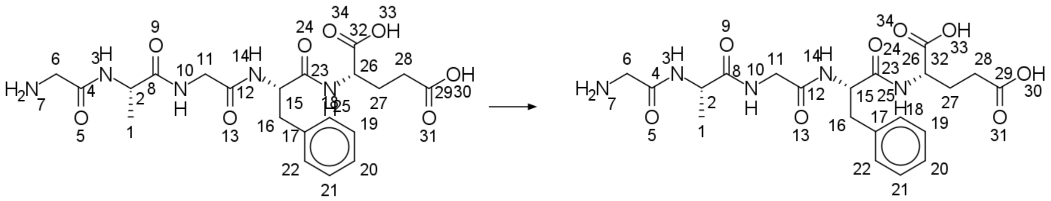
c. Next, map the atoms (assign index numbers to them). In MarvinSketch, select the structure, and in the menu, select **Structure** -> **Mapping** -> **Map Atoms**. This will assign an unique index number to all atoms. This mapping is completely arbitrary and does not represent IUPAC-conform enumeration.



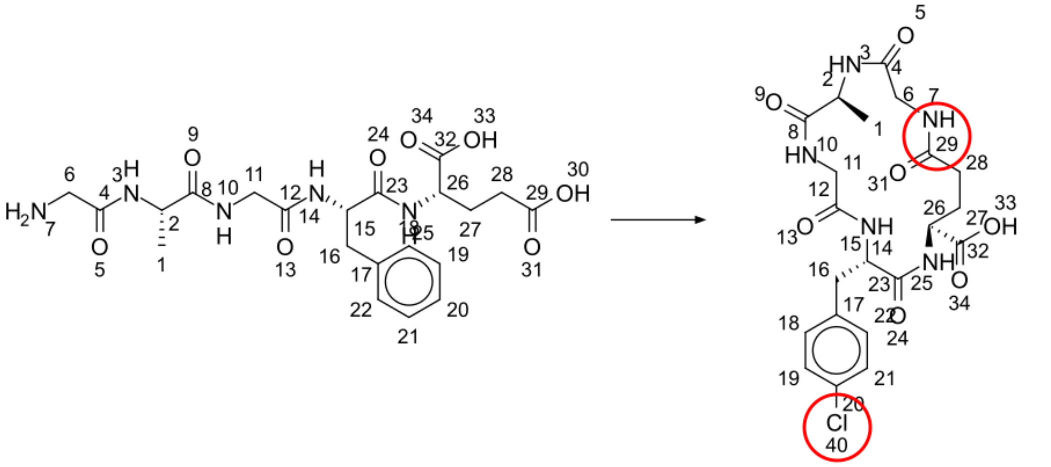
d. Next, copy the structure and draw the reaction arrow. In MarvinSketch, select the structure, and in the menu, select **Edit** -> **Copy** (or use the Ctrl+C key combination). Then, select the reaction arrow from the left-hand side toolbar, and draw an arrow from left to right.



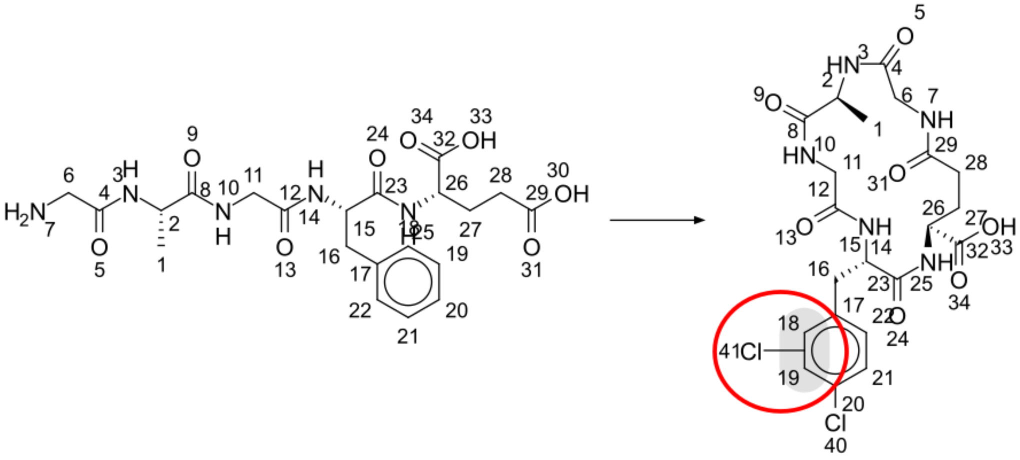
e. Next, paste your previously copied substrate on the product side (the right-hand side of the arrow).



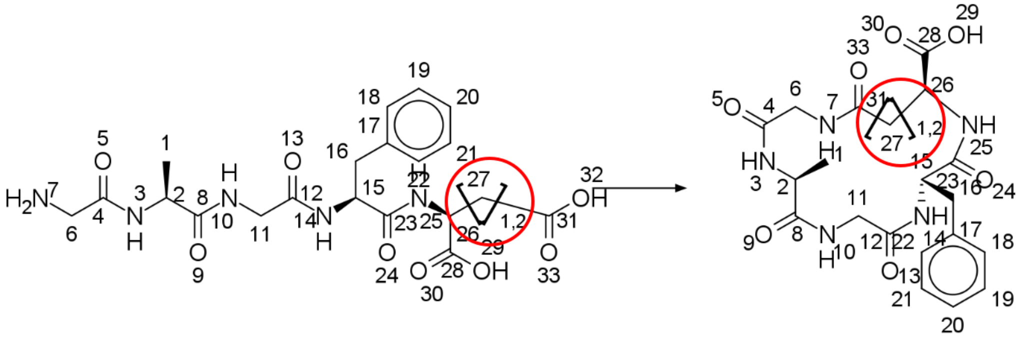
f. Next, draw the changes that are introduced by the enzymatic reaction. If this introduces any new atoms, they also have to be mapped. This can be done by selecting the newly added atom, right-clicking on the canvas, and selecting **Map -> M…** -> adding a so-far **unused** number. In our hypothetical example, we assume that the enzyme introduces both a **macrolactam cyclization** and a **chlorination**. The macrolactam cyclisation leads to a loss of water, which does not have to be accounted for. However, we have to map the new chlorine atom, and we give it the unused index *'40'*. If this concludes the drawn reaction, go on to **step i).**



g. (**Optional**: *Position Variation Bonds*): With **MarvinSketch**, we can use specific functionality to assign additional information to the reaction. One functionality is adding *Position Variation Bonds*. These specify variable locations for a functional group (e.g. variable chlorination on an aromatic ring). To add Position Variation Bonds select the atoms where the optional bond will be located. Then, in the menu, go to **Structure** -> **Add** -> **Position Variation Bond.** This will create a free floating bond and a gray border around the previously selected atoms, indicating the atoms to which the functional group will be applied. Now, add the desired atom or functional group to the outward side of the floating bond. As before, add atom mappings to the added atom/functional group. In our example, we want to indicate that there are multiple chlorinations on the phenol-ring: one in the *para*-position, and either one in the *ortho*- or *meta*-position.



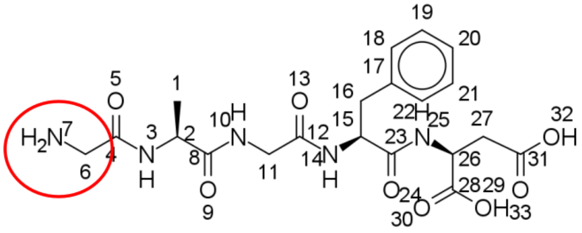
h. (**Optional**: *Frequency Variation*) With **MarvinSketch**, we can use specific functionality to assign additional information to the reaction. One functionality is adding *Frequency Variation*. This allows specifying certain repeating elements (e.g. an aliphatic carbon chain of variable length). To add Frequency Variation, select the atom(s) where the Frequency Variation label should be applied. Then, go to **Structure** -> **Group** -> **Frequency Variation**. In the pop-up menu, set *"type"* to *"Repeating unit with repetition ranges"*. Set *"repetition range"* to a fixed number of repetitions (e.g. **2** to repeat units twice) **or** a range (e.g. **2-3** to repeat units twice or thrice). Set *"Polymer repeat pattern"* to *"head-to-tail"* (no other pattern is supported) and *"bracket style"* to "*square[]*". In our example, we want to indicate that the macrocyclization can happen with the N-terminal glycine and either a C-terminal aspartic or glutamic acid (one or two C-atoms, respectively).



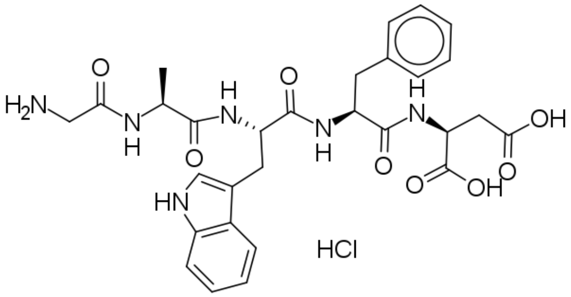
i. Finally, to export the reaction, select substrate, product, and the reaction arrow, right-click on canvas, and select **Copy As -> ChemAxon SMARTS (CXSMARTS),** which stores the SMARTS string in the clipboard. If you want to verify if the SMARTS was exported correctly, you can also try to paste it on the canvas. If everything went right, you should see the complete reaction.

2. After the reaction SMARTS/CXSMARTS was created, some additional information needs to be specified: the literature reference, evidence, and any database crosslinks; is the reaction iterative (i.e occurs multiple times exhaustively); does it contain Frequency Variation of Position Variation Bonds (see above); and finally, are there any explicit hydrogen atoms to specify. By default, SMARTS strings **do not** preserve hydrogen atoms. In case of ambiguous implicit hydrogens (e.g. primary vs secondary amine), specifying the expected number of hydrogens is important to prevent mismatches. For example, to indicate a primary amine, two hydrogens need to be explicitly specified. In our previous example, the primary amine of the N-terminal glycine needs to be specified; else, the SMARTS string would also match a pattern inside a longer peptide chain. Further, also the hydrogens on the alpha-carbon of Gly1 need to be specified explicitly; else, the pattern would match any amino acid. For the X (any) amino acid in position three, we do not specify any explicit hydrogens - this way, this position will match any amino acid. However, not all hydrogen must be specified in this way, as long as they do not introduce ambiguity. Note that only hydrogens on the substrate side (the "matching"

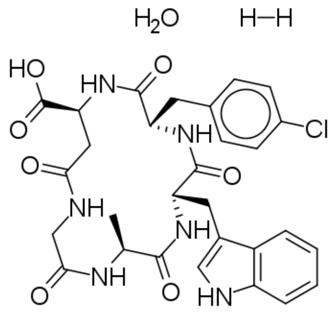
side) must be specified.



3. Next, to validate the reaction SMARTS, one or more **substrate - product pairs** need to be specified. These structures must be specified as SMILES strings. The substrate - product pair can be either a balanced, authentic reaction, or also just an example reaction (e.g. when the exact substrate and/or product is not known).
    a. Draw the substrate (or the substructure, if the exact substrate is not known) using any chemistry drawing tool. If the substrate - product pair should be balanced, also add any supplementary reaction partners or co-factors. Select all molecules, right-click on canvas, and select **Copy As** -> **ChemAxon SMILES (CXSMILES),** which stores the SMILES string in the clipboard. If you want to verify if the SMILES was exported correctly, you can also try to paste it on the canvas. If everything went right, you should see all substrates. In our example, we drew the peptide GAWFD, substituting the “any” amino acid in the third position with a tryptophan.



    b. Draw the product (or the substructure, if the exact product is not known) using any chemistry drawing tool. If the substrate - product pair should be balanced, also add any supplementary reaction partners or co-factors. Export the SMILES string as described in point 3a. Multiple products can be specified, if necessary.



   c. Next, some additional information needs to be specified: is the reaction balanced (i.e. is it stoichiometrically balanced); is the reaction authentic (i.e. not only substructures); is the reaction describing an intermediate (i.e. not the reaction step that leads to a mature product); any database cross-references and finally, a literature reference and the evidence for the reaction pair.
   d. If necessary, multiple reaction pairs can be described.

4. Next, the term that best describes the tailoring/maturation reaction must be specified. Multiple terms can be specified.
5. Additionally, some information about the tailoring enzyme/maturase must be specified. This includes the commonly used name of the protein and an optional description; cross-references to UniProt and/or NCBI GenPept as well as the primary literature reference. Also, any auxiliary enzymes that are co-forming the maturation machinery can be specified with name and database cross-references (e.g. in case of microcin J25, both McjB and McjC are required for the lasso peptide macrolactam formation, so for an entry of McjB, McjC needs to be specified as auxiliary enzyme, and vice versa).

Supplementary Figures

Figure S1

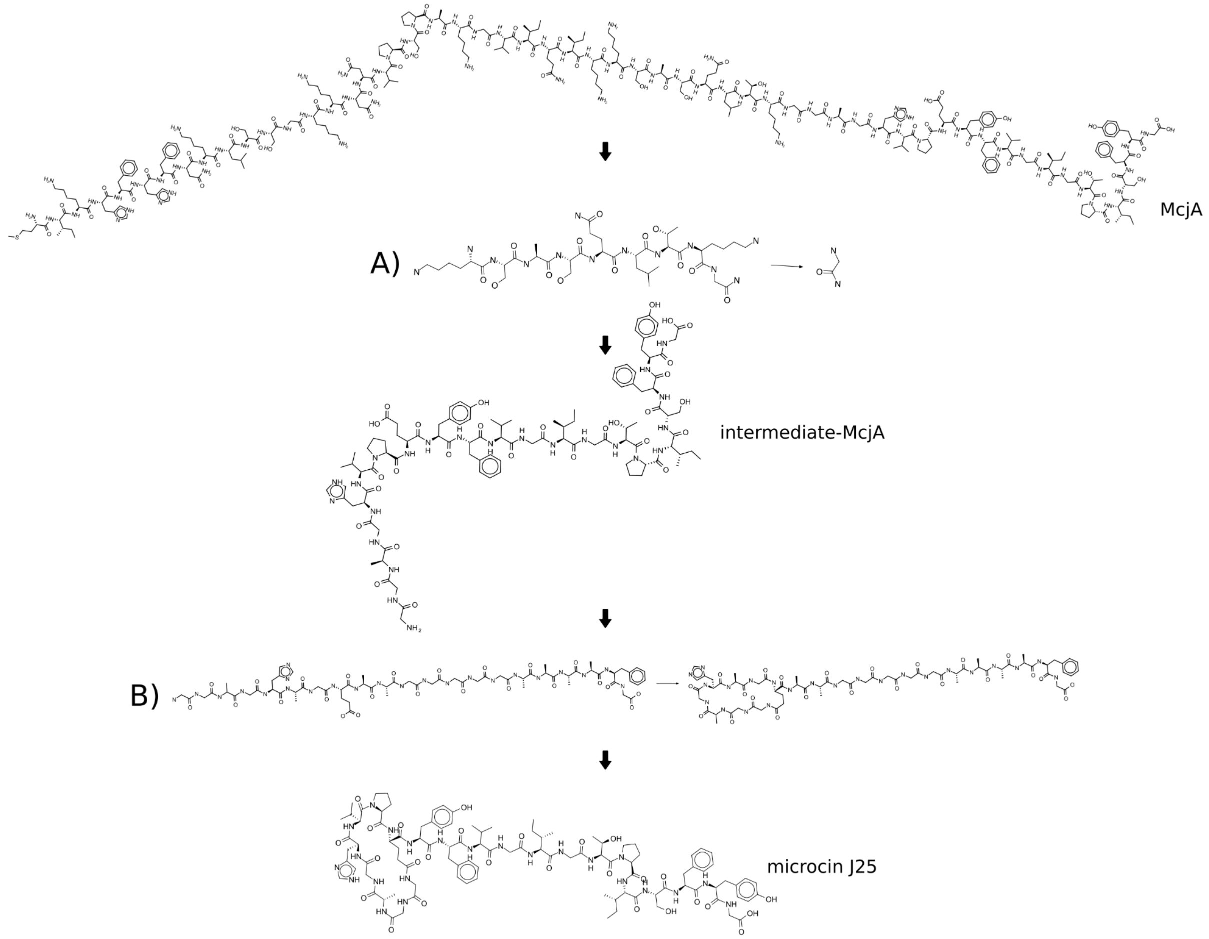


Figure S1: The biosynthetic pathway for the creation of microcin J25. The precursor peptide McjA is first transformed by reaction A, representing the reaction smarts contained in MITE0000004 (https://github.com/mmzdouc/mite-preprint-reference/blob/main/mcjB.json), leading to the cleaved linear intermediate. Next, the intermediate is transformed to microcin J25 by reaction B contained in MITE0000001 (https://github.com/mmzdouc/mite-preprint-reference/blob/main/mcjC.json).