Application of process-oriented case-based reasoning to the delimitation of mobile genetic elements in bacterial chromosomes

Toufik Hamadouche[1,2,*,†]

[1]Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
[2]Université de Lorraine, INRAE, DynAMic, F-54000 Nancy, France

Abstract

This PhD project investigates the use of process-oriented case-based reasoning (PO-CBR) to formalize and automate expert reasoning in the field of microbiology. The goal is to represent expert reasoning as structured process cases, where each case encodes a sequence of reasoning steps that can be reused to solve new problems. The method is applied to the problem of delimiting mobile genetic elements (MGEs) in bacterial genomes, a task that traditionally relies on manual biological expertise. By formalizing this reasoning using PO-CBR, we build a case base that can be reused to delimit different MGEs. This approach integrates adaptation mechanisms to handle failures and adjust reasoning. Initial application of the case base on 254 manually annotated MGEs in 124 bacterial genomes show a high success rate (96.8% of elements have been correctly delimited). This study demonstrates the feasibility of encoding biological expertise into a structured automated reasoning system that offers a reliable alternative to identify MGEs in bacterial genomes.

Keywords

PO-CBR, application to genomics, identification of mobile genetic elements, knowledge representation, expert reasoning

1. Introduction

Case-Based Reasoning (CBR [1]) is an approach that solves new problems by relying on past experiences, each of these experiences is represented by a case. In this research, I use Process-Oriented Case-Based Reasoning (POCBR [2, 3]), an extension of CBR in which these experiences are presented as sequences of steps. The aim of my thesis is to formalize the biological expertise in the form of episodes using the PO-CBR approach, thus capturing and reusing the knowledge of biologists in process cases. The final objective is to develop an explanatory system based on expert knowledge defined by biologists, which improves both the precision and efficiency of the task, while also ensuring the explainability of the decisions made.

This approach is applied to a microbiological problem: the precise delimitation (or identification) of Mobile Genetic Elements (MGEs) in bacterial chromosomes. MGEs are DNA segments capable of moving between bacteria through the mechanism of bacterial conjugation [4]. Their precise delimitation is important, as they frequently disseminate antibiotic resistance and virulence genes at high rates, with a significant impact on human health. Currently, this task relies on the manual expertise of biologists [5, 6]. Although previous work has addressed the automatic delimitation of MGEs [7], it often lacks precision in identifying the exact start and end positions of these elements in bacterial genomes and lacks explainability of the results obtained. A paper about this work has been accepted to ICCBR-2025.

ICCBR DC'25: Doctoral Consortium at ICCBR-2025, July, 2025, Biarritz, France
*Corresponding author.
toufik.hamadouche@univ-lorraine.fr (T. Hamadouche)
0009-0000-3594-0295 (T. Hamadouche)

2. Research Plan

2.1. Biological background on mobile genetic elements

Mobile Genetic Elements (MGEs) are DNA segments that are integrated into the bacterial genome. These elements can excise themselves from a bacterium, transfer, and then reintegrate into the genome of another bacteria. A microbiology expert has studied for years these elements, and through his observations, he was able to establish the knowledge required to delimit MGEs —that is, his reasoning experience to found boundaries of MGEs. This reasoning follows a sequence of steps based on genomic characteristics of these elements, such as the integrase gene (genes responsible for excision and integration of MGEs) and the target gene (or the insertion sites) where the element integrates.

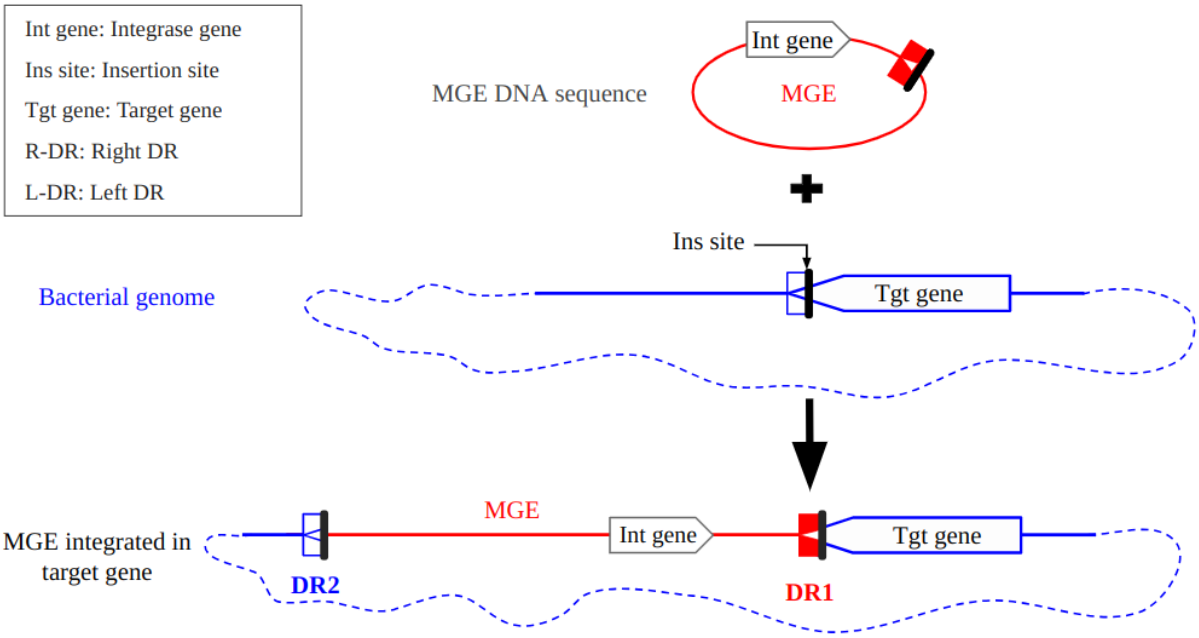

Figure 1: MGE integration process into the insertion site (ins site) of its target gene.

Figure 1 illustrates the integration process of an MGE into a bacterial genome. The MGE, carrying an integrase gene, integrates specifically at a target gene and generates two flanking direct repeats (DR1 and DR2) at the insertion site. This mechanism underlies the reasoning patterns that our system seeks to formalize.

2.2. Research Objectives

Our research aims to automate the delimitation of mobile genetic elements (MGEs) by formalizing the reasoning of a microbiology expert into reusable process cases using the PO-CBR approach. To achieve this, the following objectives have been defined:

1. **Capture expert reasoning** through the formalization of representative MGE delimitation cases as process cases, based on integration mechanisms involving integrase and target gene classes.
2. **Define a case representation model**, where each problem is described by structured attributes (e.g., integrase class, target gene class), and each solution is a sequence of reasoning steps applied by the expert.
3. **Implement a retrieval mechanism** to identify and reuse a relevant process case from the case base to solve a new delimitation problem.

4. **Develop and integrate adaptation mechanisms** that can adjust retrieved solutions when direct reuse fails (e.g., when DRs differ at the nucleotide level but match at the amino acid level).
5. **Evaluate the system** on a collection of previously annotated MGEs to measure its ability to reproduce expert-level delimitation results.
6. **Ensure explainability** at both system and user levels, by making the reasoning steps traceable and enabling the triggering of rule-based adaptations based on failure analysis.

The purpose of these objectives is to provide a system of reasoning that is structured and adaptable for precise MGE delimitation.

2.3. Approach / Methodology

Definition of the case base. The modeling of the microbiology expert's knowledge as process cases led to the establishment of the case base (CB). A *source case* (element of CB) is a pair $(\mathrm{x}^s, \mathrm{y}^s)$, where x^s denotes a source problem, and y^s is its corresponding solution. A delimitation problem x is formalized by three attributes:

- `idGenomes`: bacterial genome identifiers characterizing a bacterial strain;
- `targetGeneClass`: the target gene class where the mobile element is integrated
- `integraseClass`: the integrase class involved in the excision and integration processes of an MGE.

The solution y^s applied by the expert to solve the problem x^s corresponds to an ordered sequence of steps and knowledge about the MGE to be delimited —for example, the number of genes between the integrase and target genes, and the potential size of the MGE boundaries. The first two steps in the delimitation process consist of: (1) identifying the integrase genes specified in the attributes of the MGE delimitation problem within the bacterial genome and (2) locating the target gene, also specified in the problem attributes, within the bacterial genome. Currently, the process case solution y^s is implemented as a sequence of reasoning steps defined in Python. Each step corresponds to an operation such as locating an integrase, validating gene distance, or identifying boundary DRs. Ongoing work will extend this structure into a graph-based representation to allow conditional paths and clear process modeling.

A delimitation result for an MGE, denoted by o^s, is obtained by executing y^s on x^s: $o^s = \mathrm{y}^s(\mathrm{x}^s)$. o^s includes the genomic coordinates of the element's integrase and target gene, as well as the genomic positions of the left and right boundaries, which flanks and precisely defines the position of the element within the bacterial genome.

Each case $(\mathrm{x}^s, \mathrm{y}^s)$ is defined from discussion with the microbiology expert, who describes a concrete example of MGE delimitation. From this discussion, the problem x^s and the reasoning process he applied, represented by the solution y^s, are formalized.

Retrieval and reuse of a source case to solve a target problem. To solve a new delimitation problem $\mathrm{x}^{\mathrm{tgt}}$, a retrieval mechanism selects in CB a relevant process case $(\mathrm{x}^s, \mathrm{y}^s)$ to the problem $\mathrm{x}^{\mathrm{tgt}}$. This retrieval mechanism relies on the similarity between the attributes defining x^s and $\mathrm{x}^{\mathrm{tgt}}$.

The reuse of a process case $(\mathrm{x}^s, \mathrm{y}^s) \in \mathrm{CB}$ consists in applying y^s to $\mathrm{x}^{\mathrm{tgt}}$: $\mathrm{y}^s(\mathrm{x}^{\mathrm{tgt}})$ is computed, and then there are two possibility:

- Either the computation $\mathrm{y}^s(\mathrm{x}^{\mathrm{tgt}})$ succeeds and returns the delimitation of MGEs for the target problem (o^{tgt}).
- Or it returns a failure: the application of y^s does not succeed to solve $\mathrm{x}^{\mathrm{tgt}}$. In such cases, it may be possible to *adapt* the source case so that it solves $\mathrm{x}^{\mathrm{tgt}}$. This adaptation step is currently under consideration: a few adaptation rules have already been acquired to improve the system's reuse capabilities. Adaptation in our system applies to the solution process y^s, not to the source case itself, and is triggered only when the process fails to handle $\mathrm{x}^{\mathrm{tgt}}$.

Failures during case reuse are automatically detected at execution time. For example, if an MGEs limits is not found or the target gene is missing, a failure message is generated. These messages are used to trigger adaptation rules, when available. The adaptation rules are designed based on recurring failure patterns observed during testing, and their application is guided by the type of failure identified.

Evaluation of the system. At the start of the acquiring and modeling of process cases, the expert provided 291 informal cases, each corresponding to a distinct MGE to be delimited. From these, 37 process cases were selected to constitute the case base (CB). The selection is based on biological diversity criterion of MGEs, ensuring a broad coverage of integration mechanisms. The aim was to maximize the coverage of the remaining $291 - 37 = 254$ cases, representing the test base (TB). These informal cases refer to real MGE delimitation problems manually annotated by the expert prior to the formalization phase. They represent ground truth examples used to test the system's ability to reproduce expert-level reasoning.

A total of 246 elements (96.8%) were delimited precisely by the system, with only 8 (3.2%) exhibiting delimitation failures unrelated to the case base. However, the developed approach's key limitation is its reliance on external genomic annotations (public databases). External genomic databases (NCBI) and tools such as ICEscreen [8] are used to extract gene annotations (e.g., integrases, tRNA genes), which are required to execute the reasoning steps in each process case. The system depends on the availability and accuracy of these annotations.

3. Progress Summary

A PO-CBR system has been implemented, allowing the reuse of process cases on new delimitation problems. The case base is applicable and the reuse mechanism works efficiently on bacterial strains of the Streptococcus group. Current work focuses on the adaptation of process cases. The objective is to improve reuse on Streptococcus strains and extend the approach to other bacterial groups. This requires further acquisition and formalization of expert knowledge, specifically for adaptation rules. We are also investigating the XAI (eXplainable Artificial Intelligence) dimension of the system. We consider explainability at two levels: (1) for the user, to understand the reasoning steps behind a result, and (2) for the system itself, since adaptation rules are triggered by explanations generated from failure analysis. For example, when a process case fails on a new problem, the system displays a detailed trace of the failed step (e.g., unmatched limts or missing target gene), and uses this information to trigger an appropriate adaptation rule, if available.

4. Conclusion and Future Work

This work demonstrates the relevance of PO-CBR for automating the delimitation of mobile genetic elements (MGEs) in bacterial genomes, a task traditionally performed through manual expert analysis. The developed approach retrieves and applies process cases according to problem similarity, and integrates adaptation mechanisms where required. Currently, the PO-CBR system has shown high delimitation accuracy on bacterial strains from the streptococcus group, precisely identifying MGE boundaries while making the reasoning process explainable.

In the future, we will focus on extending the approach to other bacterial groups and new types of MGEs. A key objective is to define adaptation rules, in order to better handle cases not initially covered by the current case base. Improvements will also be made to the external data used, by integrating more comprehensive external data sources and balancing the expansion of the case base with the evolution of adaptation rules. Process cases are currently implemented as Python files. This implementation will be refined by integrating a graphical representation for intuitive modification and export to Python for execution.

Declaration on Generative AI

During the preparation of this work, the authors used Perplexity.ai exclusively as a research assistant to locate MGEs biology background information, and Reverso for grammar, spelling checks and sentence reformulation. All content was reviewed and verified by the authors, who take full responsibility for the final publicatio

References

[1] C. K. Riesbeck, R. C. Schank, Inside Case-Based Reasoning, Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1989.

[2] M. Minor, S. Montani, J. A. Recio-García, Process-oriented case-based reasoning, Information Systems 40 (2014) 103–105. URL: https://doi.org/10.1016/j.is.2013.06.004. doi:10.1016/j.is.2013.06.004.

[3] G. Müller, R. Bergmann, Generalization of Workflows in Process-Oriented Case-Based Reasoning, in: I. Russell, W. Eberle (Eds.), Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference (FLAIRS-28), AAAI Press, Hollywood, Florida, USA, 2015, pp. 391–396. URL: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10437.

[4] X. Bellanger, S. Payot, N. Leblond-Bourget, G. Guédon., Conjugative and mobilizable genomic islands in bacteria: evolution and diversity, FEMS Microbiology Reviews 38 (2014) 720–760. URL: https://doi.org/10.1111/1574-6976.12058. doi:10.1111/1574-6976.12058. arXiv:https://academic.oup.com/femsre/article-pdf/38/4/720/18147733/38-4-720.pdf.

[5] C. Ambroset, C. Coluzzi, G. Guédon, M.-D. Devignes, V. Loux, T. Lacroix, S. Payot, N. Leblond-Bourget., New insights into the classification and integration specificity of streptococcus integrative conjugative elements through extensive genome exploration, Frontiers in Microbiology 6 (2015) 1483. doi:10.3389/fmicb.2015.01483.

[6] C. Coluzzi, G. Guédon, M.-D. Devignes, C. Ambroset, V. Loux, T. Lacroix, S. Payot, N. Leblond-Bourget., A Glimpse into the world of integrative and mobilizable elements in streptococci reveals an unexpected diversity and novel families of mobilization proteins, Frontiers in Microbiology 8 (2017). URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.00443/full. doi:10.3389/fmicb.2017.00443, publisher: Frontiers.

[7] M. Liu, X. Li, Y. Xie, D. Bi, J. Sun, J. Li, C. Tai, Z. Deng, H.-Y. Ou., ICEberg 2.0: an updated database of bacterial integrative and conjugative elements, Nucleic Acids Research 47 (2018) D660–D665. URL: https://doi.org/10.1093/nar/gky1123. doi:10.1093/nar/gky1123. arXiv:https://academic.oup.com/nar/article-pdf/47/D1/D660/27437376/gky1123.pdf.

[8] J. Lao, T. Lacroix, G. Guédon, C. Coluzzi, S. Payot, N. Leblond-Bourget, H. Chiapello, ICEscreen: a tool to detect Firmicute ICEs and IMEs, isolated or enclosed in composite structures, NAR Genomics and Bioinformatics 4 (2022) lqac079. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9585547/. doi:10.1093/nargab/lqac079.