AUTOMATIC KNOWLEDGE BASE CONSTRUCTION AND HYPOTHESIS GENERATION: ANTIBIOTIC RESISTANCE MECHANISMS FOR *ESCHERICHIA COLI*

Allocation: NSF PRAC/350 Khn
PI: Ilias Tagkopoulos[1]

[1]University of California, Davis

ML

BI

EXECUTIVE SUMMARY

Antibiotic resistance is one of the leading threats to global health, food security, and development according to the World Health Organization [1]. The construction of a cohesive knowledge base for antibiotic resistance that can be a source for machine learning methods will have a broad impact in the field and eventually enable an artificial intelligence (AI) system to automate knowledge discovery in unprecedentedly efficient and unbiased ways. With this goal in mind, we built a knowledge base housing one million facts for *E. coli* antibiotic resistance mechanisms that are specifically structured for efficient machine learning. On top of this knowledge base, we trained the multilayered machine learning method for generating novel hypotheses of antibiotic resistance. The cross-validation results showed that the predictor can achieve an AUC (Area Under the Curve) of 0.868 for the ROC (Receiver Operator Characteristic) and Average Precision of 0.24, surpassing the baselines. The proposed framework is generically applicable, and we plan to make the tool publicly available so that anyone can apply it to their domain of interest.

RESEARCH CHALLENGE

Antibiotic resistance can affect anyone of any age or nationality. It is a natural phenomenon that is accelerated by our way of life and overuse of antibiotics in livestock and medicine. Aside from the global threat of antibiotic resistance, its immediate impact results in longer hospital stays, higher medical costs (estimated at $20 billion annually in the United States alone), and increased mortality [1].

Constructing a knowledge base for antibiotic resistance that can be applied to machine learning methods for generating hypotheses will accelerate the overall discovery process in unprecedentedly efficient and unbiased ways [2–4]. A large number of knowledge bases have recently been created using graph representation, which stores factual information in the form of relationships among entities. These include YAGO [5], NELL [6], Freebase [7], and Google Knowledge Graph [8] for storing general facts about people, cities, and the like. Notable examples in the biological domain include KnowLife [9] and BioGraph [10].

However, key challenges exist to enabling the creation of a machine-learnable knowledge base for *E. coli* antibiotic resistance. Although knowledge conflicts exist and are reported in the literature [11], these inconsistencies are not curated in existing knowledge bases. Even in the literature, these conflicts are quantified only in part [12]. The inconsistencies are especially prevalent between high-throughput measurements and biological networks, making it nontrivial to draw biologically meaningful conclusions in an automated way [13]. Several truth discovery methods have been proposed over the past decade, and they have been successfully applied in diverse domains.

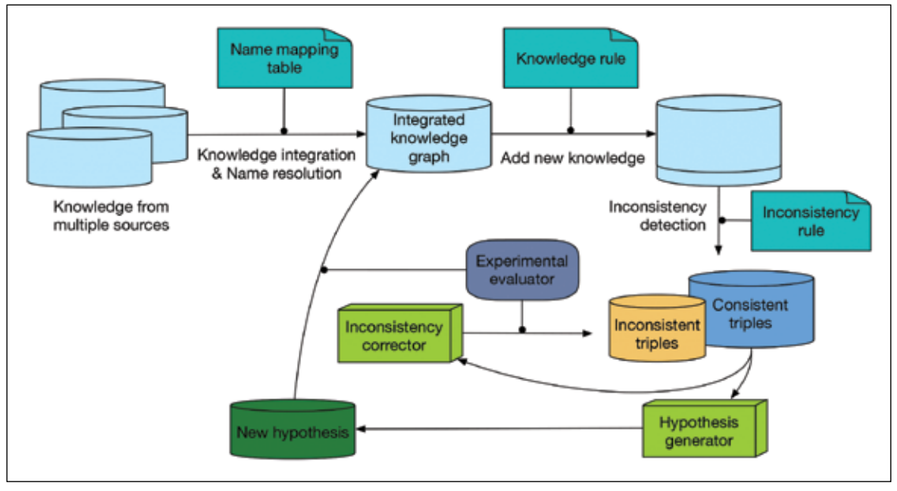

Figure 1: Schematic diagram of the proposed framework for knowledge base construction and novel hypothesis generation.

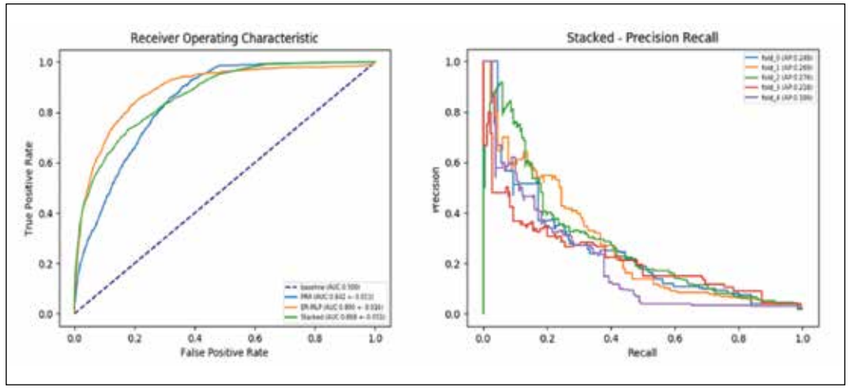

Figure 2: The ROC for each of the ER–MLP, PRA, and Stacked models (left). The Precision Recall curve for Stacked models (right).

A primary application domain involves web sourcing, where information conveyed on a web page conflicts with other web pages [7]. Many truth discovery methods were first developed to resolve web-source conflicts. These methods employ one of three possible approaches: iterative methods, optimization methods, and probabilistic methods [14,15]. Recently, the use of a link prediction method has been proposed to decide truth among conflicts [16]. Some notable recent work in the biological sciences includes inconsistency repair in the *E. coli* gene regulatory network using answer set programming [13] and inconsistency resolution in signal transduction knowledge using integer linear programming [17].

Another challenge is that negative findings are not curated as well as positive findings in existing biological knowledge bases despite their importance in training the machine learning models to classify what knowledge is likely true [16]. Finally, existing knowledge bases do not annotate temporal information about antibiotic exposure despite its importance in the emergence of antibiotic resistance [18,19], which obscures the understanding of the time series dynamics of antibiotic resistance mechanisms.

METHODS & CODES

We built a generic framework that first constructs an inconsistency-free knowledge graph for a specific domain and then trains the hypothesis generator on top of this knowledge graph that is based on the multilayered machine learning method. This method has been published elsewhere [20], and we optimized the published code to best utilize its principles for our needs.

RESULTS & IMPACT

In this work, we developed an inconsistency-resolved, machine learning-friendly, time series knowledge graph for *E. coli* antibiotic resistance. The knowledge graph incorporates a total of one million triples from 10 sources where the five distinctive durations of antibiotic exposure exist, ranging from 30 mins to seven days. Furthermore, this knowledge base houses positive and negative findings and thus facilitates training and evaluation of hypothesis generators that are built with machine-learning methods. On top of this knowledge base, we trained the multilayered machine-learning method that employs a path-ranking algorithm (PRA) and entity-relation multi-layered perceptrons (ER–MLP) for generating novel hypotheses of antibiotic resistance. The cross-validation results showed that the predictor can achieve an AUC of 0.868 for ROC and Average Precision of 0.24, surpassing the baselines. The proposed framework is generically applicable, and we believe this tool can accelerate knowledge discovery in an unbiased way by automatically generating a novel hypothesis from a knowledge base in a specific domain. This framework will be publicly available for use in other domains of interest.

WHY BLUE WATERS

Access to Blue Waters was critical to the success of this project, and project staff helped us to best utilize the Blue Waters high-performance computing resource for our needs by helping us to efficiently run large jobs in parallel and by responding to our requests in a timely manner.