[BIOINFO 2024]

Data Imputation for Connectivity Mapping Using Multivariate Normal Distribution Based Drug Correlation

Jiwon Jang[1], Seri Park[1], and Jong Wha J. Joo[1,*]

[1]*Department of Computer Science and Engineering, Dongguk University*
**Corresponding author: jwjjoo@dgu.ac.kr*

Drug repositioning is the process of using existing drugs for new disease treatments or indications. A crucial part of this process is the creation of a Connectivity Map (CMap), which is a database or tool used to compare gene expression profiles of drugs to identify potential drug-disease associations. One recent discovery of drug repositioning using CMap is the identification of drugs for the treatment of epithelial ovarian cancer (EOC). Researchers utilized gene expression profiles from clinical samples to identify genes associated with patient outcomes, such as time to recurrence (TTR). These gene signatures were then compared with CMap data to identify drugs that could potentially be repurposed for ovarian cancer treatment. Through CMap analysis, several compounds, including mitoxantrone and doxorubicin, were identified and subsequently tested in vitro, showing promising effects on ovarian cancer cell viability. Drug repositioning, as demonstrated, offers high efficiency in terms of clinical trials, cost-effectiveness, and safety, making it a promising approach for future biological and medical advancements. The initial version of CMap, however, was limited in diversity, as it focused solely on cancer cell lines. Recently, the Library of Integrated Network-Based Cellular Signatures (LINCS) significantly expanded the scope of CMap with the development of L1000 analysis. Despite this expansion, many clinically meaningful combinations remain absent from the dataset. In this study, we applied and compared five imputation methods to solve the missing data, evaluating the scalability of the Connectivity Map. To improve upon limitations of traditional methods, we employed the PhenImp method, a multivariate normal distribution-based approach originally used for imputing missing phenotype data in Genome Wide Association Studies. Using this method, we calculated drug correlations across different cells and applied them to the dataset. Notably, our approach outperformed existing methods in terms of accuracy. Additionally, we focused on computational efficiency to ensure that our method could scale well with large datasets. We incorporated Tanimoto similarity calculations to optimize data structure and improve scalability and plan to package and distribute the software for future applications.