Estimation of relationships between chemical substructures and antibiotic resistance-related gene expression in bacteria: Adapting a canonical correlation analysis for small sample data of gathered features using consensus clustering

Tsuyoshi Esaki1*, Takaaki Horinouchi2, Yayoi Natsume-Kitatani3,4, Yosui Nojima3 Iwao Sakane5, and Hidetoshi Matsui1,6

1 The Center for Data Science Education and Research, Shiga University, 1-1-1 Banba, Hikone, Shiga 522-8522, Japan
2 Center for Biosystems Dynamics Research, RIKEN, 6-2-3 Furuedai, Suita, Osaka 565-0874, Japan
3 Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, 7-8-6 Saito Asagi, Ibaraki, Osaka 567-0085, Japan
4 Center of Drug Design Research, National Institutes of Biomedical Innovation, Health and Nutrition, 7-8-6 Saito Asagi, Ibaraki, Osaka 567-0085, Japan
5 Central Research Institute, ITO EN Ltd., 21 Megami, Makinohara, Shizuoka 421-0516, Japan
6 Faculty of Data Science, Shiga University, 1-1-1 Banba, Hikone, Shiga 522-8522, Japan

*E-mail: Tsuyoshi-esaki@biwako.shiga-u.ac.jp

(Received May 7; accepted September 8; published online September 30)

Abstract

The emergence of antibiotic-resistant bacteria is a serious public health concern. Understanding the relationships between antibiotic compounds and phenotypic changes related to the acquisition of resistance is important to estimate the effective characteristics of drug seeds. It is important to analyze the relationships between phenotypic changes and compound structures; hence, we performed a canonical correlation analysis (CCA) for high dimensional phenotypic and compound structure datasets. For the CCA, the required sample number must be larger than the feature number; however, collecting a large amount of data can sometimes be difficult. Thus, we combined consensus clustering to gather and reduce features. The CCA was performed using the clustered features, and it revealed relationships between the features of chemical substructures and the expression level of genes related to several types of antibiotic resistance.

Key Words: cheminformatics, bioinformatics, canonical correlation analysis, clustering, antibiotic resistance

Area of Interest: Emerging new technology

1. Introduction

The emergence of antibiotic-resistant bacteria is a serious public health concern worldwide. To
understand the drug resistance mechanisms in bacteria, quantitative phenotypic changes associated with the acquisition of drug resistance in *Escherichia coli* have been investigated [1]. However, relationships between specific antibiotic structures and resistance acquisition by bacteria are not clear. Understanding such relationships will provide new insights to control bacterial drug resistance and support the development of novel antibiotic drugs.

To investigate relationships between compound structures and drug resistance acquisition, high dimensional datasets can be treated as two different datasets for analysis. To analyze such multiparameter datasets, we applied canonical correlation analysis (CCA) [2]. The CCA has been used to estimate the relationships between two sets of variates such as adverse-effect profiles and chemical fragments [3, 4]. For the CCA, the required sample number must be greater than the feature number; however, collecting huge datasets is sometimes difficult. Thus, effective methods that reduce similar features and retain specific features are required.

Here, we performed consensus clustering (CC) to gather features of both chemical substructures and phenotypic changes of acquired drug-resistant bacterial strains. For this purpose, we used gene expression data of antibiotic-resistant *E. coli* strains [1]. This approach enables to reveal relationships between the features of chemical substructures and expression of genes related to several types of antibiotic resistance with no prior knowledge.

2. Methods

To analyze the relationships between chemical substructures and phenotypic changes related to the acquisition of drug resistance, the mRNA expression level of 4444 genes in *E. coli* strains developed in *vitro* with resistance to 10 antibiotic compounds (compound names, antibiotic classes, and bacterial cell targets are shown in Table S1) was obtained. The dataset showing changes in gene expression was generated by collecting data from a previous study [1]. This dataset includes data of four strains resistant to each antibiotic (i.e., a total of 40 resistant strains and the wild-type parental strain); thus, average gene expression among the resistant strains for the same drug was calculated and normalized to the value of the parental strain. Structure data files (SDFs) of 10 compounds were collected from ChEMBL [5] and Morgan fingerprint with two radii was used to generate 2048 substructures from the SDFs in rdkit (ver. 2019.09.3) [6] within Python (ver. 3.7.3).

CC is a stable clustering method that is based on subsampling to select a cluster number (k) [7]. ConsensusClusterPlus (ver. 1.50.0) [8] was applied using the following parameters: resampling ratio of item and features = 0.9; subsample number = 200; binary distance for hierarchical clustering (for substructures); and 1 - Spearman’s correlation for partitioning around medoids clustering (for transcriptomes). Based on the relative increase in the total area under the cumulative distribution function curve, the appropriate k value was determined to be 5 for substructure (from c_clst1 to c_clst5) and transcriptome (from g_clst1 to g_clst5; feature numbers are shown in Table S2). The average values in the same clusters were calculated to perform the CCA.

The CCA is a method used to reveal relationships between two groups by maximizing the canonical correlation of two linear transformation features for samples. In this study, we prepared a set of 10 compounds with five clustered substructures and gene expression features. In brief, a sparse CCA was performed using the PCA (ver. 1.2.1) [9] in R (ver. 3.6.1) with the following parameters: number of components = 2, and the L1 penalty of L1 norm was 0.633 for genes and compounds based on penalized matrix decomposition [10]. A schematic of the data analysis approach is shown in Figure S1.
3. Results and Discussion

From a two-dimensional scatter plot, we calculated two orthogonal canonical coefficient vectors under the correlation contributions of 0.901 and 0.822 for canonical coefficients 1 and 2, respectively (Figure 1). The results of CC and CCA are shown in Figures S1–S9. Overall, we observed clusters with large coefficient values in each axis. The results of clustering (Figure S8) showed that each cluster represents a different action mechanism of drugs; for example, quinolones (ENX and CPFX) and other drugs were divided by c_clst3, whereas cephalosporins (CPZ and CFIX) were divided by c_clst1. We investigated the type of substructure that was represented in each cluster (Figures 2, S10, and S11). Notably, paired clusters of compounds and genes were identified (e.g., c_clst5 and g_clst2; both have a high canonical coefficient 1 value). Therefore, we investigated the genes that were represented in each cluster (Table S3) and noticed several previously reported antibiotic resistance-related genes in each cluster; for example, genes related to purine nucleotide biosynthesis were enriched in g_clst2. g_clst2 included ompF and ompC (porin protein), g_clst3 included marRABC (multiple antibiotic resistance family) and mdtABCD (multidrug efflux pump), and g_clst4 included acrAB (multidrug efflux pump). These findings suggest that these clusters represent different mechanisms of resistance for each class of antibiotics.

We used CC to reduce the number of features and performed a sparse CCA to identify relationships between gene expression and compounds. The relationship between chemical substructure and drug action mechanism (mentioned in previous literature [1]) can be implied from Figure S8. Furthermore, the relationship between them can be derived from the data in Figure S9. From these results, we consider that the assumption of a relationship between chemical structure and gene expression is pertinent. acrAB was strongly related to phenotypic changes in both gene expression and Raman spectrum [11]. With this method, new insights into the relationship between antibiotic compound structures and drug-resistant phenotypes could be obtained. This method can be applied to other high-dimensional datasets such as Raman spectra.

In this study, we analyzed available data and were blinded (i.e., no prior knowledge) to relationships between features in each dataset. With such information, grouping features may help understand correlations. In this context, it is important to verify appropriate distances such as correlation coefficients and consider optimal clustering algorithms such as spectral clustering. Furthermore, sufficient experimental data must be collected because compound and gene dynamics and interactions are too complex to study with limited datasets. Although there are several challenges in revealing additional mechanisms and developing new drugs, we believe that finding the
relationships between chemical substructures and gene expression is important for compound optimization. We hope this method will help discover new drug seeds using chem-bio informatics approaches.

4. Supplementary Information

Figures S1–S11 and Tables S1–S3 are available as supplemental data.

5. Acknowledgments

This work was partially supported by a research grant for priority areas from Shiga University.

6. Conflict of Interest

The authors declare no conflict of interest.

References

[1] Suzuki, S.; Horinouchi, T.; Furusawa, C. Prediction of antibiotic resistance by gene expression profiles. *Nat Commun.* 2014, 5, 5792. doi: 10.1038/ncomms6792

[2] Harold, H. Relations between two sets of variates. *Biometrika* 1936, 28(3–4), 321–377. doi: 10.1093/biomet/28.3-4.321

[3] Pauwels, E.; Stoven, V.; Yamanishi, Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics.* 2011, 12(169), 1–13 doi: 10.1186/btq073

[4] Mizutani, S.; Pauwels, E.; Stoven, V.; Goto, S.; Yamanishi, Y. Relating drug-protein interaction network with drug side effects. *Bioinformatics* 2012, 28(18), i522–i528. doi: 10.1093/bioinformatics/bts383

[5] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; et al. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* 2012, 40(D1), D1100-1107. doi: 10.1093/nar/gkr777

[6] RDKit, https://www.rdkit.org (accessed Apr. 1, 2020)

[7] Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* 2003, 52, 91–118.

[8] Wilkerson, M. D.; Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 2010, 26(12), 1572–1573. doi: 10.1093/bioinformatics/btq170

[9] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; et al. Scikit-learn: Machine Learning in Python. *JMLR.* 2011, 12(85), 2825–2830.

[10] Witten, D. M.; Tibshirani, R.; Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 2009, 10(3), 515–534. doi: 10.1093/biostatistics/kxp008

[11] Germond, A.; Ichimura, T.; Horinouchi, T.; Fujita, H.; Furusawa, C.; et al. Raman spectral signature reflects transcriptomic features of antibiotic resistance in Escherichia coli. *Commun Biol.* 2018, 1(85), 1–10. doi: 10.1038/s42003-018-0093-8