Supporting Information: Microbial Genotype-Phenotype Mapping by Class Association Rule Mining

Makio Tamura \(^1\) and Patrik D’haeseleer \(^1\)

February 26, 2008

\(^1\)COMP CAR / CMELS Microbial Systems Biology, Lawrence Livermore National Laboratory, Livermore, CA, 94550, email: makio323@gmail.com
0.1 Supporting Materials in Discussion

0.1.1 Figures

Figure S1: (a) Frequency distribution of mutual information between phylogenetic profiles of two types of COG pairs. Red dots indicate the mutual information between pairs of phylogenetic profiles of COGs for which corresponding enzymes are connected via a compound in a KEGG pathway map, whereas blue dots indicate the mutual information between those of randomly selected COG pairs. The two distributions show distinct difference after around \( MI \geq 0.25 \), where the distribution for pairs of related enzymes has a long tail, while that of random COG pair plunges to zero. 4,224 pairs are counted for both connected and random pairs, and STRING 7.0 is used for COG and enzyme mapping. (b) Schematic diagram for the conversion from phylogenetic profile of COG to a connectivity graph. Node is a COG, and edge is assigned when mutual information between profiles of two COGs is more than a certain threshold.

0.1.2 Files

PhenotypeNameNotRecoveredCOGsList.txt

These files contain lists of COGs covered by 3-COG association rules, but which are not covered by pairwise association with relaxed FDR level of \( 10^{-1} \). The first column is the FDR of the rule in which the COG is first found, the second column shows whether the rule is positively or negatively associated with the phenotype, the third column is the COG id, and the fourth column is the annotation of the COG.

0.2 Prediction

0.2.1 Summary

We compared the mining capability of netCAR with three alternative algorithms: (I) a standard CAR algorithm, CARapriori, (II) an algorithm which generates sets of COGs that have high one-to-one correlation with the phenotype, MI-CAR, and (III) a Support Vector Machine (SVM) (Christianini and Shawe-Taylor, 2000; Burges, 1998) as a black-box classifier based on the entire genome content. We measured the \( F_1 \) score, Precision, and Recall, which are commonly used accuracy statistics in information retrieval (see Materials and methods of this supporting material), of extracted rules by 25-folds cross-validation. netCAR rules show similar or better prediction performance than rules extracted by CARapriori in all phenotypes except for facultative phenotype, and even significantly better than SVM models for endospore formation.

0.2.2 Comparison Setting

For each rule, we make a prediction of presence or absence of the phenotype based on whether a genome does or does not contain all COGs in the rules, and the extracted rules are ranked by \( F_1 \) score for the training set. We also evaluate a majority-vote consensus prediction (see Materials and methods of this supporting material) of the top \( N \) rules. Experiments I and II are designed to compare the difference of hypothesis spaces explored by the three CAR mining methods, and experiment III compares the difference of hypothesis representation used by CAR (AND combinations of small sets of COGs) against the hyperplane representation of SVM.

To test whether the use of the connectivity graph between COGs in netCAR has a positive effect on the performance of the algorithm, we compare it with a simpler algorithm, which we call MI-CAR. MI-CAR starts out by generating a set of Parents showing higher mutual infor-
Figure S3: (a) Number of positively correlated rules within FDR levels, and (b) number of unique COGs in extracted positively correlated rule within FDR levels. Blue, orange, and green lines are 1-COG, 2-COG, and 3-COG rules, respectively, and broken lines are number of uncharacterized COGs in the same colored association rules.

Figure S4: Performance of netCAR, CARapriori, MI-CAR, and SVM on the aerobic, anaerobic, facultative, endospore, motility, and Gram negativity phenotypes, binned in increments of 5 or 3 rules for visibility. Average cross-validation F1 score for the top 30 rules of each CAR algorithm, ranked by their training set F1 score. The SVM hyperplane prediction is based on a weighted sum across all COGs, so its F1 score is indicated as a constant value.

that netCAR and MI-CAR evaluate roughly the same number of sets. For example, whereas netCAR evaluates about 800,000 candidate 3-COG sets starting from 30 Parent COGs, MI-CAR evaluates all possible 3-COG sets out of 170 Parent COGs, resulting in a similar number of sets examined. As with CARapriori, rules extracted by netCAR and MI-CAR are also sorted by their F1 score on the training data. Due to the long computational time, we cannot perform the experiment CARapriori for 3- and 4-COG extraction.

Figure S4 and S5 shows the F1 scores of individual and consensus prediction of the top 30 rules extracted by netCAR, MI-CAR, and CARapriori, and that of SVM model for the six phenotypes. The Precision and Recall for all six phenotypes are available in Figure S6-S9.

In general, F1 scores of 2-COG rules are better than those of 1-COG rules by both individual and consensus prediction for the CAR mining algorithms, but 3-COG rules are not always better than 2-COG rules, and the F1 scores of 4-COG rules tend to be lower than those of 2-, and 3-COG rules for all of the six phenotypes. It takes about 10 hours for MI-CAR to extract 3-COG sets, compared to 5 minutes for netCAR. We estimate that it would take more than 1 month for CARapriori to extract 3-COG sets, so we did not include this algorithm for
3- and 4- COG extraction.

0.2.3 netCAR vs CARapriori

$F_1$ score of netCAR rules are generally higher than those of CARapriori in all six phenotypes, especially in the consensus prediction.

The performance of netCAR rules over those of CARapriori supports the effectiveness of restriction along a path of the connectivity graph instead of the exhaustive checking. For similar $F_1$ scores, rules generated by netCAR tends to show higher Precision but lower Recall, whereas CARapriori rules tend to show higher Recall but lower Precision.

We observe a distinctive difference in prediction of the rarest phenotype in our dataset, endospore formation (present in only 11 out of 155 organisms), where both individual and consensus prediction by 2-, 3-, or 4- COG rules of netCAR perform significantly better than CARapriori as well as MI-CAR rules. The netCAR algorithm may have an advantage against CARapriori and MI-CAR to mine genes for rare microbial phenotypes.

0.2.4 netCAR vs MI-CAR

Interestingly, the $F_1$ score of netCAR rules show compatible performance with those of MI-CAR, both for individual and consensus prediction, for aerobic, anaerobic, motility, and Gram negative, but the consensus predic-
There may be some genes with a negative association with the anaerobic phenotype.

### 0.2.6 COG Content in Rule

For CARapriori, the number of unique COGs used in the top $N$ rules increases linearly (Figure S10). This linearity may explain the decreasing tendency in Precision of the consensus prediction of CARapriori as the number of rules increases, compared to netCAR and MI-CAR. In contrast, the number of unique COGs in the top netCAR and MI-CAR rules increase slowly in non-linear manner after a certain threshold. This suggests that netCAR and MI-CAR may be more resistant to picking new COGs for rule generation.

### 0.2.7 Materials and methods

**$F_1$ score**

We use the micro-averaged $F_1$ score since some phenotypes, such as endospore formation, have few positive instances, resulting in the precision or recall (and thus the $F_1$ score) becoming undetermined for several of the cross validation test sets. The micro $F_1$ score is calculated as follows: $F_1 = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$ where $\text{Precision} = TP / (TP + FP)$, $\text{Recall} = TP / (TP + FN)$, and $FTP, CTP, and CFN$ is the sum of true positive, false positive, and false negative count, respectively, over the test sets.

**Consensus Prediction**

A consensus prediction $ConsPred \in [0, 1]$ of the top $N$ rules is made by $\text{Pred}_1, \ldots, \text{Pred}_N \in [0, 1]$ of the top $N$ rules such that $ConsPred = 1$ if $n_i \geq n_0$ or otherwise 0, where $n_1$ and $n_0$ is the total count of Pred$_i = 1$ for $i = 1, \ldots, N$ and Pred$_i = 0$ for $i = 1, \ldots, N$, respectively.

**Experimental parameters**

Experiments were performed under version 1.6 Java runtime environment on a 64 bit Linux machine with 7.0 gigabytes memory and 3.00 GHz CPU power. For rule mining by netCAR, the mutual information to select Parent COGs, mMI$p$, is adjusted so that the number of Parents is about 30 for each cross validation data set. The threshold mutual information, mMI$c$, is set to $1.3 \times$ average mutual information among Parents times for 2-
and 3-COG rule mining. The $F_1$ score is stable around these parameter values (Table S1) For 4-COG rule mining, these parameters would result in more than 60% of all COGs to be assigned as Children, resulting in excessive running times, so $mmIc$ is adjusted such that the total number of Children is less than 850. The threshold value to select the final rules, $mmIrp$, is set to $mmIrp \times 1.1$, or 0.25, whichever is less. The same parameters of netCAR are used to estimate the number of Parent COGs for the corresponding $MI-CAR$ rule extraction. For rule mining by CARapriori, minimum Support is set to $0.3 \times$ the fraction of positive instances across all organisms. For example, if 40% of all samples are positive, the minimum Support is set to be $40\% \times 0.3 = 12\%$. Minimum Confidence is set to 50%. The $F_1$ score is stable around these parameter values (Table S2), but we observed that the $F_1$ score decreases when the minimum confidence is set to a value larger than 60%. We use LIBSVM (Chang and Lin, 2001) package version 2.84 for the SVM experiment with default values of the linear kernel function.

### 0.3 False Discovery Rate and Mutual Information

Figure S9: The Recall of consensus prediction based on the top $N$ rules of each CAR algorithm and SVM for aerobic, anaerobic, facultative, endospore, motility, and Gram negative prediction. The Recall score of SVM is independent of the number of rules in the figure. Data is plotted every 3-rule increment starting from 3 rules.

Figure S10: The number of unique COGs represented in the top $N$ rules of each CAR algorithm. Since the top rules may involve many of the same COGs, the number of unique COGs does not linearly increase in netCAR and MI-CAR.

Table S1: 2-COG rule extraction by netCAR with various parameters and their average $F_1$ scores

| Parameter | 10 | 20 | 30 | 40 | 50 |
|-----------|----|----|----|----|----|
| 0.7       | 67.93 | 69.55 | 70.14 | 70.16 | 69.88 |
| 1.0       | 65.45 | 68.01 | 67.85 | 68.35 | 69.07 |
| 1.3       | 67.63 | 68.74 | 68.56 | 68.04 | 67.93 |
| 1.6       | NaN  | 70.37 | 69.20 | 68.88 | 68.53 |
| 1.9       | NaN  | 70.31 | 70.46 | 69.59 | 69.32 |

Each column corresponds to the number of Parent and each row corresponds the ratio of $mmIc$ to the average mutual information among the Parents. The value is average over 25-fold cross validation experiment of the mean value of the top 30 rules’ $F_1$ score extracted by netCAR with each combination for anaerobic phenotype prediction. NaN value indicates that the number of extracted rule is less than 30.
Table S2: 2-COG rule extraction by CARAPRIORI with various parameters and their average $F_1$ scores

|      | 30   | 50   | 60   | 70   | 90   |
|------|------|------|------|------|------|
| 0.1  | 69.70| 69.70| 69.70| 56.52| 30.56|
| 0.2  | 69.70| 69.70| 69.70| 56.52| 30.56|
| 0.3  | 69.70| 69.70| 69.70| 56.52| 30.56|
| 0.4  | 69.70| 69.70| 69.70| 56.52| 20.53|
| 0.5  | 69.70| 69.70| 69.70| 56.52| NaN  |

Each column corresponds to the minimum *Confidence* and each row corresponds to the ratio of minimum *Support* to positive phenotype instances across all organisms. The value is average over 25-fold cross validation experiment of the mean value of the top 30 rules’ $F_1$ score extracted by CARAPRIORI with each combination for aerobic phenotype prediction. NaN value indicates that the number of extracted rule is less than 30.
Figure S11: Relationship between False Discovery Rate and Mutual information for the COG phylogenetic profile and the six phenotype profiles. Blue, orange, and green lines are 1-COG, 2-COG, and 3-COG rules, respectively. If $N_o$ and $N_r$ are the number of rules that have a mutual information $mi$ or higher with respect to the original and randomly permuted phenotype profile, respectively, then $N_o/N_r$ is a simple estimated positive False Discovery Rate (Storey and Tibshirani, 2001) for the given mutual information $mi$. We calculated the median value of $N_o/N_r$ from 200 random permutation experiments. We can scan all pairwise association, but it takes too much time to scan all 2-COG and 3-COG associations. Therefore, we randomly selected as many 2- and 3-COG sets from the COG phylogenetic profile as possible within 15 minutes computational time: $3.6 \times 10^5$ (0.5% of all possible 2-COG combinations) and $5.7 \times 10^5$ (0.0002%), respectively. This process is repeated 500 times, and FDR for the mutual information $mi$ is calculated as an average of the median value. Error bar is the variance of the experiment. Plots are fitted by $MI = a \times \log_{10}(FDR) + b$ for aerobic, anaerobic, endospore, motility, and Gram negativity, and by $MI = a \times (\log_{10}(FDR))^2 + b$ for facultative.
Bibliography

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2), 121–167.

Chang, C. and Lin, C. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm, 80, 604–611.

Christianini, N. and Shawe-Taylor, J. (2000). An introduction to support vector machines: and other kernel-based learning methods. CAMBRIDGE UNIV PR.

Storey, J. and Tibshirani, R. (2001). Estimating the Positive False Discovery Rate Under Dependence, with Applications to DNA Microarrays. Dept. of Statistics, Stanford University.