Motivation: Inexpensive sequencing has allowed the assembly of genomes from bacterial isolates and communities to rapidly outpace experimental characterization of the living biospecimens. Accurate genome-based prediction of biochemical features of phylogenetically novel bacteria aids in taxonomic annotation, nutrient requirements, protein secretion pathway assessment, and strain isolation from complex communities.

Surprisingly, even the most fundamental biochemical prediction, the Gram stain phenotype, has not been automated. Here we describe a systematic Gram stain prediction approach and measure its accuracy.

Methods: From a training set of 1,549 bacteria that had undergone both Gram staining and genome sequencing, a two-step machine learning approach was employed for feature selection followed by classification for prediction of Gram status from genomic features alone. Multiple machine learning approaches were evaluated regarding their performance on this particular dataset. From all known bacterial PFAM domains that were annotated in the training set, 44 robust discriminative features were selected using a Logistic Regression model to build a Random Forest classifier.

Results: The classifier achieved a cross-validation accuracy above 95%. As expected, discriminative features identified by stability selection were overwhelmingly found to be associated with structure, function, or biosynthesis of either the cell membrane or cell wall.

Table 1. 1,549 genomes were downloaded from IMG in June 2016 and gene products were annotated using MUSCLE and Pfam. A total of 3,822 PFAM domains were found in at least one genome, and abundance counts for each of them were used as features for feature selection and classification steps. 44 robust discriminative PFAM features were selected using a Randomized Logistic Regression model. A random forest (RF) decision tree estimator was then trained also performed better on early classification. Cross-validation accuracy was determined by splitting the data, fitting the model, and computing the accuracy score on the off-set of the data ten times and the mean score was averaged.

Significant features

Conclusions

- Gram staining information is sparse and not easily accessible

- Existing Gram staining annotations for ~50 taxa are ambiguous

- A Random Forest classifier trained on 44 selected PFAMs yielded an accuracy of ~95%

- The classifier allows to characterize unknown genomes and metagenome bins based on PFAM profile

Acknowledgements

We thank Robert Murray for help with tracking down Gram staining information and Susan Hinterliefer for valuable input on the poster design.