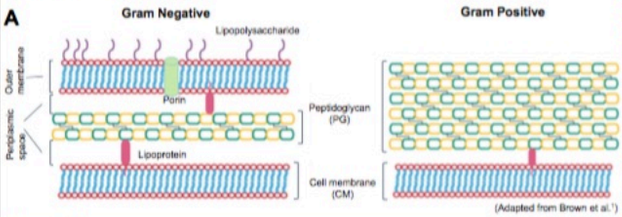
ClassyGram - prediction of Gram staining phenotype from genomic scan

Thomas Weinmaier[1], Alex Crits-Christoph[1], Todd DeSantis[1]
[1]Second Genome, Inc., South San Francisco, CA 94080, USA

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 status prediction approach and measure it's 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 discriminatory 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, discriminatory features identified by stability selection were overwhelmingly found to be associated with structure, function, or biosynthesis of either the cell membrane or cell wall.

Biology background



B

Cell appearance	Chemical effect	Step description	Chemical effect	Cell appearance
	Cell membrane remains clear	Begin with heat-fixed cells	Cell membrane remains clear	
	Cell membrane is stained with dye	Flood with crystal violet (CV) dye	PG is flooded with crystal violet dye	
	CV-iodine complex does not adhere to CM due to thin PG layer	Add iodine solution	CM traps CV-iodine complex	
	CV-iodine complex washed out of PG	Wash with alcohol	Trapped CV-iodine complex not washed out	
	Safranin stains the washed PG and CM	Counter stain with safranin	No effect by counter stain as PG remains stained by CV	

Figure 1. A: schematic cell wall structure of Gram-negative and Gram-positive bacteria. Gram-negative bacteria have a thin peptidoglycan layer between the inner and outer lipid membranes. Anchored in the outer membrane are lipopolysaccharides and it contains porins and other channels for non-vesicle-mediated transport. The cell wall in Gram-positive bacteria is formed by a single lipid membrane covered by a thick layer of peptidoglycan and lipoteichoic acid, which is anchored in the cell membrane. B: Experimental steps in the Gram staining procedure as well as the chemical effect on the cell wall and the microscopic appearance of the bacterial cells. As a result of the staining procedure Gram-positive bacteria appear purple, whereas Gram-negative bacteria are pink.

Publicly available Gram staining annotations

	NCBI	IMG	PATRIC
Total annotations	4,952	11,539	2,491
Gram-positive	1,931	4,601	849
Gram-negative	3,021	6,938	1,642
Unique species	1,524	2,274	853
Unique genera	716	964	509
Unique families	285	367	261
Unique orders	152	185	136
Unique classes	76	83	60
Unique phyla	38	37	30
Unique kingdoms	2	3	1

Table 1. Overview of Gram staining annotations available from three public resources NCBI[2], JGI Integrated Microbial Genomes system[3] (IMG) and PATRIC[4] as of September 7th 2018. 35%-40% of annotations are Gram-positive. PATRIC contains only bacteria, whereas NCBI and IMG also list archaea (NCBI:37, IMG: 131) and eukaryotes (NCBI:2, IMG:7).

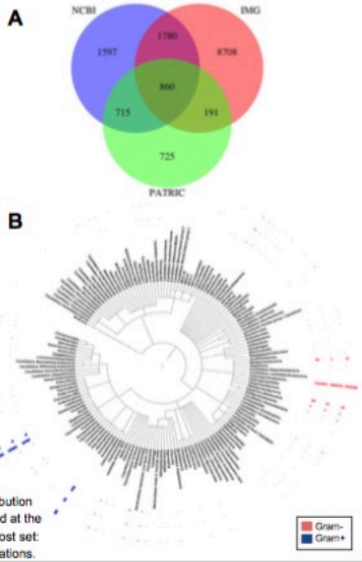


Figure 2. A:Venn diagram showing the overlap of organisms with Gram staining annotations in NCBI, IMG and PATRIC. NCBI taxon IDs were used to determine matching entries. B: Phylogenetic tree showing the taxonomic distribution of Gram staining annotations. Tree is based on the NCBI taxonomy and pruned at the class level. Bar height indicates counts of annotations within the class. Innermost set: NCBI annotations, middle set: IMG annotations, outermost set: PATRIC annotations.

Workflow for ClassyGram

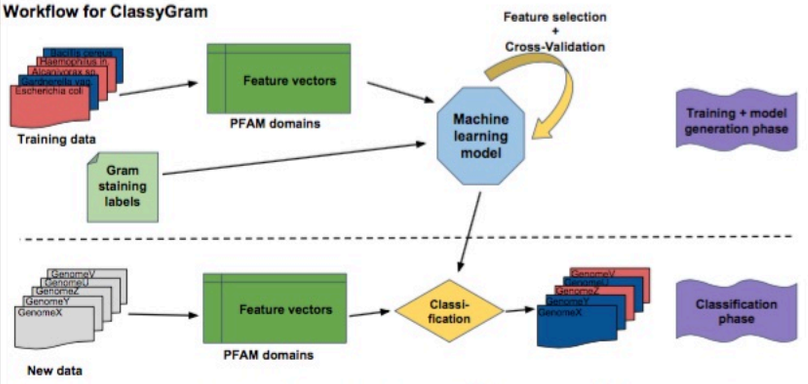


Figure 3. Schematic workflow for ClassyGram. Feature vectors of abundances of PFAM domains encoded in the genome are calculated on a set of training genomes and used along with the corresponding Gram staining labels to train a Machine Learning (ML) model. The accuracy of the model is evaluated in a 10-fold cross-validation that randomly splits the data into 90% training data and uses the remaining 10% as test data. For a new genome a feature vector of PFAM domain abundances is calculated and the ML model is used to predict the Gram staining.

ML technique	Feature set	Avg. accuracy
Random Forest (RF)	44 PFAMs	95.222%
Random Forest (RF)	13,622 PFAMs	95.224%
Support vector machine (SVM)	13,622 PFAMs	91.996%
Naïve Bayes (NB)	13,622 PFAMs	88.643%

Table 2. 1,549 genomes were downloaded from IMG in June 2016 and gene products were annotated using HMMER3 and Pfam 31. A total of 13,622 PFAM domains were found in at least one genome, and abundance counts for each of them were used as initial feature inputs to feature selection and classification steps. 44 robust discriminatory PFAM features were selected using a Randomized Logistic Regression model[5]. A random forest (RF) decision tree estimator consistently performed better or on-par with other models (SVM, NB) tested. The RF model was built with 100 estimators, quality of splits was measured using the Gini impurity, and samples were drawn with replacement. Cross-validation accuracy was determined by splitting the data, fitting the model, and computing the accuracy score on the rest of the data ten times and the mean score was averaged.

Significant features

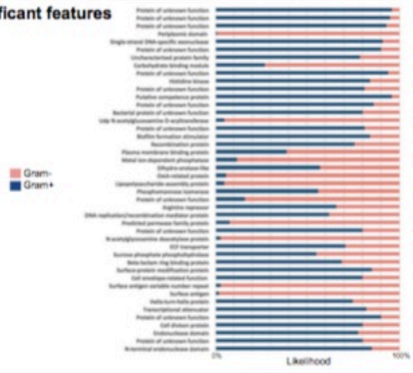


Figure 4. The RF model did not perform significantly better when trained on all original input features as compared to selective features, indicating that the 44 features selected were a near-optimally informative set out of all PFAM domains tested. Three PFAMs with high scoring Gini impurity in the random forest model were associated with a sporulation-related endonuclease domain protein, known to be highly conserved among Gram-positive bacteria and essential in the process of Gram-positive sporulation. Predictively, these PFAMs were found to be overwhelmingly represented in Gram positive species in the training dataset. A cell divison protein was also highly ranked by Gini impurity importance by the random forest and was overrepresented in Gram positive species in the training set. Differentially discriminatory PFAMs associated with Gram-negative genomes in the training set included proteins related to a bacterial surface antigen protein family known to play an essential role in outer membrane protein assembly conserved across Gram-negative species[6]. Proteins involved in lipopolysaccharide assembly were also highly predictive of Gram status.

Receiver operating characteristic curve

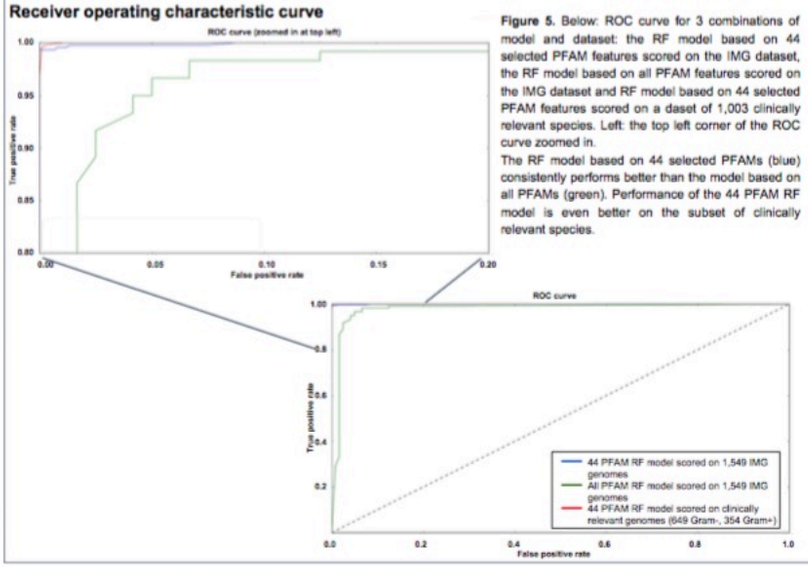


Figure 5. Below: ROC curve for 3 combinations of model and dataset: the RF model based on 44 selected PFAM features scored on the IMG dataset, the RF model based on all PFAM features scored on the IMG dataset and RF model based on 44 selected PFAM features scored on a daset of 1,003 clinically relevant species. Left: the top left corner of the ROC curve zoomed in.
The RF model based on 44 selected PFAMs (blue) consistently performs better than the model based on all PFAMs (green). Performance of the 44 PFAM RF model is even better on the subset of clinically relevant species.

Conclusions

- Available 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
- Next steps:
  - retraining the classifier on updated training data
  - consider ambiguous taxa members of a third group: → Gram+, Gram-, Unknown

References

1) Brown et al. (2015) Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi. Nat Rev Microbiol. 13(10):620-30
2) NCBI Resource Coordinators. (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44(D1):D7-19.
3) Chen et al. (2017) IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 45(D1):D507-D516.
4) Wattam et al. (2017) Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 45(D1):D535-D542.
5) Pedregosa et al. (2011) Scikit-learn: Machine Learning in Python. JMLR. 12(Oct):2825−2830
6) Voulhoux et al. (2003) Role of a highly conserved bacterial protein in outer membrane protein assembly. Science. 299(5604):262-5.

Acknowledgements

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

SECOND GENOME
THE MICROBIOME COMPANY