On the limits of computational functional genomics for bacterial lifestyle prediction

Eudes Barbosa[1,4], Richard Röttger[2], Anne-Christin Hauschild[3], Vasco Azevedo[4] and Jan Baumbach[1]

[1] University of Southern Denmark, Denmark
[2] Max Planck Institute for Informatics, Germany
[3] International Max Planck Research School for Computer Science, Germany
[4] Federal University of Minas Gerais, Brazil
eudes@imada.sdu.dk
roettger@mpi-inf.mpg.de
a.hauschild@mpi-inf.mpg.de
vasco@icb.ufmg.br
jan.baumbach@imada.sdu.dk

Abstract: We review the level of genomic specificity regarding actinobacterial pathogenicity. As they occupy various niches in diverse habitats, one may assume the existence of lifestyle-specific genomic features. We include 240 actinobacteria classified into four pathogenicity classes: human pathogens (HP), broad-spectrum pathogens (BP), opportunistic pathogens (OP), and non-pathogenic (NP). We hypothesize: (H1) Pathogens (HPs and BPs) possess specific pathogenicity signature genes. (H2) The same holds for opportunistic pathogens. (H3) Broad-spectrum and exclusively human pathogens cannot be distinguished from each other due to an observation bias, i.e. many HPs might be yet unclassified BPs. (H4) There is no intrinsic genomic characteristic of opportunistic pathogens compared to pathogens, as small mutations are likely to play a more dominant role in order to survive the immune system. To study these hypotheses, we implemented a bioinformatics pipeline that combines evolutionary sequence analysis with statistical learning methods (Random Forest with feature selection, model tuning and robustness analysis). Essentially, we present orthologous gene sets that computationally distinguish pathogens from non-pathogens (H1). We further show a clear limit in differentiating opportunistic pathogens from both, non-pathogens (H2) and pathogens (H4). Human pathogens may also not be distinguished from bacteria annotated as broad-spectrum pathogens based on a small set of orthologous genes only (H3), as many human pathogens might as well target a broad range of mammals but have not been annotated accordingly. In conclusion, we illustrate that even in the post-genome era and despite next-generation sequencing technology our ability to efficiently deduce real-world conclusions, such as pathogenicity classification, remains quite limited.

1 Background and Results

With the emergence of the so-called next-generation sequencing technology, the available data sets exploded such that we have over 27,000 registered sequencing projects at NCBI (NCBI web site, August 1, 2014). We now wonder to what extend we can deduce real-world qualitative information from this treasure of data. The aim of this paper was to investigate the power of computational functional genomics to predict bacterial lifestyles utilizing DNA sequence information only. Specifically, we asked the question if we may utilize sequence similarity to identify dedicated lifestyle-specific protein-coding genes. We restrict our report, first, to a set of 240 well-studied actinobacterial genomes and, second, to four pathogenicity lifestyles, namely: human pathogens (HP), broad-spectrum pathogens (BP), opportunistic pathogens (OP), and non-pathogenic (NP).

In [EB14], we illustrate and quantify the boundaries we face when trying to deduce a specific microbial pathogenicity class if given the genomic repertoire only, at least in the case of actinobacteria (see figure 1 below). In summary, we show that we indeed find signature genes that differentiate pathogens from non-pathogens. Only a small set of three genes for each bias, i.e. classification direction, is sufficient to reach an approximately 90% accuracy (figures 2-4). When trying to classify the different pathogenicity lifestyles though, it appears that too many confounding factors unbalance our data sets such that we cannot differentiate, for instance, a strain-specific from a lifestyle-specific gene.

We conclude that even in the post-genome era, and even for supposedly simple questions, our ability to efficiently deduce real-world implications from large-scale *de novo* next-generation sequencing data remains quite limited.

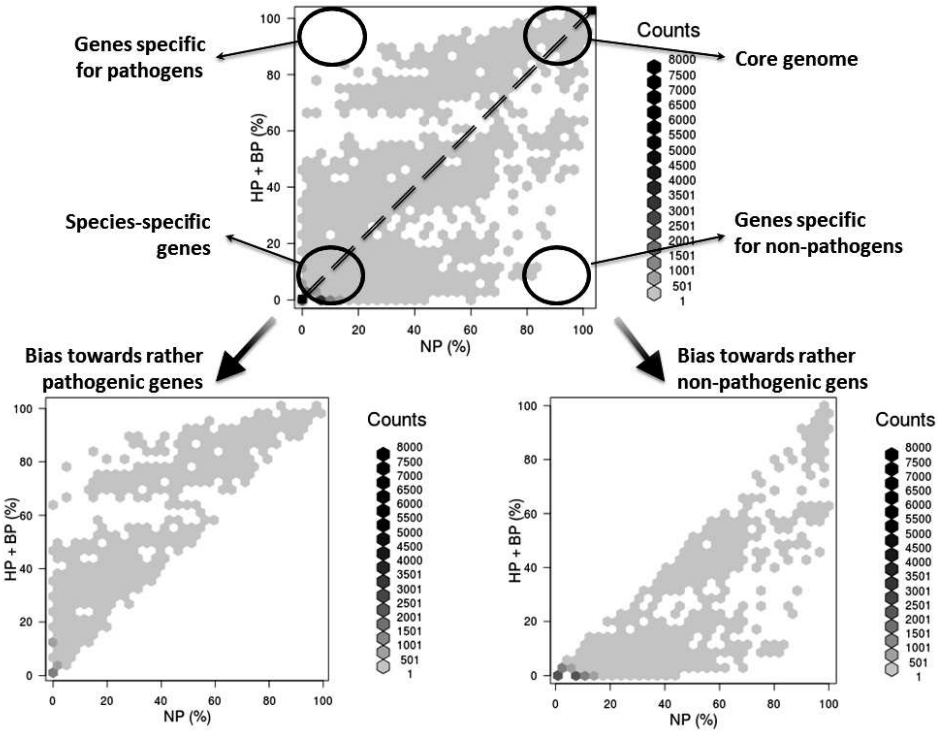

Figure 1: Illustration of our bias introduction strategy. Distribution of homologous gene clusters over two lifestyles (pathogens vs. non-pathogens) and illustration of our strategy to introduce a feature selection bias into our statistical learning pipelines. Both axes in all three plots describe the percentage of species in the respective class(es), here human pathogens (HP) and broad pathogens (BP) vs. non-pathogens (NP). The color-coding of the heat map depicts the number of clusters of homologous genes that certain percentages of pathogens/non-pathogens share. Thus, in the upper right of such a joint distribution plot, we find the core genome (homologous genes present in all species of both classes); and in the lower left, we see unique, species-specific genes. Genes close to the axis are more class specific. Genes close to the axis tails are highly class specific and, thus, the distinctive homologous gene candidates we were hoping to find. As there is no single such gene, we scanned for sets of lifestyle-distinctive genes. To find such feature genes for pathogenic lifestyles, for instance, we remove all genes that are found more often in non-pathogens (NP) than in pathogens (HP+BP), i.e. the gene clusters below the dotted line in the upper plot, such that our follow-up machine learning routines are biased towards utilizing pathogenicity-specific features (genes in the bottom left plot) for classification.

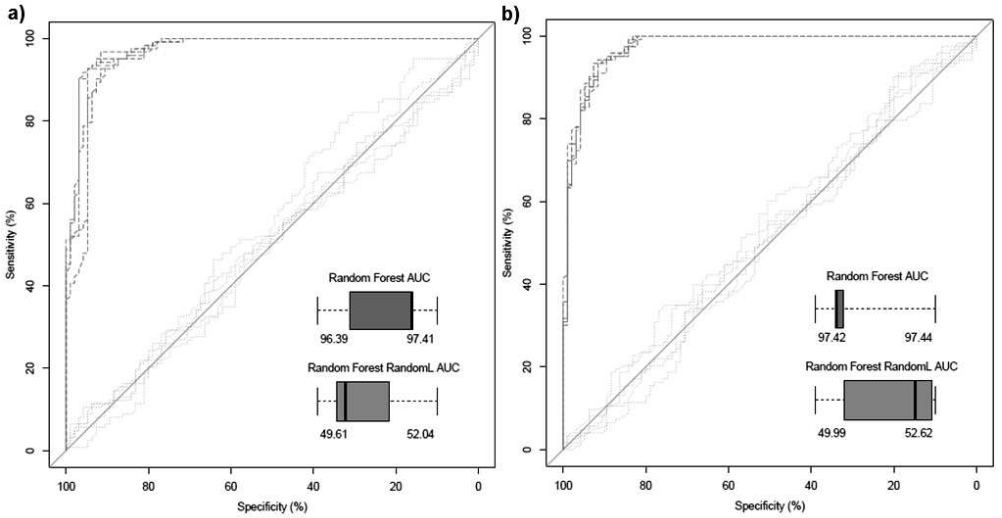

Figure 2: Classification performance non-pathogens vs. pathogens. ROC (receiver operating characteristics) plots were generated to inspect the performance of the classification models. The data was evaluated five times using different 5-fold cross validation sets to receive robust quality estimations of our classifiers. The real label classifier curves are presented in dark green dashed lines, while the random label classifier curves are given in light green dotted lines (the ones close to the base line). The variation of the AUCs (area under curve) in the cross validation was included in the figure as a box plot (bottom right). The numbers below each box plot are the lower and upper quartiles. **a)** Pathogen classifier results (NP vs. HP+BP). We biased the predictors towards using pathogen-specific genes (see Figure 1). **b)** Non-pathogen classifier results (NP vs. HP+BP) where the predicator now was biased to prefer the non-pathogen-specific genes. See text for a full description of our machine learning strategy and refer to Figure 1 regarding the "bias".

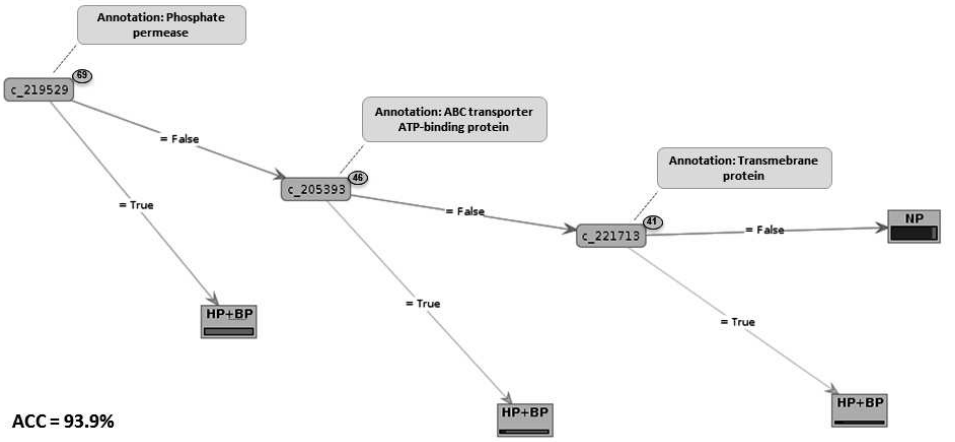

Figure 3: Decision tree created using the genes most discriminative for pathogen (HP+BP). Our classification pipeline (see full text) selected the above three genes as most representative for pathogens. We learned and visualize them as a simple decision tree by using the RapidMiner software. Nodes represent gene clusters with the following Transitivity Clustering IDs: 219529, 205393 and 221713, which are associated to the GenBank annotations "Phosphate permease" (e.g. UniprotKB AC: I6YD06 or P65712), "ABC transporter ATP-binding protein" (e.g. UniprotKB AC: D9Q9K6) and "transmembrane protein" (e.g. UniprotKB AC: G2MY46), respectively. The small circles close to the Transitivity Clustering IDs indicate the cluster size. Using only these three features the decision tree already obtains an accuracy of 93.9%.

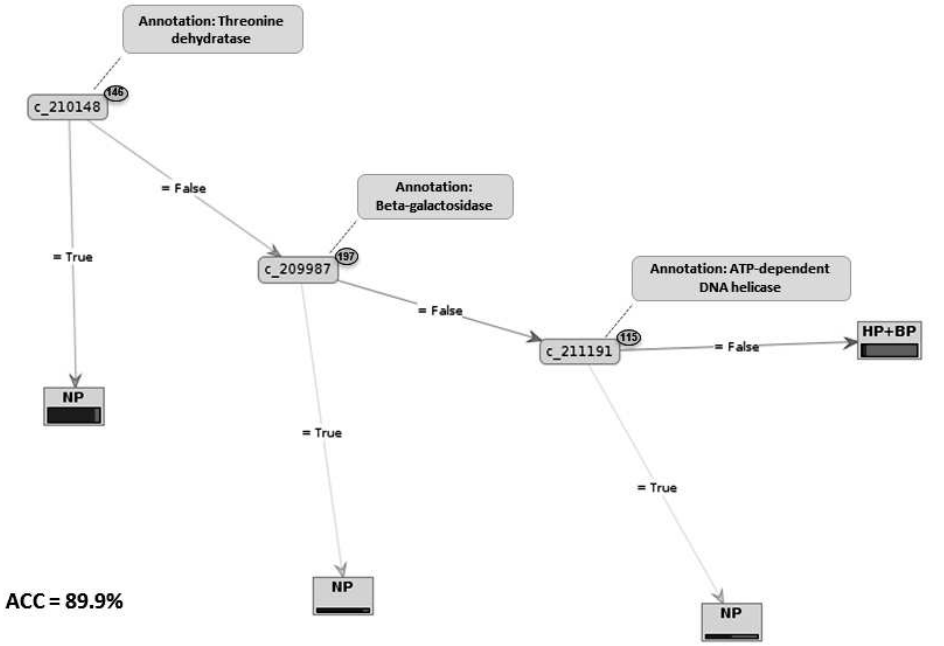

Figure 4: Decision tree created using the genes most discriminative for non-pathogen (NP). Our classification pipeline (see full text) selected the above three genes as most representative for pathogens. We learned and visualize them as a simple decision tree by using the RapidMiner software. Nodes represent gene clusters with the following Transitivity Clustering IDs: 210148, 209987 and 211191, which are associated to the GenBank annotations "Threonine dehydratase" (e.g. UniprotKB AC: E3ERF0), "Beta-galactosidase" (e.g. UniprotKB AC: D6Y6J1) and "ATP-dependent DNA helicase" (e.g. UniprotKB AC: G0FLF9), respectively. The small circles close to the Transitivity Clustering IDs indicate the cluster size. Using only these three features the decision tree already obtains an accuracy of 89.9%.

References

[EB14] Barbosa E, Röttger R, Hauschild A-C, Azevedo V, Baumbach J: On the limits of computational functional genomics for bacterial lifestyle prediction. Briefings in Functional Genomics 2014.