HL108

On the limits of computational functional genomics for bacterial lifestyle prediction

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

1 Department of Mathematics and Computer Science, the University of Southern Denmark, Denmark

2 Max Planck Institute for Informatics, Germany

3 International Max Planck Research School for Computer Science, Germany

4 Bioinformatics Department, Federal University of Minas Gerais in Brazil, Brazil

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 (HPs), broad-spectrum pathogens (BPs), opportunistic pathogens (OPs) and non-pathogenic (NP).We hypothesize: (H1) Pathogens (HPs and BPs) possess specific pathogenicity signature genes. (H2) The same holds for OPs. (H3) Broad-spectrum and exclusively HPs cannot be distinguished from each other because of an observation bias, i.e. many HPs might yet be unclassified BPs. (H4) There is no intrinsic genomic characteristic of OPs compared with pathogens, as small mutations are likely to play a more dominant role 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 NPs (H1).We further show a clear limit in differentiating OPs from both NPs (H2) and pathogens (H4). HPs may also not be distinguished from bacteria annotated as BPs based only on a small set of orthologous genes (H3), as many HPs 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.