Predicting prokaryotic incubation times from genomic features

Maeva Fincker - mfincker

Project proposal
CS 229 - Fall 2016

We have barely scratched the surface when it comes to microbial diversity: for every microorganism we can culture and study in lab, there exist at least another 50 in the environment that are labelled “unculturable”. Although most of these unculturable strains resist growing under laboratory conditions, recent advances in sequencing has given us access to their genome. One of the main questions that bioinformatics in the field of microbiology is trying to answer is the following: how can we infer function from genomic information without access to experimental validation? [1] For the class project, I will focus on one specific microbial physiological function: the growth rate of a microbe (how fast it can divide). More specifically, **I will try to estimate the growth rate of a microorganism (using incubation time as a proxy) from genomic features.** While similar attempts have been made to predict other functional characteristics such as type of metabolism [2], no study has tried to predict growth rate from genomic information to my knowledge.

The project can be divided in 2 main phases: data gathering and estimation of the growth rate via supervised learning algorithms. There are currently 7016 complete or near-complete prokaryotic (bacteria + archaea) genomes in the NCBI database. Features such as GC% content, size, and E.C. numbers can be relatively easily extracted from these annotated genomes. While genomic information is readily accessible, there is no good database recording growth rates for all sequenced prokaryotes. I will therefore need to mine the literature for growth rate values. I am hoping that a simple heuristic search will be enough to extract growth rates from the literature as I do not wish for this project to turn into an NLP one. I am aiming for a minimum of 1000 examples. The second phase of the project - prediction of the growth rate - can be further divided in two. My goal is double: to predict growth rate accurately given a genome, and to understand which features play an important role in the growth rate determination. Distinct classification algorithms are differentially suited to answer these 2 sub-aims and I expect that a more constrained model will be required to identify features of importance.

References:

[1] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, Jun. 2015.

[2] F. M. Lauro, D. McDougald, T. Thomas, T. J. Williams, S. Egan, S. Rice, M. Z. DeMaere, L. Ting, H. Ertan, J. Johnson, S. Ferriera, A. Lapidus, I. Anderson, N. Kyrpides, A. C. Munk, C. Detter, C. S. Han, M. V. Brown, F. T. Robb, S. Kjelleberg, and R. Cavicchioli, “The genomic basis of trophic strategy in marine bacteria.,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 37, pp. 15527–15533, Sep. 2009.