Predicting prokaryotic incubation times from genomic features

Maeva Fincker - mfincker

Project milestone
CS 229 - Fall 2016

Introduction

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?{Libbrecht:2015bb} For the class project, I am focusing on one specific microbial physiological function: the growth rate of a microbe (how fast it can divide). More specifically, **I am trying to estimate the growth rate of a microorganism (using incubation time as a proxy) from genomic.** While similar attempts have been made to predict other functional characteristics such as type of metabolism {Lauro:2009uu}, no study has tried to predict growth rate from genomic information to my knowledge.

Dataset

Predicting growth rate, a continuous variable, from genomic features is inherently a regression problem. The growth rate of a microorganism describes how long it takes for a cell to divide. When a new organism is isolated, microbiologists report growth rates in various ways, which make direct comparisons difficult, if they even measure such a physiological trait at all. Since mining scientific literature for growth rates values could be a machine learning project on its own that I did not wish to tackle, I decided to use **incubation time** as a proxy for growth rate. Incubation times are reported by strain collections and represent how long approximately one has to wait to observe full growth of a microbe. The BacDive database [ref] contains the incubation times of 2312 microorganisms in 6 different buckets of time: 1-2 days, 2-3 days, 3-7 days, 8-14 days, 10-14 days and > 14 days. Out of these 2312 organisms, 783 have a complete genome (or ordered scaffold genome assembly) available on NCBI [ref] but only 605 of these genomes have been properly annotated. Therefore, my dataset is composed of 605 genomes and their associated incubation time.

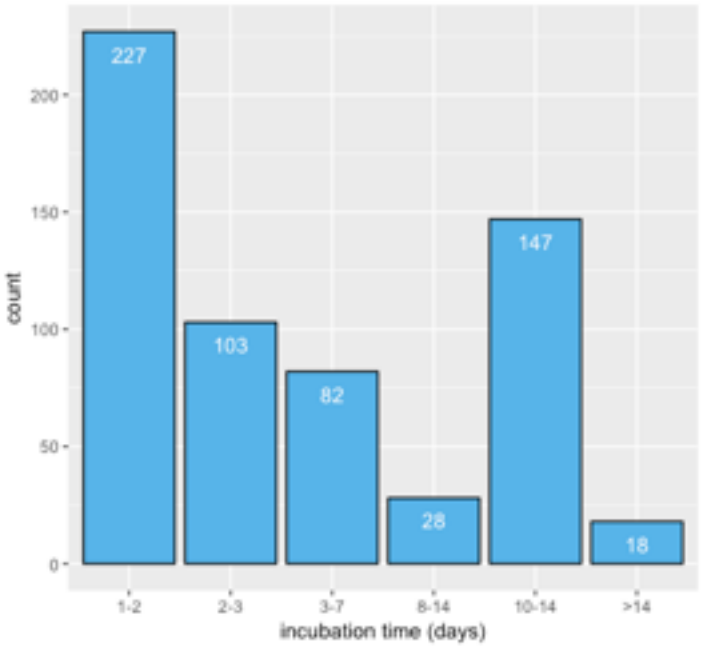

Fig. 1 : Incubation time histogram - whole dataset

Given the hypothesis that incubation time is dictated by metabolic capabilities and functions, I focused on extracting metabolically-relevant features from the genomes: number of proteins in represented Pfam families [ref] and the length of the genome. The Pfam families are clusters of homolog proteins which share similar functions. Given a genome, I counted the number of proteins belonging to each Pfam families. The main drawback of using Pfam families as features is that this genome preprocessing yielded a total of 14618 features, which is unreasonable when I only have access to 605 examples in total. Keeping Pfam features that appear in at least 3 genomes lowered the total number of features to 7549. This number is still very high but I am hoping that feature selection methods and L1-regularization (to keep some of the features' weight to 0) will help me identify significant features. (see **future work**). In retrospect, I should have started with feature selection before trying to implement classifiers, but instead I tried to implement a softmax classifier as well as a one-vs-all SVM. As one could have expected, these attempts did not yield significant results but let me identify issues that needs to be addressed to before I can get a potentially useful classifier.

Preliminary analyses

All of my features but one (the genome length) are low counts (no greater than ~100), whereas genome length is in millions of base pairs. I therefore normalized each feature vector by its total sum so that values are now between 0 and 1. This may not be the best way for normalizing my dataset and I will experiment with other normalization. For my preliminary implementation of classifiers, I split the data into a training set (535 observations) and a test set (70 observations).

Softmax regression

My first approach was to implement a simple softmax regression with 6 classes corresponding to my 6 incubation time bins with stochastic gradient descent on batches of 10 training examples. This did not perform well and reaches an accuracy of 61% on the test set. I did not expect it to work well as my data is extremely noisy with a potentially high number of features that are irrelevant.

SVM

I then tried to implement a one-vs-all SVM classifier. I first trained 6 binary-SVMs, one for each label class where I considered the training examples labelled from one class as positive and training examples from all other classes as negative. This did not work at all: the 6 SVMs classified all examples as being negative and were not able to separate my data. This might be due to the fact that my data is not linearly separable (the use of a kernel might help here). Another issue might be the imbalance between the number of positive and negative examples during training.

What's next

Preliminary attempts at classification did not yield usable results. My dataset is highly-dimensional and I will focus on feature selection in priority. Given the way Pfam families are built, I suspect that many of my features are redundant and highly correlated. Removing highly correlated variables should decrease the total number of feature. Additionally, using score-base

feature selection, such as the correlation score between each feature and the incubation time, should highlight the most significant features to use. Once the number of variables has been lowered enough, I will implement different classifiers and compare them (SVM with one-vs-all or one-vs-one strategies, RF). In addition, I will test different regularization terms as a regularizer can help me keep weights close to zero even with high-dimensionality.

I realize that there is a lot more to do for this project and I would like to have more time to work on it but unfortunately as a 4th year PhD student in environmental microbiology, my research takes precedence over classes. Please, I would appreciate if, in your feedback, you could let me know how much work is needed on the project to get credit for the class, which I am taking pass/fail (given that I got good grades on the midterm and the homeworks so far). Thank you.