Random subsampling techniques for sea bass mortality prediction

Giovanni Gaio, Simone Moretti

June 26, 2025

1 Introduction

The use of genomes to predict the effects of disease has been largely celebrated. But actually finding such links is usually not easy, and an automatic data-driven technique to predict disease outcome would be very useful. Given the size of genomes it is not always easy to understand the genes that have an phenotype related to the disease effects. The size of the datasets hinders the efficacy of standard machine-learning techniques, as few example genomes are available but each is composed of hundreds of thousands of bases, if not many millions.

The scope of this work is to experiment with some simple random subsampling techniques. The idea is to keep only a subset of the genome bases while keeping all the individual genomes sequenced. The hope is that this will allow a standard machine-learning algorithm to better understand the structure and importance of the gene with respect to the disease outcome.

2 Methodology

We tried to evaluate the effectiveness of random subsampling of genes in improving predictions of mortality.

2.1 Dataset

The main dataset we used was a table filled with the genome of 990 individual sea basses. The sequenced genome at our disposal was made up of 6072853 individual genes.

In addition to the set of genes for some individuals, we had a set of annotated genes. For a subset of genes some further information is known: a function (either `Open_chromatin`, `Enhancer` or `Promoter`) and a tissue number (between 0 and 25).

2.2 Subsampling techniques

We always subsampled selecting a subset of genes in the whole dataset.

The idea is to help our estimator to not get lost in the $\sim 10^6$ genes, but allowing it to work with only a smaller number of genes at a time.

Once we subsampled the dataset, we used the `XGBoost`[1] library to construct a predictor in a fast and easy manner.

Uniform subsampling. The first and simplest thing we tried was to randomly and uniformly subsample the genes on the entire genome. We selected a given fraction of the genes, keeping or discarding each gene with fixed and uniform probability.

Uniform subsampling on chromosomes. A second possibility is to uniformly and randomly sample a fixed number of genes on each chromosome.

2.3 Annotated genes

Using the additional information at our disposal we used only a subset of annotated genes. Among these we further subsampled to obtain a small feature set.

Subsampling using annotations. Some genes have been linked to specific organs, tissues or functions. This information can be used to select and sample genes with only some function or organ and subsample among these.

3 Results

In this section we'll look briefly at the results we got from each of the different techniques outlined in section 2.

3.1 Uniform subsampling

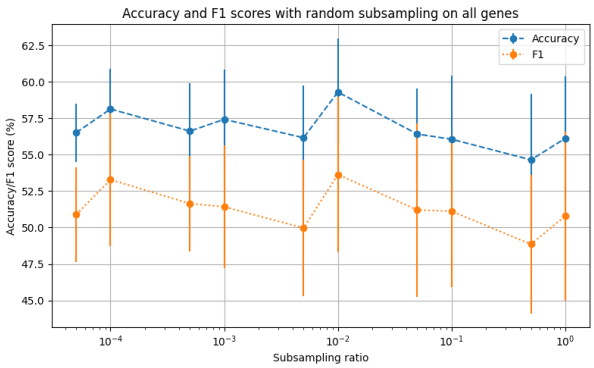

Figure 1: Plot of the accuracy and F1 scores when subsampling uniformly on the whole genome.

The first simple test was done selecting a random subset of genes. The results for each subsampling rate are shown in Figure 1.

3.2 Subsampling uniform on chromosomes

3.3 Annotated genes subsampling

4 Conclusions

References

[1] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

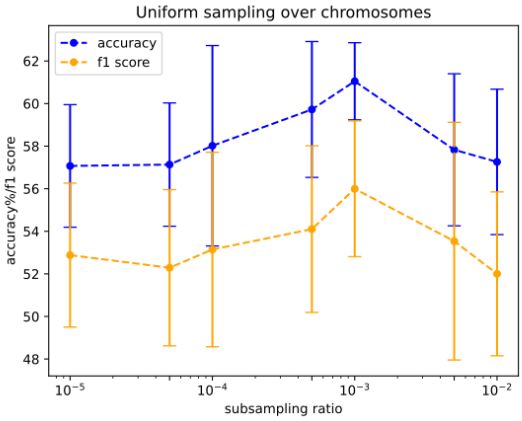

Figure 2: Plot of the accuracy and F1 scores when subsampling uniformly on each chromosome.

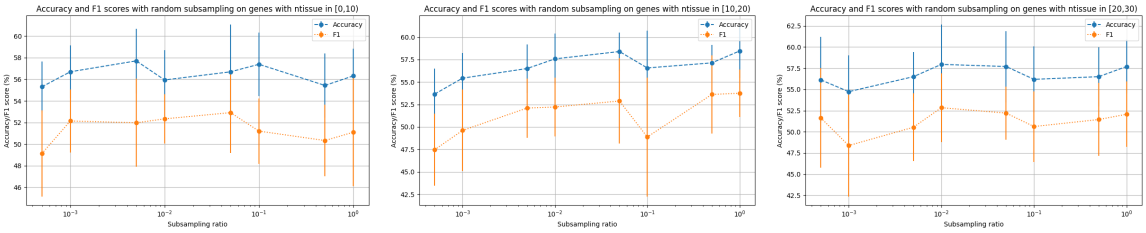

Figure 3: Plots of the accuracy and F1 scores while using only genes with given tissue number, on the x-axis there is the subsampling rate.