acidoseq: a tool for studying unclassified Acidobacteria reads

A Preprint

Samantha C. Pendleton [1,*]

1 Department of Mathematics, Physics, and Computer Science; Aberystwyth University, UK

* samanfapendle@outlook.com

Last Updated: 10 December 2019

Abstract

Motivation: Acidobacteria's many recovered sequences are labelled as 'unclassified' due to the lacking knowledge of this recently discovered phyla. Acidobacteria's subdivision presence in soil is dependant on pH. Furthermore, subdivisions have sequences with a consistent GC content. In order to predict some unclassified sequences into subdivisions we present, acidoseq, a tool to sort the unclassified sequences into subdivisions based on GC content and pH of the soil sample.
Results: A data-set of Nanopore sequenced reads, from an Aberystwyth soil-sample, was annotated and analysed with tools: Kaiju and acidoseq - using a pH of 6.25 as a future prediction from a trend of the Countryside Survey: Topsoil Soil pH, UK Soil Observatory from 2007, 1978, and 1998 respectively. We found from a BLAST job a presence of subdivisions 6 and 4. As expected, the unclassified reads were placed into subdivision 6 and 4, which are highly abundant in this pH score.
Availability: acidoseq, a Python package, is available, open-source, on GitHub under the MIT license, with documentation and installation instructions: https://github.com/sap218/acidoseq. Also available on PyPI: https://pypi.org/project/acidoseq/ for other operating systems.

Keywords: sequence analysis, acidobacteria, bioinformatics

1 Introduction

Acidobacteria was only recently recognised as a phylum [1] in 2012 [2]. Despite newly discovered, Acidobacteria is one of the most abundant phyla in Earth soil and one of the most diverse bacteria [3]. Whilst diverse, a majority of Acidobacteria reads recovered are unclassified due to lack of knowledge.

This phylum has many class groups, known as subdivisions. These subdivisions exist in certain soil pH; subdivisions 4 and 6 are highly abundant as pH increases, whilst 1 and 3 have a negative correlation [4]. Furthermore, the GC content of sequences in the same subdivision have a pattern; e.g. subdivision 5 sequences, on average, have a GC content above 60\% [5].

Our research question consists of understanding unclassified reads and gaining more knowledge from them: studying sequence similarity.

We produced, acidoseq. The first package designed specifically for Acidobacteria, acidoseq, is a Python v3.5 package that relies on three inputs.

2 Method

The sample for our study was extracted from soil in Aberystwyth. This data-set was provided from Aberystwyth University; Institute of Biological, Environmental and Rural Sciences. Supplied from a BBC Radio 4 podcast event: showing real-time sequencing from the portable MinION. The data we studied was a collection of Nanopore reads that were converted into FASTA format.

One input includes a list of Acidobacteria species with corresponding NCBI taxonomy ID (this is provided via the Git repository). The other two required inputs are csv files:

- an output from Kaiju. Kaiju classifies metagenomes and returns a list of sequences, labelling those classified or unclassified (the NCBI taxonomy ID for those classified), plus more information. acidoseq only needs two columns from this output: seq ID and NCBI taxon ID for those classified.
- item the full FASTA file. acidoseq requires this file in order to extract the reads that are Acidobacteria (via the seq IDs from Kaiju which the NCBI taxon IDs are the same as the list of Acidobacteria corresponding ID file).

To understand the average GC content within subdivisions, we looked at current, available Acidobacteria genomes. As of 2016, a total of 10 fully published genomes were available [6] via NCBI. As of Tuesday 25th September (2018), 159 partial genomes were available. These full and partial, classified, genomes were downloaded and analysed: specifically looking into the GC content within subdivisions, see table 1.1, in order to create the prediction of subdivisions within pH. We analysed full genomes alone, partial separately too, then combined to gain an average.

Class	**Subdivison**	**Full**	**Partial**	**Other**
Acidobacteriia	1	58.13	57.95	
	2*			57.6
Solibacteres	3	61.9	62.17	52.75
Blastocatellia	4	61.31	58.87	
	5			65.43
	6*	67.22		
Holophagae	8		66.84	
	10*			
	13*			58.5
	22*			67.15
	23*			63

Table 1.1. of the various subdivisions of Acidobacteria and the mean GC content (figures rounded two decimal places) from various sources: NCBI *full* genomes, Latest Refseq *partial*, and *other sources* (papers and NCBI brief descriptions). *unclassified subdivisions.

3 Features

acidoseq is the main script where the analysis occurs, however, a user must know the pH of the soil in order to run it. First a user is recommended to use acidomap. acidomap is a command line script and a major of this package. acidomap is provided for users who don’t know the pH of the soil sample: a

user simply inputs a city/town for a graphical visualisation. The plot is based on the Countryside Survey: Topsoil – Soil pH 2007 data. There a map of the UK with soil pH scale shaded across, with the entered city/town marked on in a transparent circle in order to visualise the possible soil pH.

Despite stating acidoseq requires a Kaiju output, it actually could use other tool output files: as long as there remains a column for sequence ID and corresponding annotated NCBI taxon ID.
The Github repository includes instructions about cutting columns of a Kaiju output file, however a user could use a Kraken (a software for assigning taxonomic labels to DNA sequences) output file if they were to cut columns correctly.

A user inputs, via a command line interface, if they wish to gain information of all Acidobacteria content or only looking into the unclassified reads. Looking at all the classified and unclassified reads together provides limited information due to knowledge of subdivision location for those classified. However, if the user chose to study only the unclassified reads, further analysis is provided based on sorting these unclassified reads into subdivisions: based on GC and pH.

- Prints the coverage percentage of total sequences which were classified via Kaiju from the whole FASTA file.
- Outputs a file which includes all of the unclassified Acidobacteria sequences - available if users wish to do further analysis.
- Statistical information of the list of unclassified reads: information including maximum/minimum, and mean of AT and GC content, plus read-length.
- Plots of the AT and GC comparison (to observe the DNA stability of the reads [7]). Another GC histogram with regions of subdivisions average GC highlighted (only certain subdivisions are present on the plot - based on pH). The style of these plots are chosen by the user via command line option, explained below.
- Separate files are returned which contains the reads that resided in the predicted subdivisions, allowing for further user analysis - we used these files to run a BLAST job.
- As mentioned, a user can choose the design of their plots. This is a minor feature of acidoseq which was implemented for user preference: some users prefer style consistency in their work. A command line option requires the user to input a plot style, such as `ggplot' or `seaborn' or `classic' - a list of styles are provided via the repository. If the user ignores this parameter, or misspells, then the script uses a random style.

4 Results

With acidomap's results, see figure 1.1, we used a pH of 6.25, including a prediction from a trend of 2007, 1978, and 1998 respectively (Countryside Survey: Topsoil Soil pH, UK Soil Observatory).

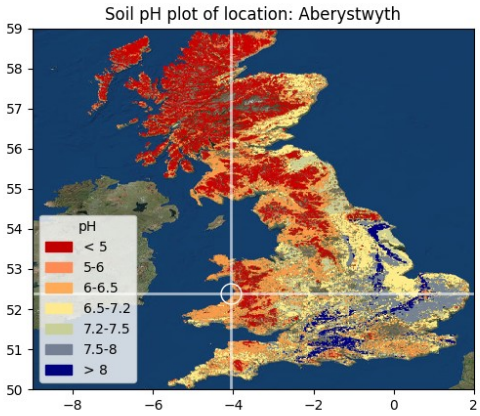

Figure 1.1. acidomap output of pH of soils across the UK where a particular location is highlighted based on text input.

acidoseq's output files of unclassified reads included subdivisions 6, 4 and 23 - as one would expect with this pH score, see figure 1.2. Using Blast2Go, we ran a BLAST job to study the content of these sequences further, we can see that there is definitely proof of subdivision dependency on pH and a similarity in GC content.

- Subdivision 6 output: highest sequence similarity was 90.91\% Luteitalea pratensis, which resides in subdivision 6 of Acidobacteria.

- Pyrinomonas methylaliphatogenes strain type strain:K22; species resides in subdivision 4.
- An unclassified species was identified: Acidobacteriaceae bacterium URHE0068; this result can lead to further understanding and potential further investigation.

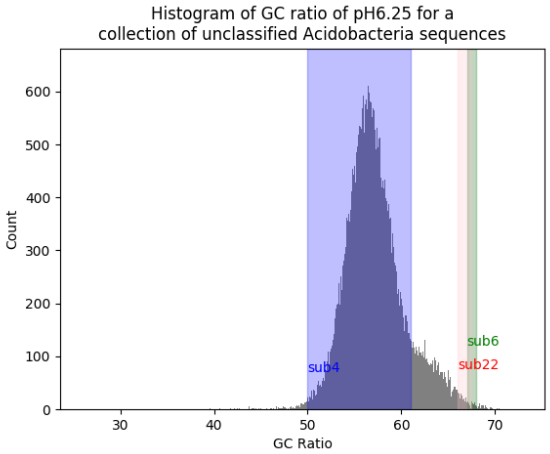

Figure 1.2. acidoseq output of displaying which unclassified sequences residing in the subdivison best suited.

5 Discussion

We were able to extract the unclassified reads from Kaiju and predict them into subdivisions based on pH and GC. After concluding our BLAST job, our results display evidence that our technique may reveal sequence similarities with these predicted subdivisions: sequence similarities, subdivision similarities via GC content.

Acknowledgements

The Countryside Survey: Topsoil Soil pH, UK Soil Observatory (2007) data is supplied by Natural Environment Research Council - data of soil pH resource is available under the Open Government Licence and relevant key publications: Digital Object Identifiers. This research study explained that their data/information could be used with enough citation/references to owners, which their requirements have been met. We would like to acknowledge them for the intellectual engagement.

References

[1] Anna M Kielak, Cristine C Barreto, George A Kowalchuk, Johannes A van Veen, and Eiko E Kuramae. The ecology of acidobacteria: moving beyond genes and genomes. Frontiers in Microbiology, 7:744, 2016.
[2] J Cameron Thrash and John D Coates. Phylum xvii. acidobacteria phyl. nov. In Bergey's Manual R © of Systematic Bacteriology, pages 725–735. Springer, 2010.
[3] Susan M Barns, Elizabeth C Cain, Leslie Sommerville, and Cheryl R Kuske. Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. Applied and environmental microbiology, 73(9):3113–3116, 2007.
[4] Andrea K Bartram, Xingpeng Jiang, Michael DJ Lynch, Andre P Masella, Graeme W Nicol, Jonathan Dushoff, and Josh D Neufeld. Exploring links between ph and bacterial community composition in soils from the craibstone experimental farm. FEMS microbiology ecology, 87(2):403–415, 2014.
[5] Achim Quaiser, Torsten Ochsenreiter, Christa Lanz, Stephan C Schuster, Alexander H Treusch, Jürgen Eck, and Christa Schleper. Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. Molecular microbiology, 50(2):563–575, 2003.
[6] Naomi L Ward, Jean F Challacombe, Peter H Janssen, Bernard Henrissat, Pedro M Coutinho, Martin Wu, Gary Xie, Daniel H Haft, Michelle Sait, Jonathan Badger, et al. Three genomes from the phylum acidobacteria provide insight into the lifestyles of these microorganisms in soils. Applied and environmental microbiology, 75(7):2046–2056, 2009.
[7] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. Nucleic acids research, 34(2):564–574, 2006.