MinCE meets kallisto: fast quantification of large metagenomics datasets along with species and strain abundance

Student: Thorhallur Audur Helgason

Mentor: Professor Lior Pachter

The following is the project plan for my work this summer at Caltech. The field of research is Bioinformatics / Computational Biology, and my work will be a continuation and expansion of a software tool I started to develop last summer, when working for Professor Páll Melsted at the University of Iceland.[1] The working name of the method and software is MinCE, and before detailing my goals for this summer, I describe the current function of MinCE and the problems it is designed to solve.

MinCE

A genome is an immensely information-rich molecule but that information is completely coded with only 4 types of nucleotides; A,C,G,T for DNA and A,C,G,U for RNA. The human genome consists of $\sim$ 3 billion base pairs of DNA, and due to a bottleneck in the human population approximately 10,000 years ago, the genomes of every pair of individuals are highly similar. Genomes of other organisms vary in length and divergence in genome sequence generally correlates with divergence in phenotype. This presents various problems (and some opportunities), from which stems the field of Bioinformatics. Two such problems are comparison between genomes and identifying the species and/or variant of an unknown individual solely from sequencing data. Both problems have to deal with the issue that no single, whole genome can be said to represent the species as a whole, since variation within species is always present. As a result, a direct comparison of 'letter-by-letter' between genomes is not only extremely slow, but in fact doomed to fail.

Therefore another approach is required. By processing every overlapping subsequence of length 31 within the genome with a hash function, the genome can be represented as a collection of the lowest N hash values resulting from the process, where N is a fixed size. This collection is referred to as a sketch and was first proposed in Mash, a bioinformatics software tool which MinCE builds upon.[2] This sketch is effectively a random subset of the sequences comprising the genome and can be used as a quick substitute for comparison between genomes, the reliability of which is mathematically supported. With a database of such sketches, raw sequencing data of unknown organisms can be processed into a sketch and paired to its closest sketch match.

However, this approach is insufficient when comparing very related genomes, as the sketches will prove identical for multiple species or variants within species. MinCE proposes a solution to this, by sketching a large database of genomes and cataloging the specific clusters of nearly identical sketches which result from this process. By pre-processing these clusters, we can build a deBruijn graph from each cluster and algorithmically select specific sequences of nucleotides from the graph. The sequences are selected, such that they collectively distinguish between every individual within the cluster. By systematically searching for these specific sequences in some unidentified sequencing data, we can extend the capabilities and the specificity of this approach. MinCE currently runs on a database of roughly 285,000 *Eubacteria* and *Archaea* genomes, with a bare-bones command line interface to query fastq or fasta files. Most aspects have been translated into C++ but certain algorithms, which are particularly complicated, still only run through Python. The program can currently be downloaded and run locally, with the entire program and the pre-processed database only taking about 3Gb of space.

[1] https://github.com/mannaudur/MinCE

[2] https://github.com/marbl/Mash

Phases and time table

My work this summer can be split into three phases, detailed below. The total time I estimate for these objectives is 8-9 weeks.

1. Completely finish work on MinCE in its current form. (Estimate 1-2.5 weeks)

My estimate for this phase is extended into the possibility of 2.5 weeks to account for any set-up time related to my work and getting oriented. The completion of MinCE refers to a reworking of certain key-algorithms, a refining of the search-index architecture (which can be drastically improved with respect to speed) and translating the whole program into C/C++. Each of these problems is isolated and I expect they can be resolved efficiently. Finally, there is a need to break some of the larger clusters into smaller, overlapping sets. The union-find approach to cluster identification used presently can result in a domino-effect where large clusters have a tendency to grow even larger. As a result, a few clusters are impractically large. This is a cost of preferring fewer, non-overlapping cluster sets, rather than many overlapping sets of a fixed metric radius. A tentative solution to this problem could be to break these few largest sets into smaller sub-clusters which in fact do overlap and are each funneled into a deBruijn graph. These sub-clusters would most likely be identified and arranged with implementation of the Leiden algorithm.

2. Integrate MinCE into kallisto. (Estimate 4 weeks)

kallisto is a program for quantifying abundances of transcripts from bulk and single-cell RNA-Seq data, written by Lior Pachter and Páll Melsted.[3] Single-cell RNA-Seq can in particular be used to pair unique transcripts of certain genes to their corresponding cells and thus shed a light on cell- and organ-wise specific translation of the genome. Thus the methodology of representing species or strains within species as sketches, with deBruijn graph extensions where appropriate, could also be applied within an organism, with sketches representing either unique transcripts, certain cell groups or a broader set of equivalence groups between the two. I will work with Prof. Pachter's student Delaney Sullivan on implementing this bridge between the two software tools and subsequent work on realizing the potential of the merge.

3. Run the kallisto-integrated MinCE on metagenomic datasets with subsequent revision of MinCE's structure based on results. (Estimate 3 weeks)

Metagenomics is the study of genetic material recovered directly from environmental samples, defined by Kevin Chen and Lior Pachter as "the application of modern genomics technique without the need for isolation and lab cultivation of individual species".[4] Raw sequencing data from real-life samples, such as a swab from a restaurant kitchen counter or a fecal sample from a human gut, should be expected to contain multitudes of species of organisms. The raw data from standard sequencing techniques is output as small snippets of read genetic material and is completely unorganized with respect to its order within a genome and respective species-origins - it's effectively a soup of data. MinCE has the capacity to receive such data directly and identify multiple matches within the pre-processed database it works upon. This has yet to be verified with real metagenomic data and the last phase would kick off with tests to confirm this feature. If these tests prove successful, I expect they will shed light on various opportunities for the software tool. This point is admittedly vague but with so many moving parts leading up to this final phase, I expect that the focus will become clearer as it draws nearer, in particular with Prof. Pachter's and Delaney's insight.

[3] https://pachterlab.github.io/kallisto/about

[4] https://en.wikipedia.org/wiki/Metagenomics

What I will need

I will work with graduate student Delaney Sullivan and with Prof. Pachter to obtain the data I need, as well as for advice on algorithmic questions. I will also benefit from a connection to a remote server to work on, as constructing the deBruijn graphs and extracting the necessary sequences is a slow process and is best left to simmer while other work goes on. Such a server will be provided by the Pachter lab.

Project value

The utilization potential of this program is enormous. Raw sequencing data can be piped directly into MinCE, without the need for any aligning of reads or extra processing, and the result will in theory detail every species and/or strain from the database found therein. MinCE will therefore allow for quantification of large metagenomics datasets along with species and strain abundance, a task that is currently extremely challenging. For the food industry, public health sectors and medical diagnoses, this has the potential to radically expedite processes which until now have been very time consuming. With the integration of kallisto, academic utilization within genetic research, developmental biology and medicine is also promising. The full scope of this combination between kallisto and MinCE is sure to reveal itself further through the summer's work.

Thorhallur Audur Helgason

Reykjavík, Iceland

May 16th 2022

References

1. https://github.com/mannaudur/MinCE
2. https://github.com/marbl/Mash
3. https://en.wikipedia.org/wiki/Metagenomics
4. https://pachterlab.github.io/kallisto/about