Studying three-dimensional genome organization using Hi-C

HARRIS A. LAZARIS
BMI Final Project
April 29, 2014

Summary

Morbi tempor congue porta. Proin semper, leo vitae faucibus dictum, metus mauris lacinia lorem, ac congue leo felis eu turpis. Sed nec nunc pellentesque, gravida eros at, porttitor ipsum. Praesent consequat urna a lacus lobortis ultrices eget ac metus. In tempus hendrerit rhoncus. Mauris dignissim turpis id sollicitudin lacinia. Praesent libero tellus, fringilla nec ullamcorper at, ultrices id nulla. Phasellus placerat a tellus a malesuada.

Keywords: Hi–C , 3D genome organization , evaluation

Introduction

Background information

Background information goes here ...

Motivation

Motivation goes here ...

This statement requires citation [**?**]; this one does too [**?**]. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean dictum lacus sem, ut varius ante dignissim ac. Sed a mi quis lectus feugiat aliquam. Nunc sed vulputate velit. Sed commodo metus vel felis semper, quis rutrum odio vulputate. Donec a elit porttitor, facilisis nisl sit amet, dignissim arcu. Vivamus accumsan pellentesque nulla at euismod. Duis porta rutrum sem, eu facilisis mi varius sed. Suspendisse potenti. Mauris rhoncus neque nisi, ut laoreet augue pretium luctus. Vestibulum sit amet luctus sem, luctus ultrices leo. Aenean vitae sem leo.

Nullam semper quam at ante convallis posuere. Ut faucibus tellus ac massa luctus consectetur. Nulla pellentesque tortor et aliquam vehicula. Maecenas imperdiet euismod enim ut pharetra. Suspendisse pulvinar sapien vitae placerat pellentesque. Nulla facilisi. Aenean vitae nunc venenatis, vehicula neque in, congue ligula.

Pellentesque quis neque fringilla, varius ligula quis, malesuada dolor. Aenean malesuada urna porta, condimentum nisl sed, scelerisque nisi. Suspendisse ac orci quis massa porta dignissim. Morbi sollicitudin, felis eget tristique laoreet, ante lacus pretium lacus, nec ornare sem lorem a velit. Pellentesque eu erat congue, ullamcorper ante ut, tristique turpis. Nam sodales mi sed nisl tincidunt vestibulum. Interdum et malesuada fames ac ante ipsum primis in faucibus.

Materials-Methods

Results

Complexity–Timing Analysis

Focus area

I focused on the use of Python packages (Numpy, Pandas) as well as on visualization (with matplotlib and Pandas).

Conclusion-Discussion

It is clear from the analysis presented in this study that even state-of-the-art Hi–C analysis techniques do not give extremely reproducible results. When two different restriction enzymes are used with the same sample as source, while the correlation is relatively high, it is not ideal. Moreover, when the resolution is increased (128kb bins instead of 4096kb bins), the correlation even in the case of technical replicates treated with the same enzymes is very low. Thus, new more robust analysis techniques that lead to more reproducible results, are required. This is going to be a large part of my PhD thesis work.

While I optimized the code for speed, by using Numpy instead of native Python to deal with matrices and list comprehension insted of for loops when possible, the real bottleneck is the I/O operations (reading input and writing output). This is critical as the matrix files that I have to deal with, are really large. One solution to the problem would be to minimize I/O operations by using Numpy for everything, but this would require much

better knowledge of Numpy than I currently have. Moreover, Numpy may be not as versatile or efficient as R in certain circumstances which means that I may be unable to fully replace R with Numpy. Another approach would be to use rpy or rpy2 for R/Python integration but I have been always facing problems with rpy/rpy2 installation on my system.

As far as the visualization is concerned, both Pandas and matplotlib look interesting but:

1. They need time to learn.
2. There are many requirements (dependencies etc) and in many cases it is difficult to run code using these packages on the cluster.
3. The resulting graphs, especially in the case of Pandas, do not seem to be extremely customizable. Moreover, I had to write more code to achieve the same result I would get with writing less code in R.

I will certainly try to explore Numpy, matplotlib, Pandas and other packages further, as it is always possible that the drawbacks I see right now compared to R may be just due to luck of expertise on all these Python packages.

References