Building a NCBI web scraper and deep-learning model for inferring host pathogen genomic interactions

Jacob M. Green

Accessing and applying genomic information is one of the greatest challenges in the era of next generation sequencing. An immense amount of genomic information is stored on the National Center for Biotechnology Information (NCBI) which includes DNA and RNA sequences, protein information, chromosome level structuring, gene expression, sequence homology, and a large archive of associated research papers. NCBI is also connected through every major bioscience database giving it unparalleled access to a myriad of other resources. Through this database researchers can access many different types of genomic information and apply bioinformatic analyses to answer biologically relevant questions. The ability to data mine this resource has become increasing more important as sequencing is more accessible and the cost and quality of generating this data has significantly improved since the human genome was sequenced in 2001. A combination of machine-learning and data mining offer pattern-based approaches to resolve some of the heuristic challenges when applying algorithms to accessing and processing genomic information.

Host-pathogen genetic interactions can fundamentally reshape host genomic architecture. Host pathogen interactions are the foundation for germ theory, which states that microorganisms or "germs" can infect a host and cause disease. The ability for a pathogen to cause disease is predicated on many factors including the genetic content of the pathogen. How a host fights these pathogens, experiences a diseased state, and recovers from cellular or physiological damage is dictated by the host immune response. Both the innate and adaptive immune system are supported by the expression of genes that can recognize pathogens, transmit a signal to the cell or surrounding cells, and fight the pathogen. Therefore, the ability for a pathogen to infect a host and the ability to a host to defend against a pathogen is based on the genetic material of each organism, which is an indirect interaction. In many ways the host-pathogen genomic interaction is indirect and in others much more direct. In some pathogens such as the herpesvirus, genetic content from the virus can undergo introgression, a process where genetic information is transferred from one species to another. This introgression of pathogenic genetic information leaves artifacts in the host genome and can contribute to the diversification and development of gene families such as the superfamily of immunoglobulins. Few tools have been developed to identify the level of pathogen introgression into host genomes.

For this project I propose to design an NCBI web scraper and deep-learning model for accessing host and pathogen genomic information and apply an alignment-based method to infer levels of pathogen introgression. I will apply this workflow to known human pathogens and identify different levels on introgression in the human genome. This project is solely software based and will leverage NCBI and associated databases for data mining, open source tools available in Rstudio packages such as Rentrez and bioconductor, Anaconda environments to build program structures for alignment, and programming languages like Python and Unix to develop the working script. This project will be headed by myself with the assistance of Dr. Brice Loose for the CSC-593 course. Outside resources such as access to the school Blue Waves server and other computational skills support in networking languages, script design, and assistance with deep-learning models would be necessary to meet the goals of this project. To assess project completion, I will develop a working script that can alignment pathogen and host genomic information. This tool will also provide information on the level of introgression.

Jacob, this sounds like a fascinating topic. Are you focusing on a particular host/pathogen pair to develop a benchmark? This might be the most straight-forward approach, given that this would allow you to compare whatever scraping+learning you do against an existing knowledge base, before you journey into the outer unknown.

I think it will also be important to put some tighter bounds on what you want the machine learning model to do for you. In a few weeks, I will ask for an update. Hopefully you will have some of these questions and some more detail at that point. Also note the Brown HPC could be a useful tool?

10/10