Discovery and Annotation of Repeats, Signatures, and Patterns in Genomic Sequences

Michael Robinson, S.S. Iyengar, Fellow IEEE, Puneeth Iyengar* and Ning Xie

School of Computing and Information Sciences, Florida International University, Miami, FL 33199

*Department of Radiation Oncology, UT Southwestern Medical School, 5323 Harry Hines Boulevard

Dallas, Texas 75390

MOTIVATION

In this modern age of genomics, proteomics, and many other "-omics," vast quantities of information are generated but are associated with uncertainty in significance. Biologists do not have the wherewithal to identify computationally the portions of the genome that are significant to biologic disease processes and the information science experts do not have enough knowledge of these biologic processes to accurately model how to extract the relevant information. This difficulty is present in a setting of ever increasing production of data. Therefore, any efforts between biologists/clinicians and computational experts to better understand genomic data will be of significant potential benefit ultimately to patient care and survival.

We have offered several medical scenarios in which better understanding of the patterns and signatures within genomes would offer invaluable knowledge that would directly guide patient management and improve patient survival.

APPROACH

We have developed a new computational framework that finds patterns and signatures among genomic data files. This framework compares genomes of any size, locating exact sub-sequences found in the source genome and not in the target genomes. We first extract DNA, RNA or amino acid from genomic data files of multiple formats, including NGS. We then partition this data into sub-sequences of any length, creating new data structures. This new framework have many uses, among them:

1 - Creation of genomic fingerprints/signatures that are predictive or/and prognostic for specific biologic events or treatment outcomes

2 - Identification of sources and types of bacteria or viruses that spread in hospital or community settings,

3 - Confirmation of the presence of either new primary lesions or metastatic disease in all solid tumors, aiding in staging of cancer and treatments strategies,

4 - Tracking HIV/TB mutational changes in patients, offering the possibility for new treatment paradigms.

ALGORITHMS

We developed multiple algorithms to create our framework and the current applications, some of them are:

1 - Create sub-sequences from multiple genomic file formats

2 - Determine Minimum RAM space needed

3 - Determine Minimum Secondary storage needed

4 - Size of Sub-Sequences Data Files with Duplicates

5 - Determine Big O value for each step

6 - Validate our results in every step of the process

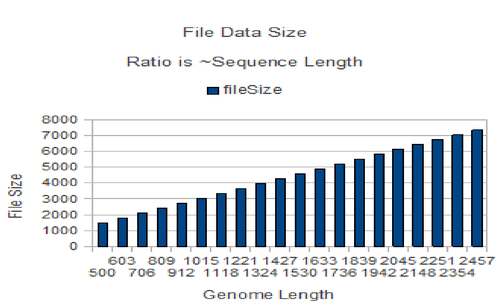


Figure 1. Using genomes of different lengths (from 500 to 2457), and extracting sub-sequences of the same length (3) producing files that are linear in size. The file size is a ratio of the fileSize and the genome length/size, which is the sub-sequence's length.

IMPLEMENTATION

We first obtain genomes that will serve as sample and reference genomes. At the present time we are using the human genome as a reference genome. We have also used the genomes of 5 Pseudomonas aeruginosa bacteria as sample and reference genomes. Any genome (bacteria, viruses, plants, any animal, etc.) can be used as a reference and/or sample genome.

Figure 1: Multiple biological life forms : Animal, Virus, Bacteria, Plant

Our framework extracts the DNA, RNA or Amino Acids from multiple data formats including Next Generation Sequencing (NGS) data files of any size. Utilizing a user selected size, we create new data structures containing unique sub-sequences and their correspondent indexes/locations including all repeats found in the genomes. In some cases these repeats are in the millions.

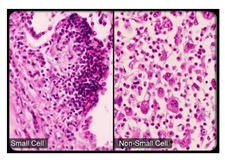
Faced with the constraints of compiling and understanding the unique and potentially clinically relevant aspects of extremely large genomic data sets, our group set out to develop new tools for data analysis emphasizing novel pattern recognition agents. Fundamentally, our new computational framework can process data files of any size, limited by traditional storage size and operating systems limitations.

The National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), houses GenBank, a database that serves as an archive for all publicly available DNA sequences. As of 15 April 2012, GenBank release 189.0 has 151,824,421 loci, 139,266,481,398 bases, from 151,824,421 reported sequences. GenBank has grown at an exponential rate, doubling in size every 18 months.

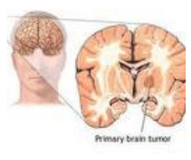
The GenBank database includes additional data sets which are constructed mechanically from the main sequence data collection, and therefore are excluded from this count. GenBank additionally houses over 1.2 Petabytes of sample data produced by machines from companies like Illumina.

ACCOMPLISHMENTS

Our first application is an approach to help with quality of life and improve the survival of Cancer related cachexia (CC) patients includes identifying a genomic signature that could predict the specific type of solid tumors that can induce CC, thereby allowing for earlier pharmacologic intervention. Any new knowledge regarding cachexia could be of high impact and importance for other associated diseases including rheumatoid processes, AIDS, tuberculosis, etc.

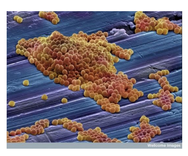


Lung Cancer



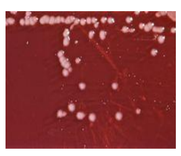
Brain Cancer

A second significant application of new tools for comparing genome data in elucidating signatures patterns, is in making fundamental bacterial, viral comparisons to understand spread of etiology of infections in hospital and community clinic settings.



B0006626 MRSA

Annie Cavanagh. Wellcome Images



Pseudomonas

Our third case study demonstrates a need for new computational tools that can help decipher patterns and signatures in vast quantities of genomic information. We conclude that a comparative genomic analysis would be far more precise than microscopic analysis or computed tomography (CT) scans, in determining if the cancer in one lung has spread to the other lung or if it has spread via the bloodstream from another part of the body.

VALIDATION

Validation processes were conducted in every step. The data validation process proved that the implementation of the data extraction programs worked correctly, and the search validation process proved that there were no false-positives.

PREVIOUS WORK

1 - Identification of Genomic Signatures for the design of essays for the detention and monitoring of Anthrax threats.

This work done by US Army Medical Research Institute of Infectious Diseases, et al. was able to distinguish B. anthracis from its close relatives B. cereus and B. thuringiensis. It is similar to our framework but it has limitations that we do not have.

In our research we examined 11 related works. The closest one found is the one mentioned above. All related works have limitations, see list below.

Related Work Limitations:

- Designed to be used for bacteria or viruses only
- Short sequences analysis only.
- Limitations in the size of the genome,
- Sequences < 20 bp get ignored
- Sequences >18bp are ignored
- Only sequences <10 or <11 are used
- Sequences from 15bp to 100bp are used
- Results have 18% error
- Use binary for 2 bit per base, therefore it only uses the A,C,G,T or ACGU bases.
- Some of the related works recognize that their results are not complete and that further work is needed.

FUTURE_WORK

Our framework allows us to develop many other useful applications. By providing these data structures to end-users, additional research can be done with easy to use tools such as Spread Sheets and Word Processors. These data structures can also be further analyzed with programs written in any programming languages that can access text files.

1 - User Batch File. User will determine where to obtain the raw data from and where to write results, including local servers and/or Cloud locations.

2 - Advanced Repeats. Find all sub-sequences that have repeats, including amounts, locations, distance between them, genes containing repeats.

3 – Analyze repeat properties found in all signatures such as: Are repeats in parts of the genomes where we have not found genes (areas known as junk).

4 - Find exact matches in multiple genomes.

5 - SearchPatterns. Find sub-sequences of any length.

6 - SearchSignatures. Signatures found with locations.

7 – Create DNA Fingerprint Libraries.

8 - Website. Accept/process users sequences.

9 - Software for end-user in-house use.

10 - Create customized software for end-users needs.

11 - Align Signatures to find sequences with mutations.

12 - Given a sequence (e.g. a tumor) find all tumors in a genome.

ACKNOWLEDGMENTS

We thank Dr. Shu-Ching Chen for his great support, and the institutions that provided the cancer tumor data samples used in our experiments.

We also thank Pilar, Mark and Daniel Robinson for their patience and great encouragement.