Poster: Towards Finding Unknown Genes: the GenomePro Framework

Michael Robinson and Naphtali Rishe
School of Computer & Information Sciences
Florida International University
Miami, Florida, USA
michael.robinson@cs.fiu.edu

***Abstract*—**We present a data processing framework aimed at facilitating the discovery of unknown genes in any genome, extracting DNA, RNA or Protein sub-sequences of any length. Using our GenomePro framework, we process raw input data files, of any size, in multiple formats such as NGS, fasta, and GBK, extracting all sub-sequences, of lengths selected by the end user. The only limitations are the computer storage size and/or operating system restrictions. Our framework can be applied to any life form genome.

Keywords – Genes, DNA, RNA, Proteins, Genome, Sub-Sequence, Sequence, Repeats, chromosomes, Dark Matter

I. MOTIVATION

The Human DNA contains approximately 20,000–25,000 known genes. The Human Genome contains about 3.2 billion bases/nucleotides. In these 3.2 billion bases, genes have been identified in about 2% of the genome; the remaining 98% is known as the dark matter. Examining dark matter areas leads to predicting the location of currently unknown genes.

A characteristic of the Human Genome data is its large size. The Genome Project started in 1990 and was completed in April 2003, for the first time sequencing the human genome and producing files of about 3.2 Gigabytes in size. During the next 10 years, a new sequencing method, called Next Generation Sequencing (NGS) was created, producing files of over 1.4 terabytes. NGS has changed genomics research allowing scientists to perform experiments that previously were not possible or affordable. NGS experiments generate unprecedented volumes of data, which present challenges and opportunities for data management, storage, and analysis.

II. APPROACH

Using our GenomePro framework, we process raw input data files, of any size, from multiple formats such as NGS, Fasta, and GBK, extracting all sub-sequences, of lengths selected by the end user. The framework can be applied to any life form genome.

The GenomePro framework includes new data structures containing genome sub-sequences of any length, with all their repeat locations and additional information. This allows to compare genomes from different species and from different times, detecting differences, changes, and similarities.

III. VALIDATION

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

IV. IMPLEMENTATION

Data Input: GenomePro uses genomic data text files of multiple formats and sizes as input, as seen in the sample data tables below. Some of these input files can be larger than one terabyte in size.

Fasta input files, shown in Table 1, differ in size from 1759 bases in the Porcine Circovirus Type 1, to millions of bases in bacteria, to billions of bases in the human genome. We obtain these files from NCBI (National Center for Biotechnology Information).

Table 1: Fasta Genomic Data Sample

Fasta Format: >gi|9629360|ref|NP_057850.1| Pr55(Gag) [Human immunodeficiency virus 1] MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHI VWASRELERFAVNPGLLETSEGCRQILGQLQPSLQ TGSEELRSLYNT

NGS input files, shown in Table 2, differ in size and can be greater than one terabyte. We obtained these files from The Galaxy Project and other sources.

Table 2: NGS Genomic Data Sample

NGS Format: @ERR030881.107 HWI-BRUNOP16X_0001:2:1:13663:10 ATCTTTTGTGGCTACAGTAAGTTCAATCTGAAGTCA + 5.544,444344555CC?CAEF@EEFFFFFFFFFFFFFFFEFFF

GBK format input files, shown in Table 3, contain large amounts of information, such as genes' names, locations and

lengths, annotations, locus tags, protein translations and others. We obtain these files from NCBI.

Table 3: GBK Genomic Data Sample

GBK Format:	
LOCUS	NC_001489 7478 bp ss-RNA linear VRL 08-DEC-2008
DEFINITION	Hepatitis A virus, complete genome.
ACCESSION	NC_001489
VERSION	NC_001489.1 GI:9626732

Output Data Structures: Our GenomePro framework produces, as shown Table 4, a data structure containing one sub-sequence per record/line, aaaa and aaac; each data structure can contain billions of records. Sub-sequences can be of any length (length of 4 in this example). In the next column, we show the total amount of sub-sequences found in the original genome, in this case 3, for sub-sequence 'aaaa' and 4 for sub-sequence 'aaac'. Finally, we find the sorted locations of each unique sub-sequence, in this case 12 1003 and 1143 for 'aaaa' and 1 11 23 and 101 for 'aaac'. Finding the same sub-sequence at different locations in the same genome is referred to as "finding repeats".

Table 4: GenomePro Output Files

aaaa	3 12 1003 1143
Aaac	4 1 11 23 101

Users of the Framework provide batch files, as shown in Table 5, containing the names and locations of the files to be processed by GenomePro. This is a typical user-supplied batch file with chromosome names and locations:

Table 5: User Input Batch File

~/hs/hs_alt_HuRef_chr1.mapC
~/hs/hs_alt_HuRef_chr2.mapC
~/hs/hs_alt_HuRef_chr3.mapC

Batch files are used to process hundreds of genomes per job, creating a new output file for each controlling genome; these new output files contain subsequences that are found in multiple genomes. Each record in each output file allows to determine which subsequences are inside one or more gene(s) in its correspondent genome. This information allows to determine which subsequences are found in known genes of large genome groups, allowing the creation of Subsequence Fingerprint Files, as shown in Table 6. Fingerprints are then searched for in the dark matter areas. We conclude that fingerprints that are found in dark matter areas could belong to unknown genes.

Table 6: Fingerprints

agctat 6 hs_ch2 hs_ch12 hs_ch16 …
ctgtaa 4 hs_ch3 hs_ch15 hs_ch16 …
Tataga 7 hs_ch1 hs_ch10 hs_ch14 …

Locations where the sub-sequences do not belong to any currently known gene, are in the Dark Matter areas and therefore could be in unknown genes, see Table 7.

Table 7: Possible Unknown Genes Locations

agctat 234 23144 456234 7923887 …
ctgtaa 1542 6401 94523 6723871 ...
tataga 2314 456234 792380 8767 ...

V. ALGORITHMS

We have implemented the following functions within the Framework:

1 - Create sub-sequences from multiple genomic file formats.

2 - Determine Minimum RAM space needed.

3 - Determine Minimum Secondary storage needed.

4 - Determine size order of magnitude in each step.

5 - Validate results in every process step.

6 - Map all data into files with duplicates.

7 - Reduce duplicate files creating one file with unique subsequences and the locations of all of its repeats in the genome.

VI. CONCLUSION AND FUTURE WORK

Output files for each controlling genome contain common sub-sequences found in multiple chromosomes. Common sub-sequences that belong to known genes in its correspondent genome, allow us to create Fingerprint Files.

The sub-sequences in the Fingerprint Files that are found in the Dark Matter areas can then be used by researchers to further study genomes' dark matter areas and possibly find new genes.

We are deploying a web service wherein users would upload raw genomic data in multiple formats, such as Fasta, NGS, and GBK, and obtain resultants data on Repeats, Signatures, Equivalencies, and find subsequences that perform critical functions in genes.