AN INFORMATION THEORETIC FRAMEWORK FOR GENOMIC DATA ANALYSIS

Aaron McKenna[1] and Gil Alterovitz[2,3,4,5]

[1]Bioinformatics Program, Boston University, Boston, MA. [2]Division of Health Sciences and Technology, Harvard Medical School and Massachusetts Institute of Technology, Boston, MA. [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA. [4]Children's Hospital Informatics Program, Boston, MA. [5]Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston, MA

ABSTRACT

The breadth of biological data collected in the last decade has far outstripped the methods available to process it. To effectively investigate and explore this abundance of data, novel automated collection and analysis approaches must be devised. We have developed a new open software framework, the Open Genomic Analysis Platform (OGAP), to aid in the analysis of genomic data. It is capable of analyzing a variety of data source, and focuses on using information theory to characterize data. The frameworks has is capable of import a variety of genome tied data, and provides custom analysis and visualization of results. We then demonstrate the use of this framework analyzing the Prochlorococcus Marinus organism. We show a strong correlation between the information content of sequence data and up regulation of gene expression during lytic infection.

1. INTRODUCTION

We present the Open Genomic Analysis Platform (OGAP) as a novel information theory driven platform, with significant advantages over current tools. OGAP's distinction is its genome centric organization, allowing end users to quickly associate new genomic data with a specific organism. Data associated in this way can be quickly and efficiently analyzed and visualized, allowing end users to quickly and effectively test novel hypotheses'.

A variety of tools are currently available to statistically analyze biological data. Tools like Bioconductor[1] are highly capable, but require the user to be comfortable with command line tools and R programming, interacting with the data at a much lower level. In comparison, OGAP offers the user an intuitive graphical interface to manage data manipulation, while maintaining powerful computational methods. By providing the source code though an open source license, along with clearly structured interfaces, the OGAP platform is designed to allow users to add new data sources with a minimal amount of programming.

2. MATERIALS AND METHODS

The Open Genomic Analysis Platform was designed to open and extensible, and was developed using the Java language. Java virtual machines are available for most common computing platforms, making OGAP portable to a variety of systems. The OGAP framework provides a variety of software interfaces to extend its capabilities and allow the end user to add new data sources. Each data source associates itself with the genome with a specific tag, where by each value can then be retrieved by the its associated tag. The object hierarchy is detailed in figure 1.

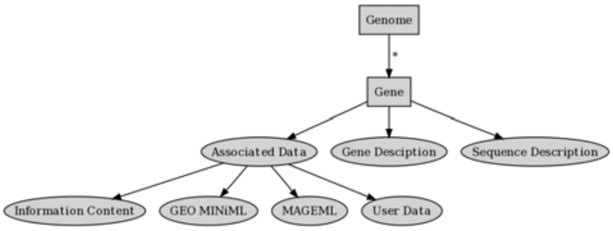

Fig. 1. Genome object hierarchy. The gene and it's associated data is the focus

OGAP was designed as a framework, with the goal of leveraging as much functionality as possible from currently available tools. Biojava[1] is an open source biology toolset, that we integrated to process sequence files. Basic statistical analysis functions are provided by Apache Commons math package. The platform also provides a Java math interface, that end users can implement to tie in various commercial tools. For instance, the spearman correlation values used in the P. Marunis experiment were determined by Wolfram's Mathematica. This interface can be extended for commercial tools that provide either a Java API or command line functionality.

2.1. Processing Genome Information

Incorporating sequence information requires flexibility; although some formats provide superior annotation and feature information, many sequences are available in less verbose formats. OGAP supports the most common format FASTA, though the more annotated Genbank format is also supported and is a better option when available. OGAP stores gene annotations, sequences, and information content calculations for each gene targeted on specific organisms.

OGAP currently supports the most common gene expression format provided by the NCBI Gene expression omnibus and the European Bioinformatics Institutes ArrayExpress databases. Files provided in MINiML format are easily incorporated, along with data in comma separated value format.

OGAP also provides for an automated discovery of associated data set from the National Center for Biotechnology Information (NCBI), using their publicly available ENTREZ toolset. Searches can be conducted using taxonomy identifiers or a less strict keyword search. OGAP will attempt to retrieve all associated sequence and gene expression data publicly available from the NCBI databases that match the target organism.

2.2. Information Content

OGAP was designed with the goal of strongly incorporating information theory concepts into the heart of the application. Information theory has a well-established role in biological analysis, and has been used in a variety of different ways to describe and categorize biological data. The relative weights of ontology terms in an organism, or the relative abundance at a specific binding site are two common examples [2]. To calculate the information content of a sequence, OGAP uses the following formula [3]:

$$I(A_n) = -\log_2 p(A_n)$$

Where $p(A_n)$ is the probability of seeing a particular nucleotide at the specific location in the sequence. OGAP calculates the information content of each gene's nucleotide encoding sequence, using the organisms overall nucleotide abundance to determine the background probability for each nucleotide.

Along with calculating the sequence information content, the Shannon entropy is also calculated and stored for each sequence. We calculate this value using the following equation:

$$H = \sum_{i=1}^{m} p(A_n)\log_2 p(A_n)$$

Where $p(A_n)$ is again the probability of seeing a specific nucleotide at a specific sequence location.

3. RESULTS

To evaluate the effectiveness of the OGAP framework, we used the platform to investigate the relationship between sequence entropy values and gene expression levels. We used the OGAP framework to data mine all associated sequence and gene expression data from NCBI within our specified constraints. We focused on organisms where a complete genetic sequence was available, and where the sequence length was less than thirty million base pairs. This was done using the NCBI Entrez tool wrappers provided by the OGAP framework. Approximately ten gigabytes of data were obtained over a multiday period. This represented 49 unique species: their genomes and gene expression profiles from a variety of experimental conditions.

The OGAP toolset was then used to generate Spearman rank correlation values between the two data sets. This value is determined using the follow equation [4]:

$$r = 1 - 6\sum \frac{d^2}{N(N^2 - 1)}$$

Where d represents the distance between paired values, and N represents the number of joint data points. This calculation is useful in quickly determining if the two data sets are correlated, regardless of their distribution. Figure 2 shows the results of calculating Spearman rank correlation values between gene expression levels and total gene sequence entropy.

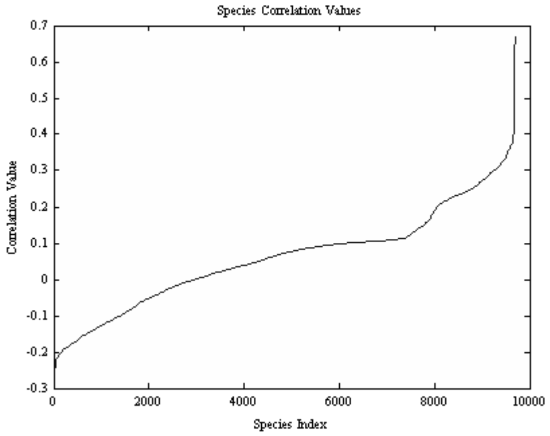

Fig. 2. Spearman rank value across all retrieved species, sorted by correlation value. The majority showed little to no correlation between gene expression values and gene information content.

We chose then to focus on the Prochlorococcus Marinus (P. Marinus) results, since it showed a high favorable correlation between gene expression and sequence entropy, with a P value of 1x10^{-5}. This P value is calculated using the Spearman rank correlation value along with the sample size, in this case the gene count for the organism. P. Marinus is a highly abundant photosynthetic bacterium in the world's oceans. [5] The specific strain of interest in our experiment was Prochlorococcus

Marinus MED4, which was sequenced by the Department of Energy's Joint Genome Institute. [6] Since many gene expression experiments were associated with the species; we used the OGAP platform to enumerate each gene expression experiment against the total entropy values of the gene sequences.

Each set was analyzed to determine which experimental conditions contributed to the high Spearman rank. We found that the gene expression values from a data set associated with lytic infection of the bacteria were responsible for the strongest correlation.

4. CONCLUSION

This article introduces a new framework for the analysis and investigation of genomic data, the Open Genome Analysis Platform, with an emphasis on information theoretic methods. It allows researches to incorporate and mine a large variety of publicly available data sources, while removing the burden of excess knowledge of the underlying implementation. OGAP has been designed with the open source community in mind, allowing users to quickly expand the OGAP tool to process new genome related information. Finally, it incorporates visualization methods that allow end users to quickly analyze and evaluate experimental results. The OGAP framework allows biological researchers to quickly test complex and computationally expensive hypotheses across vast amounts of genomic data in minimal time, which should allow new and previously impossible theories to be explored.

5. REFERENCES

[1] Mangalam H. *The Bio* toolkits--a brief overview*. Briefings in Bioinformatics, 2002 Sep;3(3):296-302.

[2] Alterovitz G, Xiang M, Ramoni M. *An Information Theoretic Framework for Ontology-based Bioinformatics*. Information Theory and Applications Workshop. Jan. 29 2007

[3] Strait BJ, Dewey TG. *The Shannon information entropy of protein sequences*. Biophys J. 1996 Jul;71(1):148-55.

[4] Weiss O, Jiménez-Montaño MA, Herzel H *Information content of protein sequences* J Theor Biol. 2000 Oct 7;206(3):379-86.

[5] Partensky F, Hess WR, Vaulot D. *Prochlorococcus, a marine photosynthetic prokaryote of global significance*. Microbiol Mol Biol Rev. 1999 Mar; 63(1):106-27.

[6] Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW. *Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation.* Nature. 2003 Aug 28;424(6952):1042-7.

[7] Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T, Kettler G, Sullivan MB, Steen R, Hess WR, Church GM, Chisholm SW. *Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution*. Nature. 2007 Sep 6;449(7158):83-6