2010 IEEE International Conference on Bioinformatics and Bioengineering

Comparison of Gene Prediction Programs for Metagenomic Data

Non Yok,Gail Rosen
ECE Department, Drexel University
Philadelphia, PA, USA
ng39@drexel.edu, gailr@ece.drexel.edu

***Abstract*— This manuscript presents the most rigorous benchmarking of gene annotation algorithms for metagenomic datasets to date. We compare three different programs: GeneMark, MetaGeneAnnotator (MGA) and Orphelia. The comparisons are based on their performances over simulated fragments from hundred species of diverse lineages. We defined three different types of fragments: one type from the intra-coding region and the other types are from the gene edges. The general observation was that performances of all these programs improve as we increase the length of the fragment. On the other hand, intra-coding fragments of our data show a low annotation error in all of the programs if compared to the genes edges.**

Keywords- Metagenomic; Orphelia; MGA; GeneMark; fragments, sensitivity; specificity; error

I. INTRODUCTION

Metagenomic analysis is defined as the characterization of microbial genomes via the direct isolation of genomic sequences from the environment without prior cultivation (1). There are three major software programs used in gene prediction in metagenomic: GeneMark (3, 4), MGA (6) and Orphelia(2). GeneMark heuristic uses 3-periodic zero order Markov model that works with codon frequencies table to predict genes in metagenomics (3). MGA is an upgrade version of another software package, called MetaGene(MG) (5)which is used in gene prediction in metagenomic sequence data. In MGA genes are predicted in two stages: First stage all ORFs are scored by their base compositions and lengths using the dicodon regression models (Bacteria, Archae, Phage, Self) (6). Second, an optimal (high-scored) combination of ORFs is calculated using additional statistics (6). However, Orphelia algorithm uses artificial neural network to predict genes in metagenomic reads (2). In this paper, we will benchmark these programs using artificially fragmented reads extracted from 100 species of diverse lineages. Hoff et al. used only 12 species in their comparison (2); therefore, their sample is too small to represent an environmental sample. Also, no predecessor has separately examined fragments that contain gene edges as opposed to intra-coding regions. Because our metagenomic data is diverse and fragmented to represent Sanger's reads and the next generation sequencing, the results of our experiments reveal some important issues that our predecessors did not discover.

II. METHOD

A. Fragment Types

We created artificial metagenomic fragments out of 100 organisms and grouped them according to their lengths: 100 to 700 bp fragment groups. Next, we defined three different types of fragments of equal length from each group based on the order of coding and flanking regions in the fragment. We named them Type A, Type B and Type C. For instance, in 700 bp fragments, Type A fragment consists of flanking region of variable length, 300-400 bp followed by a coding region of a variable length 300-400 bp. The length of the flanking and the coding regions is determined randomly, but the length of the whole fragment must equal to 700 bp. Type B fragment is different from Type A, in that it consists purely of coding sequence and it is picked randomly from within a gene region.

B. Performance Metrics

The objective of this paper is to compare the performances of these software packages in gene prediction. The means of comparison are measures of sensitivity, specificity, and f-measure. In addition to two newly defined measures known as annotation and prediction errors.

$$annotationErr = \frac{|Lp - Lgb| + |Rp - Rgb|}{|Fgb|} \quad (1)$$

Where Lp stands for the left end index of the gene annotation of the program, while Lgb stands for the left end index of the GenBank annotation. Rp stands for the right end index of the fragment annotation by the program, but Rgb stands for the right end index of the fragment annotation according to the GenBank. Fgb denotes fragment length according to the GenBank annotation.

$$predictionErr = \frac{Gm}{Gt} \quad (2)$$

Where Gm is the number of the missed genes by the program and Gt is the total number of genes

III. RESULTS

978-0-7695-4083-2/10 $26.00 © 2010 IEEE
DOI 10.1109/BIBE.2010.58

Tables 1 to 6 along with graph of f-measure and are the major results of the paper. These results reflect the performances of the software tools utilized in the experiment. Each table contains five different measured values with exceptions of Type B fragments, their tables contain only three measures and the 100 bp fragments their tables does not contain annotation error. The measure is found insignificant due to the shortness of the fragments.

measure	GeneMark	MGA	Orphelia
Annotation Err	33.38	36.31	35.78
Sensitivity	86.36	98.91	100
Specificity	68.468	65	79.66
Prediction Err	12	1	0
F-measure	76.38	78.44	88.67

Table 1: Performances of the three programs:GenMark, Orphelia and MGA over Type A fragments of 700 bp lengths. On this table Orphelia misses no genes; therefore, its sensitivity is 100 %

measure	GeneMark	MGA	Orphelia
Annotation Err	17.06	9.41	7
Sensitivity	87.77	98.9	98.9
Prediction Err	11	1	1

Table 2: Performances of the three programs:GenMark, Orphelia and MGA over Type B fragments of 700 bp lengths. On this table, the prediction error of both Orphelia and MGA is only 1%

measure	GeneMark	MGA	Orphelia
Sensitivity	60.86	87.9	90.47
Spec ificity	62.22	60.6	59.37
Prediction Err	36	11	8
F-measure	61.53	71.7	71.69

Table 3: Performances of the three programs: Orphelia, MGA and GeneMark over fragments of Type C of lengths 700 bp. The sensitivity of Orphelia as well as MGA is 100 %

measure	GeneMark	MGA	Orphelia
Annotation Err	23.37	13.6	22.83
Sensitivity	88.29	100	100
Spec ificity	83	85.8	88.18
Prediction Err	11	0	0
F-measure	85.56	92.3	93.71

Table 4: Table 4: Performances of the three programs: Orphelia, MGA and GeneMark over fragments of Type A.

measure	GeneMark	MGA	Orphelia
Sensitivity	17.71	32.1	48.48
Specificity	40.90	64	44.44
Prediction Err	79	57	39
F-measure	26.56	36	46.15

Orphelia's sensitivity is the highest.

Table 5: Performances of the three programs: GeneMark, Orphelia and MGA over Type B fragments of 100 bp lengths.
Table 6: Performances of the three programs: GeneMark,

measure	GeneMark	MGA	Orphelia
Sensitivity	62.36	95.4	97.2
Prediction Err	35	4	2

Orphelia and MGA over Type B fragments of 100 bp lengths.

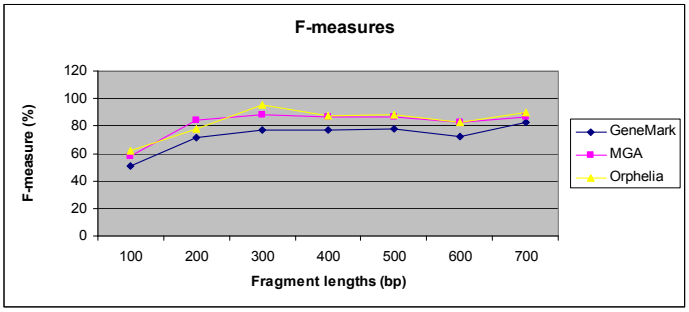

Figure 1: The graph of the f-measure graph of the three programs: Orphelia, MGA and GeneMark for fragments of lengths 100 to 700 bp.

IV. CONCLUSION

The performances of GeneMark, MGA and Orphelia worsen with ultra short length fragments like those produced by Illumina Solexa. Overall, Orphelia performs the best for next-generation length reads.

V. Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. #0845827

V1. REFERENCES

[1] Thomas Lingner Rolf Daniel Burkhard Morgenstern Katharina J Hoff, Maike Tech and Peter Meinicke. Gene prediction in metagenomic fragments: A large scale machine learning approach. BioMed Central, page 14, 2008

[2] Peter Meinicke Katharina J. Hoff, Thomas Lingner and Maike Tech. Orphelia: predicting genes in etagenomic sequencing reads. Nucleic Acids Research, page 5, 2009.

I3] John Besemer and Mark Borodovsky. Heuristic approach to deriving models for gene finding. Nucleic Acids Res, page 10, 1999New York: Academic, 1963, pp. 271–350.

[4] Lukashin A. and Borodovsky M., GeneMark.hmm: new solutions for gene finding, NAR, 1998, Vol. 26, No. 4, pp. 1107-1115.

[5] Jungho Park Hideki Noguchi and Toshihisa Takagi. Metagene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res, Vol. 34,:No. 19 5623–5630, 2006.

6] Takeaki Taniguchi Hideki Noguchi and Takehiko Itoh. Metageneannotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA RESEARCH, page 10, 2008.