ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

A Study of gene prediction program of metagenomic data on various gene prediction software

Meenakshi Sharma[1] and Pankaj Bhambri[2]

1. Prusing Mtech of computer science,Guru nanak dev engineering college Ludhiana,Punjab, Punjab, India.
2. Asisstant professor of computer science, Guru nanak dev engineering college Ludhiana, Punjab, India.

Manuscript Info

Manuscript History:

Received: 11 August 2013
Final Accepted: 25 August 2013
Published Online: September 2013

Key words:

Abstract

This thesis work presents the most rigorous approach of gene algorithims for metagenomic datasets to data .we differeniate four different programs:glimmer,metagenmark(MGK),prodigal and orphilla.The differentiation are based on their performances over simulated fragments from many species of diverse lineages.we defined four different fragments type:two types come from coding regions and two come on non-coding region and other types are from the gene edges.In this we use 20 species to differeniate every program with their performance.In this we use various formula to predict the performance of various program the formula that we are:specificity,sensitivity,prediction error,predictionaccuracy,annotation errorand F-measure .so that sample is examined easily and manually by calculation based and no predecessors has separately examined fragments that contain gene edge as opposed to non –coding region in our results are 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 low annotation error in all of the programs if compared to the gene edge fragments. Overall, we found an upper-bound performance by combining all the methods.

Introduction

Metagenomic analysis is explain the study of metagenomes, genetic material deal from environment samples.The broafield as environmental genomics, ecogenomics or community genomics. Metagenomic analysis is defined as themicrobial genomes via the direct isolation or deal with genomic sequences from the environment. An important thing of metagenomics is to identify genes in fragmented sequences. Accurate gene annotation for environmental samples is needed so that genes can be classified to their correct functions. Gene prediction programs can be divided into two different parts. First, the ab initio programs that employ a training set with known annotation for training the parameters of their models are widely used in gene prediction . There are a large number of gene-finding programs of this kind e.g. GENIE, GENEID, ORPHILLA, GENMARK, GENSCAN and GLIMMER. The second group ofgene prediction programs, known as similarity-based programs,use external information about the input sequence topredict genes. Some popular similarity-based programs are GENEWISE, AGenDA and the well known BLASTwhich searches the input sequence against a database ofknown protein sequences. In this we use some conventional approaches to metagenomics is that they are based on the identification of open reading frames (ORFs), which begin with a start codon and end with an in-frame stop codon. Due to the short sequence length of metagenomic fragments, there are incomplete ORFs of at least 100 bp input length that lack a start and/or stop codon, thus conventional programs cannot beapplied to metagenomic. Similarly, the homology based approaches for gene predictions are only applied to genes with known homology. Therefore, both of these categories do not work well for metagenomic fragments which areabout 1000 and < 400 bp when produced by Sanger's andnext generation sequencing, respectively [8]. This paper will examine metagenomic programs for gene prediction: Orphelia, MGA, prodigal, glimmer and metagenmark in artificially simulated fragments from 100 species .

Material and Methods

2.1. Types of Fragments

In artificial metagenomic fragments out ofa hundred genomes and grouped them according to their lengths: 1000 bp fragment groups. We defined four different types of fragments of equal length from each group based on the order of coding region and flanking area of the gene in the fragment. We named them case A, case B, case C and case D fragments. For instance, in 1000 bp fragment groups, case A we used sometine codon part and other end we used non-condon part of 1000 bp lengthfragment and length is determine by using select part they show range of sequence that we used in .Case B fragment is different from case A, in that it consists purely of coding sequence and it is picked randomly from within a gene region. Some of case B fragments may contain start or stop codons. Case C fragment is similar to case A fragment with the exception that the coding region comes before the flanking region in the fragment. However, case D fragment is from non-coding region of the DNA. Main reason for having artificially simulated fragments is: We want to control the amounts of the different types of fragments (and not only benchmark the results)like as is usually done in the literature. True metagenomic may have unknown organisms and genes, with simulated data, we can control and know the ground truth in order to benchmark it.

2.2. Parameters of the Programs

In this we have required options on web submission windows for some of these programs to give gene prediction. In GeneMark's submission window we used the following method to examine our samples: For the option of "Kingdom" givenon its web page, we used "Mixture of bacteria and archaea". This option should be used for all environmental samples as well as for human and other microbiomes. The option allows for using bacterial and archaeal heuristic models concurrently. It also helps in achieving high sensitivity with somewhat lower specificity. For the second option, "Model" we used "Codon Polynomial fitting order 3" for this option no temperature parameters are needed. In the output options we selected "nucleotide" and "HMM" .Orphelia provides two models for scoring open reading frames in metagenomic reads

2.3. Performance Metrics

After simulating case A, B, C and D fragments, we examined the entire fragments for their gene annotations by employing the software tools mentioned above. Next, we analyzed the results of the predictions using sensitivity,specificity, and harmonic-mean(f-measure) measures.The sensitivity measure estimates the program capability of detecting annotated genes.Next, we calculated the specificity measures from results of gene predictions of the four programs.

This measure quantifies the reliability of gene prediction by the programs. In addition, we used a newly defined measure that reflects the annotation accuracy of the software programs. It also conveys the annotation error using a reference annotation of the GenBank. In this we used various formulas to predict the program

analytically as follows:

AnnotationError =jLp □ Lgbj + Rp □ Rgbj jFgbj ; (1)

where Lp stands for the left end index of the gene annotation of the software program. Lgb stands for the left end index of the GenBank's annotation. Rp stands for the right end index of the fragment annotation of the program, while Rgb stands for the right end index of the fragment annotation of the GenBank. And jFgbj stands for the fragment length according to the GenBank annotation. Moreover, we calculated the percentage of the genes that were missed and named the measure as prediction error.

specificity= TP/TP+FP
sensitivity= TP/TP+FN

- Predicted number of positives (PP) =TP+FP
- Predicted number of negatives (PN) = TN+FN

PredictionError = Gm/Gt

where Gm is the number of the missed genes by the program, and Gt is the total number of genes examined by the program.

Prediction Accuracy$=TP+TN/TP+FN+FP+TN$
$PPV=TP/TP+FP$
Annotation Error$=|Lp-Lgb|+|Rp-Rgb||Fgb|$
$F-measure=2*Sensitivity*Specificity/Sensitivity+Specificity.$

Results

Tables 1 to 4 along with graphs of the f-measure, the prediction error and the receiver operating characteristic (ROC), are the results of the paper. They reflect the performances of the software tools utilized in the experiments. Each table contains five different measured values for each program with exception of case B fragments; their tables contain only four measures.

Measure	Metagenmark	Orphelia	Glimmer	Prodigal
specificity	60%	64%	51%	51%
sensitivity	100%	100%	80%	100%
Predicted number of positives(PP)	33	31	31	31
Predicted number of negative(PN)	7	9	9	9
Correlation-Coefficient (CC)	0.46	0.538	0.0598	0.16
PredictionError	0	0	0.2	0
Prediction Accuracy	67%	72%	52%	52%
Positive predictive value(PPV)	60%	64%	51%	51%
F-measure	100%	78%	62%	67%

Table 1: Performances of the four programs:metagenmark, Orphelia,glimmer and prodigal over fragments of case A of the 1000 bp group. In this we taken coding regionof length 1000bp.Thelength of flanking and coding regions are determined randomly between 1000bp for each fragment. In this case Orphelia misses no genes; therefore, its sensitivity is 100 (%) and also its specificity is the highest.

Measure	Metagenmark	Orphelia	Glimmer	Prodigal
Specificity	100%	100%	100%	100%
Sensitivity	100%	100%	80%	100%
Predicted number of positives(PP)	20	20	16	20
Predicted number of negative(PN)	0	0	4	0
Correlation-Coefficient (CC)	0	0	0	0
PredictionError	0	0	20	0
Prediction Accuracy	100%	100%	80%	100%
Positive predictive value(PPV)	100%	100%	100%	100%
F-measure	100%	100%	88%	100%

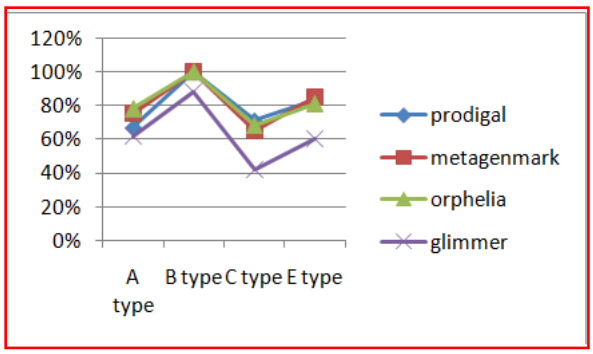
Table 2: Performances of the four program metagenmark, Orphelia,glimmer and prodigal over case B fragments of the 1000 bp group .In this we taken coding and non-coding part both of 1000bp length. This fragment case is purely from gene region. The annotation errors onthis table are lower than those of case A fragments (Table 1).

Measure	Metagenmark	Orphelia	Glimmer	Prodigal
specificity	53%	54%	45%	55%
sensitivity	85%	95%	45%	100%
Predicted number of positives(PP)	32	35	20	38
Predicted number of negative(PN)	8	5	20	2
Correlation-Coefficient (CC)	0.12	0.22	0.10	0.22
PredictionError	15	5	55	0
Prediction Accuracy	55%	57%	45%	55%
Positive predictive value(PPV)	53%	54%	45%	52%
F-measure	65%	68%	45%	71%

Table 3: Performances of the four programs: metagenmark, Orphelia,glimmer and prodigal over fragments of case C of the 700 bp group. Thelength of flanking and coding regions are determined randomly between 300-400 bp for each fragment. Both prodigal and orphelia miss no genes in this case.

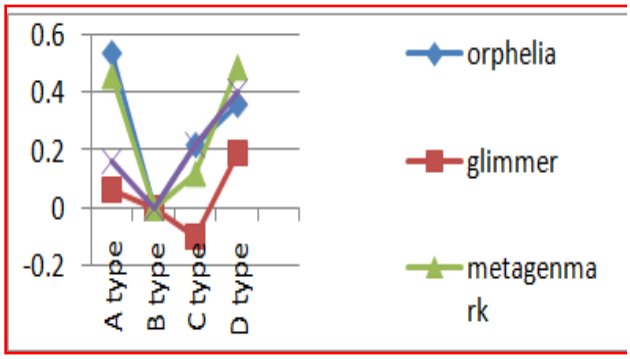
Measure	Metagenmark	Orphelia	Glimmer	Prodigal
specificity	76%	76%	77%	74%
sensitivity	97%	87%	50%	95%
Predicted number of positives(PP)	51	46	26	51
Predicted number of negative(PN)	9	14	34	9
Correlation-Coefficient (CC)	0.49	0.36	0.19	0.396
PredictionError	2.5	12	50	5
Prediction Accuracy	78%	73%	56%	75%
Positive predictive value(PPV)	76%	76%	77%	75%
F-measure	85%	81%	60%	83%

Table 4: Performances of the four programs: metagenmark, Orphelia,glimmer and prodigal over fragments of case C of the 700 bp group.In this we taken non-codind and coding part both of 1000bp length Thelength of flanking and coding regions are determined randomly between 300-400 bp for each fragment. Both prodigal and orphelia miss no genes in this case.

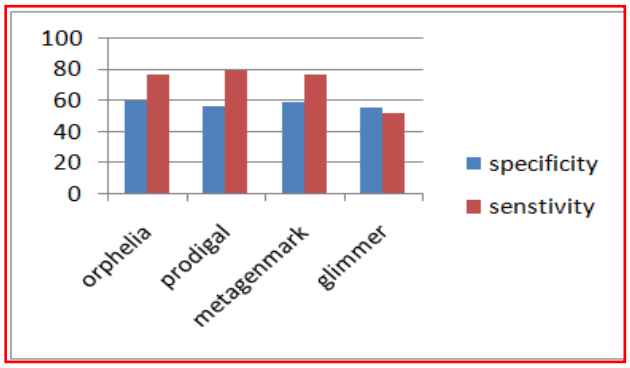


Graph. 1 F- distribution chart of all programs

Graph 1: The graph of the f-measures of the four programs: metaganmark, Orphelia, glimmer and prodigal in fragments of lengths 100 to 1000 bp. It was generated with 33 (%) fragments from case A, 34 (%) fragments from case B and 33 (%) fragments from case C. Gene- Mark's and Orphia's specificities are close in values.



Grap .2 Correlation coefficient chart of all program



Graph.3 Average sensitivity and specificity of all program

Graph 3: The graph of the sensitivity of the four programs: metaganmark,Orphelia,glimmer and prodigal in fragments of lengths 100 to 1000 bp. It was generated with 33 (%) fragments from case A, 34 (%)fragments from case B and 33 (%)fragments from case C. Gene-Mark's and Orphia's specificities are close in values.

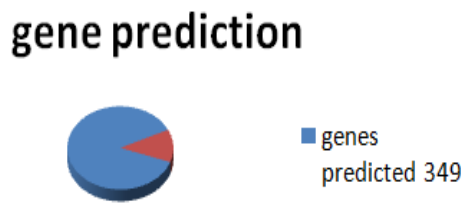


Chart.1 missed and predicted genes by all program

Conclusion

In conclusion, gene-edge type fragments have a higher annotation error than intra-coding regions. The performances of all algorithms worsen for shorter reads. Usage of all the four programs together in upper- bound analysis enhance the accuracy of gene prediction as it is apparent from Table 7 data.

References

1)'http://topaz.gatech.edu/genemark/metagenome/training/'.

2)Rastogi s.c,mendiratta namita , rastogi parag,"bioinformatics""gene prediction".

3)BurgeChris and Samuel Karolin(1997)," Prediction of Complete Gene Structures in Human Genomic DNA, Department of mathematics and Stanford university, Stanford CA94305, USA. Volume: 4, Pages: 241-251.

4)Christopher B. Burge. Identification of genes in human genomicdna.ph.d. thesis. Stanford University, Stanford, CA,USA., 1997.

5) John Besemer and Mark Borodovsky. Heuristic approach to deriving models for gene finding. Nucleic Acids Res, page 10, 1999.

6)Garg ashutosh, kasif Simon, pavlovic Vladimir (2001)," A Bayesian framework for combining gene predictions", bioinformatics program Boston university, vol.18 no.1 2001 page 19-27.

7) Gens Parra, Enrique Blanco, and Roderic Guig. Geneid in drosophila. Genome Res., pages 10: 511–515, 2000.BIOINFORMATICS, vol 19:pages 15751577, 2003.

8) Ewan Birney, Michele Clamp, and Richard Durbin. Genewiseand genomewise. Genome Res., page 14: 988995,2004.

9)Yang Weng; Yunmin Zhu (2006)," Combining Gene-Finding Programs by Using Dempster-Shafer Theory of Evidence for Gene Prediction," Computational Intelligence and Security, vol.16 no.1 2006 page 1-8.

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

11) Elaine R. Mardis. The impact of next-generation sequencing technology on genetics. Trends in Genetics, volume 24:133– 141, 2008.

112) Hideki Noguchi, Takeaki Taniguchi, 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.

13) Katharina J. Hoff, Thomas Lingner, Peter Meinicke, and Maike Tech. Orphelia: predicting genes in metagenomic sequencing reads.Nucleic Acids Research, page 5, 2009.

14) Pop, M (2011)."Invited: Challenges in metagenomic Assembly". Dept. of Comput. Sci., Univ. of Maryland, College Park, MD, USA, IEEE, Volume122, issue date: Feb 2011, pa9-9.

15) Charuvaka, A. Rangwala (2011), H. ."Evalution of short read of metagenomic assembly".Bioinformatics and Biomedical IEEE (2011).Volume:132Issue date: May 2011, page 1-9.

16)Liu, Yongchu Guo, Jiangtao Zhu, Huaiqiu(2011) ,"Gene Prediction of metagenomic fragment by SVM algorithim".Volume-3,issue date 2011,page no 1738-1743.