GENFOCS – *a comparative tool on gene finding with sensitivity and specificity*

Lijo Anto M.A*

M.Phil Scholars, Centre for Bioinformatics, University of Kerala, Thiruvananthapuram.
Supervisors
Prof. Indira Ghosh, Director, Bioinformatics Centre, University of Pune, Pune.
Dr. Achuthsankar S.Nair, Hon.Director, Centre for Bioinformatics, University of Kerala, Thiruvananthapuram.

ABSTRACT

The latest developments in DNA sequencing techniques paved the way for tremendous increase in biological databases. Once the whole DNA sequence of an organism is known, the next big task is to predict all the potential genes present in that sequence. This idea, known as gene finding, gave birth to number of computational gene finding programs. But the most critical thing with biological studies is their accuracy in predicting exact genes. So there should be some method to compare these different tools and also to find out their accuracies. The tool named GENFOCS is a novel comparison tool on existing gene finders. The objective of this tool is not necessarily to find whose prediction is better. But this work implements a system for comparing the predictions of selected gene finders, by generating exact replica of individual predictions over the queried DNA sequence. The tool encompasses different ways of displaying the predictions of individual tools. GENFOCS displays all the individual predictions in a single window, which gives an easy method for comparing different individual predictions. It also finds out various accuracy measures like sensitivity and specificity at different levels for selected organisms.

1 INTRODUCTION

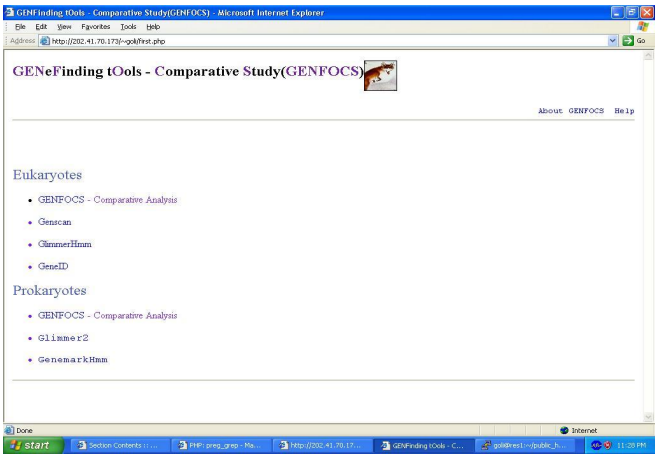
The aim of this thesis work is to build a useful software tool in the field of bioinformatics. Before developing any tool, a good ground work is needed. So in the first part of this work, a good repository of bioinformatics tools is created covering the major fields. The area selected for the detailed study is gene finding tools and their comparison. This area is one of the most important fields in biological study. Various computing techniques/tools are existing for finding genes in a genomic sequence. But the accuracy of these tools has great significance in biological studies. The results from first part studies shows that only very less number of tools are there for simultaneously looking into the results of these gene finding tools. An integrated tool which will display the outputs of these gene finders under a single roof will be very useful for the academic as well as research purpose. The new tool should be unbiased to any techniques/tools that already exist. This idea gives birth to the new tool named GENFOCS. GENFOCS comes with a lot of new features. Using this tool, user can run the individual gene finders with all the options. User can also select a comparative run of the supporting gene finders for selected organisms (now includes *Arabidopsis thaliana, Human, Rice*). The outputs generated by individual gene finders can be visualized in the same window. User can also visualize the outputs in tabular as well as graphical forms. For selected organisms (now only for *Arabidopsis thaliana*), user can compare the sensitivity as well as specificity of gene finders at various levels.

1.1 GENFOCS – Introduction

GENFOCS comes with a lot of new features. Using this tool, user can run the individual gene finders with all the options. User can also select a comparative run of the supporting gene finders for selected organisms (now includes *Arabidopsis thaliana, Human, Rice*). The outputs generated by individual gene finders can be visualized in the same window. User can also visualize the outputs in tabular as well as graphical forms. For selected organisms (now only for *Arabidopsis thaliana*), user can compare the sensitivity as well as specificity of gene finders at various levels.

GENFOCS is a comparative analysis tool for both prokaryotic as well as eukaryotic gene finding tools. All the gene finding tools that GENFOCS implements are free to download. Under prokaryotes, GENFOCS implements Glimmer2 gene finder program while under eukaryotes the list includes Genscan (C. Burge et.al, 1997), Glimmerhmm (M. Pertea et.al, 2004) and Geneid (M. Burset et.al, 1996). A comparison of Glimmer2 (A.L Delcher, 1999) output with that of Genemark.hmm is also possible for the whole genome even though there is no provision for running Genemark.hmm due to nonavailability for downloading. The main page of GENFOCS is given below.

To Whom correspondence should be addressed



The main feature of GENFOCS is allowing the user to run different gene finders, under each category, simultaneously and compare their outputs for a given DNA sequence. The tool dynamically collects the various gene related information generated by these gene finders and displays in different representations for better user understanding. The tool also finds out various sensitivity and specificity measures of the gene finders under eukaryotes. This feature allows the user to verify the accuracy of individual gene finders. In the case of eukaryotes, the features like number of genes, DNA strand (forward/backward), number of exons per gene, type of exons (Initial/Intermediate/Terminal),type of gene(single exon/multiple exon), starting and ending nucleotide positions of each exon, length of individual exons in nucleotides etc, are displayed. Under prokaryotes the features displayed are number of ORFs, starting and ending nucleotide positions of ORFs, length of ORFs etc. The user can view the individual outputs generated by respective gene finders during the simultaneous run for a given input sequence.Users can also run the individual gene finders, by changing the allowed parameters of individual gene finders. GENFOCS also comes with a comprehensive help for the users. It includes what is GENFOCS, How to run GENFOCS, How to run individual gene finders and How to install the offline versions of these gene finders under Linux platform. It also provides some sample DNA sequences for the users who are new to this area.

2. METHODOLOGY

The GENFOCS tool has mainly two divisions, one for eukaryotes and the other for prokaryotes. Under each division, there is comparative run as well as individual run options. If the user selects the individual run for a particular gene finder, then a new page for the selected gene finder is displayed. First the user has to select one of the organisms from Human, Arabidopsis or Rice. GENFOCS provides only those organisms which are common to one particular category since it is a comparative run. Then the user can paste or upload the corresponding DNA sequence. Since Genscan allows only a maximum of 1 million base pairs, GENFOCS also restrict the size of the DNA sequence entered to this limit. GENFOCS also provides direct links to NCBI and DDBJ web sites for downloading the required sequence of interest. For viewing the various sensitivity and specificity measures, the user has to select the particular option for that. Due to time limitation, GENFOCS provides these measures only for *Arabidopsis thaliana*. The details like starting nucleotide, ending nucleotide and chromosome number are required for the accuracy measurement. The feasibility of final output depends on the computing power available to GENFOCS for running.

2.1 GENFOCS - Algorithm for parsing outputs

The algorithm used for parsing the individual eukaryotic gene finder outputs is as follows:

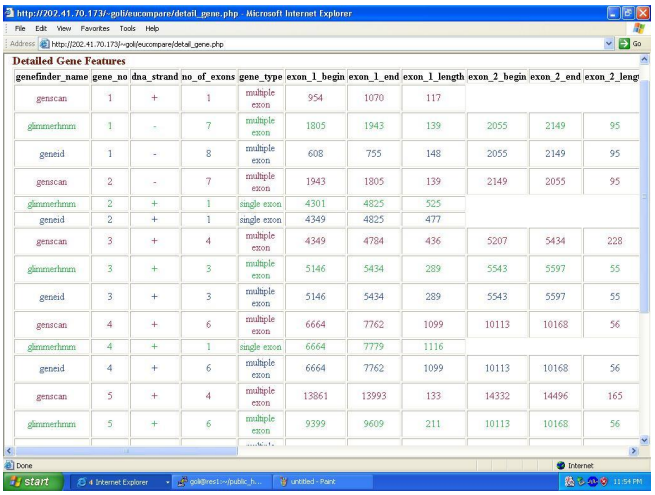
1. Open the output file.
2. Read data line-by-line from the opened file and store temporarily in array1.
3. Split each line using space as the separator and store each output fields in array2.
4. Search for the substring a space followed by bp in array1, the number before that is taken as the number of base pairs.
5. From array2, using different index values, find the details like gene number, number of exons in that gene, dna strand on which that gene is predicted, type of each exon, exon begin, exon end and exon length.
6. Store the above details in the database for further use.

2.2 GENFOCS - Visual Representations

Simultaneously looking into individual output fields of each gene finder is very difficult. User has to either memorize or write down separately each of them, which may not be feasible for large genome sequences. So there should be some new method of displaying all the same fields of different gene finder's output. The GENFOCS option **GENFOCS Results** eliminates this difficulty. On selecting, this option displays a table which gives overall information like gene finder name, the number of base pairs and the number of genes predicted by each.

Users can have a detailed tabular view of the common features with the option named **View Tabular Output**. The tabular form of output displays all the common features pertaining to gene number 1, if any, of Genscan in the first row, followed by that of Glimmerhmm, then by that of Geneid. Then the details of gene number 2, if any, in the order described above. This gives an easy way of comparing individual output features from different gene finders. Separate colors are used for different gene finder's details which makes less error prone while reading. On placing mouse over each value displayed, GENFOCS gives a small explana-

tion of what that value is. This feature helps users by not scrolling back to top of the table for checking the feature name. This is also a feature that makes GENFOCS unique. The following figure shows these features.



Detailed Gene Features

genefinder_name	gene_no	dna_strand	no_of_exons	gene_type	exon_1_begin	exon_1_end	exon_1_length	exon_2_begin	exon_2_end	exon_2_leng
genscan	1	+	1	multiple exon	954	1070	117			
glimmerhmm	1	-	7	multiple exon	1805	1943	139	2055	2149	95
geneid	1	-	8	multiple exon	608	755	148	2055	2149	95
genscan	2	-	7	multiple exon	1943	1805	139	2149	2055	95
glimmerhmm	2	+	1	single exon	4301	4825	525			
geneid	2	+	1	single exon	4349	4825	477			
genscan	3	+	4	multiple exon	4349	4784	436	5207	5434	228
glimmerhmm	3	+	3	multiple exon	5146	5434	289	5543	5597	55
geneid	3	+	3	multiple exon	5146	5434	289	5543	5597	55
genscan	4	+	6	multiple exon	6664	7762	1099	10113	10168	56
glimmerhmm	4	+	1	single exon	6664	7779	1116			
geneid	4	+	6	multiple exon	6664	7762	1099	10113	10168	56
genscan	5	+	4	multiple exon	13861	13993	133	14332	14496	165
glimmerhmm	5	+	6	multiple exon	9399	9609	211	10113	10168	56

The GENFOCS option named **View Graphical Comparison** gives the overall comparison of the outputs generated. This option displays two graphs. The first graph shows a bar diagram of Gene Finders Vs Number of predicted genes. The second graph shows pie charts that gives an idea about the percentage of coding and non coding nucleotides present in the given DNA sequence. This graph shows percentage of coding nucleotide in one color and that of non coding nucleotide in another color. The percentage values are also printed on the graph for easy understanding.

The percentage values are calculated by:

Coding% = (*total exon length/total number of base pairs*) * 100
Non coding% = 100 - *Coding%*

Pie charts are shown for Genscan, Glimmerhmm and Geneid. These charts give an overall idea about the amount of coding/non coding nucleotides identified by individual gene finders. The following figure shows the graphs generated by GENFOCS.

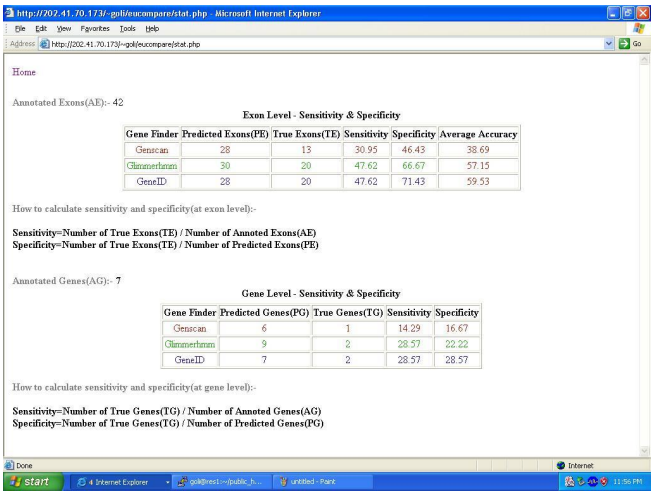
2.3 GENFOCS - Sensitivity and Specificity

The following algorithm is used for finding sensitivity and specificity by GENFOCS.

1. Open the GFF file for the specified organism and chromosome.
2. Split the GFF file line-by-line and store in array1.
3. Split each above line, with space as the separator and store in array2.
4. Look for the substring *gene* in each GFF line, using array1.
5. If (match found) {*if* (begin and end nucleotides within the allowed limits) {
(a) Store gene start, gene end positions in separate arrays.
(b) Store the exon details of current gene, like exon begin and exon end, in separate arrays.
}}
6. For each gene finder, do the following:
7. For each gene finder, check the starting and ending of their exons with the exons got from GFF file. If both begin and end exactly matches, they are taken as True Exons.
8. The total number of exons, within the given range, from the GFF file is taken the number of Annotated Exons.
9. The total number of exons predicted by individual gene finders is taken as Predicted Exons.
10. If for any genes predicted by individual gene finders, whose first exon's starting position and last exon's ending position are equal with the starting and ending positions of genes from GFF file, then they are taken as True Genes.
11. The total number of genes predicted by individual gene finders is taken as Predicted Genes.
12. The total number of genes, within the given range, from the GFF file is taken the number of Annotated Genes. The sensitivity and specificity are calculated by:

Sn_gene = (*TG/AG*) * 100%
Sp_gene = (*TG/PG*) * 100%
Sn_exon = (*TE/AE*) * 100%
Sp_exon = (*TE/PE*) * 100%
Avg_gene = ((*Sn_gene* + *Sp_gene*)/2) * 100%
Avg_exon = ((*Sn_exon* + *Sp_exon*)/2) *100%

where TG is True Genes, AG is Annotated Genes, PG is Predicted Genes, TE is True Exons, AE is Annotated Exons and PE is Predicted Exons. The following figure shows the GENFOCS page displaying sensitivity and specificity measures.



3. RESULTS AND DISCUSSION

For evaluating the accuracy of eukaryotic gene finders, the following data is first collected from the NCBI website.

Organism selected: *Arabidopsis thaliana*

Sequence details

1. Chromosome: 1 Contig: NC 003070.5 start: 1 stop: 20000
2. Chromosome: 2 Contig: NC 003071.3 start: 1 stop: 20000
3. Chromosome: 3 Contig: NC 003074.4 start: 1 stop: 20000
4. Chromosome: 4 Contig: NC 003075.3 start: 1 stop: 20000
5. Chromosome: 5 Contig: NC 003076.4 start: 1 stop: 20000
6. Strand: plus

GENFOCS is allowed to run in the comparative mode for each case and the various sensitivity and specificity measures are calculated:

For chromosome: 1 the value of AE=25. The values of PE and TE for the three eukaryotic gene finders are:

1. Genscan: PE=12, TE=5
2. Glimmerhmm: PE=14, TE=8
3. Geneid: PE=12, TE=8.

The following table shows the exon level accuracy of each eukaryotic gene finders for chromosome 1.

Gene Finder	Sn_exon	Sp_exon	Avg_exon
Genscan	20.00	41.67	30.84
Glimmerhmm	32.00	57.14	44.57
Geneid	32.00	66.67	49.34

The following table shows the average accuracy of each eukaryotic gene finders for the five set of sequences selected.

Gene Finder	Avg_exon	Avg_gene
Genscan	28.83	20.00
Glimmerhmm	41.70	60.00
Geneid	44.63	40.00

A complete run of the entire genome will give more accurate values of sensitivity and specificity. But the time and resource available for GENFOCS was not enough to run an entire genome. The study shows that all the eukaryotic gene finders considered are less accurate in gene level sensitivity and specificity compared to exon level. Mostly they all fail in predicting the starting nucleotide of first exon.

CONCLUSION

The GENFOCS tool described in this work has a novel approach. Even within the time and resource limitations, GENFOCS implements a lot of new features. From the studies it is found that the sensitivity and specificity of gene finders increases from gene level to exon level. It also shows an increase from exon level to nucleotide level. The integrated approach of the tool shows a new way of studying gene finding related issues. The tool now itself is very good for academic purpose under the heading *Gene Finding*. Once the proposed enhancements are also completely implemented, the tool can be used for research as well.

ACKNOWLEDGEMENTS

I express my gratitude to Prof. Indira Ghosh, Director, Bioinformatics Centre, University of Pune, Pune, Dr. Oommen V Oommen, Head of the Dept, Dept of Zoology, University of Kerala and Dr. Achuthsankar S. Nair, Director, Centre for Bioinformatics, University of Kerala.

REFERENCE

C. Burge and S Karlin: Prediction of complete gene structure in human genomic DNA, J.Mol.Biol, 1997, 268:78.

M. Burset and R. Guig.: Evaluation of gene structure prediction programs. Genomics, 1996, 353:357.

A. L. Delcher et al: Improved microbial gene identification with glimmer, Nucleic Acids Research, 1999, 4636:4641.

S. Salzberg et al: Microbial gene identification using interpolated markov models, Nucleic Acids Research, 1998, 544:548.

M. Pertea W. H. Majoros and S. L. Salzberg: Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders, Bioinformatics, 2004, 2878:2879.