International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 11, November - 2017

Protein Function Prediction from Genome Sequence based on PFP Algorithm

Rejwana Haque
Dept. of Computer Science and Engineering
Bangladesh University of Business and Technology
Dhaka, Bangladesh

Abstract **— Protein function prediction is the central problem of bioinformatics. Now its importance is increasing because of the rapid improved computer algorithm makes large amount of accumulation of biological data waiting for characterization. For characterization of genome the function prediction methods first translates genome sequence into protein and than classify proteins into classes of functions. This approach is comparatively slower for predicting protein function from genome sequence. In this paper we address the problem of classifying genome sequence into Gene Ontology without translating the sequence into protein. For this we use Human genome sequence. We use sequence based function prediction method PFP for this classification.**

Keywords— PFP;GO Term; GOA; Scoring GO Terms; Function Association Matrix (FAM) ; Raw Score

I. INTRODUCTION

The advancement of new technologies resulted the rapid growth of new sequence of genomes. It means that genome sequence data are being produced at much greater rate than they are experimentally characterized. So it is need to characterize new gene sequence in a faster way to synchronize with the availability of new sequence. That's why automatic computer based methods are needed to be developed. Standard procedure for genome characterization are literature based annotation and electronic annotation.

The literature-based annotations are typically two types. Annotate genes from paper by paper perspective and annotate on Gene-by-Gene basis [1]. Whereas sequence based manual annotation carried out in TIGR is predicting protein-coding genes and than translating it to protein sequence followed by sequence based protein classification approach [1].

Electrical annotation mainly focuses on protein sequence to function prediction. New methods are developed those can be applied to proteins those are not only highly similar but also can be applied on weekly similar proteins as a source of functional annotation. These methods are based on the realization that weekly similar sequences may also share some functional similarity. Such methods include those use BLAST or PSI-BLAST search results systematically by applying algorithmic techniques and making use of the Gene Ontology (GO) vocabulary structure[3]. These methods include PFP [3, 4], ESG [5], Gotcha [6], GOPET [7], Onto-Blast [8], GOFigure [9]. These methods speeds up the assignment of protein function reduce the gap of new sequence being available and assigning function to them but needs genomes to be translated to protein sequence.

II. METERIALS & MATHODOLOGY

A. Dataset

We used Genome sequence of Chromosome 1 of HUMAN Chromosome of Genebank [15] in gff format, SwissProt i.e. experimentally reviewed sequence for protein sequence in list format and FASTA format and HUMAN GOA[11,12] for gene ontology annotation. To format dataset the we used the cds of to get the protein coding genome sequence. From the reference chromosome sequence and gff file we formatted a genome sequence file as fasta file where the headers include the ids of the protein that is coded by the gene.

TABLE I. EXAMPLE ENTRY OF THE FORMATTED FASTA FILE

Header	Sequence
ID=cds0;Name=NP_001005484.1;Parent=rna8;Dbxref=CCDS:CCDS30547.1,GeneID:79501,Genbank:NP_001005484.1,HGNC:14825,HPRD:14974;gbkey=CDS;gene=OR4F5;product=olfactory receptor 4F5;protein_id=NP_001005484.1	ATGGTGACTGAATTCATTTTTCTGGGTCTCTCTGA TTCTCAGGAACTCCAGACCTTCCTATTTATGTTGT TTTTTGTATTCTATGGAGGAATCGTGTTTGGAAAC CTTCTTATTGTCATAACAGTGGTATCTGACTCCCA CCTTCACTCTCCCATGTACTTCCTGCTAGCCAACC TCTCACTCATTGATCTGTCTCTGTCTTCAGTCACA GCCCCCAAGATGATTACTGACTTTTTCAGCCAGCG CAAAGTCATCTCTTTCAAGGGCTGCCTTGTTCAGA TATTTCTCCTTCACTTCTTTGGTGGGAGTGAGATG GTGATCCTCATAGCCATGGGCTTTGACAGATATAT AGCAATATGCAAGCCCCTACACTACACTACAATT ATGTGTGGCAACGCATGTGTCGGCATTATGGCTGT CACATGGGGAATTGGCTTTCTCCATTCGGTGAGCC AGTTGGCGTTTGCCGTGCACTTACTCTTCTGTGGT CCCAATGAGGTCGATAGTTTTTATTGTGACCTTCC TAGGGTAATCAAACTTGCCTGTACAGATACCTAC AGGCTAGATATTATGGTCATTGCTAACAGTGGTGT GCTCACTGTGTGTTCTTTTGTTCTTCTAATCATCTC ATACACTATCATCCTAATGACCATCCAGCATCGCC CTTTAGATAAGTCGTCCAAAGCTCTGTCCACTTTG ACTGCTCACATTACAGTAGTTCTTTTGTTCTTTGG ACCATGTGTCTTTATTTATGCCTGGCCATTCCCCA TCAAGTCATTAGATAAATTCCTTGCTGTATTTTAT TCTGTGATCACCCCTCTCTTGAACCCAATTATATA CACACTGAGGAACAAAGACATGAAGACGGCAAT AAGACAGCTGAGAAAATGGGATGCACATTCTAGT GTAAAGTTTTAG

For annotating GO to genome sequence we use the protein ids as we do not had direct database for genome to gene ontology. For our benchmark evaluation here, we have used three HUMAN Gene Ontology Annotations for experimentally reviewed Human proteins from UniProt for the BLAST database. GOA annotation sets were retrieved from the Gene Ontology Annotation (GOA) project [14] on human protein at the European Bioinformatics Institute (EBI). We used BLASTn for searching similar nucleotide sequence. As we

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 11, November - 2017

don't have direct annotation database of genome to gene ontology, we use the protein id that a genome codes for mapping the genome to gene ontology.

B. Algorithm

We used a modified version of protein function prediction algorithm PFP [3, 4] for classification of genome. We used the raw score [3] for classifying genomes.

C. FAM Matrix

Our method incorporates functional association likewise PFP by using Function Association Matrix (FAM). The FAM describes the probability at which two GO terms occur together in the same sequence by calculating the co-occurrence of each pair of annotations within UniProt HUMAN sequences. Fig. 1 shows a graphical representation of FAM. All GO terms associated with HUMAN sequence are aligned on the both axis of the matrix, and the association between GO pairs is shown in a gray scale. The numbers indicates the hash index to GO terms that we assigned to each GO term for convenience of our work. White spots indicate non zero association of two HUMAN GO terms in FAM.

D. Scoring GO Terms

Our method uses BLASTn to obtain similar sequence hits from human sequence database of a target sequence and calculates raw score of each go term against the given genome sequence. This classification is based of raw score of the GO terms. The score is calculated as PFP raw score [3].

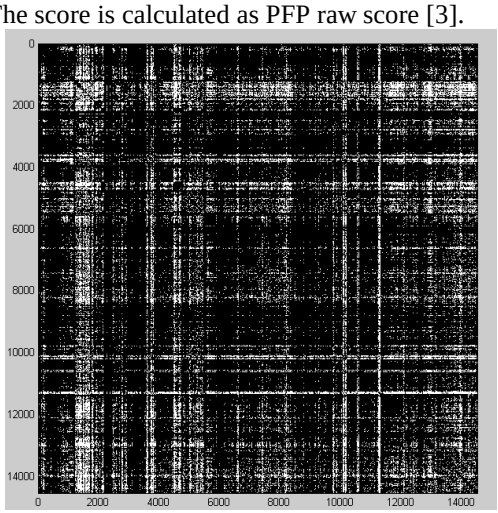

Fig. 1. UniProt Human FAM

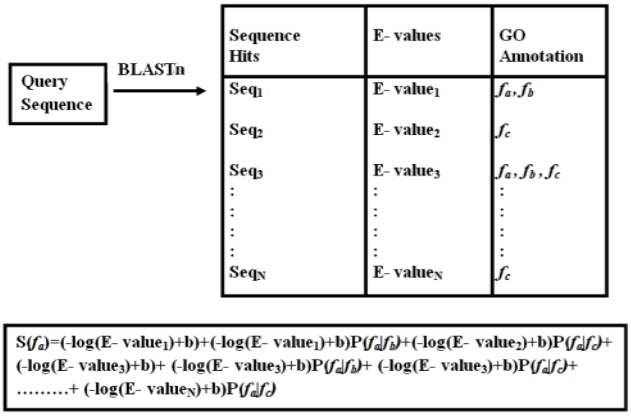

Fig. 2. Example of score computation for a GO term f_a

The score of a GO term is calculated as:

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{Nfunc(i)} \left(\left((-\log(Evalue(i)) + b \right) P\left(f_a | f_j\right) \right) \quad (1)$$

where $s(f_a)$ is score of GO term f_a , N is the number of similar sequences retrieved by BLASTn, $Nfunc(i)$is the total number of GO terms annotated to retrieved sequence i, $Evalue(i)$ is the Expect value given to the sequence i, f_j is a GO term annotated to ith retrieved sequence, and b is the constant value, that keeps the score positive. For our scoring we take b=10. $P\left(f_a | f_j\right)$ is the conditional probability for f_a from the function association matrix (FAM) given f_j is obtained to be annotated with ith sequence.

E. Top scored GO Terms

After computing raw score for a GO term the top 1% terms (that have maximum raw score) of three different aspects are annotated to the input genome sequence. And other GO terms are not annotated to the sequence. Predicted molecular functions with their raw score of the test sequence are included in the paper. Biological process and cellular components are also classified in the same way.

III. RESULTS

A. Accuracy of Prediction

For each sequence the number of GO terms predicted correctly to be annotated and not to be annotated was calculated. We analyzed the average degree of correctness for all the test sequences. The terms that are actually annotated and predicted to be annotated to the input sequence and the terms that are not annotated and also predicted to be annotated are considered as correct prediction. We computed average accuracy of the prediction method for the test sequences. To analyze the prediction performances of the method, we also computed precision and recall.

$$Pricision := \frac{TP}{TP + FP}$$

$$Recall := \frac{TP}{TP + FN}$$

Where, TP denote true positive, FP denote false positive and FN denote false negative.

To analyze the performance of our algorithm we applied confusion matrix to the predicted and actual GO terms for each test sequence. Target class 1 represents GO terms that are actually annotated to the input sequence and Target Class 0 represents GO terms that are actually not annotated to the input sequence. Output class 1 represents GO terms that are predicted to be annotated to the input sequence and Output class 0 represents GO terms that are predicted to be not annotated to the input sequence. It is noticed that 98.9% of the GO terms are correctly classified not to be annotated with the test protein. And 0.2% of the GO terms that are classified to be annotated with the test protein are actually annotated with the protein. It is also seen in the confusion matrix that almost 99.1% of the GO terms are predicted to be annotated of not to be annotated correctly.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 11, November - 2017

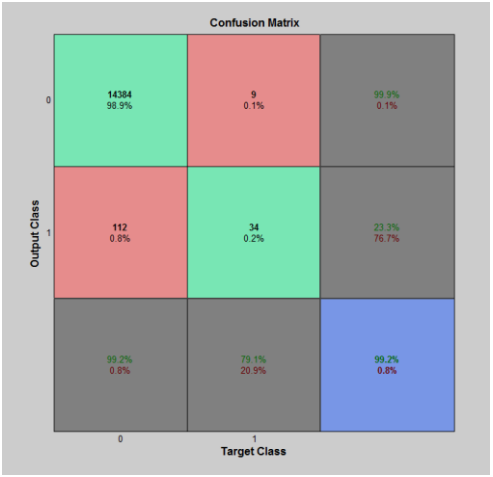

Fig. 3. Confusion matrix

The diagonal green cells in Fig.3 show the number of GO terms that are correctly classified for the input sequence. The red cells shows the GO terms that are misclassified i.e. not correctly predicted for the input sequence. The blue cell represents the total percentage of correct predictions (in green) and incorrect predictions (in red).

Table I shows molecular functions predicted to be annotated with the test genome. The raw scores of actual molecular function annotations that are annotated are also shown. For simplicity only few molecular functions those are annotated with the test sequence are included in the table.

TABLE II. PREDICTED MOLECULAR FUNCTION ANNTATIONS FOR TEST GENOME THAT CODES PROTEIN P31946

Protein ID and Name	Molecular Function Predicted to be Annotated	Score of GO term correctly predicted	GO terms actually annotated
P31946	'GO:0005515'	17695.4520828359	'GO:0005515'
	'GO:0005524'	3669.03249981943	'GO:0019904'
	'GO:0019904'		
	'GO:0044822'		
	'GO:0019899'	2492.52752259378	'GO:0019899'
	'GO:0019901'		
	'GO:0042802'		
	'GO:0042803'		
	'GO:0046982'		
	'GO:0046872'		
	'GO:0003677'		
	'GO:0003700'		
	'GO:0008270'		
	'GO:0044325'		
	'GO:0008134'		
	'GO:0051219'	1383.49966859759	'GO:0051219'
	'GO:0004674'		
	'GO:0032403'	1252.55926067018	'GO:0032403'
	'GO:0042826'	1210.96659763875	'GO:0042826'
	'GO:0005509'		
	'GO:0008022'	1124.97934066400	'GO:0008022'
	'GO:0031625'		
	'GO:0003714'	992.920909962544	'GO:0003714'
	'GO:0003779'		
	'GO:0003682'		
	'GO:0043565'		
	'GO:0005525'		
	'GO:0050815'	821.903824623834	'GO:0050815'

IV. DISCUSSION

In our function prediction method we just take top scored GO terms to be annotated with the input genome sequence. In future we intend to calculate Z-score and P value of each GO terms. We have to evaluate the performance of the prediction method more accurately and using more test sequence.

REFERENCES

[1] http://www.geneontology.org/page/go-annotation-standard-operating-procedures#elect
[2] Meghana Chitale, Ishita K Khan, Daisuke Kihara : In-depth performance evaluation of PFP and ESG sequence-based function prediction methods in CAFA 2011 experiment. BMC Bioinformatics 2013. 14(Suppl 3):S2
[3] Hawkins T, Luban S, Kihara D: Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Science 2006, 15:1550-1556.
[4] Hawkins T, Chitale M, Luban S, Kihara D: PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins: Structure, Function, and Bioinformatics 2009, 74 :566-582.
[5] Chitale M, Hawkins T, Park C, Kihara D: ESG: extended similarity group method for automated protein function prediction. Bioinformatics 2009, 25:1739-1745.
[6] Martin D, Berriman M, Barton G: GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 2004, 5:178-194.
[7] Vinayagam A, del Val C, Schubert F, Eils R, Glatting KH, Suhai S, et al: GOPET: a tool for automated predictions of Gene Ontology terms. BMC Bioinformatics 2006, 7:161-167.
[8] Zehetner G: OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res2003, 31 :3799-3803.
[9] Khan S, Situ G, Decker K, Schmidt CJ: GoFigure: Automated GeneOntology annotation. Bioinformatics 2003, 19:2484-2485.
[10] http://www.uniprot.org/uniprot/
[11] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/
[12] http://www.ebi.ac.uk/GOA/human_release
[13] Huntley RP, Sawford T, Martin MJ, O'Donovan C. (2014) Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. Gigascience. 2014;3(1):4.
[14] Mutowo-Meullenet P, Huntley RP, Dimmer EC, Alam-Faruque Y, Sawford T, Jesus Martin M, O'Donovan C, Apweiler R. (2013) Use of Gene Ontology Annotation to understand the peroxisome proteome in humans. Database.
[15] https://www.ncbi.nlm.nih.gov/genome/guide/human/