Computational Biology and Bioinformatics
2019; 7(1): 1-4
http://www.sciencepublishinggroup.com/j/cbb
doi: 10.11648/j.cbb.20190701.11
ISSN: 2330-8265 (Print); ISSN: 2330-8281 (Online)

Transcription Direction Patterns of Adjacent Genes in Mycobacterium Tuberculosis Using GENAVIS

Angela Uche Makolo

Department of Computer Science, University of Ibadan, Ibadan, Nigeria

Email address:
aumakolo@gmail.com

To cite this article:
Angela Uche Makolo. Transcription Direction Patterns of Adjacent Genes in Mycobacterium Tuberculosis Using GENAVIS. *Computational Biology and Bioinformatics*. Vol. 7, No. 1, 2019, pp. 1-4. doi: 10.11648/j.cbb.20190701.11

Received: February 2, 2019; **Accepted**: March 22, 2019; **Published**: April 18, 2019

Abstract: The understanding of the relationship of a gene with other genes in its neighbourhood and the implication of this relationship on the biochemical activities of the entire genome need an efficient computational tool to unfold the relationship. GENAVIS: (GENe Adjacency VIsualization Software) is an open source, platform independent web-based software for modeling neighborhood genes as binary codes. We also incorporated the feature of having an interactive visual representation of patterns of the binary code for a specific gene family in multiple microbial genomes. The concept of using binary code for representation is derived from computational thinking techniques which models problems using computer logic of applying abstraction and pattern matching to extract hidden patterns aimed at knowledge discovery. The result provides an insight into the analysis of transcriptional unit with more than one gene and genes encoding for universal stress protein, which also allows for a comparative analysis of multiple genomes as the basis for biosynthetic pathways and multi-gene function prediction.

Keywords: Neighbourhood, Adjacency, Transcription Direction, Universal Stress Protein, Binary Code

1. Introduction

The availability of computational tools for Knowledge-building, sense-making and decision making on gene neighborhood in post genomics era remains a challenge to Biomedical scientist. Several studies have been conducted to understand the effect of location of genes in the entire genome on their biochemical activities. Gene-to-Gene analysis has shown that the biochemical activities within a region in DNA sequence are functions of contributions of individual gene within the neighbourhood. As disclosed that neighboring genes are often expressed in similar patterns suggests the involvement of chromatin domains in the control of genes within a genomic neighborhood [10]. That is, the genomic location has some impact on gene expression which generally has influence on the gene function within a framework of expression defined by that neighbourhood. In a theoretical study, gene neighbourhood was listed as one of the factors that affect gene expression but was quick to assume that the existence of gene expression neighbourhoods is not necessary for the correct and coordinated expression of genes that have the same expression profiles [6]. Hence, it is suggested that gene neighbourhood await the creation of more powerful tools to reveal their purpose.

The understanding of gene neighbourhood has been applied in several studies; it was used as entropy measure under the frame of neighborhood rough sets to develop a novel gene selection method for tackling the uncertainty and noisy of gene expression data [3]. Their gene selection model was applied on tumor classification for the discovery of compact gene subsets with improved accuracy. Besides, a novel gene-ranking method based on neighborhood rough set reduction for molecular cancer classification based on gene expression profile was developed [4].

They claimed that the method shows that only few top-ranked genes could achieve higher tumor classification accuracy, which are found to play a crucial role in the occurrence of tumor. Similar direction to gene neighbourhood is in the area of genome integrity where orientation and localization of the genome constituents affect the biochemical activities they are involved in, [8] and [11].

The search for appropriate computational tools for this biomedical knowledge discovery in genes for Universal Stress Protein (USP) led to the development of a software for the analysis of neighborhood genes and modeling them as binary codes. Universal stress proteins are proteins of interest

since they are widely spread in nature and in stress conditions such as heat shock, nutrient starvation, the presence of oxidants or DNA-damaging agents which may arrest cell growth, they constitute a natural biological defense mechanism. The three-dimensional (3D) structure of USP using a combined proteomic and bioinformatics approach was predicted [12]. The concept of gene neighbourhood to investigate the improvement of plant performance under abiotic stress conditions by studying the effects of transcription directions of transgenes and the gypsy insulators on the transcript levels of transgenes in transgenic Arabidopsis is now used [5]. Knowledge of the mechanisms of mycobacterium tuberculosis is still poor and insights provided by genome sequence data enables a better understanding of this mechanism.

The binary representation enables easy pattern matching of the different gene component and the comparative analysis of multiple genomes and the prediction of transcriptional units which are the basis of biomolecular network or biosynthetic pathways. Users also have opportunities of analyzing a large genome at the same time providing an opportunity for comparative genome analysis using the binary coded gene adjacency.

Our model encodes neighborhood genes in binary accession based on their transcription direction, this enables the prediction of stress response equipped biological functions such as the potential chromosomal encoded pathways. In Williams et al It was reported that the conservation of gene neighborhood of the 140 aa universal stress protein in the *B. cereus* genomes led to the identification of a predicted plasmid-encoded transcriptional unit that included a USP gene and a sulfate uptake gene in the soil-inhabiting *Bacillus megaterium*. In addition, Gene neighborhood analysis combined with visual analytics of chemical ligand binding sites data provided knowledge-building biological insights on possible cellular functions of *B. megaterium* universal stress proteins. These functions included sulfate and potassium uptake, acid extrusion, cellular energy-level sensing, survival in high oxygen conditions and acetate utilization [2].

The information deduced that gene neighborhood can be modeled in terms of transcription direction and function of adjacent gene led to the development of GENAVIS, a software, encompassing an algorithm for searching through the genome file, encoding the transcription direction in binary and generating an interactive visualization.

2. Materials and Methods

The Gene Adjacency Software is designed to determine the transcription direction of neighboring genes in an organism using binary encoding. Comparing a gene with its two adjacent neighbors, the binary digit of 1 is assigned if the transcription direction of the two genes is the same and the digit 0 is assigned otherwise. USP column is the reference gene used in the comparison. The output from the algorithm is a three-digit binary code with each digit representing the reference gene and each of the two adjacent neighbors.

The architecture of our software is depicted in figure 2. The processes involved in the algorithm are as stated:

i Input Genome file are downloaded from http://brcdownloads.patricbrc.org/patric2/genomes for the organism of interest. The files with RefSeq.cds.tab extension are used because they are information on transcription direction;
ii The transcription direction is stored in the identifier 'STRAND' which is either a positive or negative;
iii A linear search module searches through the file comparing the value of STRAND identifier;
iv It encodes the USP of interest with a 1 and the adjacent gene to it a 1 if it is the same direction and 0 otherwise;
v Generate a visual representation of the binary output file for all the genome analyzed.

The genomic file containing the transcription information of the organism contains a column under the heading Strand with records of either a + sign or a – sign representing the transcription direction of the corresponding gene. A change in sign from + to – indicates a change in the direction of transcription between the two genes. This change thus results in a switch in the binary code from 0 to 1 or otherwise.

The binary accession allows the prediction of stress response equipped biological functions. A binary accession of 111, 110, 011 are potential chromosomal encoded pathways. The potential pathways are further verified using transcriptional units provided by the BioCyc Collection of Pathway and Genome Databases. In BioCyc (http://biocyc.org/), a transcription unit is defined as a set of one or more genes that are transcribed to produce a single messenger RNA. We were interested in a transcriptional unit with more than one gene and included a gene encoding at least a universal stress protein.

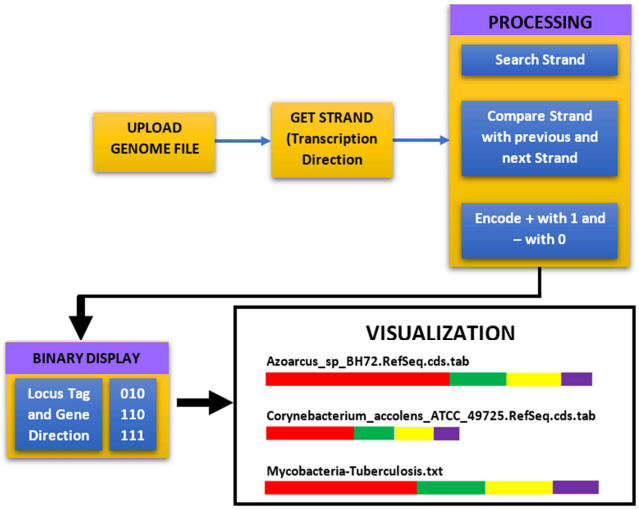

Figure 1. *Architecture of GENAVIS Program.*

2.1. The GENAVIS Algorithm

Input: The set of input Files refseq.cds.tab.
i For each file in Files
ii For each genomes in file

iii Extract GeneObjects (Genomename, locustag and strand) from file
iv Compare current geneobject with previous and next geneobject (based on strand)
v Return any of 111, 011, 110, 010 based on comparison conditions
vi Generate a graphical visualization of the binary output

2.2. Analysis of the GENAVIS Algorithm

Depending on the number of input files, we could have a best case and worst case scenario.

2.2.1. Best Case Scenario

When only one input file is supplied: The outer loop runs only once and while the inner loop goes through each genome objects in the input file. The inner loop depends on the number of genomes in the input file. This could run to over 2500 or more. In the best case scenario, the algorithm runs n times, therefore the growth is O(n). Note that Extraction of Geneobjects, and Comparison are constant.

2.2.2. Worst Case Scenario

When two or more input file is supplied: The outer loop runs at least twice, which is equally n times. The inner loop runs through each files stepped into by the outer loop, the inner loop also runs n times. The growth order in this scenario is therefore $O(n^2)$

2.3. Data Structure

The algorithm uses the List data structure, which is an abstract data type to hold each record of the input file as an object. The list just like an array data structure holds a finite sequence of objects that can be manipulated using sequential access and can further be converted to an array for more flexibility with the operations. This data structure becomes the most suitable for this work since the number of items to be put in each list is unknown prior to the run of the program and the list size grows dynamically.

3. Results

The input to our tool is the genomic features of Mycobacterium tuberculosis BTB09-453 containing genome_name, accession_no, annotation, feature_type, na_feature_id, locus_tag, start_max, end_min, strand, na_length and the gene_product of individual gene in the genome. The gene adjacency investigation here is not a comparative study but an analysis of transcription direction of a single genome. The output of the program is a graphical visualization of the three-digit binary code showing the direction of transcription of a gene with respect to the gene preceding it and the one immediately after it – its adjacent genes. Being the pivot gene for comparison, the second digit in the binary code is always 1 while the first and third digits will be a 1 only if the transcription direction is the same as that of the pivot gene or 0 otherwise.

Figures 2 and 3 shows the visualization of the transcription direction patterns of adjacent genes in Mycobacterium tuberculosis BTB09-453, where the number of genes having the same transcription direction with their adjacent genes is higher in this genome (i.e. 111=>1777). The number of the genes having transcription direction in either side of the adjacent genes is the same and lesser than that of the same direction (i.e. 110 => 829, 011 => 829), whereas genes that do not exhibit the same transcription direction with the two adjacent genes are few (i.e. 010 => 526).

Figure 2. *Genome Visualization of the Transcription Direction of Adjacent Genes in Mycobacterium tuberculosis BTB09-453.*

Comparing a gene with its two adjacent neighbors, the binary digit of 1 is assigned if the transcription direction of the same and the digit 0 is assigned otherwise.

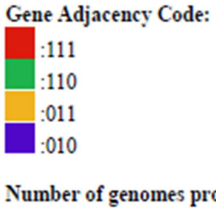

Figure 3. *Visualization Result for Transcription Direction Patterns of Adjacent Genes in Mycobacterium tuberculosis BTB09-453.*

4. Discussion

The prediction of transcription direction enabled us to identify functions of USP which are unknown using neighborhood gene information. Information about transcription direction of genes is useful in predicting gene functions. Our result shows that the genome has constituent that may encode more for universal stress protein because a transcription unit is defined as a set of one or more genes that are transcribed to produce a single messenger RNA. The result provides an insight into the analysis of transcriptional unit with more than one gene and genes encoding for universal stress protein.

GENAVIS was used for the analysis of the transcription direction and its modeling using binary code. This can allow a comparative analysis of multiple genomes using their transcriptional units that further form the basis of bio-molecular network or biosynthetic pathways and multi-gene function prediction.

5. Conclusion

The best case and worst-case scenarios of GENAVIS algorithm are dependent on the number of one or more input

files in use. These factors of one or more files triggers the number of outer loops to be performed on the algorithm operation. Hence, the number of genomes in the input file is related to the number of inner loops runs. The growth order is O(n) and O(n^2) for the best- and worst-case scenarios of the algorithm respectively. In the GENAVIS algorithm, the data structure of unknown prior list of infinite sequence was adopted through manipulation of sequential access and conversion to flexible array operations. Thus, from the results obtained using the input of genomic features of Mycobacterium tuberculosis BTB09-453, the output is a graphical visualization aimed at extracting hiding patterns with a three-binary code each representing the transcription direction, gene preceding and adjacent genes. The transcription direction is vital for identifying functions of universal stress protein and predicting gene functions.

References

[1] B. C. Resendea, A. B. Rebelatoa, V. D'Afonsecab, A. R. Santos b, T. Stutzmanb, V. A. Azevedob, L. L. Santos A, A. Miyoshi B, D. O. Lopes. DNA repair in Corynebacterium model. Gene 2011 482: 1-7.

[2] Baraka S. Williams, Raphael D. Isokpehi, Andreas N. Mbah, Antoinesha L. Hollman, Christina O. Bernard, 1Shaneka S. Simmons, Wellington K. Ayensu, and Bianca L. Garner. Functional Annotation Analytics of Bacillus Genomes Reveals Stress Responsive Acetate Utilization and Sulfate Uptake in the Biotechnologically Relevant Bacillus megaterium. BioinformBiol Insights. 2012; 6:275-86. doi: 10.4137/BBI.S7977. Epub 2012 Nov 21.

[3] Chen Y., Zhang Z., Zheng J., Maa Y. and Xue Y. (2017). Gene selection for tumor classification using neighborhood rough sets and entropy measures. Journal of Biomedical Informatics 67 (2017) 59–68. http://dx.doi.org/10.1016/j.jbi.2017.02.007.

[4] Hou M., Wang S., Li X. and Lei Y. (2009). Neighborhood Rough Set Reduction-Based Gene Selection and Prioritization for Gene Expression Profile Analysis and Molecular Cancer Classification. Journal of Biomedicine and Biotechnology. Volume 2010, Article ID 726413, 12 pages doi:10.1155/2010/726413.

[5] Jiang W., Sun l., Yang X., Wang M., Esmaeili N., Pehlivan N., Zhao R., Zhang H. and Zhao Y. (2017). The Effects of Transcription Directions of Transgenes and the gypsy Insulators on the Transcript Levels of Transgenes in Transgenic Arabidopsis. Scientific Reports | 7: 14757 | DOI:10.1038/s41598-017-15284-x.

[6] Jones R (2010) There Goes the (Gene Expression) Neighbourhood Theory. PLoS Biol 8(11): e1001002. doi:10.1371/journal.pbio.1001002.

[7] L. Martinez-Martinez et al., Clinical significance of Corynebacterium striatum isolated from human sample, Clin Microbiol Infect. 1997 Feb; 3(6):634-639.

[8] Larkin J. D., Cook P. R., and Papantonis A. (2012). Dynamic Reconfiguration of Long Human Genes during One Transcription Cycle. Molecular and Cellular Biology. Volume 32 Number 14. p. 2738–2747.

[9] Makolo Angela and Isokpehi Raphael (2015): Interactive Visual Representations of Gene Transcriptional Direction Patterns in Microbial Genomes. European Molecular Biology Organization Conference on Visualization of Biological Data, VISBI 2015, USA. http://vizbi.org/Posters/2015/B10.

[10] Oliver B., Parisi M. and Clark D. (2002). Gene expression neighborhoods. Journal of Biology 2002, Volume 1, Issue 1, Article 4. http://jbiol.com/content/1/1/4.

[11] Srivatsan A, Tehranchi A, MacAlpine DM, Wang JD (2010) Co-Orientation of Replication and Transcription Preserves Genome Integrity. PLoS Genet 6(1): e1000810. doi:10.1371/journal.pgen.1000810.

[12] Tremonte P., Succi M., Coppola R., Sorrentino E., Tipaldi L., Picariello G., Pannella G. and Fraternali F. (2016) Homology-Based Modeling of Universal Stress Protein from Listeria innocua Up-Regulated under Acid Stress Conditions. Front. Microbiol.7:1998. doi: 10.3389/fmicb.2016.01998.