In Silico Investigations Of Virulent Gene Transfers In Xanthomonas Oryzae: A Study On Rice Bacterial Leaf Blight Disease.

Dharmendra Kashyap[1*]

[1*]Assist Professor, Department of Microbiology & Bioinformatics, Atal Bihari Vajpayee Vishwavidyalaya, Bilaspur (C.G.) Pin 495009, kashyapdk97@gmail.com

***Corresponding Author:** Dharmendra Kashyap
*Assist Professor, Department of Microbiology & Bioinformatics, Atal Bihari Vajpayee Vishwavidyalaya, Bilaspur (C.G.) Pin 495009, kashyapdk97@gmail.com

Abstract:
Rice (Oryza sativa L.) is a major global food crop, providing nutrition for nearly half the world's population. India ranks second in both rice production and acreage, with rice contributing nearly 70% of calories in the Indian diet. The rice is susceptible to various diseases caused by fungi, bacteria, nematodes, and viruses, leading to significant crop losses. Bacterial leaf blight (BLB), is a widespread disease caused by the plant pathogenic bacterium Xanthomonas oryzae pv. oryzae, with reports of frequently gaining genes from non-ancestral origins through methods like conjugation and transduction from other species and genera. These laterally transmitted genes (LTGs) enhance the bacterium's adaptability, pathogenicity, and ability to resist host defences. The present study integrates multiple computational methods to find and analyze genes with potential lateral transfer and abnormal properties, providing insights into the evolution and adaptability of Xanthomonas oryzae pv. oryzae. In the present study, a workflow of computational algorithms to identify horizontally transferred genes (HTGs) in bacterial chromosomes was employed. The SeqWord Gene Island Sniffer program predicted 12 genomic islands (GIs) containing genes with non-ancestral features, characterized by decreased GC content and potentially fast-evolving DNA regions. The DFAST server annotated 248 protein-coding sequences from the identified islands, and NCBI BLAST+ executables matched 225 of these proteins with those of Xanthomonas oryzae $PXO99^{A}$ proteome. MP3 tool predicted 80 pathogenic proteins using the SVM method for analysis. A locally created database of putative horizontally transmitted proteins consisting of nearly 1.3 lakh sequences revealed 20 proteins potentially involved in lateral transfer. Dark Horse web server validated 13 genes of it, and CodonW software assessed anomalous gene nature by correspondence analysis, examining G+C, GC3, and ENc values for 13 anticipated genes compared to overall organism values.

Keywords: Oryza sativa L., Xanthomonas oryzae, Bacterial Leaf Blight, Laterally Transferred Genes, SeqWord Gene Island Sniffer

Introduction

Rice (Oryza sativa L.) is one of the world's prime food crops, consumed by billions of people globally. It is estimated that rice production reaches 470.60 million tons annually, cultivated on over 157 million hectares with an average yield of 4.6 tons per hectare. India ranks among the top rice producers in the world, holding second place to China in both production and cultivated area. Rice is a vital part of the Indian diet, serving as the primary source of calories for over 70% of the population. Within India, rice is grown on approximately 42.70 million hectares with a yield of 3.61 tons per hectare. The Rice Production and Area Under Cultivation in India for the year 2022-23 is estimated at a record 1308.37 lakh tons (130.84 million tons) and the Area under cultivation is nearly 47.832 million hectares. (GOI, 2024; USDA, 2024)

Rice is susceptible to various pathogens, like bacteria, nematodes, fungi and viruses, which reduce both the quantity and quality of the crop. The severity of these diseases can be from moderate to intense, with fungal, nematode and bacterial infections posing the greatest threat. The rapid multiplication and spread of bacteria give them an advantage over other pathogens. These diseases can significantly reduce production, exceeding 50% in some cases, resulting in substantial financial losses estimated at billions of rupees. (Mew 1987)

Pathogenic bacteria, plant or animal, rely on specialized gene clusters called pathogenicity islands (PAIs) and metabolic islands (MAIs) containing specialized genes to successfully infect and colonize their hosts. These genes allow the bacteria to recognize their host, evade the host's immune defences, multiply within the host, and eventually spread to new hosts. PAIs and MAIs have been identified in the complete genomes of some plant pathogenic bacteria. (Ochman et al. 2000 and Juhas et al. 2009) The genes within these islands have a mosaic-like structure, with differences in nucleotide composition and codon usage compared to the rest of the bacterial genome. This suggests that the PAI and MAI genes were acquired from other organisms through a process called horizontal gene transfer (HGT) or Lateral

Gene Transfer (LGT), rather than being inherited vertically from the bacterial ancestor. (Hacker et al. 1997 and Dobrindt et al. 2004) HGT and LGT involves the movement of genetic material between different species and between members of the same species respectively. Both horizontal and lateral gene transfer events have contributed to the fitness and adaptability of plant pathogens. (Hacker et al., 2001) Through reductive evolution, plant pathogens have evolved PAIs that optimize the host-pathogen relationship. LGT has been shown to significantly shape the genome architecture of many bacterial plant pathogens. (Nakamura et al. 2004) Laterally transferred genes can be distinguished from vertically inherited genes based on differences in compositional traits like nucleotide and codon usage. This allows for the identification of recently acquired genes, like those found in PAIs and MAIs, in the genomes of plant pathogenic bacteria. (Dufraigne et al. 2005) Lateral gene transfer, including transformation, transduction, and conjugation, is a mechanism through which DNA is transferred between organisms without sexual reproduction. (Lawrence et al. 2002) Detecting laterally transferred genes can be challenging, and existing methodologies have limitations. Two approaches, known as methodologies, are commonly used, the first approach concentrates on revealing genes that may have been passed down from one generation to the next, while the second approach looks for genes with unusual similarities between organisms. The second approach includes phylogenetic methods, which recognize the abnormal similarity of a gene or distribution between organisms. (Azad and Lawrence 2012) Transformation, transduction, and conjugation are all methods for lateral gene transfer. The mechanisms of natural transformation and conjugation, as well as their different barriers to gene flow, have been well described. (Thomas and Nielsen 2005) Identification of HTG and LTG is quite a complex task as it involves a huge amount of information. Moreover, various pathogens like Xanthomonas oryzae have not been investigated significantly for such types of genes and a gap exists for computational methods of identification. This gave us a motivation to identify such genes. With the availability various available software and databases, we were able to devise a relatively easy method for the identification of HTG with a virulent nature with a very high accuracy. The current workflow is a novel methodology integrating various tools and databases that can be utilized for the identification of HTG in Gram-negative bacteria like Xanthomonas oryzae and others.

Materials and methodology

Sequence Retrieval and processing

The sequence of Xanthomonas oryzae pv. oryzae PXO99^{A}, with NCBI RefSeq ID NC_010717.2, BioProject ID PRJNA224116, BioSample ID SAMN02603061, and taxonomy ID 360094, contains a DNA sequence with a length of 5238555 base pairs, 4959 coding genes and proteins. The FASTA file containing the coding genes and protein sequences were taken from the NCBI server for analysis. **(**Salzberg et al. 2008**)**

Genomic Island Detection

The SeqWord Gene Island Sniffer (SWGIS) program, written in Python, was developed to automatically detect genomic islands that may contain horizontally transferred genes in bacterial and plasmid DNA sequences. The updated version, SWGIS 2.0, is optimized for identifying Mobile Genetic Elements and Laterally Transferred Genes. The software examines compositional biases in tetra-nucleotide distribution throughout the genome, revealing regions that contradict significantly from the general oligonucleotide usage pattern, which are considered genomic islands. The software, accessible at seqword.bi.up.ac.za, operates with default parameters, including a sliding window size of 8000 bp. When used on a raw DNA sequence file, SWGIS aids in identifying genomic islands, which are regions with unique features indicative of a lateral gene transfer in the organism. **(**Bezuidt et al. 2009**)**

Gene Annotation

The DNA Fragment Analysis and Annotation Tool (DFAST), a software program, designed for annotating genes in DNA sequences of Prokaryotes. It is available in both standalone and online server versions, compatible with Macintosh and Linux systems. The tool is used to predict proteins for Xanthomonas oryzae pv oryzae PXO99A. The source code for DFAST is accessible under the GPLv3 license at https://github.com/nigyta/dfast_core/, and an online version can be found at https://dfast.nig.ac.jp/. DFAST is used for annotating genes and their protein products, predicting genes, proteins, and RNA encoded by genomic islands. The server also conducts statistical analyses, including GC content percentage, number of CDS, average protein length, coding ratio percentage, number of rRNAs encoded, and the number of CRISPRs. (Tanizawa et al. 2016; Seemann et al. 2014)

Protein Comparison

The research utilized the NCBI Blast+ executable version 2.12.0 to compare proteins identified by the DFast server with the proteins actually encoded by Xanthomonas oryzae PXO99A obtained from NCBI. Proteins in Xanthomonas oryzae PXO99^{A} that showed an E-value close to zero and 100% query coverage were considered as predicted by DFast. Subsequently, the identified proteins underwent a local BLAST analysis using the NCBI Blast+ program. This analysis was based on the Xanthomonas oryzae proteome to identify proteins encoded by the organism itself. The NCBI Blast+ executables can be downloaded from https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. The default parameters of the BLAST executables were applied to locate protein sequences encoded by the organism. **(**Altschul et al. 1990 and Altschul 1990**)**

Pathogenicity Prediction

The MP3 software is designed to predict pathogenic proteins in Genomic and Metagenomic Data using three techniques: Support Vector Machine, Hidden Markov Model, and a Hybrid method that combines both. It is specifically used to predict the pathogenicity of proteins found in Xanthomonas oryzae pv oryzae PXO99A. The standalone edition of the software is available for Linux operating systems and can be used to evaluate the pathogenicity of a specific protein sequence. The software was downloaded from http://metagenomics.iiserb.ac.in/mp3/download.php. To examine the virulent characteristics of identified proteins, the MP3 software is installed on Linux operating systems, and protein sequences in FASTA format files are analyzed using the software's SVM, HMM, and Hybrid approaches with specific commands for calculating pathogenic proteins. (Gupta et al. 2014)

Horizontal Gene Transfer Analysis

A database of horizontally transferred genes was compiled by identifying proteins from literature sources, particularly those predicted computationally by researchers Hyeonsoo Jeong and Arshan Nasir at the University of Urbana-Champaign, IL, USA. This database contained nearly 130,000 protein sequences and was utilized to detect horizontally transferred sequences among pathogenic proteins in Xanthomonas oryzae pv oryzae PXO99^{A}. Proteins showing virulent traits underwent local BLAST analysis using this curated database of horizontally/laterally transferred gene products. This database, curated from literature sources, was tailored to suit the specific research requirements. The local BLAST analysis focused on proteins identified through MP3 analysis utilizing the SVM method, with attention to criteria such as E-value and Query coverage. These criteria ensured accurate identification of protein function and characteristics, with thresholds set at E-values below 0.05 and identity percentages exceeding 30. (Jeong and Nasir 2017)

HGT source Identification

The DARKHORSE tool, available at http://darkhorse.ucsd.edu/tutorial.html, was employed to identify genes with potential horizontal transfer. It utilized an algorithm for rapid and automated identification and classification of phylogenetically atypical proteins. The algorithm selected potential orthologous matches from a reference database, calculating LPI scores for each genome protein. Phylogenetically atypical proteins, indicative of potential horizontal gene transfer candidates, were identified by selecting matches with low LPI scores (e.g., MAX < 0.6). A low LPI of 0.6 or lower was used for gene identification, with the search encompassing 73 taxonomic groups of Archaea and 1383 taxonomic groups of Bacteria. **(**Podell et al. 2007**)**

Codon Usage Analysis

CodonW had been a software tool that John F. Peden developed in the lab of Paul Sharp at the Department of Genetics, University of Nottingham, UK. CodonW had aided in gaining insights into the genetic code and its evolutionary aspects. The genes identified had undergone CodonW analysis, a package crafted to simplify multivariate analysis of codon usage. CodonW had also been able to analyze sequences encoded by genetic codes that did not adhere to the universal code. In the study, CodonW had been utilized to compute GC content, GC at the 3rd position, and ENc (effective number of codons) for all computationally predicted genes, encompassing the total genome of Xanthomonas oryzae. (Sharp et al. 1987 and Sharp et al. 1994)

Results and Discussion

Results from the SWGIS 2.0 computational tool

The SWGIS v2.0 computational tool (standalone) using Linux OS was employed for detection of
Genomic Islands in the genome sequence of Xanthomonas oryzae pv. oryzae PXO99A. The SWGIS v2.0 scanned the sequence with default parameters for identification purpose.

The program predicted 12 Genomic Islands as given in table no 1.

Total number of residues in Genome file is 5238555 residue and the % of Adenine is 18.23, % of Thymine is 18.14, % of Guanine is 31.71 and % of Cytosine is 31.92. The identified GI contains the ATGC in % as follows 19.96, 19.95, 29.46 and 30.63. So the identified GI have slightly lower GC content then that of the overall GC percentage. For G the content is 2.25% lower while for C it is 1.29%.

Table no 1:: Result of output of The SWGIS 2.0

S/no o Islands	Total length (in basepairs)	%t of A	% of T	% of G	% of C
1	17599	20.61	19.47	29.65	30.28
2	17600	19.66	20.57	29.46	30.31
3	23500	20.57	20.70	29.53	29.20
4	43500	19.42	19.43	30.28	30.87
5	22100	19.26	21.28	29.16	30.30
6	22100	20.07	19.25	29.15	31.53
7	22100	19.36	16.95	30.15	33.54
8	26100	21.31	20.68	28.45	29.56

9	25100	21.24	19.91	28.28	30.57
10	27000	19.95	22.94	28.55	28.57
11	22600	19.08	18.60	29.98	32.35
12	28000	18.97	19.69	30.86	30.48

Annotation from the DFAST (DDBJ Fast Annotation and Submission Tool) server

The DFAST server predicted the following results. Of the 12 Genomic Islands, a total of 258 genes were predicted. The annotation produced 248 protein coding sequences (CDS) and thereby 248 protein sequences were predicted. Besides this 6 t-RNA and 4 r-RNA were also predicted. The GC % was 60.1 for all the input sequences. The coding ratio was 74.9%.

Result of local BLAST with Xanthomonas oryae proteome

The predicted 248 proteins were BLASTed locally against the Xanthomonas oryae proteome which consist of 3907 protein sequences. The BLASTing was done with parameter viz E-value 0.0 and Query coverage of more then 70%. This resulted in identification of 225 protein sequences in Xanthomonas oryae pv oryae PXO99^{A}. This resulted in identification of 225 protein sequences in Xanthomonas oryae pv oryae PXO99^{A}. These sequences were the actual proteins that were predicted by to be present in predicted Genomic Islands. The list is given in table no 2 along with NCBI accession no.

Table no 2 :: List of 225 identified proteins after local blast with Xanthomonas oryae proteome

S/No	Accession no.	S/No	Accession no.	S/No	Accession no.	S/No	Accession no.	S/No	Accession no.
1	WP_012443646.1	**51**	WP_041182581.1	**101**	WP_115877367.1	**151**	WP_011408522.1	**201**	WP_027703597.1
2	WP_070808022.1	**52**	WP_115877377.1	**102**	WP_115877390.1	**152**	WP_011408397.1	**202**	WP_011407606.1
3	WP_048488785.1	**53**	WP_041182721.1	**103**	WP_115877362.1	**153**	WP_041182297.1	**203**	WP_011257669.1
4	WP_041182656.1	**54**	WP_041182575.1	**104**	WP_115801888.1	**154**	WP_011409560.1	**204**	WP_161798565.1
5	WP_041182681.1	**55**	WP_041182787.1	**105**	WP_115877354.1	**155**	WP_012444488.1	**205**	WP_012446212.1
6	WP_041182569.1	**56**	WP_041182834.1	**106**	WP_012444745.1	**156**	WP_012444637.1	**206**	WP_161798564.1
7	WP_012445727.1	**57**	WP_011407237.1	**107**	WP_012446283.1	**157**	WP_041182823.1	**207**	WP_053077065.1
8	WP_012444515.1	**58**	WP_041182856.1	**108**	WP_115877363.1	**158**	WP_041182696.1	**208**	WP_041182807.1
9	WP_012443687.1	**59**	WP_041182574.1	**109**	WP_011258802.1	**159**	WP_011409552.1	**209**	WP_082356982.1
10	WP_012445091.1	**60**	WP_041182619.1	**110**	WP_011258803.1	**160**	WP_041182294.1	**210**	WP_099051307.1
11	WP_011257146.1	**61**	WP_041182667.1	**111**	WP_011258529.1	**161**	WP_041182589.1	**211**	WP_041182803.1
12	WP_011257143.1	**62**	WP_041182821.1	**112**	WP_012445118.1	**162**	WP_012443995.1	**212**	WP_041182804.1
13	WP_011257145.1	**63**	WP_041182615.1	**113**	WP_012444055.1	**163**	WP_012445480.1	**213**	WP_041182810.1
14	WP_115877337.1	**64**	WP_011407176.1	**114**	WP_011258188.1	**164**	WP_082356957.1	**214**	WP_012446246.1
15	WP_109181928.1	**65**	WP_041182948.1	**115**	WP_012443955.1	**165**	WP_011259250.1	**215**	WP_011257439.1
16	WP_115877338.1	**66**	WP_041182607.1	**116**	WP_011257031.1	**166**	WP_011408205.1	**216**	WP_012446249.1
17	WP_115877399.1	**67**	WP_115877353.1	**117**	WP_012445397.1	**167**	WP_011408197.1	**217**	WP_011257436.1
18	WP_115877379.1	**68**	WP_012444351.1	**118**	WP_011259046.1	**168**	WP_012445305.1	**218**	WP_011257438.1
19	WP_115877371.1	**69**	WP_012443634.1	**119**	WP_012446083.1	**169**	WP_011408196.1	**219**	WP_099051333.1
20	WP_109182027.1	**70**	WP_109182069.1	**120**	WP_041182641.1	**170**	WP_011258527.1	**220**	WP_012446253.1
21	WP_115877384.1	**71**	WP_115877351.1	**121**	WP_011408069.1	**171**	WP_027703763.1	**221**	WP_011257243.1
22	WP_115877393.1	**72**	WP_115877341.1	**122**	WP_011257570.1	**172**	WP_027703410.1	**222**	WP_011257245.1
23	WP_115877345.1	**73**	WP_115877380.1	**123**	WP_012446347.1	**173**	WP_027703492.1	**223**	WP_011407318.1
24	WP_115840174.1	**74**	WP_115877360.1	**124**	WP_011258442.1	**174**	WP_012445310.1	**224**	WP_011257238.1
25	WP_115877339.1	**75**	WP_115877350.1	**125**	WP_012445901.1	**175**	WP_012444311.1	**225**	WP_011257239.1
26	WP_115877336.1	**76**	WP_109182089.1	**126**	WP_012445230.1	**176**	WP_012445676.1		
27	WP_109182067.1	**77**	WP_115877352.1	**127**	WP_012445794.1	**177**	WP_012445678.1		
28	WP_115877400.1	**78**	WP_115877343.1	**128**	WP_012443979.1	**178**	WP_012445675.1		
29	WP_115877391.1	**79**	WP_115877347.1	**129**	WP_011257854.1	**179**	WP_048488806.1		
30	WP_115801902.1	**80**	WP_115862280.1	**130**	WP_011257310.1	**180**	WP_011259947.1		
31	WP_041182661.1	**81**	WP_115877386.1	**131**	WP_012445035.1	**181**	WP_012445677.1		
32	WP_115862292.1	**82**	WP_115862274.1	**132**	WP_012445394.1	**182**	WP_115877375.1		
33	WP_109181963.1	**83**	WP_109181915.1	**133**	WP_012444074.1	**183**	WP_012445722.1		
34	WP_041182775.1	**84**	WP_011407626.1	**134**	WP_012444121.1	**184**	WP_012445723.1		
35	WP_115877372.1	**85**	WP_115877383.1	**135**	WP_011408067.1	**185**	WP_075239694.1		

36	WP_041182780.1	**86**	WP_115877355.1	**136**	WP_011407587.1	**186**	WP_011258207.1
37	WP_041182741.1	**87**	WP_115877369.1	**137**	WP_082357007.1	**187**	WP_011258200.1
38	WP_041182545.1	**88**	WP_115862279.1	**138**	WP_082356999.1	**188**	WP_012445724.1
39	WP_115877397.1	**89**	WP_115877373.1	**139**	WP_012444473.1	**189**	WP_075240152.1
40	WP_115862289.1	**90**	WP_115877382.1	**140**	WP_012444487.1	**190**	WP_011407948.1
41	WP_115801893.1	**91**	WP_115877361.1	**141**	WP_012445479.1	**191**	WP_011258198.1
42	WP_011257851.1	**92**	WP_115877389.1	**142**	WP_012445373.1	**192**	WP_011407947.1
43	WP_115877388.1	**93**	WP_115877364.1	**143**	WP_012444638.1	**193**	WP_011407610.1
44	WP_115877401.1	**94**	WP_115877346.1	**144**	WP_012444489.1	**194**	WP_011407603.1
45	WP_115840195.1	**95**	WP_115877358.1	**145**	WP_012444636.1	**195**	WP_012446112.1
46	WP_041182571.1	**96**	WP_115877348.1	**146**	WP_012444640.1	**196**	WP_011257665.1
47	WP_041182811.1	**97**	WP_115877340.1	**147**	WP_012444644.1	**197**	WP_011257668.1
48	WP_041182722.1	**98**	WP_115877370.1	**148**	WP_027703472.1	**198**	WP_011407601.1
49	WP_109182045.1	**99**	WP_115877359.1	**149**	WP_011407609.1	**199**	WP_011407605.1
50	WP_041182719.1	**100**	WP_115877398.1	**150**	WP_082325607.1	**200**	WP_011407612.1

Result of local BLAST with Xanthomonas oryae proteome

The predicted 248 proteins were BLASTed locally against the Xanthomonas oryae proteome sequences. The BLASTing was done with parameter viz E-value 0.0 and Query coverage of more then 70%. This resulted in identification of 225 protein sequences in Xanthomonas oryae pv oryae PXO99^{A}.

Laterally Transferred Sequence Analysis

The predicted proteins with pathogenic nature were BLASTed against the locally prepared database of HTG/LTG products. The local blasting was done with E-value as selection criteria keeping its value 0.05 or less. The result of BLAST against the local database of predicted Laterally transferred sequence resulted in identification of 20 protein sequences which may be the result of Lateral transfer. The predicted result is given here in tabular form. It is clear from the result that most of the identified protein sequences were with percentage identity in a range of 23 to 33, only two sequences viz WP_011257146.1 and WP_011408205.1 showed identity of 100%. A list of LTG's are given in table no 3.

Table no 3 :: Summary of results from predicted protein sequences after local blasting with database of Laterally Transferred Gene products

S no	Query accession no	Subject accession no	% identity	Alignment length	Mismatches	Gap opens	Query beginning	Query ending	Subject beginning	Subject ending	E-value	Bit score
1	WP_011257143.1	gi|328452382|gb|AEB08211.1|	27.795	331	201	11	76	401	56	353	5.02E-24	103
2	WP_011257145.1	gi|209960010|gb|ACJ00647.1|	28.52	277	158	9	124	385	144	395	2.60E-13	72.4
3	WP_011257146.1	gi|84365735|dbj|BAE66893.1|	100	436	0	0	1	436	1	436	0	851
4	WP_011257239.1	gi|151360998|gb|ABS04001.1|	28.788	132	84	6	55	182	47	172	1.81E-04	45.1
5	WP_011257668.1	gi|83655156|gb|ABC39219.1|	26.364	220	123	6	42	252	53	242	3.68E-08	56.2
6	WP_011407318.1	gi|257474042|gb|ACV54362.1|	27.823	248	163	8	11	248	10	251	1.91E-12	70.1
7	WP_011407612.1	gi|308739720|gb|ADO37380.1|	27.168	346	215	11	5	328	16	346	3.93E-21	94.7
8	WP_011408069.1	gi|109627148|emb|CAJ53630.1|	23.913	276	139	9	207	432	15	269	1.54E-04	45.8
9	WP_011408196.1	gi|260651540|emb|CBG74663.1|	30.769	91	51	3	54	137	97	182	0.015	40
10	WP_011408205.1	gi|188521810|gb|ACD59755.1|	100	889	0	0	1	889	1	889	0	1809

11	WP_012444311.1	gi|218173178|gb|ACK71911.1|	31.519	698	461	9	9	696	301	991	2.79E-95	318
12	WP_012444489.1	gi|157912763|gb|ABV94196.1|	28.704	108	54	3	429	535	368	453	1.40E-06	53.9
13	WP_012444636.1	gi|157912763|gb|ABV94196.1|	28.704	108	54	3	429	535	368	453	1.45E-06	53.9
14	WP_012445310.1	gi|254947271|gb|ACT91971.1|	23.103	290	183	9	70	326	66	348	1.34E-07	55.1
15	WP_012445675.1	gi|91716355|gb|ABE56281.1|	24.427	131	90	2	445	575	408	529	1.32E-05	50.4
16	WP_012445677.1	gi|91716355|gb|ABE56281.1|	24.427	131	90	2	76	206	408	529	7.37E-06	
17	WP_012445724.1	gi|427362785|gb|AFY45507.1|	28	100	69	2	346	445	198	294	2.19E-04	44.7
18	WP_041182810.1	gi|291583868|gb|ADE11526.1|	33.028	654	374	14	92	716	170	788	1.45E-94	314
19	WP_161798564.1	gi|442799364|gb|AGC75169.1|	22.865	363	228	10	54	377	73	422	1.85E-07	54.7
20	WP_161798565.1	gi|428686769|gb|AFZ46629.1|	25.738	610	414	15	49	632	383	979	5.35E-40	159

Results from Dark Horse

From the result of local BLAST, 20 sequences were subjected to **Dark Horse analysis.** The search parameters were as following. The search was performed against 73 Archaea and 1383 Bacteria taxonomical group. LPI (Lineage Probability Index) index of 0.6 or less (High scoring matches are found in close phylogenetic relatives, and unlikely to be horizontal in nature while low scoring matches are potential HGT candidates.) E-value of 0.00005 or less and Phylogenetic Granularity level must be the strain. Out of the 20 protein sequences 13 were turned out with positive result. The study identified several sequences with potential horizontal gene transfer (HGT) events. For sequence ID WP-011257143.1, Desulfobacterium autotrophicum is the most likely candidate for contributing to the HGT event, with a normalized LPI value ranging from 0.461 to 0.564. Similarly, for sequence ID WP_011257145.1, Nitrococcus mobilis is identified as the probable organism for the HGT event, with an LPI value of 0.561. For WP_011257146.1, Opitutus terrae PB90-1 is the most likely contributor, with an LPI value ranging from 0.464 to 0.563. Blastopirellula marina DSM 3645 is identified for WP_011257239.1 with an LPI value of 0.42. Geobacillus sp. Y412MC10 is associated with WP_11257668.1 with an LPI value of 0.43. Rhizobium etli Brasil 5 is linked to WP_011407318.1 with an LPI value of 0.435. Methylobacillus flagellatus KT is linked to WP_011407612.1 with an LPI value of 0.5. Bacteroides fragilis NCTC 9343 is associated with WP_11408069.1 with an LPI value of 0.45. Sphingomonas sp. SKA58 is identified for WP_011408196.1 with an LPI value of 0.47. Thermobifida fusca YX is linked to WP_011408205.1 with an LPI value of 0.406. Herpetosiphon aurantiacus is identified for WP_012444311.1 with an LPI value of 0.477. Clostridium beijerinckii is associated with WP_041182810.1 with an LPI value of 0.45. Lastly, Petrotoga mobilis SJ95 & Leeuwenhoekiella blandensis is linked to WP_161798565.1 with an LPI value of 0.01, which had the highest number of associated organisms, 41. Notably, WP_011257145.1, WP_011257146.1, and WP_012444311.1 only showed two associated organisms each contributing to potential HGT events. A list is given in table no 4.

Table no 4 :: Summary of the results from Dark-Horse

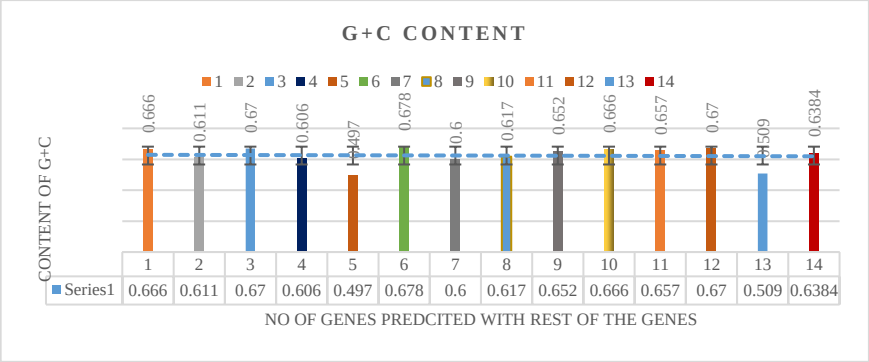
S no	**SEQ ID**	**Query_species**	**Query_id**	**Norm_LPI**	**Query_description**	**Bestmatch_id**	**Bestmatch_species**	**Bestmatch_descripton**
01	WP-011257143.1	Clostridium_beijerinckii_NCIMB_8052	gi|150017724|ref|YP_001309978.1|	0.461	Hypothetical protein	gi|224371068|ref|YP_002605232.1|	Desulfobacterium autotrophicum HRM2	ArcE [Desulfobacterium autotrophicum HRM2]
02	WP_011257145.1	Dehalococcoides_sp_VS	641417637	0.561	Protein of unknown function DUF214	gi|88810831|ref|ZP_01126088.1|	Nitrococcus mobilis Nb-231	Hypothetical protein NB231_17163 [Nitrococcus mobilis Nb-231]gi|88792461|gb|EAR

03	WP_011257146.1	Shewanella_woodyi_ATCC_51908	gi|170726376|ref|YP_001760402.1|	0.464	Hypothetical protein	gi|182413330|ref|YP_001818396.1|	Opitutus terrae PB90-1	Hypothetical protein Oter_1512 [Opitutus terrae PB90-1]gi|177840544|gb|ACB74796
04	WP_011257239.1	Bacteroides_ovatus_ATCC_8483	641383049	0.42	Hypothetical protein	gi|87309129|ref|ZP_01091266.1|	Blastopirellula marina DSM 3645	Cellulase [Blastopirellula marina DSM 3645]
05	WP_11257668.1	Pseudomonas_fluorescens_Pf-5	gi|70729409|ref|YP_259147.1|	0.43	Glycosyl transferase, group 2 family protein	gi|192812792|ref|ZP_03041459.1|	Geobacillus sp. Y412MC10	Glycosyl transferase family 2 [Geobacillus sp. Y412MC10] gi|192798600|gb|EDV7511
06	WP_011407318.1	Bacillus_thuringiensis_sv_israelensis_ATCC_35646	638641705	0.435	Tetracycline resistance protein	gi|218508856|ref|ZP_03506734.1|	Rhizobium etli Brasil 5	Tetracycline efflux transporter protein [Rhizobium etli Brasil 5]
07	WP_011407612.1	Thermosynechococcus_elongatus	gi|22298001|ref|NP_681248.1|	0.5	dTDP-glucose 4,6-dehydratase	gi|91776360|ref|YP_546116.1|	Methylobacillus flagellatus KT	dTDP-glucose 4,6-dehydratase [Methylobacillus flagellatus KT]
08	WP_11408069.1	Pseudomonas_putida_W619	gi|170722075|ref|YP_001749763.1|	0.42	Sulfatase	gi|60682678|ref|YP_212822.1|	Bacteroides fragilis NCTC 9343	Putative sulfatase [Bacteroides fragilis NCTC 9343]
09	WP_011408196.1	Clostridium_acetobutylicum	gi|15895779|ref|NP_349128.1|	0.47	Beta galactosidase	gi|94496174|ref|ZP_01302752.1|	Sphingomonas sp. SKA58	Beta-galactosidase [Sphingomonas sp. SKA58] gi|94424353|gb|EAT09376.1| beta-gala
10	WP_011408205.1	Bacillus_licheniformis_ATCC_14580	gi|52081845|ref|YP_080636.1|	0.406	Putative glycoside hydrolase family 3	gi|72162008|ref|YP_289665.1|	Thermobifida fusca YX	Exo-1,4-beta-glucosidase [Thermobifida fusca YX]
11	WP_012444311.1	Anaeromyxobacter_Fw109-5	gi|153005977|ref|YP_001380302.1|	0.477	ABC transporter related	gi|159900258|ref|YP_001546505.1|	Herpetosiphon aurantiacus ATCC 23779	ABC transporter related [Herpetosiphon aurantiacus ATCC 23779]
12	WP_041182810.1	Prosthecochloris_aestuarii_DSM_271	gi|194333102|ref|YP_002014962.1|	0.45	Multi-sensor hybrid histidine kinase	gi|150017373|ref|YP_001309627.1|	Clostridium beijerinckii NCIMB 8052	Multi-sensor hybrid histidine kinase [Clostridium beijerinckii NCIMB 8052]
13	WP_161798565.1	Pyrobaculum_aerophilum	gi|18313526|ref|NP_560193.1|	0.01	Iron (III) ABC transporter ATP-binding protein,	gi|86143028|ref|ZP_01061450.1|	Leeuwenhoekiella blandensis MED217	ABC transporter, ATP-binding protein [Flavobacterium sp. MED217] gi|85830473|gb|
		Thermococcus_AM4	scf3_TAM4_1155	0.01	Glutamine ABC transporter, ATP-binding protein	gi|160903092|ref|YP_001568673.1|	Petrotoga mobilis SJ95	ABC transporter related [Petrotoga mobilis SJ95]

Results from CodonW

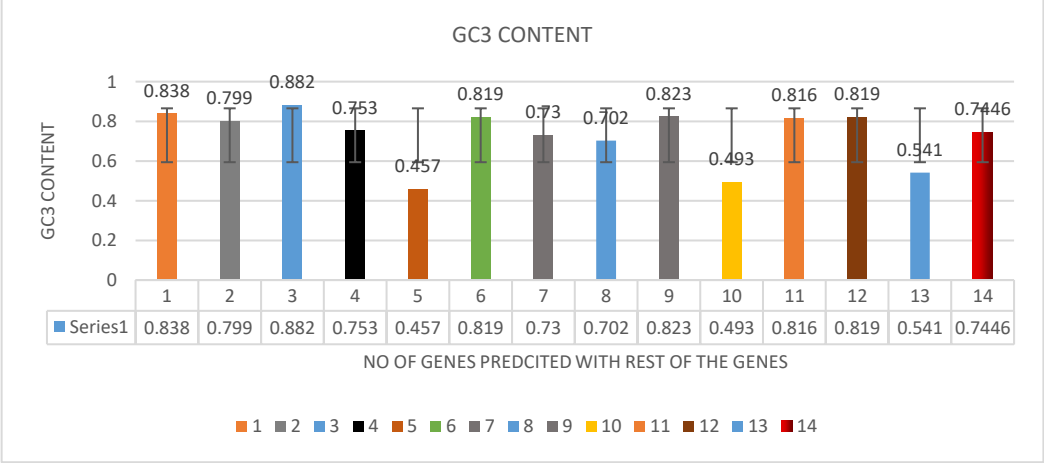
GC content of all 13 computationally predicted against the rest of genes of the X. oryzae. The GC% of all genes of X. oryzae is shown at position no 14, and the rest of the genes are represented by 1 to 13. we can see that GC content of genes range from 49.7% for PXO_RS27070 position 4635405 to 4636559 to 67.8% for PXO_RS23315, mean was 63.84. The figure no 1 is given here containing the GC content analysis of genes.

Figure no 1 :: GC content analysis of all 13 computationally predicted against the rest of genes of the X. oryzae.



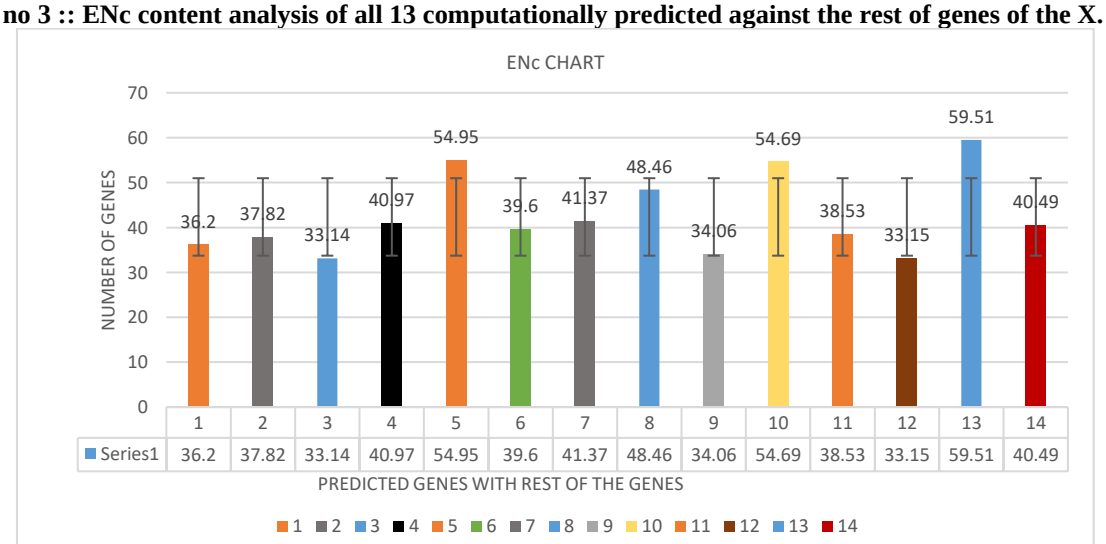
The GC content at 3rd position of all genes of X. oryzae is shown at position no 14, and the rest of the genes are represented by 1 to 13. we can see that GC3 content of genes range from 0.457 for PXO_RS27070 to 0.882 for PXO_RS00680, mean was 0.7446. The figure no 2 contains the GC3 content analysis of all 13 computationally predicted genes.

Figure no 2 :: GC3 content analysis of all 13 computationally predicted against the rest of genes of the X. oryzae.



The ENc content of all genes of X. oryzae is shown at position no 14, and the rest of the genes are represented by 1 to 13. highest number of ENc was shown by PXO_RS21965 and least by PXO_RS00680. mean was 40.49. The results gave an idea that the predicted genes are showing deviation from mean values predicted for all genes. The figure no 3 contains the analysis of all 13 computationally predicted against the rest of genes of the X. oryzae.

Figure no 3 :: ENc content analysis of all 13 computationally predicted against the rest of genes of the X. oryzae.



Finally we were able to predicted 13 gene and protein sequence with a strong evidence of LTG origin. A list is given in table no 5 below.

Table no 5:: List of gene and protein sequence finally predicted

S no	Locus tag	location	Protein ID	Protein Product
1	PXO_RS00655	140332..141585	WP_011257143.1	Efflux RND transporter periplasmic adaptor subunit
2	PXO_RS00675	143511..144668	WP_011257145.1	ABC transporter permease
3	PXO_RS00680	144681..145991	WP_011257146.1	ABC transporter permease
4	PXO_RS23345	5097859..5098989	WP_011257239.1	Glycoside hydrolase family 5 protein
5	PXO_RS27070	4635405..4636559	WP_011257668.1	Glycosyltransferase
6	PXO_RS23315	5091458..5092721	WP_011407318.1	TCR/Tet family MFS transporter
7	PXO_RS21085	4625599..4626654	WP_011407612.1	dTDP-glucose 4,6-dehydratase
8	PXO_RS06905	1490476..1492065	WP_011408069.1	Phosphoethanolamine transferase
9	PXO_RS14870	3274434..3276275	WP_011408196.1	Beta-galactosidase
10	PXO_RS14820	3257300..3259969	WP_011408205.1	glycoside hydrolase family 3 C-terminal domain-containing protein
11	PXO_RS06015	1285459..1287618	WP_012444311.1	peptidase domain-containing ABC transporter RaxB
12	PXO_RS22165	4852212..4854521	WP_041182810.1	ATP-binding protein
13	PXO_RS21965	4800059..4801975	WP_161798565.1	ATP-binding cassette domain-containing protein

Conclusion

The study employed a combination of bioinformatics tools to identify and characterize laterally transferred genes (LTGs) in the genome of Xanthomonas oryzae pv oryzae PXO99A. The SeqWord Gene Island Sniffer program was used to predict genomic islands with a high probability of containing atypical genes, which may have been acquired through horizontal gene transfer (HGT). The program identified 12 genomic islands with a lower GC content compared to the overall GC percentage of the genome, suggesting that these regions may be fast-evolving DNA segments. The DDBJ Fast Annotation and Submission Tool (DFAST) was used to annotate the genes and gene products from the identified islands, resulting in the prediction of 248 protein-coding sequences. The NCBI BLAST+ executable was employed to compare the identified proteins with those of Xanthomonas oryae PXO99^{A}, resulting in the identification of 225 protein sequences. The MP3 software was used to predict pathogenic proteins among the identified sequences, and 80 protein sequences were selected based on their SVM scores. The selected protein sequences were then blasted against a locally made database of computationally predicted horizontally transferred genes, resulting in the identification of 20 protein sequences that may have been acquired through HGT. The Dark Horse web server was used to identify the probable sources of HGT for these sequences, and 13 sequences were found to have positive results. The analysis suggested that these sequences may have been acquired from various organisms, including Desulfobacterium autotrophicum, Nitrococcus mobilis, Opitutus terrae PB90-1, and others. The study also analyzed the GC content, GC content at the 3rd position (GC3), and the ENC (effective number of codons) content of the predicted genes. The results showed that the predicted genes deviated from the mean values predicted for all genes of X. oryzae, suggesting that they may have distinct evolutionary histories. Overall, the study demonstrates the power of combining multiple bioinformatics tools to identify and characterize LTGs in bacterial genomes. The results provide insights into the evolution of Xanthomonas oryzae pv oryzae PXO99^{A} and suggest that HGT may have played a significant role in shaping its genome. The identified genes may be used for analysis and may be subjected to further validation. The identified genes may be further subjected for various genomic analysis like regulation and expression related one. The identified proteins may again be subjected for further analysis like identification of mass, physio chemical properties and interaction with other proteins in the organism with special focus on pathogenicity related pathways etc.

Acknowledgments

The author thanks the Dept of Microbiology and Bioinformatics, Atal Bihari Vajpayee Vishwavidyalaya, Bilaspur, Chhattisgarh for funding and providing all the necessary infrastructure to carryout the research work.

Conflict of Interest

No conflict of interest exist.

References

1. Altschul, S. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1006/jmbi.1990.9999
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
3. Azad, R. K., & Lawrence, J. G. (2012). Detecting laterally transferred genes. In Methods in Molecular Biology (pp. 281–308). Totowa, NJ: Humana Press.
4. Bezuidt, O., & Reva, O. (2009). SeqWord Gene Island Sniffer: A program to study the lateral genetic exchange among bacteria. World Academy of Science, Engineering and Technology, 58, 1169–1174.
5. Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology, 2(5), 414–424. https://doi.org/10.1038/nrmicro884
6. Government of India. (2024). Second advance estimates of production of major crops released [Press release]. Retrieved April 29, 2024, from https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1899193
7. Gupta, A., Kapil, R., Dhakan, D. B., & Sharma, V. K. (2014). MP3: A software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLOS ONE, 9(4), e93907. https://doi.org/10.1371/journal.pone.0093907
8. Hacker, J., & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity: A Darwinian view of the evolution of microbes. EMBO Reports, 2(5), 376–381. https://doi.org/10.1093/embo-reports/kve097
9. Hacker, J., Blum-Oehler, G., Mühldorfer, I., & Tschäpe, H. (1997). Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution. Molecular Microbiology, 23(6), 1089–1097. https://doi.org/10.1046/j.1365-2958.1997.3101672.x
10. Jeong, H., & Nasir, A. (2017). A preliminary list of horizontally transferred genes in prokaryotes determined by tree reconstruction and reconciliation. Frontiers in Genetics, 8. https://doi.org/10.3389/fgene.2017.00112
11. Juhas, M., Van Der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., & Crook, D. W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiology Reviews, 33(2), 376–393. https://doi.org/10.1111/j.1574-6976.2008.00136.x
12. Mew, T. W. (1987). Current status and future prospects of research on bacterial blight of rice. Annual Review of Phytopathology, 25(1), 359–382. https://doi.org/10.1146/annurev.py.25.090187.002043
13. Nakamura, Y., Itoh, T., Matsuda, H., & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Genetics, 36(7), 760–766. https://doi.org/10.1038/ng1381
14. Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature, 405(6784), 299–304. https://doi.org/10.1038/35012500
15. Podell, S., & Gaasterland, T. (2007). DarkHorse: A method for genome-wide prediction of horizontal gene transfer. Genome Biology, 8(2), R16. https://doi.org/10.1186/gb-2007-8-2-r16
16. Salzberg, S. L., Sommer, D. D., Schatz, M. C., Phillippy, A. M., Rabinowicz, P. D., Tsuge, S., Furutani, A., Ochiai, H., Delcher, A. L., Kelley, D., Madupu, R., Puiu, D., Radune, D., Shumway, M., Trapnell, C., Aparna, G., Jha, G., Pandey, A., Patil, P. B., Bogdanove, A. J. (2008). Genome sequence and rapid evolution of the rice pathogen Xanthomonas oryzae pv. oryzae PXO99A. BMC Genomics, 9(1), 204. https://doi.org/10.1186/1471-2164-9-204
17. Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153
18. Sharp, P. M., & Li, W.-H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research, 15(3), 1281–1295. https://doi.org/10.1093/nar/15.3.1281
19. Sharp, P. M., & Matassi, G. (1994). Codon usage and genome evolution. Current Opinion in Genetics & Development, 4(6), 851–860. https://doi.org/10.1016/0959-437X(94)90070-1
20. Tanizawa, Y., Fujisawa, T., Kaminuma, E., Nakamura, Y., & Arita, M. (2016). DFAST and DAGA: Web-based integrated genome annotation tools and resources. Bioscience of Microbiota, Food and Health, 35(4), 173–184. https://doi.org/10.12938/bmfh.16-003
21. Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology, 3(9), 711–721. https://doi.org/10.1038/nrmicro1234
22. U.S. Department of Agriculture. (2024). India rice area, yield and production. Retrieved April 29, 2024, from https://ipad.fas.usda.gov/countrysummary/Default.aspx?id=IN&crop=Rice