In silico analysis of mercury resistance genes extracted from Pseudomonas spp. involved in bioremediation

Duguma Dibbisa (chemduguma2013@gmail.com)
Haramaya University
Gobena Wagari
Oda Bultum University

Research Article

Keywords: Bioremediation, Contamination, Heavy metal, and Transcription factors

Posted Date: May 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1630105/v1

Abstract

Background: Microbial gene and gene production were diverse and beneficial for heavy metal bioremediation from the contaminated sites. Screening of genes and gene products plays an important role in the detoxification of pollutants. Understanding of promoter region and its regulatory elements is a vital implication of microbial genes. To the best of our knowledge, there is no *in silico* analysis report so far on *mer* genes families used for heavy metal bioremediation.

Results: The motif distribution was observed densely upstream of the TSS from +1 to -400bp and sparsely distributed above -500bp, according to the current study. MEME identified the best common candidate motifs of TFs binding with the lowest e value (7.2e 033) and is the most statistically significant candidate motif. The EXPREG output of the 11 TFs with varying degrees of function as activation, repression transcriptions, and dual purposes was thoroughly examined. Data revealed that transcriptional gene regulation in terms of activation and repression was observed at 36.4% and 54.56% respectively. This shows that the vast majority of TFs are involved in the transcription gene repression rather than activation. Likewise, EXPREG output revealed that transcriptional conformational modes, such as monomer, dimer, tetramer, and other factors, were also analyzed. The data indicated that the majority of transcriptional conformation mode was dual which accounts for 96%. CpG island analysis using online and offline tools revealed that the gene body had fewer CpG islands as compared with the promoter regions.

Conclusion: Understanding the common candidate motifs, transcriptional factors binding sites, and regulatory elements of the mer operon gene cluster using a machine learning approach could help us better understand gene expression patterns in heavy metal bioremediation.

Introduction

Worldwide human populations are increasing at alarming rates. It has been estimated this population will reach nine billion by the year 2050 [1, 2]. Population growth contributes to the degradation of natural resources. Thus, environmental protection is imperative for a functioning and balanced ecosystem. Several environmental pollutants cause multifaceted degradation and affect ecosystem components, particularly soil, water, and the entire biodiversity. Heavy metals chemically refer to a class of a specific subdivision of elements marked with metallic properties. It is the most significant atmospheric contaminant discharged from natural and anthropogenic activities. Metals are everywhere but in different concentrations. Exceeding the required concentration will result in contamination [3]. The density at $5gcm^{-3}$ and the concentrations of heavy metals present in the environment are highly toxic to biodiversity [4].

The availability or entry of heavy metals into the ecosystem comes from various sources, either naturally or human-induced activities. The natural sources of heavy metal contamination include geological weathering, volcanic eruptions, industrial effluents, and chemicals widely used in the agricultural sectors namely: pesticides herbicides, and insecticides are sources of anthropogenic activities [5]. Our natural environment is also contaminated by heavy corrosion, metal ions, heavy metal leaching, resuspension, and household wastes released into the soil and groundwater. Gold mining and other metal industries are the main causes of soil contamination from mercury. Mercury is a unique important heavy metal extensively used in the developing and developed world for income. Nonetheless, the trend in developing countries is significantly lower; for example, in Ethiopia, it is a common practice in some areas [6]. Heavy metal like mercury is essential for living organisms in certain concentrations; however, its excessive concentrations are significantly carcinogenic and toxic. The toxicity of these heavy metals can cause severe illness in humans [3].

The removal of heavy metals from the environment has become an extremely pertinent issue in the current scenario. The use of different methods to remove or reduce the harmful effects of heavy metals contamination is physical evacuation, chemical cleaning and stabilization of metals at the site, and use of the biological entities as bioremediation

[7]. Using microbial biomass as a platform for heavy metal ion removal is an alternative method of bioremediation. It is a biological phenomenon in which microbes use genes and gene products to take up and accumulate metal ions in the intracellular space for use in cellular processes [8]. Heavy metal ions can be absorbed and accumulated by microorganisms in their intracellular space and used for a variety of purposes. So far, various studies focus on the cost-effective and environmentally friendly applications of bioremediation in heavy metal removal. Transcription factors (TFs) that recognize specific DNA sequences near promoter regions and transcription-binding sites associated with genes that play key roles in the structure and function of genes and the region of promoter of genes in mercury bioremediation have not yet been studied. Therefore, the objective of this study was to identify the promoter region, transcriptional factors with corresponding binding sites, CpG islands involved in the regulation of the expression, as well as its regulatory elements, to provide baseline information for working mercuric bioremediation and environmental applications.

Material And Methods

Determination of TSS and Promoter Regions

The *Pseudomonas spp*. genes sequences responsible for mercuric bioremediation were retrieved from the NCBI genome browser that is available at https://www.ncbi.nlm.nih.gov/gene in March 2022 as in Table 1. For the current study, about ten protein-coding sequences were extracted after checking the search results in sequence databases. To analyze the specific gene further, the presence of the starting coding sequences was predicted whether they were found on positive or negative strands. The region of the transcriptional start site (TSS) was determined by extending sequences from the genomic coordinate regions. The FASTA file format of query sequences was used for further analysis. The prepared 1kb upstream sequences from the transcriptional start site (TSS) were taken to Neural Network Promoter Prediction (NNPP version 2.2) (https://www.fruitfly.org/seq_tools/promoter.html) tools to obtain potential core promoter [9]. The NNPP version 2.2 toolset was used with a minimum standard predictive promoter score with a default cut-off value of 0.8 for prokaryotic cells and intended to eliminate zero counts by 80% from the query sequences before the transformation.

Table 1
Mercury bioremediation genes and their general function and genome coordinates

S.N	Gene Id	Gene symbol	Genome coordinate	Start codons	Gene function
1.	69751970	*merA*	c33607-31961	ATG	Mercury(II) reductase
2.	66762507	*merB*	c3805546-3806184	ATG	Organomercurial lyase
3.	66762509	*merC*	c3808349-3807915	ATG	Organomercurial transporter
4.	69747981	*merD*	188629–188994	ATG	Mercury resistance co-regulator
5.	69751968	*merE*	c31582-31346	ATG	Broad-spectrum mercury transporter
6.	69751971	*merF*	c33849-33604	ATG	Mercury resistance system transport protein
7.	69751974	*merR*	34565–34999	ATG	Hg(II)-responsive transcriptional regulator
8.	69751972	*merP*	c34127-33852	ATG	Mercury resistance system periplasmic binding protein
9.	69747978	*merT*	186216–186566	ATG	Mercuric ion transporter
10.	46432416	*merG*	5771173–5771826	ATG	Phenyl mercury resistance protein

Determination of Common motifs and TFs *Pseudomonas spp.* genes

The promoter sequence region identified based on the established criteria were imported and studied using (MEME) released 5.4.1 version via web a server hosted by the National Biomedical Computational Resource (https://meme-suite.org/meme/tools/meme) [10] to look for common candidate motifs that serve for the binding sites of transcriptional factors that regulate the expression of heavy metal accumulate genes. MEME-Suit searches for statistically significant candidate motifs in the sequence were imported. The MEME output is in the form of XML, text, MASTHML, MASTXML, MAST text, and HTML showing the candidate motifs as local multiples of the input promoter sequences. The MEME-suit predicted and discovered gene sequences with novel motifs (fixed-length repetitive patterns) were submitted to online tools. This technique determined the occurrence of the common motifs that serve as binding sites for the transcription factors expected to regulate expression levels of heavy metal bioaccumulation. MEME-Suit was used to perform motif prediction and discovery, motif alignment analysis, motif scanning, and motif comparison [11]. Before starting the search for typed sequences, the basic search parameters for the motif distribution menu were set, including the distribution of motif locations, the zero option or more occurrences per sequence, while keeping the number of motifs and the remaining motif width as the default. After the MEME searches were completed, the search results page was linked to the MEME output in HTML format. This stage was a fundamental initial point of view for the expected value (e-value). The smaller the e values, the better the agreement [11]. At the bottom of the MEME HTML output, one or all of the candidate motifs can be forwarded for further analysis and the identical motifs can be further characterized by other web server programs. In these cases, the TOMTOM webserver was used to search for sequences that matched the identified motif for its respective TFs. TOMTOM output includes LOGOS representing the alignment of the candidate motif and TF with the p-value and q-value (a measure of the false discovery rate) of the match and links back to the parent transcription database for a more detailed sequence match information [12, 11].

Search for CpG islands for *Pseudomonas spp.* encoding genes

A 1kb query sequence in FASTA format from the upstream of the TSS and the entire sequence of the gene bodies were prepared for all ten *Pseudomonas spp.* protein-coding sequences. The regulatory region, CpG islands representing regions of a sequence, was examined with two algorithms. The first algorithm was the offline tool CLC Genomic Workbench version 20.0.40, CLC Bio, Aarhus, Denmark) used to search the restriction enzyme sites *MspI*, with fragment sizes between 40 and 220bp parameters). The second tool was the Takai and Jones algorithm with search criteria in GC contents of ≥ 55% and observed CpG/expected CpG ratio of ≥ 0.65%, and a length of ≥ 500bp [13]. The CpG island search tool available at the web link (http://dbcat.cgm.ntu.edu.tw/) was used for this purpose.

Results

Determination of Transcriptional Start Sites (TSS)

Understanding a regulatory gene is one of the most difficult challenges in the entire genome. Therefore, identification of the TSS is key information for gene expression. Transcription start sites (TSS) are the first nucleotides of DNA sequences where transcription has been started. On the other hand, it is where the RNA polymerase enzyme binds upstream of the start site. The online Neural Network Promoter Prediction (NNPP) version 2.20 databases were used to find the TSS for the gene extracted from *Pseudomonas spp.*, which is widely used for mercury bioremediation. The promoter region located upstream of 1kb of the TSS was characterized on the assumption that the functional gene elements of the promoter can be found within the region. The predicted values for each of the coding sequences of *mer* operon gene varieties in mercury bioremediation have been summarized and presented in Table 2. Accordingly, the *mer* operon gene variety has several TSS values ranging from 1 to 4. Interestingly, about six identified genes (*merA*, *merB*, *merD*, *meer*, *merF*, and *merP*) have the same TSS values and *merC* has only one TSS value as can be seen from Table 2. The current studies show that the promoter region of almost all sequences had multiple TSS values, showing a similar investigation of genome-wide identification of TSS, promoter, and TF binding sites in *E. coli* [14].

TSSs were located at various distances from the start codon, as observed in Table 2. This shows that promoter sequence plays an important role in enhancing or hindering transcription initiation and gene regulation in response to environmental changes. The genes indicated by *merD*, *merG*, and *merR* were the highest values observed for positive-strand localization, respectively. While *merB* and *MerE* were the highest values that have been among the other TSS found on the negative strands. However, the majority of the TSS of *mer* operon genes were found on the negative strand, while few of them were on the positive strands. Therefore, knowing the wide application of TSS such as gene function and its structure determination, predicting the promoter region, and gene regulation have been perceived in the current scenario of gene prediction.

Table 2
TSS number, its promoter predictive score values, and distance from the start codon of *mer*-operon genes associated with mercury bioremediation

SN	Gene Id	Gene symbol	No of predictive promoter	No of TSS Identified	The predictive score value cut off at 0.80	Distance From ATG	Orientation of Complementary Strands
1.	69751970	*merA*	2	2	0.97, 0.91	-929	-ve
2.	66762507	*merB*	2	2	0.85,0.82	-1951	-ve
3.	66762509	*merC*	1	1	0.85	-686	-ve
4.	69747981	*merD*	2	2	0.93,0.89	2921	+ve
5.	69751968	*merE*	2	2	0.86,0.94	-1361	-ve
6.	69751971	*merF*	2	2	0.97,0.91	-687	-ve
7.	69751974	*merR*	4	4	0.97, 0.89, 0.89,0.86	865	+ve
8.	69751972	*merP*	2	2	0.97,0.91	-409	-ve
9.	69747978	*merT*	3	3	0.92,0.93.0.89	663	+ve
10.	46432416	*merG*	3	3	0.89,0.94,0.85	2217	+ve

Determination of Common motifs and TFs

The five candidate motifs were predicted and investigated by the MEME algorithm as shown in Table 3. Algorithm-generated the five most promising candidate motifs concerning the ten imported thousand-length gene sequences. The predicted motifs and proportion of promoters containing common motifs for the *mer* operon gene were evaluated. The data show that the best common motifs (motif_1) with the lowest e-values have 100% binding sites. The predicted candidate motifs have the lowest (motif_5) and highest (motif_1) e-values (7.2e-033 and 7.3e-074), respectively. Therefore, the most likely candidate (motif 1) has the highest binding sites compared to the other candidate motifs. As it was presented in Table 3, the two common candidate motifs (motif_2 and motif_3) shared common binding sites and had common motif width by variation in the e-values. It could be hypothesized that these transcription factors activate gene regulatory roles in the bioremediation of environmental pollutants by mercury (II) reductase in the case of the *merA* gene, organomercury lyase (*merB*), mercury transporter genes (*merC*, *merE*, *merF*, and *merT*), transcription regulators (*merR*), and finally mercury-resistant genes (*merF*, *merP*, and *merG*) as revealed in Table 3. The motif patterns in the promoter region, which operates the binding sites of the transcription factors, have enhanced gene regulation [15].

For TSS, we checked distribution from position + 1 of the upstream to position − 1kb (Fig. 1). Using the present analysis, the motif distributions (75% on the positive complement strands) and (25% on the negative complement strands) are presented in Fig. 11. They were distributed at each site according to the transcriptional start site (ATG). Additionally, the data indicates that the dense distribution of the common candidate motif lies around the + 1kb region, while few of them are distributed around the − 100kb region relative location and spatial distribution of these motifs in the promoter regions were constructed by MEME and the created logos of common motifs, resulting in different characteristics of the column's motif orientations, with the height of the letter illustrating how frequently that nucleotide is expected to be observed in that particular position of the two strands (Fig. 2). It has been suggested that the motif, found in a large number of promoter regions, could provide a significant amount of information [16].

Table 3
List of predicted motifs and the number and proportion of promoter-containing motifs

S.N	Predicted and Discovered Candidates motifs	No of the promoter for each of the motifs in %	E-value	Motif width	No of the Binding sites
1.	Motif_1	10 (100%)	7.3e-074	50	10
2.	Motif_2	7 (70%)	1.1e-046	50	7
3.	Motif_3	7 (70%)	2.0e-048	50	7
4.	Motif_4	9 (90%)	1.4e-046	50	9
5.	Motif_5	7 (70%)	7.2e-033	41	7

A candidate common motif with the lowest e-value (7.e-033) represents a statistically significant and functionally significant motif imported into TOMTOM versions 5.4.1 for further analysis (https://meme-suite.org/meme/doc/tomtom-output-format.htmll), which is a publicly available database for transcription factors prediction that could be similar to known regulatory motifs. TOMTOM provides LOGOS representing the alignment of the known motifs with the candidate transcription factors. The TOMTOM output from the databases includes links to the parental TF databases for more information such as activation, repression, and dual regulatory roles of the matched motifs. Again, there was also other conformational information associated with the TF databases such as monomer, dimer, tetramer, and unidentified as well as other factors. The binding types associated with the databases were also predicted. As indicated in Table 3. The motif_5 with the lowest e values (7.2e-033) and statistically significant with 11 matched TF from 84 collected databases with matched e values thresholds less than 10 or less as screened and observed from the TOMTOM database. The forward and reverse strands of the statistically significant strands are depicted in Fig. 1.

Motifs have been revealed to be extremely beneficial in identifying genetic regulatory networks and interpreting specific gene activities. Regulatory motif discovery analysis has advanced significantly attributable to our current computational capabilities, and it remains at the forefront of genomic investigations of bacteria employed in environmental remediation. According to the current studies, the identified candidate motif was widely dispersed between + 1 and − 400bp, sparsely distributed between − 400 and − 800bp, and less distributed above − 800bp as illustrated in Fig. 2. The distribution was on both positive and negative strands, with transcription start sites as a reference. Only one candidate motif was found on the positive complementary strands in the gene identified by gene id (66762507). Approximately 75% and 25% of candidate motifs were located on the positive and negative strands respectively. This indicates the majority of the candidate motif was discovered on the positive strands. The variation of motif distribution that is resulted from the difference in nucleotides sequences of the identified genes.

Identification of transcription factors are essential regulators of gene expression, determining, where, and to what extent genes are expressed in molecular biology. As observed in Table 5, eleven transcriptional factors matching the candidate motif were discovered, each with different regulatory activities. From the commonly identified transcriptional factors four [*PhhR* (90%), *VqsM* (7%), *CcpA* (1%) and *LrP* (1%)] have activation regulatory roles with differences in degrees. This study also revealed that only one *CtrA* (9.09%) and two namely *CRP* and *GlxR* (18.18%) TF identified from *C.crescentus, Y.pestis*, and *C.glutamicumorganismsm* have a dual and repression regulatory functions respectively. The majority of the TFs (*CodY*, *EspR*, *MatP* antoin some extent *VqsM*, *Fur*, *Lrp* as well as *CtrA*) have been found for activation of

transcription for mercuric bioremediation have not yet been described, therefore, additional wet-lab based research might be needed in the future.

Table 4
Lists of matching candidates from EXPREG transcription factor (TF)

S.N	Candidate of TF	Strains showed motif sequence binding	GC (%)	Regulatory Elements				Statistical Significance
				Activation (%)	Repression (%)	Dual (%)	Not specified (%)	
1.	*CRP*	*Y.pestis*	46.88	0	100	0	0	2.11e + 00
2.	*PhhR_*	*P.putida*	46.67	90	10	0	0	2.29e + 00
3.	*VqsM_*	*P.aeruginosa*	59.33	7	0	0	92	3.43e + 00
4.	*CodY*	*B.anthracis*	20.41	0	0	0	100	3.99e + 00
5.	*Fur*	*P.syringae*	40.25	0	13	0	85	4.88e + 00
6.	*EspR*	*M.tuberculosis*	52.83	0	0	0	100	5.95e + 00
7.	*MatP*	*E.coli*	47.23	0	0	0	100	6.75e + 00
8.	*CcpA*	*C.difficile)*	26.32	9	36	0	53	6.87e + 00
9.	*GlxR*	*C.glutamicum*	46.55	0	100	0	0	7.38e + 00
10.	*Lrp*	*E.coli*	40.00	1	1	0	97	7.91e + 00
11.	*CtrA*	*C.crescentus*	28.95	0	0	20	80	9.29e + 00

Transcription factors regulate some sets of gene regulation, and conformational factors and flexibility of genes lead to an effective and selective assembly of co-regulatory proteins to regulate the target genes. This indicates that the transitory interactions between TF and site-specific DNA sequences are common and important in a variety of biological functions. Accordingly, the transcriptional factors confirmation mechanism of eleven *mer* genes employed in mercury bioremediation was studied. According to the current results, no regulatory role has been assigned to the whole set of candidate TF as monomers, tetramers, or other conformational modes as indicated in Table 5. Approximately four of these (*PhhR*, *Fur*, *EspR*, *MatP*, and *Lrp*), discovered TF candidates, have 100% and 96% dimer conformational roles in co-regulating genes respectively. The current investigation revealed that about 54.54% of the identified common candidates for TF conformational mechanisms' function were not identified in Table 5. The conformational flexibility of TF binding proteins maximizes gene regulatory efficiency.

Table 5
Lists of match candidates from EXPREG transcription Confirmation Factor (TCF)

S.N	Candidate of TF	Strains that show motif sequence binding	GC (%)	TF Confirmation Mode				Not Specified (%)	Statistical Significance
				Monomer (%)	Dimer (%)	Tetramer (%)	Other (%)		
1.	*CRP*	*Y.pestis*	46.88	0	0	0	0	100	2.11e + 00
2.	*PhhR*	*P.putida*	46.67	0	100	0	0	0	2.29e + 00
3.	*VqsM*	*P.aeruginosa*	59.33	0	0	0	0	100	3.43e + 00
4.	*CodY*	*B.anthracis*	20.41	0	0	0	0	100	3.99e + 00
5.	*Fur*	*P.syringae*	40.25	0	100	0	0	0	4.88e + 00
6.	*EspR*	*M.tuberculosis*	52.83	0	100	0	0	0	5.95e + 00
7.	*MatP*	*E.coli*	4723	0	100	0	0	0	6.75e + 00
8.	*CcpA*	*C.difficile)*	26.32	0	0	0	0	100	6.87e + 00
9.	*GlxR*	*C.glutamicum*	46.55	0	0	0	0	100	7.38e + 00
10.	*Lrp*	*E.coli*	40.00	0	96	0		3	7.91e + 00
11.	*CtrA*	*C.crescentus*	28.95	0	0	0	0	100	9.29e + 00

CpG islands are DNA methylation sites in promoter regions that are utilized as gene regulation tools by silencing a related gene during transcription. For this study, two algorithms, offline CLC Genome Workbench version 22.0.10 and online database search tools were used. The two regions (promoter and gene body) were analyzed in FASTA format from the upstream of the TSS as well as the whole gene body sequences. Using online database searching tools, the analysis revealed that CpG islands exist in approximately 30% of the gene body and 40% of the promoter regions respectively. The gene body sequences with gene IDs 46432416, 66762507, and 69751970 were among the genes with one CpG island each when compared to other genes. Similarly, 46432416, 69747978, 69751968, and 69751974 had one CpG island of the promoter regions as depicted in Table 6 Further investigations were done offline using CLC Genome Workbench version 22.0.10 to analyze the CpG islands. The restriction enzyme *MspI* was used in the second alternative, which revealed the presence of CpG islands in both promoter regions and gene bodies, as shown in Table 7. As shown in Table 7, the restriction enzyme *MspI* was used to cut fragments between 40 and 220bps in the promoter region rather than the gene body. In general, the nucleotide cutting position of the promoter region was higher than the gene bodies. This indicated that the poorer CpG islands were observed in the gene body than in the promoter regions.

Table 6
CpG islands Identified for both promoter and gene body regions

S.N	Gene ID	Start	End	length	No of the CpG island (s) were found in both regions						
					Gene body	GC%	start	End	Length	Promoter regions	GC%
1.	46432416	8	631	624	1	57	1	974	974	1	66
2.	66762507	1	631	631	1	50	–	–	–	–	–
3.	66762509	–	–	–	–	–	–	–	–	–	–
4.	69747978	–	–	–	–	–	1	971	971	1	50
5.	69747981	–	–	–	–	–	–	–	–	–	–
6.	69751968	–	–	–	–	–	1	978	978	1	63
7.	69751970	1	1639	1639	1	53	–	–	–	–	–
8.	69751971	–	–	–	–	–	–	–	–	–	–
9.	69751972	–	–	–	–	–	–	–	–	–	–
10.	69751974	–	–	–	–	–	1	979	979	1	50

Discussions

Bacterial genomes contain a wide range of genes, each with its function, composition, structure, replication, and transcription, which are used in molecular biology research [17]. Identifying the TSSs from upstream of the gene as well as identifying the promoter region can play an important role in understanding gene regulation mechanisms in microbial cells [18]. Ten common gene sequences used in mercuric bioremediation were retrieved from NCBI databases in March 2022 for the current study. The result showed that the genes encoding mercury bioremediation were different in the TSS [19, 20].

The present study revealed that the distribution of TFs genes encoding mercuric bioremediation was found between + 1bp and − 400bp, as observed in Table 1. Promoter regions were found to share the same patterns of motifs that function as binding sites for transcriptional factors (TF) to facilitate the gene regulation mechanism. If transcription is correctly initiated, the regulatory elements present upstream of the transcribed region are eventually required to determine gene regulation. For the current study, about 11 transcriptional factors that facilitate gene regulation in mercuric bioremediation were investigated and presented in Table 3. *PhhR*, *VqsM*, *CcpA*, and *Lrp* were discovered to be involved in gene regulation activation among the TFs identified using Uniprot data. According to various studies, the transcription analysis of the *PhhR* TF was important for controlling four putative transcriptional units such as *phhA*, *hpd*, *hmgA*, and *dhcA.* According to the current analysis, the transcriptional activation of the PhhR gene in Pseudomonas aeruginosa was responsible for the transcriptional activation of genes for phenylalanine degradation and phenylalanine homeostasis [21].

From the analyzed results, transcriptional factors such as *CcpA*, *GlxR*, and *CRP* were widely used for transcriptional repressions. The current findings were consistent with the catabolic repression mediated by *CcpA* in *B. subtilis* reported by [22], the negative regulation of the *sycO-ypkA*, *ypoJ* operon in *E.coli* by cyclic AMP [23], and the *GlxR* involved in repression of *aceB*, which codes for malate synthase [24]. In the presence of a cAMP binding motif, *GlxR* TF shares common functions with the *CRP* from *E.coli*.

Table 7
MspI cutting sites and fragment sizes both in promoter and gene body regions.

Region	Corresponding sequences	Nucleotide positions of *MSPI* Sites	Fragment between 40 and 220bps
Promoter region	Prom_69751970	8(32, 51, 415, 458, 865, 871, 899, 921)	43
	Prom_66762507	7(8, 329, 345, 364, 545, 584, 747)	181,163
	Prom_66762509	4(183,745,851,907)	106,56
	Prom_69747981	3(175,209,224)	–
	Prom_69751968	5(214, 414, 535, 588, 769)	200,121,53,181
	Prom_69751971	8(32, 51, 415, 458, 865, 871, 899, 921)	43
	Prom_69751974	10(18, 28, 76, 198, 216, 227, 475, 656, 709, 830)	48,122,181,53,121
	Prom_69751972	8(32, 51, 415, 458, 865, 871, 899, 921)	43
	Prom_69747978	3(330, 364, 379)	–
	Prom_46432416	8(23, 33, 467, 473, 523, 684, 756, 856)	50,161,72,100
Gene bodies	ORF 69751970	17(77, 198, 251, 432, 680, 691, 709, 831, 879, 889, 1001, 1043, 1163, 1211, 1238, 1273, 1477)	121,53,181,122,48, 112, 42, 120, 48, 204
	ORF 66762507	3(54, 279, 319)	40
	ORF 66762509	1(403)	–
	ORF 69747981	5(21,38,91,210,337)	53,119,127
	ORF 69751968	4(47,119,131,179)	72,48
	ORF 69751971	1(119)	–
	ORF 69751974	4(50,72,100,106)	–
	ORF 69751972	1(85)	–
	ORF 69747978	3(17,210,232)	193
	ORF 46432416	1 (636)	–

The *MspI* restriction enzyme was used to search for CpG islands in both the promoter and gene body regions in Table 7. The promoter region sequences, Prom_69751970, Prom_69751971, and Prom_69751972 had the same *MspI* cleavage sites and fragments length in the current analysis. However, the location of the TSS of each promoter sequence was different. The highest and lowest cutting sites of *MspI* were found in Prom-69751974 and Prom_6974798, respectively. In the gene body region, the highest and lowest *MspI* cutting sites were represented ORF_69751970 and ORF_66762509, ORF_69751972, ORF_46432416, and ORF_69751971 respectively. The results of the *MspI* restriction enzyme digestion revealed that the promoter region had more CpG islands than its counterpart as seen in Table 7. The current finding agreed with the finding of gene expression in the promoter-associated CpG islands in human methylome reported by [25].

The *mer* genome consists of ten essential *mer* gene clusters that play an imperative function in mercuric bioremediation. The mainstream of the *mer* gene sequences found in bacterial strains belongs to gammaproteobacteria, followed by alphaproteobacteria. Those gene groups were also discovered in beta proteobacteria, firmicutes, and

actinobacteria to varying degrees. Each group of *mer* genomes performs a specific function. One of the major applications of *merA* was in reducing mercury from Hg^{2+} to Hg^{0}, a process widely used in bioremediation. While *merB* and *merC*, *merT* were important for organomercurial lyase and transporter respectively. On the other hand, *merB* and *merE* were broad-spectrum *mer* operons found in both gram-positive and gram-negative bacteria used in mercuric bioremediation. The *merD* and *merR* were among the *mer* gene clusters used in transcriptional regulation and co-regulation in the mercuric resistance in bioremediation respectively, as depicted in Table 7. The *mer* genomes and their cluster genes, in general, have played a crucial role in the current scenario of environmental contamination control mechanisms. This study was in agreement with the study conducted on biogeochemistry and bioremediation of mercury by bacteria [26].

Conclusions

The current investigation and characterizations of promoter regions of the *mer* genome and its gene clusters encoding mercuric heavy metal resistance as a means of mercuric bioremediations is very important for understating the regulatory elements and control of its expression. The current finding revealed that eleven transcriptional factors and their conformational mode identified in the promoter region of the *mer* operon gene clusters could play a major application in the heavy metal bioremediation such as mercury. By contributing to improving environmental concerns caused by global climate change, the current study contributes to improving the environment. However, additional experimental studies will be required to confirm the role of the identified TF and their shared binding locations in the regulation of the *mer* gene encoding for heavy metal bioremediation by using advanced bioinformatics tools to improve the effectiveness of the *mer* gene clusters.

Abbreviations

TSS	Transcriptional start site
NNPP	Neural network promoter prediction
TFs	Transcriptional factors
CpG	Cytosine phosphate guanine
NCBI	National Center of Biotechnology Information

Declarations

Acknowledgment

Not Applicable

Authors' contributions

DD designed, performed the experiment, analyzed data, and wrote the manuscript. GW analyzed data and edited the manuscript. The authors have read and approved the final manuscript.

Funding

The authors didn't receive any funds for this study.

Availability of data and materials

The datasets used in this study can be obtained from the corresponding author.

Declarations

Ethical approval and consent of participate

Not applicable

Consent of publication

Not applicable

Competing of interest

The authors declare that they have no competing interests.

Authors' information

[1]School of Biological Sciences and Biotechnology, Haramaya University, Dire Dawa, Ethiopia [2]Department of Animal Science, Oda Bultum University, Chiro, Ethiopia

References

1. Danan Gu, Kirill Andreev MED. Major Trends in Population Growth Around the World. China CDC Wkly. 2021;3(28):604–613. doi:10.46234/ccdcw2021.160
2. Estrada A, Garber PA, Chaudhary A. Current, and future trends in socio-economic, demographic, and governance factors affecting global primate conservation. PeerJ. 2020;8(e9816):1–35. doi:10.7717/peerj.9816
3. Emenike CU, Jayanthi B, Agamuthu P, Fauziah SH. Biotransformation and removal of heavy metals: A review of phytoremediation and microbial remediation assessment on contaminated soil. Environ Rev. 2018;26(2):156–168. doi:10.1139/er-2017-0045
4. Ali H, Khan E, Ilahi I. Environmental chemistry and ecotoxicology of hazardous heavy metals: Environmental persistence, toxicity, and bioaccumulation. J Chem. 2019;2019(Cd). doi:10.1155/2019/6730305
5. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and Health Impacts of Air Pollution: A Review. Front Public Heal. 2020;8:1–13. doi:10.3389/fpubh.2020.00014
6. Getaneh W, Alemayehu T. Metal contamination of the environment by placer and primary gold mining in the Adola region of southern Ethiopia. Environ Geol. 2006;50(3):339–352. doi:10.1007/s00254-006-0213-5
7. Rehan M, S. Alsohim A. Bioremediation of Heavy Metals. In: Environmental Chemistry and Recent Pollution Control Approaches.; 2019:145–158. doi:10.5772/intechopen.88339
8. Diep P, Mahadevan R, Yakunin AF. Heavy metal removal by bioaccumulation using genetically engineered microorganisms. Front Bioeng Biotechnol. 2018;6(OCT). doi:10.3389/fbioe.2018.00157
9. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012;13(4):233–245. doi:10.1038/nrg3163
10. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. AAAI Press. 1994;2:28–36.
11. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. 2015;43(May):39–49. doi:10.1093/nar/gkv416
12. Bailey TL, Boden M, Buske FA, et al. MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. 2009;37(SUPPL. 2):1–7. doi:10.1093/nar/gkp335

13. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A. 2002;99(6):3740–3745. doi:10.1073/pnas.052410099
14. Mendoza-Vargas A, Olvera L, Olvera M, et al. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. PLoS One. 2009;4(10):1–19. doi:10.1371/journal.pone.0007526
15. Mahdi RN, Rouchka EC. RBF-TSS: Identification of transcription start site in humans using radial basis functions network and oligonucleotide positional frequencies. PLoS One. 2009;4(3):1–6. doi:10.1371/journal.pone.0004878
16. Michalowski JS, Galante PAF, Nagai MH, et al. Common promoter elements in odorant and vomeronasal receptor genes. PLoS One. 2011;6(12):1–10. doi:10.1371/journal.pone.0029065
17. Koonin E V., Wolf YI. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36(21):6688–6719. doi:10.1093/nar/gkn668
18. Dinka H and AM. Infection, Genetics and Evolution Unfolding SARS-CoV-2 viral genome to understand its gene expression regulation. 2020;84(March):1–6. doi:10.1016/j.meegid.2020.104386
19. Bantihun G, Kebede M. In silico analysis of promoter region and regulatory elements of mitogenome co-expressed trn gene clusters encoding for bio-pesticide in entomopathogenic fungus, Metarhizium anisopliae: strain ME1. J Genet Eng Biotechnol. 2021;19(1):1–11. doi:10.1186/s43141-021-00191-6
20. Aman Beshir J, Kebede M. In silico analysis of promoter regions and regulatory elements (motifs and CpG islands) of the genes encoding for alcohol production in Saccharomyces cerevisiae S288C and Schizosaccharomyces pombe 972h-. J Genet Eng Biotechnol. 2021;19(1). doi:10.1186/s43141-020-00097-9
21. Palmer GC, Palmer KL, Jorth PA, Whiteley M. Characterization of the Pseudomonas aeruginosa transcriptional response to phenylalanine and tyrosine. J Bacteriol. 2010;192(11):2722–2728. doi:10.1128/JB.00112-10
22. Moreno MS, Schneider BL, Maile R, Weyler W, Jr MHS. Catabolite repression mediated by the CcpA protein in Bacillus subtilis: novel modes of regulation revealed by whole-genome analyses. Arch Microbiol. 2001;39(5):1366–1381.
23. Zhan L, Yang L, Zhou L, et al. cyclic AMP receptor protein (CRP) in Yersinia pestis. BMC Microbiol. 2009;9:1–9. doi:10.1186/1471-2180-9-178
24. Letek M, Valbuena N, Ramos A, Gil A. Characterization and Use of Catabolite-Repressed Promoters from Gluconate Genes in Corynebacterium glutamicum †. J Bacteriol. 2006;188(2):409–423. doi:10.1128/JB.188.2.409
25. Du X, Han L, Guo AY, Zhao Z. Features of methylation and gene expression in the promoter-associated CpG islands using human methylome data. *Comp Funct Genomics*. 2012;2012. doi:10.1155/2012/598987
26. Priyadarshan M, Chatterjee S, Rath S, Dash HR, Das S. Cellular and genetic mechanism of bacterial mercury resistance and their role in biogeochemistry and bioremediation. J Hazard Mater. 2022;423:1–20. doi:10.1016/j.jhazmat.2021.126985

Figures

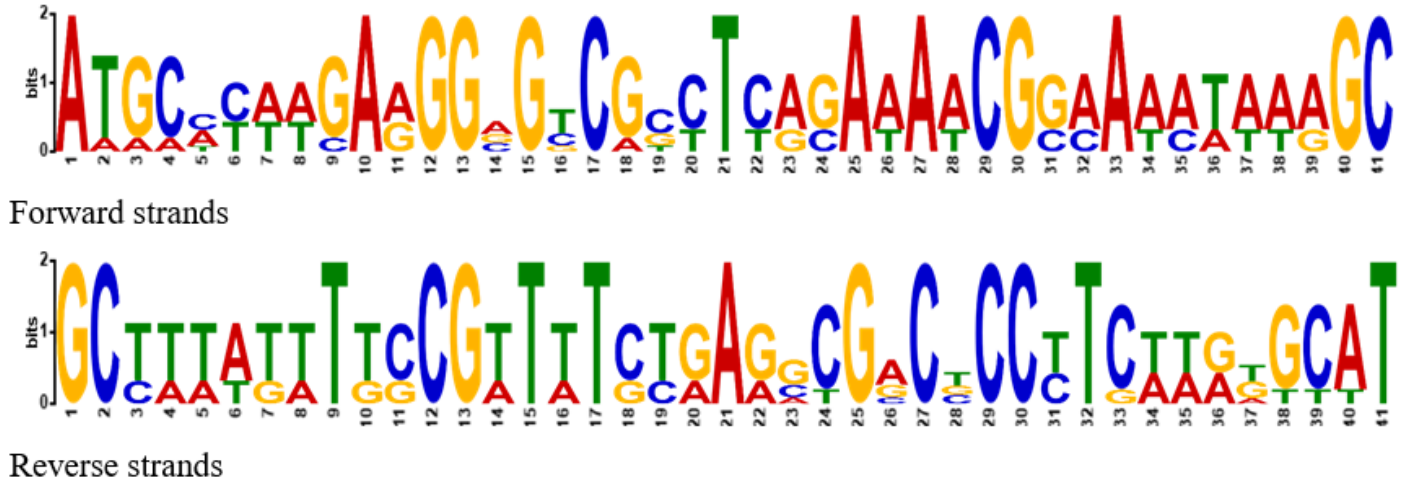

Figure 1

Sequence logos for the best mercuric bioremediation motif promoter regions

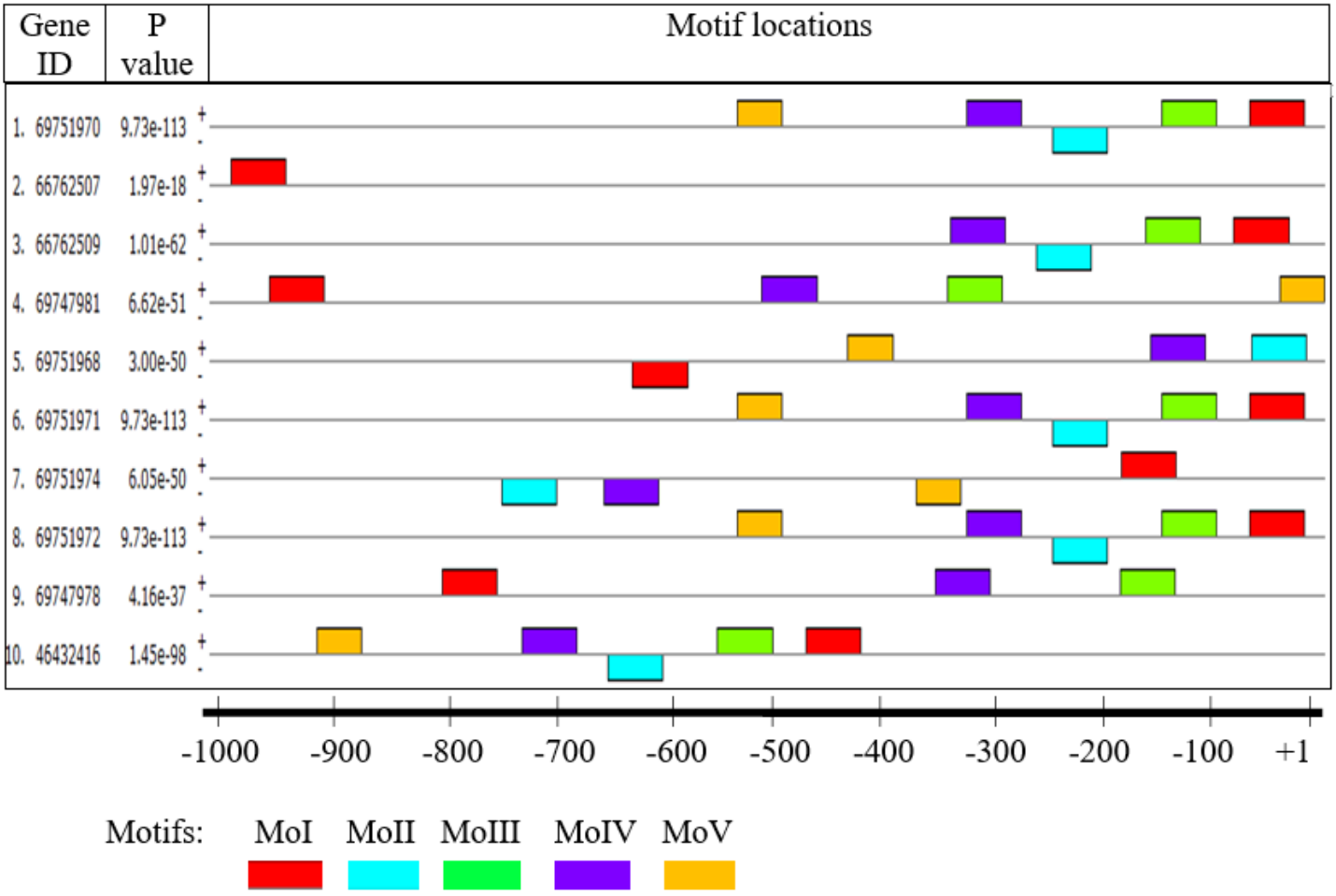

Figure 2

The relative locations of potential motifs in the promoter region relative to TSSs are illustrated in block diagrams. The nucleotide locations in the promoter region for *mer* genes encoding for mercury bioremediation are indicated at the bottom of the graph, ranging from +1 (start of TSSs) to upstream 1kb.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- WholeGenebodysequencedata.txt
- wholepromoterregionsequence.txt