BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net **Hypothesis**

Volume 8(4)

An approach to delineate primers for a group of poorly conserved sequences incorporating the common motif region

Mousumi Sahu, Jagajjit Sahu, Smita Sahoo, Budheswar Dehury, Kishore Sarma, Ranjan Sarmah, Priyabrata Sen, Mahendra Kumar Modi, Madhumita Barooah*

Agri-Bioinformatics Promotion Programme, Department of Agricultural Biotechnology, Assam Agricultural University, Jorhat-785013,Assam, India; Madhumita Barooah – Email: m17barooah@yahoo.co.in; Phone: +91-(376)-2340001 (O); FAX: (376)-2340001, 2340101; *Corresponding author

Received February 09, 2012; Accepted February 11, 2012; Published February 28, 2012

Abstract:
Glutathione synthetase (*gshB*) has previously been reported to confer tolerance to acidic soil condition in *Rhizobium* species. Cloning the gene coding for this enzyme necessitates the designing of proper primer sets which in turn depends on the identification of high quality sequence similarity in multiple global alignments. In this experiment, a group of homologous gene sequences related to *gshB* gene (accession no: *gi-86355669:327589-328536*) of *Rhizobium etli* CFN 42, were extracted from NCBI nucleotide sequence databases using BLASTN and were analyzed for designing degenerate primers. However, the T-coffee multiple global alignment results did not show any block of conserved region for the above sequence set to design the primers. Therefore, we attempted to identify the location of common motif region based on multiple local alignments employing the MEME algorithm supported with MAST and Primer3. The results revealed some common motif regions that enabled us to design the primer sets for related *gshB gene* sequences. The result will be validated in wet lab.

Keywords: Glutathione synthetase, Rhizobium etli CFN 42, Multiple Global Alignments, T-coffee, Degenerate Primers, Multiple Local Alignments, MEME, Primer3, MAST.

Background:

Classical methods for degenerate primer design include software applications such as Gene Fisher **[1]**, CODEHOP **[2]** or PrimaClade **[3].** These methods usually rely on the identification of clear blocks of conserved regions in multiple global alignments. Therefore, alignment quality should be very high among the sequences in order to utilize the software. However, in many cases, primer design using multiple global alignments is unsuccessful **[4],** due to poor conservation among the sequences under study. Here a method is analyzed which suggests the common motif region could be used to design degenerate of primers for sequences with poor global similarity or without well-conserved blocks in multiple global alignments. The technique is based on multiple local alignments, using the Multiple Expectation – Maximization for Motif Elicitation (MEME) algorithm **[5]**, in order to search for conserved regions long enough to serve as primers. The results of MEME were then compared with Primer3 **[6]** result. Using MAST **[7]** two well conserved motifs sites were generated from the common motif regions. The potential primer properties of these motifs will be verified further in wet lab.

We employed this technique to design primer from the common motif regions for the sequences related to glutathione synthetase gene (*gshB*) of Rhizobium etli CFN 42. We demonstrate that the amplification of sequences with similarity to genes reported in other species (possibly homologues) is

© 2012 Biomedical Informatics

possible, even when the species in question is poorly characterized at the molecular level.

Methodology:

Employing BLASTN **[8]** homologous sequences of the *gshB* gene (Accession number: *gi|86355669:327589-328536)* of *Rhizobium etli* CFN 42, were extracted from nucleotide sequence databases of NCBI (http://www.ncbi.nlm.nih.gov) **[9]** and are summarized in **Table 1 (see supplementary material).** One database of collected nucleotide sequences was constructed.

Original query (*gshB gene*) and its group of related sequences were aligned using T-COFFEE **[10]** to determine global similarity. A lack of block conservation was observed in all cases in the multiple global alignments which discarded this technique to design degenerate primers for considered sequences. Then multiple local alignments were performed using MEME to generate common motif regions for the sequences. The conditions of these alignments included a minimum motif length of 18 and a maximum of 50, and minimum sites to find was two. Then two motifs were generated using MAST by considering and combining evidences of all p-values associated with motifs resulted by MEME run.Primer3 was employed to generate primers for every individual sequence, which were then compared with the common motif portions of that sequence **Table 2 (see supplementary material)** suggested by the first MEME result. Again another MEME run was conducted to detect motifs of minimum width 18 and maximum width 22 (criteria of good primer) with all the previous parameters remaining unchanged for the similar sequence set and two motifs were generated employing MAST.

Discussion:

To design degenerate primer for this group of sequences, it was necessary to predict block of conserved regions present in the sequences, which indeed needed sequence alignment study of the sequence set. Therefore, a multiple global alignment algorithm of the T-coffee software was performed for the set of *gshB* gene homologous sequences. From the result of T-coffee no significant block conservation was detected. So the idea of designing of degenerate primer from a simple multiple global alignments of these sequences was dropped. In order to predict common motif regions for the sequences, an attempt was made for a multiple local alignment using MEME. From the MEME result common arrangement of motifs among the analyzed sequences were detected which proved the existence of two well-conserved regions among these sequences. Then combing evidence of all the p-value of the motifs, two motifs were generated using MAST server. The predicted common motifs were *(1) AAGATCTTCGTCACCGAATTTCCCGATCTGA TGCCGAAGAC (+), GTCTTCGGCATCAGATCGGGAAATTCGGTGACGAAGATCTT (-) and (2) ACGTGATCGGCGATTACATGACCGAGATCAACGTCAC (+), GTGACGTTGATCTCGGTC ATGTAATCGCCGATCACGT (-)* Then an attempt was made to predict primers form each sequence using Primer3. From the first MEME run and Primer3 test it was detected that sequences having accession number gi|86355669: 327589-328536, gb|CP000133.1|, gb|CP001191.1|, gi|115254414, gb|CP000628.1|, gb|CP001389.1|, gb|CP000738.1, gb|CP000781.1|, gi|288909149, gb|CP001029.1|, and gb|CP001510.1| shared same site of common motif regions with primers. It indicates that in many sequences, primer can be generated from these common motif regions **Table 2 (see supplementary material).**

The resulted motifs of the first MEME run were too long to consider or analyze as primers. Therefore, another MEME run was conducted to detect motifs of minimum width 18 and maximum width 22 (criteria of primer) with all the previous parameters remaining unchanged in this test.In this study, more than half of the results exhibited a common arrangement of motifs between the analyzed sequences, indicating a global similarity that could not be well resolved with common algorithms of multiple global alignments. This indicates the existence of well-conserved regions and similarity in the analyzed sequences, which can be useful for primer design.

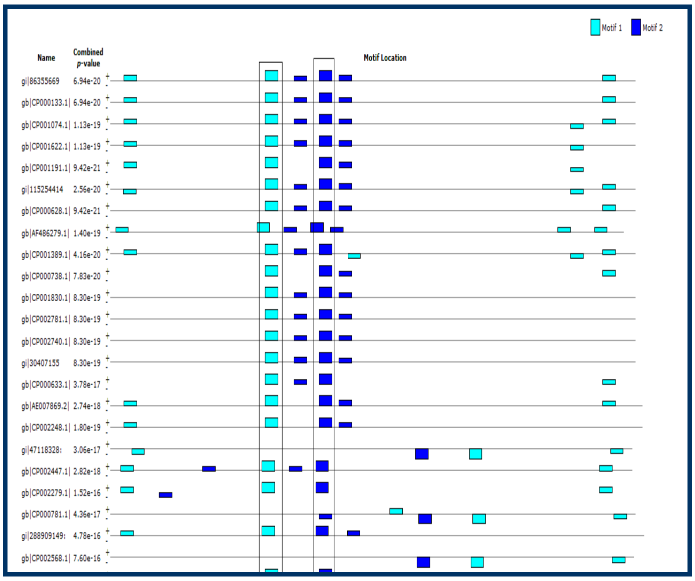

Figure 1: Second MEME run showing combined block diagram for all motifs with corresponding GENE ID and combined P-value

The second MEME run predicted two distinct similar regions of motifs **(Figure 1).** Then two specific motifs were generated employing MAST algorithm from the resulted motifs of the second MEME run. The motifs were (1) TTCGACATGGCCTATATCACCT ***(+),*** AGGTGAT ATAGGCCATGTCGAA ***(-)*** *and* (2) AAAAGATCTTCGTCACCGAATT ***(+),*** AATTCGGTG ACGAAGATCTTTT ***(-).*** The pair will be further verified in wet lab and with other in-silico primer analysis techniques. If these resulted primers will fail to be applied for any sequence, the motifs of individual sequence obtained in second MEME run can be extracted and ordered according to the score given in the program. Then by pairing those in all possible ways, pair of primers may be generated and analyzed.

Conclusion:

Our analysis attempts to design degenerate primers from common motif regions of a group of less conserved homologous gene sequences related to *gshB* gene of *Rhizobium etli* CFN 42. These sequences have a low sequence similarity in multiple global alignments. So the applied technique in this present study can be used for sequences those have very few known homologues, or to confirm them, and to design degenerate primers when the classical methods do not work. However, the experimented method in this study needs to be improved, for instance to reduce the time-consuming steps and avoiding sequences with low similarity patterns. It is also needed further investigation to apply it in the gene coding analysis study as this technique may deform the functional region of the gene in order to predict a common motif region. Therefore the predicted technique is needed to be validated in wet lab.

References:

[1] Giegerich R *et al. Syst Mol Biol.* 1996 **4**: 68 [PMID: 8877506].
[2] Rose TM *et al. Nucleic Acids Res.* 2003 31:3763 [12824413].
[3] Gadberry MD *et al. Bioinformatics* 2005 **7**: 1263 [PMID: 15539448]
[4] Kwok S *et al. PCR Methods Appl.* 1994 **4**: S39 [PMID: 8173508].
[5] Bailey TL *et al. Nucleic Acids Res.* 2006 **34**: W369 [PMID: 16845028].
[6] Rozen S & Skaletsky HJ, *Methods Mol Biol.* 2000 **132**: 365 [PMID: 10547847].
[7] Bailey TL & Gribskov M, *Bioinformatics.* 1998 **1**: 48 [PMID: 9520501].
[8] Altschul SF *et al. J. Mol. Biol.*1990 **215**: 403 [PMID: 2231712].
[9] http://www.ncbi.nlm.nih.gov/gene/3890970.
[10] Notredame C *et al. J Mol Biol.* 2008 **1**: 205[PMID: 10964570].

.

Edited by P Kangueane

Citation: Sahu *et al*. Bioinformation 8(4): 181-184 (2012)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Supplementary material:

Table 1: Showing homologous sequences of gshB gene of Rhizobium etli CFN 42 with respective accession number, organism name, query coverage and E value.

SL. NO.	Accession	Organism	Query coverage	E value
1	CP000133.1	Rhizobium etli CFN 42	100%	0.0
2	CP001074.1	Rhizobium etli CIAT 652	100%	0.0
3	CP001622.1	Rhizobium leguminosarum bv. Trifolii WSM1325	100%	0.0
4	CP001191.1	Rhizobium leguminosarum bv. Trifolii WSM2304	100%	0.0
5	AM236080.1	Rhizobium leguminosarum bv. Viciae	100%	0.0
6	CP000628.1	Agrobacterium radiobacter K84	99%	0.0
7	AF486279.1	Rhizobium tropici glutathione synthetase gene	97%	0.0
8	CP001389.1	Sinorhizobium fredii NGR234	100%	0.0
9	CP000738.1	Sinorhizobium medicae WSM419	96%	0.0
10	CP001830.1	Sinorhizobium meliloti SM11	96%	0.0
11	CP002781.1	Sinorhizobium meliloti AK83	96%	0.0
12	CP002740.1	Sinorhizobium meliloti BL225C	96%	0.0
13	AL591688.1	Sinorhizobium meliloti 1021	96%	0.0
14	CP000633.1	Agrobacterium vitis S4	98%	0.0
15	AE007869.2	Agrobacterium tumefaciens str. C58	98%	0.0
16	CP002248.1	Agrobacterium sp. H13-3	98%	0.0
17	BA000012.4	Mesorhizobium loti MAFF303099 DNA	90%	0.0
18	CP002447.1	Mesorhizobium ciceri biovar biserrulae WSM1271	90%	3e-179
19	CP002279.1	Mesorhizobium opportunistum WSM2075,	90%	1e-178
20	CP000390.1	Chelativorans sp. BNC1,	75%	1e-102
21	CP000781.1	Xanthobacter autotrophicus Py2	96%	8e-90
22	AP010946.1	Azospirillum sp. B510 DNA,	67%	5e-87
23	CP002568.1	Polymorphum gilvum SL003B-26A1	96%	9e-75
24	CP000699.1	Sphingomonas wittichii RW1	72%	9e-75
25	CP001349.1	Methylobacterium nodulans ORS 2060	96%	7e-71
26	CP000463.1	Rhodopseudomonas palustris BisA53	77%	7e-61
27	CP000394.1	Granulibacter bethesdensis CGDNIH1	67%	2e-46
28	CP000908.1	Methylobacterium extorquens PA1	72%	7e-46
29	CP001298.1	Methylobacterium chloromethanicum CM4	72%	2e-42
30	CP001029.1	Methylobacterium populi BJ001	72%	7e-41
31	CP001510.1	Methylobacterium extorquens AM1	72%	7e-36

Table 2: Showing sequence accession number and associated primers at common motif sites resulted by first MEME run

SL. NO	ACCESSION NO.	LEFT PRIMER	START	RIGHT PRIMER	START	MOTIF SITES
1	gi|86355669:327589-328536	TTTTCGTCACCGAATTTTCC	383			331-371, 379-419, 449-485, 821-857
2	gb|CP000133.1|	TTTTCGTCACCGAATTTTCC	383			331-371, 379-419, 449-485, 821-857
3	gb|CP001191.1|	ACGCTGGTAGTCAACGATCC	334			218-254(-), 331-371, 379-419, 821-857
4	gi|115254414			GGAGAATTCGGTGACGAAGA	402	331-371,-379-419,-449-485,-821-857
5	gb|CP000628.1|	CGGAAAAGATCTTCGTCACC	374			76-116(-), 331-371,379-419,449-485,821-857
6	gb|CP001389.1|	CGGAAAAGATCTTCGTCACC	374			161-197, 331-371, 379-419, 821-857
7	gb|CP000738.1|	ATCTTCGTGACCGAATTTGC	382			161-197, 331-371, 379-419, 821-857
8	gb|CP000781.1|			TTCGGTTCGCTCTACGATCT	411	215-251, 325-365, 373-413, 815-851
9	gi|288909149			GGCAGGTATTTCTGCACGAT	605	98-134(-), 467-507(-),536-576(-), 584-624(-),626-662(-)
10	gb|CP001029.1|	AGCTGTTCGTCCTCGACTTC	374			373-413, 625-665, 815-851
11	gb|CP001510.1|			CGGAGAACAGGTCGAACATC	565	373-413, 559-599(-), 625-665, 815-851