Identification Direct Repeat Dyad Symmetrical PhoB Using (PHA) family Bacteria

Mukesh Chandra, Dept of Biotechnology IET, mtechmukesh@gmail.com
Atul Katiyar Department of Biotechnology IET Lucknow atulietbio@gmail.com
Krishna Kumar Verma Dept of Biotechnology IET Lucknow kisan_lifesc@yahoo.co.in
Dr. B.N. Mishra (Prof.) Dept of Bio-information IET Lucknow probnmishra@rediffmail.com
Dr. Amod Tiwari Assist Prof, Dept Computer Science & Engineering PSIT Kanpur amodtiwari@gmail.com

Abstract

The present work is an attempt in the direction of Bio-information and it describe in genome as wide identification of direct repeat type dyad symmetrical regulatory motif or PhoB (pho box) in polyhydroxyalkanoic acid (PHA) family bacteria. PHA are a family of biopolymers (polyester) synthesized by a wide range of eubacteria. Due to potential commercial exploitation as biodegradable plastics & packaging material have attracted significant attention. The best studied PHA is poly-beta-hydroxybutyrate (PHB).

Keywords: *Symmetrical regulatory, PhoB, Polyhydroxyalkanoic acid, Poly-beta-hydroxybutyrate*

1. Introduction

The recent increase in the number of microbial sequenced genomes and the amount of genome scale experimental expression data allows the use of computational algorithms to investigate regulatory DNA motifs responsible for gene or set of genes regulation through weight matrix method *Roulet* [1] .A regulatory region could be characterized by the presence of a TFBS that typically varies from 6 to 25 bp in length. In parallel with the experimental wet lab work computational analysis were undertaken to explain the amount of information necessary for a gene regulatory system, thus save the time & wet lab economy too *Stormo* [2]. Identification of such binding sites is not only relevant for locating the promoter of a gene. But they may also allow the prediction of specific regulated gene expression pattern and responsiveness to known biological signaling pathways *Staden* [3]. Now statistical mechanical theory has been already established for TFBS prediction through weight matrix based approach, which at present remains dominant because position specific scoring matrices offer a sensitive way to represent the specificity of trans-factor and DNA interfaces.

The present work is an attempt in the same direction and it describes in silico genome wide identification of direct repeat type dyad symmetrical regulatory motif or PhoB TFBS (*pho-box*) in polyhydroxyalkanoic acids (PHA) family bacteria. PHAs are a family of biopolymers (polyesters) synthesized by a wide range of eubacteria due to potential commercial exploitation as biodegradable plastics & packaging material PHAs have attracted significant attention. In most organisms, PHB is synthesized under conditions of nutrient limitation in the presence of an excess carbon and energy source via a 3-step pathway which involves the condensation of two molecules of acetyl coenzyme-A (acetyl-CoA) to acetoacetyl-CoA via a β-ketothiolase (PhaA or PhbA), reeducation of acetoacetyl-CoA-CoA-CoA-CoA to β-hydroxybutyrate-CoA via NADPH-dependent acetoacetyl-CoA reductase (PhaB or PhbB) and polymerization to PHB via a PhA synthase (PhaC or PhbC). In alcaligenes eutrophus, all three PHA biosynthetic enzymes are synthesized constitutively. The enzyme level regulation is achieved as a result of inhibition of the β-ketothiolase (PhaA) by free coenzyme-A. Since most of studied bacteria have not detection of more such PhoB regulated PHB biosynthetic pathways and their regulatory network through PFM method is expected in different bacteria. In addition, it has also been found that many species sequenced so far possess multiple pho-box loci which are due to the existence of evolutionary conserved 7bp repeat unit of pho-box sequence. Evidence of repeat units also revealed accuracy of our predictions. In the present study, we first statistically evaluated the prediction accuracy of all matrices on known test data set (both true positive & true negative) of E. coli promoters through studied PhoB matrices constructed with the help of different algorithms based existing web tools for motif discovery viz. (i) Consensus algorithm based tool 'CONSENSUS' & (ii) Gibbs Sampler algorithm based tool 'GIBBS SAMPLER' at RSAT webserver *Helden* [4]

2. About Poly (3-hydroxyalkanoates) (PHAs)

Polies (3-hydroxyalkanoates) (PHAs) are a class of microbially produced polyesters that have potential applications as conventional plastics, specifically thermoplastic elastomers. A wealth of biological diversity in PHA formation exists, with at least 100 different PHA constituents and at least five different dedicated PHA biosynthetic pathways. The many different PHAs that have been identified to date are primarily linear; head-to-tail polyesters composed of 3-hydroxy fatty acid monomers. In these polymers, the carboxyl group of one monomer forms an ester bond with the hydroxyl group of the neighboring monomer. Two types of PHAs are:

(i) Short-side-chain PHAs (ssc-PHAs) (ii) Medium-side-chain PHAs (msc-PHAs)

2.1 PHA biosynthesis pathways

The loci encoding the genes for PHA formation have been characterized from 18 different species so far and we are looking to identify similar metabolic pathways in rest completely sequenced bacteria. Genes specifying enzymes for ssc-PHA formation are designated phb and those specifying enzymes for msc-PHA formation are designated PHA. Not all pathways have completely been elucidated in these strains. The emerging picture is that pha and phb genes are not necessarily clustered and that the gene organization varies from species to species. Other genes possibly related to PHA metabolism may be linked to the essential PHA and phb genes. Genes organization can be classified as: (i) complete phb CAB operands, (ii) interrupted phb loci, (iii) incomplete phb loci, (iv)phb loci from organisms that encode two subunit P(3HB) polymerases, (v) the phbCJ locus of A. *caviae* involved in P(3HB-3HH) formation and (vi) pha loci for msc-PHA formation in pseudomonas *Lawrence*[5] PHA can be classified into:

(i) Pathways for short side chain (ssc)-PHA or PHB Formation
  - (i) PHB Biosynthetic Pathway (most commonly occurring pathway)
  - (ii) PHA synthesis with an enoyl-CoA hydrates
  - (iii) P(3HB-3HV) formation from sugars by the methylmalonyl-CoA pathway

(ii) Pathways for medium side chain (msc)-PHA Formation
  - (i) Msc-PHAs from fatty acids (ii) Msc-PHAs from carbohydrates

2.3 PHB biosynthetic pathway

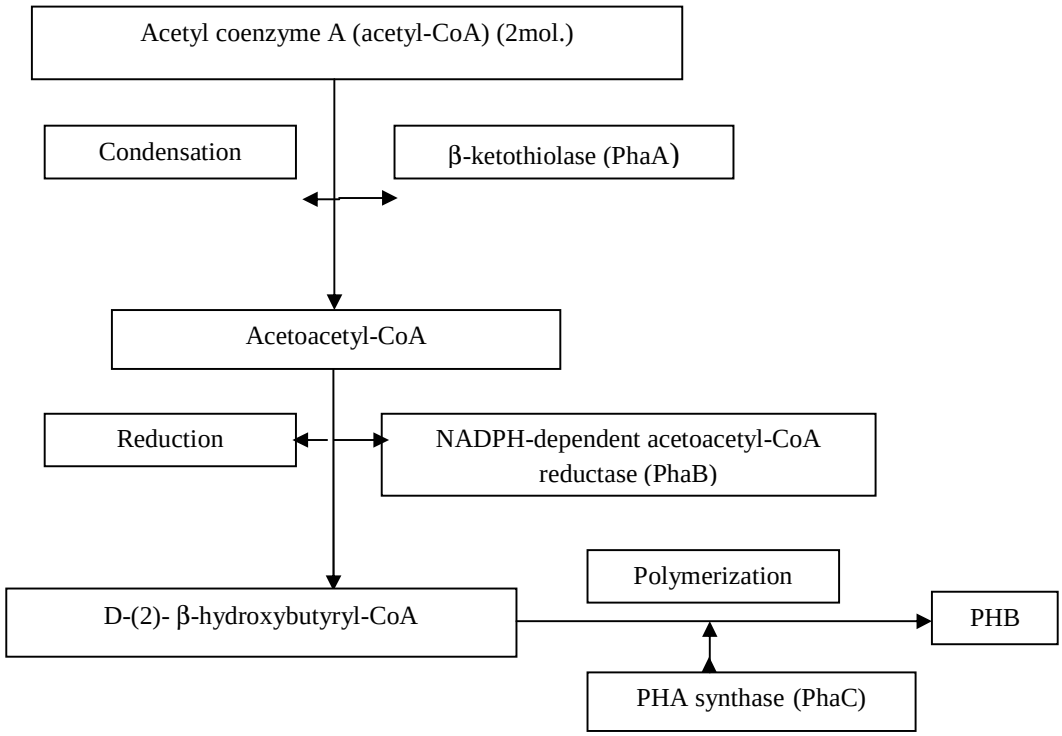


Figure1: Three step biosynthetic pathway of PHB

3. Material and Method

A data set of known 8 promoters namely, phnC, phoA, phoB, phoE, phoH, pstS, asr, ugpB of E.coli and a single phbB promoter (Acc. No. L37761.1, GI: 576782, Locus ACCPHAABC with reported length-1 to-110 bp) of Acinetobacter species *Benson* [6] were used to derived new PhoB matrix. However, information of experimentally known pho-box motifs for phoB and related promoter sequences were compiled from the biological literature and were retrieved from RegulonDB; a database on transcriptional regulation and operon organization in E. coli and GenBank at NCBI webserver

3.1 Data set of orthologous phbB promoters

A data set of 31 orthologous genes upstream sequences of 23 bacterial genomes were selected from bacterial upstream sequence database at RSAT webserver and MBGD *Uchiyama*[7] Selected PhbB product orthologs were verified further through COG database Out of studied 190 bacteria only 21 orthologous species were selected while 2 additional species were detected through BLAST search analysis. The genomes of bacteria used in this study (abbreviations used are given in parenthesis) were: coulobacter crescentus (ccr), Bradyrhizobium japonicum (bja), Rhodopseudomonas palustris (rpa), Mesorhizobium loti (mlo), Agrobacterium tumefaciencs (atc), Sinorhizobium meliloti (sme), Rickettsia conorii (rco)

3.2 Construction of PhoB matrix

A data set of 9 promoters was used for construction of modified phoB matrix through programs viz. CONSENSUS &GIBBS SAMPLER at RSAT and MEME. Data set includes formally known 8 E. coli promoters namely phoA, phoB, phoE, phoH, pstS, asr and ugpB with 400 bp upstream promoter sequences and a single phbB promoter of Acinetobacter sp. Encoding acetoacetyl-CoA reductase. Finally the 18 bp conserved pattern of the aligned sequences was converted into a table of base frequencies from which the position specific weights were calculated according to PATSER algorithm *Helden* [8]

3.3 Statistical Validation

Receiver operator characteristic (ROC) plot analysis was used to further establish optimal thresholds and identify parameters that best predicted true positives. ROC curves were generated by plotting the sensitivity for predicted true positives versus one minus the specificity (1-Sp) for predicted false positives ratio.

- The performance measurements used in this paper are defined as Equation
- Sensitivity (sn) = TP/ (TP +FN)
- Specificity (SP) = TN/ (TN + FP)
- False positive ratio = (1-Sp)

Where TP is the number of true positives (experimentally verified TFBSs which are also predicted as TFBSs), TN is the number of true negatives (experimentally verified no-TFBSs, predicted as non- TFBSs), FP the number of false positives (experimentally verified non- TFBSs, predicted as TFBSs) and FN is the number of false negatives (experimentally verified TFBSs, predicted as non-TFBSs).

3.4 Matrix based pho-box prediction

To identify regulated phbB genes, the orthologous phbB promoter data set was analyzed for detection of high-scoring matches to the known pho-box using matrix matching tool PATSER implemented at RSAT.

4. Identification of PhoB orthologs

The evidence of PhoB was verified by the help of BLAST tool at NCBI and orthologous protein clustering tool at MBGD. Similar conserved protein domains were detected in all as identified by CDD search tool *Marchler*[9] Moreover, detected orthologous PhoB proteins were multiple aligned and then phylogenetic tree was studied for evolutionary relationship evidence. Multiple alignments were done through ClustalW and the phylogenetic unrooted tree was generated by DRAWTREE program of phylip package

4.1 Calculation of motif similarity

To measure the prediction accuracy of PhoB Consensus matrix , predicted motifs were analyzed in terms of relative sequence similarity percentage was calculated as follows For known PhoB TFBSs,
Maximum score (S_{max}) = 17.71

Minimum score (S_{min}) = -39.98
Range of score = Maximum score- Minimum score
Rknown = (S_{max}- (S_{min})= 17.707-(-39.980) = 57.687
Range of predicted score (predicted motif) = predicted score ([redacted, ptof)-Minimum score (know PhoB TFBSs)
$R_{pred}=S_{pred}-S_{min}$
Motif Similarity Percentage (P %) = (Range of predicted score/Range of score) x100 p%= (R_{pred}/R_{known}) x 100%

4.1 Diagrammatic representation of conserved motif (Logo)

The evidence of pho-box motif conservation was diagrammatically represented as logo by the help of Weblogo *Crooks* [10] (http//wblogo.berkeley.edu/). Comparison of known & predicted sequence logo of pho-box motif (18 bp) derived through RegulonDB, MEME, CONSENSUS & GIBBS SAMPLER programs (e). in the figure, motif logo showing similar position specific alignment and base frequencies of predicted motifs. In bit map, the bit value of each position specific nucleotide is proportional to its relative conservation (relative frequency) in that position. The total height of all the residues in the specific position was proportional to their relative conservation (information content). Bit map showing that expect 1, 2, 3, 4, 8, 12, 14& 15 less conserved position; all other positions were almost conserved for a specific nucleotide.

5. Result & Discussion

Prediction accuracy of newly derived PhoB matrices were statistically evaluated and compared with E. coli Known PhoB TFBS data at RegulonDB Except PhoB CONSENSUS matrix none of the matrices showed sharp increase in accuracy of prediction and found matched with the experimental binding sits; a good correlation on such is shown in ROC analysis plot

5.1 STATISTICAL EVALUATION OF PREDICTION ACCURACY BY NEWLY CONSTRUCTED PHOB MATRICES

PhoB	PhoB Transcriptional dual regulator
Synonyms:	PC00031
Conformations:	PhoB- Phosphorylated transcriptional dual regulator
Site length:	17
Site Symmetry:	Direct
Notes:	The protein belongs to the two-component family. Phosphate regulon transcriptional regulatory protein PhoB is a member of the two -component regulatory system PhoR/PhoB system is involved in the regulation of the phosphate regulon gene expression. Under conditions of phosphate limitation the PhoB protein is Phosphorylated by phoshpo-PHoR. In this Phosphorylated state phosphors-PhoB acts as a transcriptional activator of the Pho regulon. There are also two phosphate-independent signal transduction pathways that control the pho regulon. These controls are regulated by the carbon an de energy source. The catabolite sensor Kinase - phosphotransferase CreCcan can phosphorylate PhoB in one pathway. In the other acetyl phosphate ins required as well as another sensor kinase (Pub MED): 90133909, 95369736, 93163134, 97055429)
Protein Sequence Reference:	MEDLINE 87060980 SWISSPROT P08402
Gene:	phoB
Gene Function	Positive DNA -Binding transcriptional regulator for pi uptake, response regulator in two-component regulatory system with PhoR (or CreC), regulates pho regulon (and asr gene).
Matrix ID:	ECK12H 008035
Matrix Name:	PhoB matrix
Method :	All the know DNA- Binding sites for this transcriptional factor were aligned using consensus. The First matrix of the cycle was used to score the sites, using Patser.
Parameters Used:	Consensus: -alphabet='A:0.25G:0.25 C:0.25' -L= Size of the protein

	Paster: - alphabet='A: 0.25 T:0.25 G:0.25 C:0.25' -t= only the top score
Reference :	Escherichia Coli strain K-12 Regulon DB Data Base v4.0, 25-AUG-05

Table .1- Details of E.Coli Documented known PhoB *trans*-factor and their direct repeat type dyad symmetrical *cis*-element (or pho-box)

Test data set of Promoters	No. Of Predicated know promoters (in %)			
	Regulon DB PhoB matrix	Consensus PhoB matrix	MEME PhoB matrix	Gibbs Sampler PhoB matrix
9 (E.Coli 8 promoters from Regulon DB and Single Acinetobacter Sp. Promoter from GenBank)	8 (88.89%)	9 (100%)	7 (77.78%)	7 (77.78%)
9 (E.Coli. 8 Promoter from RSAT and Single Acinetobacter sp promoter from GenBank)	8 (88.89%)	9 (100%)	7 (77.78%)	8 (88.89%)

Tables 2- Evaluation of PhoB matrix prediction performance on know test data set of total nine promoters belonging to E.coli & Acinetobacter species. Above data revealed the number of predicted promoters analyzed by different PhoB matrices derived through deferent web based tools viz. RegulonDB, Consensus, MEME & Gibbs sampler and their predication accuracy in percentage value.

In contrast, through ROC statistical analysis, we noted the following interesting trends: (i) all predictions at threshold weight sore 0 to 13, resulted maximum true positive sensitivity range of 91% of to 27% respectively for PhoB Consensus matrix, which is more or less accommodated within a line and thus correlate will with the experimental binding affinities, (ii) predicted weight scores outside of this range of value appear to be distributed somewhat at random and they do not correlate with affinity data (iii) those sequences that could not be predicted correctly did not correspond to extreme affinity values but are rather scattered over the whole range of experimentally determined affinities and (iv) after PFM based motif discovery analysis, newly constructed PhoB CONSENSUS matrix showed high sensitivity and lower false positives ratio at 'O' weight score threshold and predicted all the know E.coli PhoB binding sites successfully. Evidence of PhoB orthologs

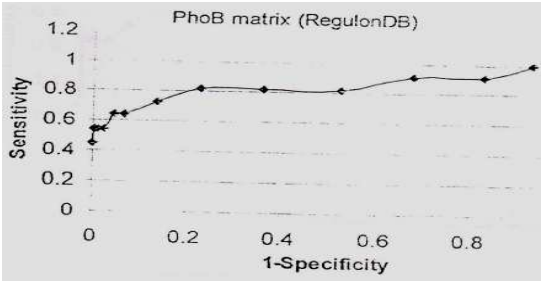
We collected experimental data of PhoB, for which binding sites and their pho-regulon has been already determined in E. coli. PhoB is a positive DNA –Binding transcriptional regulator for pi (inorganic phosphate) uptake, response regulator in two-component regulatory system with PhoR (or CreC), regulates pho operons (and asr gene) in E.coli (Table 1). Conserved domain analysis showed a characteristic similar type of 3 domain namely, (i) gnl/CDD/5334 cd00156 as REC, Signal receiver domain (113 amino acid residues in length) (ii) gnl/CDD/27993 cd00383 as trans_reg_C, Effectors domain of response regulator (95 amino acid residues) and (iii) gn1/CDD/10613 COGO745 as OmpR, Response regulators consisting of a CheY-like receiver domain and a winged- helix DNA binding domain (229 amino acid residues) in all orthologs of PhoB regulatory protein (data not shown).

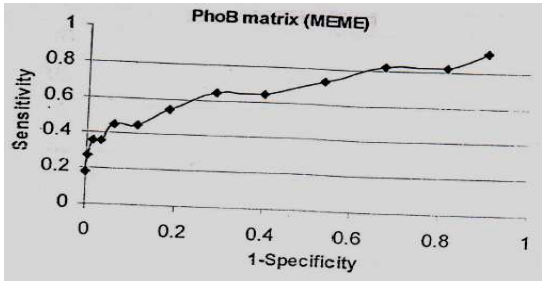
	Position																
Nucleotide	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0	0	0	2	2	6	0	7	5	7	2	1	1	0	0	3	9
C	6	0	1	1	6	1	0	1	1	1	0	7	0	0	2	6	0
G	0	0	5	0	1	1	2	1	1	1	2	0	0	7	0	0	0
T	3	9	3	6	0	1	7	0	2	0	5	1	8	2	7	0	0
Sum	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Consensus (pho-box)	C	T	G	T	C	A	T	A	A	A	T	C	T	G	T	C	A

	Position																	
Nucleotide	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

A	0	2	5	0	0	1	0	7	5	7	2	1	1	7	2	4	2	4
C	2	2	0	6	1	3	0	1	1	1	0	7	0	0	6	1	0	2
G	0	2	0	1	4	0	2	1	1	1	2	0	0	1	0	0	3	0
T	7	3	4	2	4	5	7	0	2	0	5	1	8	1	1	4	4	3
Sum	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
Consensus (pho-box)	T	T	A	C	G/T	T	T	T	T	T	A	T	T	T	C	A/T	T	A

Table 3 Comparison of known and predicted pho-box motifs and their consensus logo detected through derived matrices. (a) Know pho-box matrix of Regulon DB, (b) Derived pho-box matrix of MEME





(a) (b)

Figure2: Comparison of known and predicted pho-box motifs and their consensus logo detected through derived matrices

5.2 References:

[1] Dynamics and design principles of a basic regulatory architecture controlling metabolic pathways PLoS Biol. 2008 6(6):e146. C. Chin, V. Chubukov, E. Jolly, J. DeRisi, H. Li

[2] Papp, P. P., Chattoraj, D. K. & Schneider "Selection of DNA binding sites by regulatory proteins, statistical-mechanical theory and application to operators and promoters" 1987

[3]Rodger Staden, David P. Judge and James K. Bonfield. *Managing Sequencing Projects in the GAP4 Environment. Introduction to Bioinformatics. A Theoretical and Practical Approach.* Eds. Stephen A. Krawetz and David D. Womble. Human Press Inc., Totawa, NJ 07512 (2003)

[4]B. P. HOLDEN[3,4] Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637;

[5]Lawrence PO, Baranowski RM, Greany PD, Nation JL. 1976. Effect of host age on development of *Biosteres (Opius) longicaudatus*, a parasitoid of the Caribbean fruit fly, *Anastrepha suspensa*. *Florida Entomologist* 59: 33-39.

[6] Benson, D., Boguski, M., Lipman, D., Ostell, J., Ouellette, B., Rapp, B., & Wheeler D. (1999). Genbank. *Nucleic Acids Res*, 27, 12-7.

[7]Perform your original search, uchiyama i.2003, in *Nucl. Acids Res.* Search Nucleic Acids Research 2006 34(2):647-658; doi:10.1093/nar/gkj448

[8]Van Helden J, André B, Collado-Vides J Yeast 16 (2000), 177-187 (computational analysis of **yeast** regulatory sequences)

[9] Marchler-Bauer A, et al. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* (2002) 30:281–283

[10]Crooks et al. (2004) Gavin E Crooks; Gary Hon; John-Marc Chandonia & Steven E Brenner: WebLogo: a sequence logo generator**.** Genome Res**,** *14*, 1188-1190