ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

PREs-Clustered motifs in *Drosophila melanogaster*

Sabahuddin Ahmad[1], Abuzar Hamza[2], and Khalid Raza[1]*

[1]Department of Computer Science, Jamia Millia Islamia, New Delhi, India.

[2]Blue Star Infotech Ltd., Bengaluru, Karnataka, India.

ABSTRACT

Understanding the motif pattern is the prerequisite for understanding the Polycomb Response Elements (PREs). Being recruiter of the Polycomb Group (PcGs) Proteins, Polycomb Response Elements are playing fundamental role in the several biological processes of living beings like cell identity, cellular differentiation etc. The study of PREs is tougher in case of mammals. The diversity of the mammalian PcGs make it difficult for the scientist to establish a relation in between the PREs found in the *Drosophila melanogaster* to those of mammals. The previous studies suggest that all the PREs are not alike and there is no specific criteria still defined to say that one PRE is different from the other one. Also, the DNA fragments occupied by these elements are very small portion of the genome, making the related study much complex. Though it is known that single motif cannot act as PRE, but clusters of two or more motifs can do so. Till date, there is no statistical report, which supports this fact. Here, we utilised the available techniques in search of the known PRE motifs, in the upstream sequence of the experimentally verified genes of *Drosophila melanogaster* and performed statistical analysis by applying the concept of clustering of different types to create the clusters of PRE motifs and score these clusters using Sørensen–Dice coefficient which supported the fact that 'clusters of single motifs do not define PREs'. We expect that this method could help in devising a common signal to predict new PRE genes.

Keywords: PcG, PRE, Motif, DNA, *Drosophila melanogaster*

Correspondence author

ISSN: 0975-8585

INTRODUCTION

Biological processes like specific cell type differentiation are under epigenetic control of certain gene expressions. The quest to understand the phenomenon of the epigenetics control eventually led to the discovery of the homeotic (HOX) genes in the *Drosophila melanogaster* that maintain particular cell identity, during the cell division. It has been found that the group of the proteins are employed by these genes that regulate certain gene expression level and distinct sets of active and inactive gene to be transmitted to the daughter cells. These protein complexes mediate its job through 'gene silencing', giving the 'dividing cell' a particular cell identity. These proteins are very specific for target. These group of proteins known as Polycomb Group Proteins (PcGs) form a complex of two or more proteins and are commonly known as Polycomb repressive complexes, which are primarily of three types: Polycomb Repressive Complex 1 (PRC1), Polycomb Repressive Complex 2 (PRC2) and recently discovered PhoRC, which is a DNA binding protein [1]. There are other types too like PCL- PRC2, dRAF, which are important but studies on them are still going on to understand them better.

These protein repressive complexes PRC1, PRC2 and PhoRC are recruited by very small DNA fragments called, Polycomb Response Elements (PREs). The Polycomb Group Proteins are so named 'Polycomb' because of their comb shape. Cell pattern formation at developmental stages is the basic work of the genes on which these protein act. There are about 16 proteins, identified in fruit flies as the repressor of the homeotic genes; they are encoded by Polycomb Group (PcG) Genes. The fundamental question is how these proteins are recruited, or how these proteins form a complex, and identifies the near PRE sites in the genome. It has been a long time since when related researchers are focussing their work on this theme. According to a review related to polycomb proteins, by 2005, more than 300 research papers and 100 reviews were published [2].

The PREs have the ability to remember their activity throughout many cell generations. PREs can be best referred as 'shift workers', they take part in regulating the expression patterns of same genes at different stages of embryonic development. The lack of knowledge in context to PREs is the main reason because of which the understanding the activity of PcG proteins during mammalian embryo development remains unclear [3]. Polycomb proteins appear to silence the transcription by modifying the chromatin structure (present in both the promoter and coding regions of target genes), the silencing depends on the continuous presence of PREs and the Polycomb protein complexes that bind to them. The binding of PcGs with PREs is set of complex interaction, not a simple interaction in between the static structures. The interaction is too complex in between the protein surface and DNA helix, which may or may not be linked to a particular nucleotide sequence [4]. **Fig. 1** is a brief outline of how the polycomb complexes impart the silencing effect on to the genes.

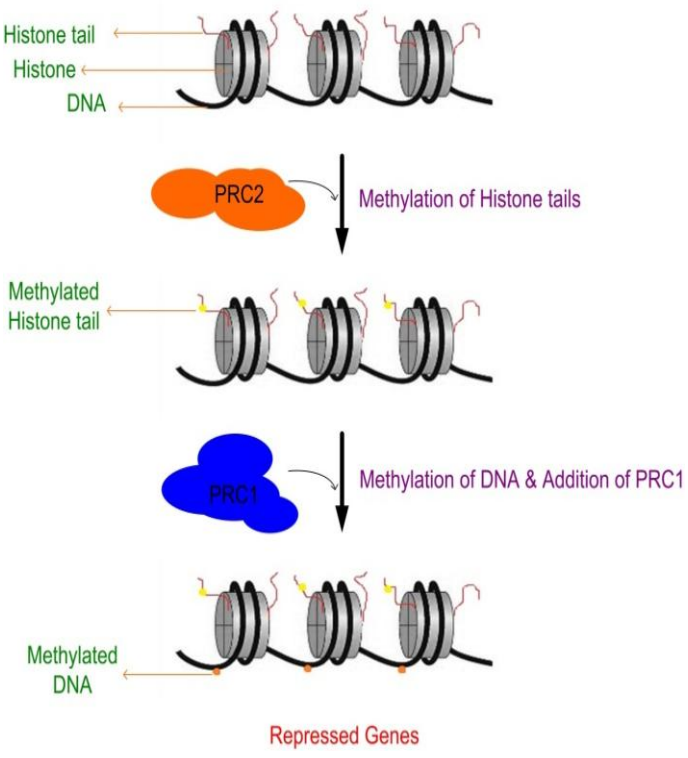

Fig. 1 Silencing of genes by Polycomb Complexes

ISSN: 0975-8585

If we consider present research scenario specifically in the case of PREs. Only two PREs such as PRE-kr and D11.12 has been acknowledge in mammalian[5-6], around dozen of *Drosophila* PREs[7-16] have been found and experimentally confirmed. Furthermore, only 9 *Drosophila* transcription factors were confirmed to be involved in the recruitment of PcG proteins. Among these, only 2 have mammalian homologues PHO and vertebrate GAF [17-18].

In this study, using available computational approaches we tried to statistically prove an important fact of the PRE motifs that 'clusters of single motif do not define PREs', i.e. for motifs to show PRE activity, they must be present in a group of two or more motifs of same or other types. This study helps us in finding the linear arrangements of motifs and their combination that can act as PRE in *Drosophila melanogaster*. We expect that scoring this cryptic arrangement may help in finding out new genes under control of these motifs.

MATERIALS AND METHODS

The genome sequence of the *Drosophila melanogaster* (version dmel_r5.49_FB2013_01) genome was obtained from Flybase [19]. Further, the experimentally verified gene data of functional assay such as transgenic analysis or chromatin immune-precipitation was obtained and curated from reported publications [20-22]. The 7 PRE motifs reported in *Drosophila melanogaster* (refer to Table I) were included for linear motif search. All the PREs located within 900 upstream gene was mapped on the genome and were taken in to consideration, as these PREs have been reported to be located tens of kilobases away from the promoter they regulate [23]. Reported experiments suggest that the cluster of single motifs do not define PREs [24]. Therefore, we considered that these signal motifs are composed of basic PRE units in cluster of k-tuple 2 and 3, with a distance of separation 220 bp between them. For each k-tuple, Sørensen–Dice coefficient [25-26] was calculated. Later on scoring, we also found that simply clustering two or more motifs without defining any parameters do not provide significant results. Applying the concept of permutation too did not provide any change after application of the filter, stating it not useful for forming the clusters. One of the methods for clustering which we applied is finding possible unique combinations between motifs. The total number of possible unique combinations (pairs) would be $n(n+1)/2$, where n is the number of motifs. Let A, B, C, and D are four motifs, then the possible number of unique combinations would be 10 of the tuple size 2 and that would be AA, AB, AC, AD, BB, BC, BD, CC, CD, and DD. This concept

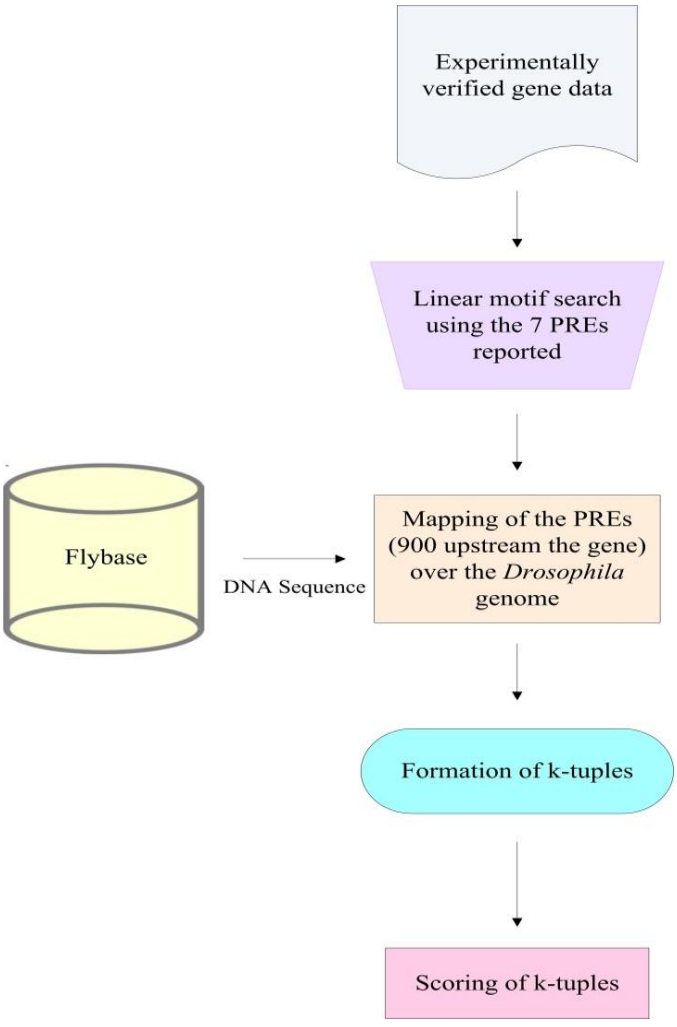

Fig. 2 The work flow of the Experiment: Methodology

was obtained from research published in support of the software tool called "jPredictor" for prediction of PREs [27]. **Fig. 2** is a brief work flow of our work.

Table I List of seven PRE motifs considered in our experiment

Alias	Type	Motif Sequence
G	engrailed 1	GAGAG
G10	long GAF	GAGAGAGAGA
PS	GAGA factor binding site	GCCAT
PM	PHO consensus 1	CNGCCATNDNND
PF	PHO consensus 2	GCCATHWY
EN1	PHO core motif	GSNMACGCCCC
Z	Zeste binding site	YGAGYG

RESULTS AND DISCUSSIONS

A. *Genome abundance of 'single motifs' count does not correspond to the motif around experimental verified gene:*

In this study, we found that the motifs when present in single clusters do not give signals for the PREs. As compared to the total count, the presence of motif was too less. The highest percentage found, was of G10, a regular expression motif. For experimentally verified genes it comes to be 0.86% and for non-experimentally verified genes it comes out to be 10.87%. Refer **Table II** for the total count, the counts of PREs in the experimentally verified and non-experimentally verified genes. The distribution of the motif is visualised in the **Fig. 3**, which is the graphical representation of individual motifs (non-clustered) for given set of experimentally verified and non-verified genes present in the window size -900.

TABLE II DISTRIBUTION OF THE PREs AT WINDOW SIZE -900

PREs	Sequence	Count in Whole Genome	Count in Non Exp. Genes	Count in Exp. Genes
En1	GSNMACGCCCC	21860	1920	75
G10	GAGAGAGAGA	38980	4239	335
GAF	GAGAG	215182	21822	1081
PF	GCCATHWY	70790	6376	188
PM	CNGCCATNDNND	45006	4434	146
PS	GCCAT	321391	30507	809
Z	YGAGYG	213667	22624	887

ISSN: 0975-8585

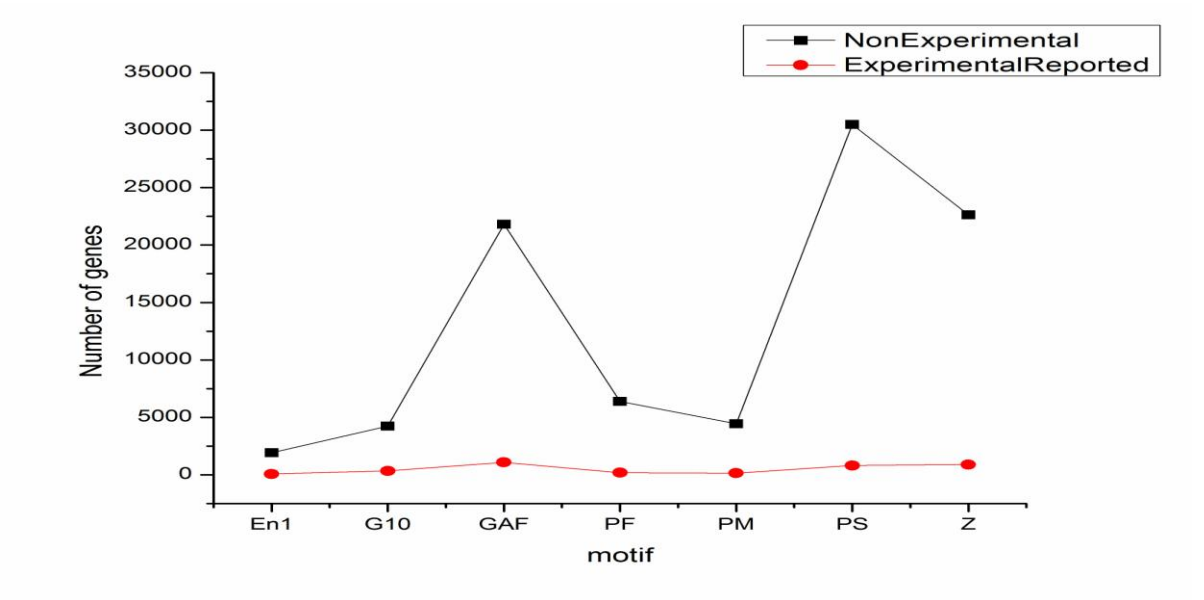

Fig. 3 Distribution of the non-clustered motifs /PREs in the window size -900

B. *Selection of En, PM and PF presence count gene surrounding indicate strong rule for PRE controlled gene:*

Unique count of the gene around the gene upstream reports that the total count is too high, but the percentage of the motif falling in the stream is too less **(See Table II)**. Clearly indicating that the parameter based on the single motif to act as a PRE would not be successful.

C. *PS-En1 motif pair is strong signal for PRE controlled gene:*

PS-En1 motif pair that was formed on the basis of the k-tuple shows the highest score among all the pair formed between the 7 PREs/motif pairs using the concept of the k-tuple. The **Table III** shows the top 10 pairs with their scores in form of frequency of appearing in the cluster to act as PRE.

TABLE II FREQUENCY BASED CALCULATION FOR MOTIF CLUSTER

Tuple/Motif Cluster	Score with filter
PS-En1	0.0083618396
En1-PS	0.0060813379
GAF-En1	0.0053211707
En1-Z	0.0041809198
Z-En1	0.0041809198
En1-GAF	0.0030406689
PF-En1	0.0026605853
En1-PF	0.0022805017
En1-En1	0.0015203345
En1-G10	0.0015203345
En1-PM	0.0011402509
G10-En1	0.0007601672
PM-En1	0.0003800836

ISSN: 0975-8585

D. *No significant change in the ranking clustered pair ranking changed after application of the filter rule:*

We tried to understand the difference that could be, after the application of the filter. But we came to the conclusion that there were no significant changes we were expecting, from this experiment. For tuple=2, the **Fig. 4** represents the frequency of the motifs when we applied filter, and the **Fig. 5** represents the frequency of the motifs without the filter.

The results were similar in the case, when we took tuple=3, for our experiments. There was no significant change after the application of the filter (See **Fig. 6** and **Fig. 7**).

E. *GAF, Z, PS and G10 major PRE unit forming cluster:*

Among all the clusters that were formed, the most occurring PREs were GAF, Z, PS, G10. These motifs were not so frequent when alone, but when they were in clusters, there appearance was frequent in the genome (See **Fig. 4-7)**.

F. *Change in Ranking:*

If we carefully look at the plots generated, we can find that, the GAF-GAF and GAF–GAF-GAF combination was the most frequent. There was steep change in the score of the pairs for k-tuple 2 (motif occurring in pairs), but the same was gradual in the case of the k-tuple 3 (motif occurring in group of three).

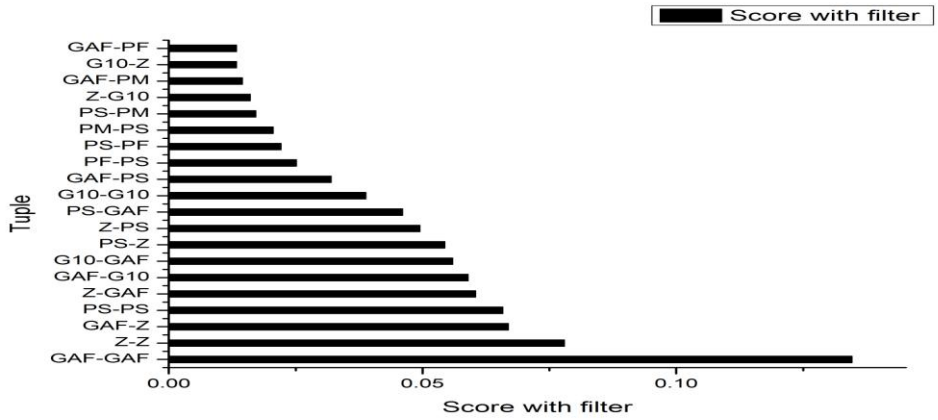

Fig. 4 Score using the filter for k-tuple=2 (motif in pairs)

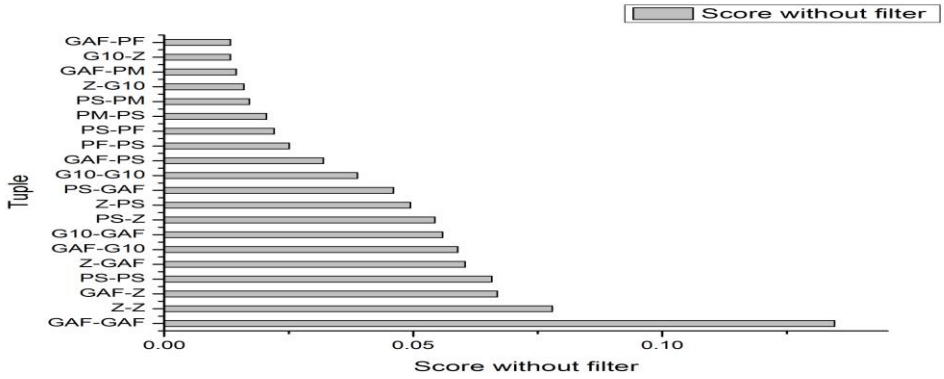

Fig. 5 Score without filter for k-tuple=2 (motif in pairs)

ISSN: 0975-8585

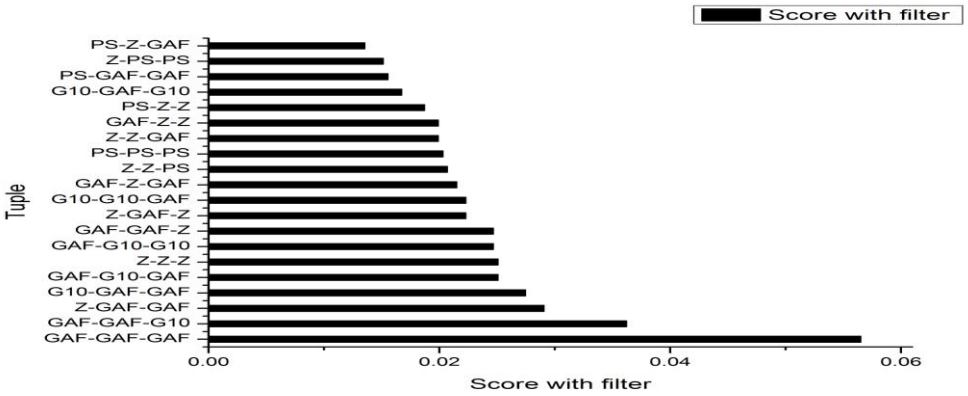

Fig. 6 Score with filter for k-tuple=3 (motif in group of 3)

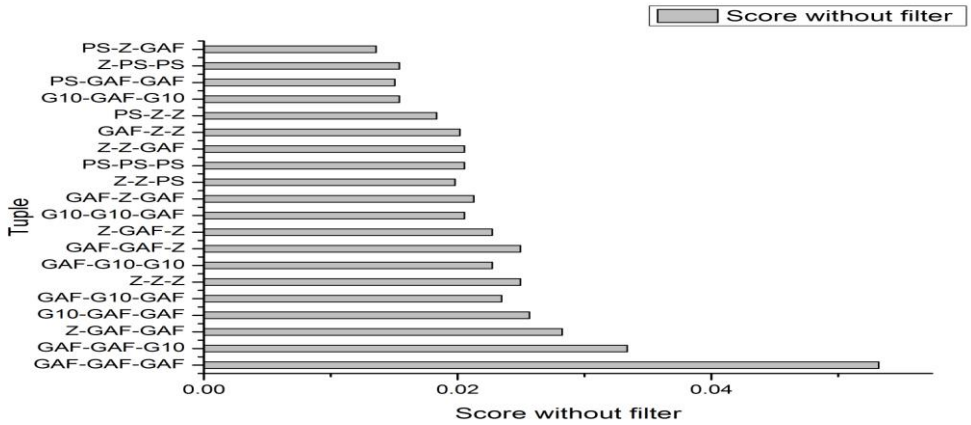

Fig. 7 Score without filter for k-tuple=3 (motif in group of 3)

G: k-tuple formation does not give significant results:

If we go through the scores plotted for k-tuple 2 and 3, there are no changes after applying the filter supporting that clustering by k-means and forming k-tuples simply without defining any parameter would not solve our problem.

H: Concept of permutation did not give positive results too:

In our experiment, we also tried to analyse the results using the concept of permutation. Permutation is the statistical technique and is helpful to select random elements. The results for the count using filter and without filter were the same, hence using the concept of permutation is useless in our case (See **Fig. 8** & **Fig. 9).**

ISSN: 0975-8585

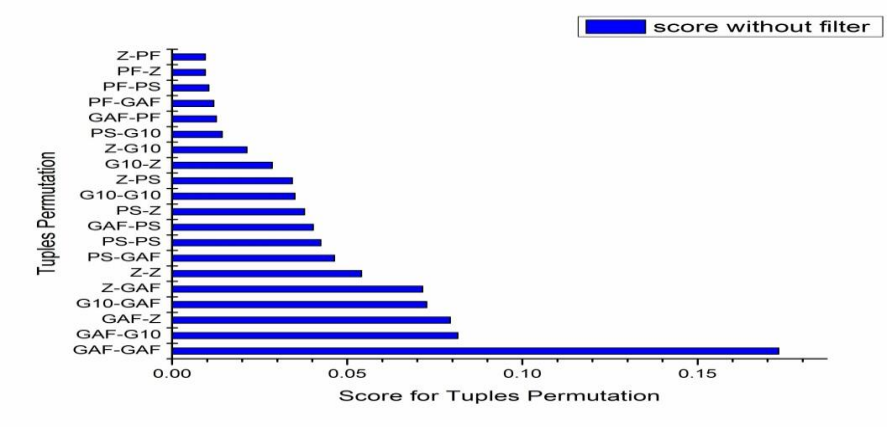

Fig. 8 Score without filter for k-tuple=2 of random motifs using permutation

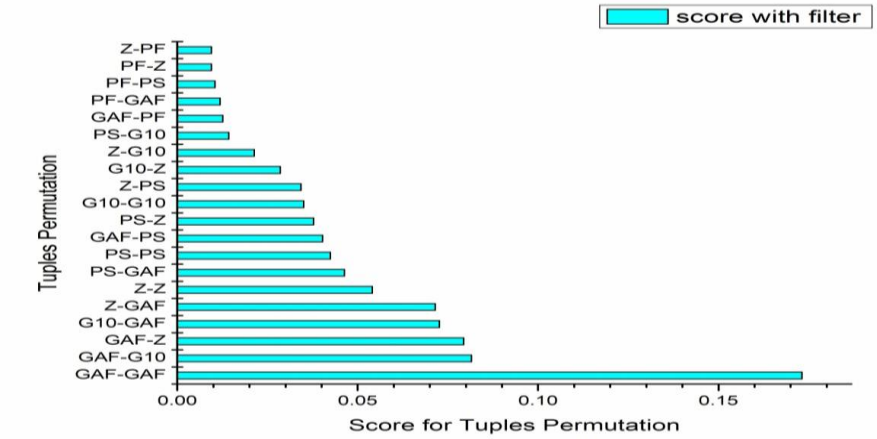

Fig. 9 Score with filter for k-tuple=2 of random motifs using permutation

I: Using the formula n(n+1)/2 for finding possible number of clustered (paired) motifs gave positive results, suggesting it as significant method:

If there are n motifs, then total number of possible unique combinations (pairs) would be $n(n+1)/2$. Let A, B, C, and D be four motifs, then the possible number of unique combinations would be 10 of the tuple size 2 and that would be AA, AB, AC, AD, BB, BC, BD, CC, CD, and DD. In this case, there were different results obtained before applying the filter and after applying the filter including experimental genes. There were two cases; one was for filter and other for non filter. For filter, the cluster of GAF-GAF was found maximum, followed by the group of Z-Z, GAF-Z and so on, in particular window size, (i.e. -900) refer to **Fig. 10**. When there was no filter, we found again that count of GAF-GAF was maximum, followed by the Z-Z cluster, but in this case of not using the filter, the third cluster found with maximum score was GAF-G10, supporting it to be possible PRE motif (Refer to **Fig. 11**)

ISSN: 0975-8585

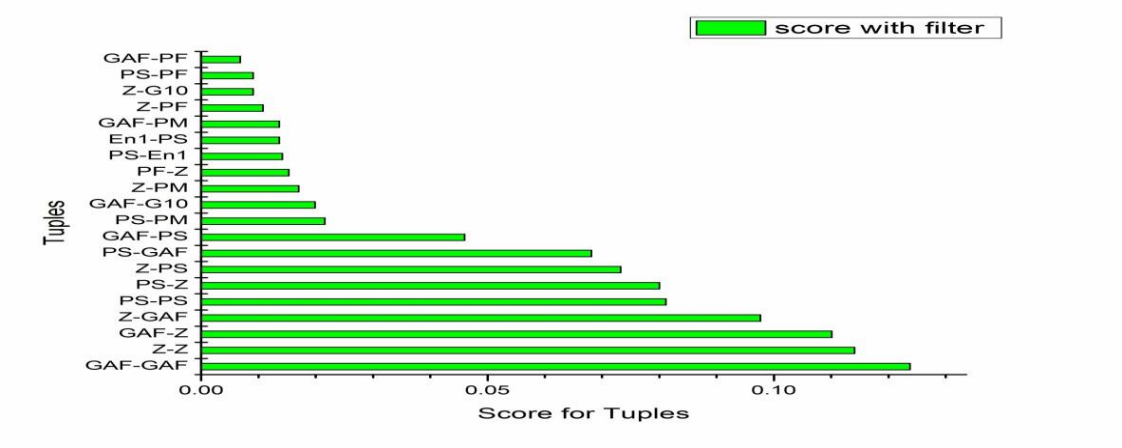

Fig. 10 Score with filter for k-tuple=2 of random motifs

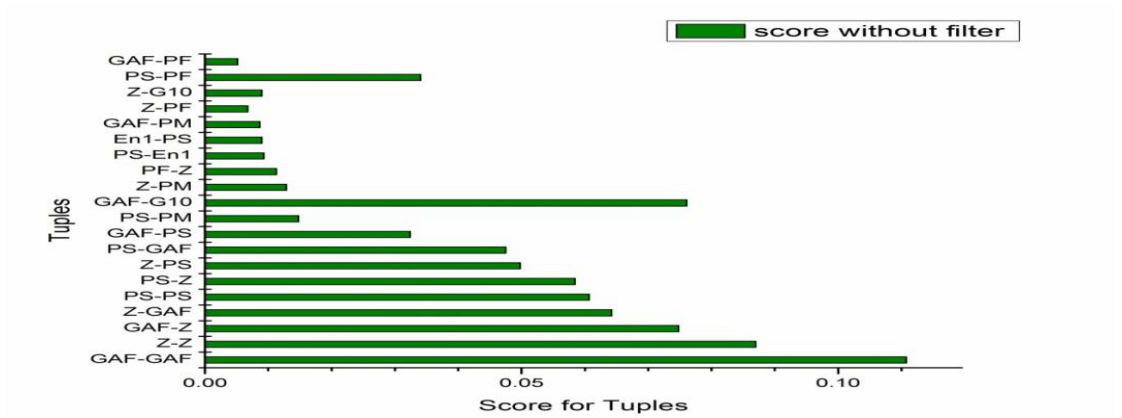

Fig. 11 Score without filter for k-tuple=2 of random motifs

CONCLUSIONS

Polycomb Response Elements (PREs) are playing a vital role in our life by recruiting the Polycomb proteins, involved in important epigenetic activities of life like cellular differentiation and cell identity. Understanding the concept how the polycomb proteins silence the gene expression is known, but exactly how the PRE motifs help in the recruiting the polycomb proteins is still not clear. It is expected that the time when we will have clear picture of the theme, how several motifs in groups act as PREs, we will be able to define set of rules, how the PREs are formed and then we can be helped in searching PREs in mammals, which yet are undiscovered, almost.

Although our experiment not reporting any novel concepts, or ideas, but we had in-hands certain conclusions. Our statistical analysis supports the information, which the scientists in related field have exploited in past few decades. We utilised the discovery that the 'PREs can be located tens of kilo-bases away from the site they regulate'. Therefore, considering window sizes like -200, -300, -400, and so on, could not be helpful for us and we used large window size, i.e. -900 for our experiment. The huge abundance of PRE motifs in this region supports that these motifs, which are responsible PRE signals are located at distinct places in the whole genomic content of the fruit flies. Further the clustering task proved an important fact; the clusters of single motifs do not define PREs. Clustering by various means suggests us that the motifs though act as PRE in group of more motifs, but there are specific set of rules followed, in which these motif clusters with other motifs and signals for PRE activity. We found that distance between the two pairs of PRE motifs i.e. 220bp favors the formation of PRE motif pairs and

hence supporting PRE activity. One of the most frequently occurring motif cluster throughout the genomic content was GAF-GAF. Overall significance concludes that the clustered motifs do give signals for PRE genes. The motif like GAF, G10, occurring more commonly, which are part of the Pho, found in the Drosophila melanogaster is the homologue to the Ying Yang 1 (YY1) found in mammals. It has been reported in experiments that Pho/YY1 are very crucial and necessary for the binding of the PRE but alone they are useless. In other words, the Pho protein requires other motifs too, in combination for proper functioning of the PREs in the recruiting process of the Polycomb Group Proteins. GAF are GAGA factor binding site and regular expression motif too. The GAGA protein is a nuclear protein, which remains attached with the DNA, during the mitosis and thereby responsible for its activity during the gene silencing. Though there are enormous number of results available through the studies related to PREs, but still the understanding are not clear; scientist are trying to device new rules and parameter, particular to PREs and PRE related motifs already discovered, to predict new PREs in the Drosophila as well in mammals.

REFERENCES

[1] Mihaly, J., Mishra, R.K. and Karch, F. (1998) A Conserved Sequence Motif in Polycomb-Response Elements. *Mol. Cell,* 1, 1065–1066.

[2] Ringrose, L., & Paro, R. (2007). Polycomb/Trithorax response elements and epigenetic memory of cell identity. Development, 134, 223–232.

[3] Cuddapah S, Roh T-Y, Cui K, Jose CC, Fuller MT, et al. (2012) A Novel Human Polycomb Binding Site Acts As a Functional Polycomb Response Element in Drosophila. PLoS ONE 7(5): e36365.

[4] Kobe Florquin, YvanSaeys, Sven Degroeve, Pierre Rouze´ and Yves Van de Peer. (2005) Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Research,* Vol. 33, No. 13 4255–4264.

[5] Sing,A., Pannell,D., Karaiskakis,A., Sturgeon,K., Djabali,M., Ellis,J., Lipshitz,H.D. and Cordes, S.P. (2009) A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. *Cell*, 138, 885–897.

[6] Woo,C.J., Kharchenko,P.V., Daheron,L., Park,P.J. and Kingston,R.E. (2010) A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell*, 140, 99–110.

[7] Kassis,J.A. (1994) Unusual properties of regulatory DNA from the *Drosophila* engrailed gene: three ''pairing-sensitive'' sites within a 1.6-kb region. *Genetics*, 136, 1025–1038.

[8] Gindhart,J.G. Jr and Kaufman,T.C. (1995) Identification of Polycomb and trithorax group responsive elements in the regulatory region of the *Drosophila* homeotic gene Sex combs reduced. *Genetics*, 139, 797–814.

[9] Americo,J., Whiteley,M., Brown,J.L., Fujioka,M., Jaynes,J.B. and Kassis,J.A. (2002) A complex array of DNA-binding proteins required for pairing-sensitive silencing by a polycomb group response element from the *Drosophila* engrailed gene. *Genetics*, 160, 1561–1571.

[10] Bloyer,S., Cavalli,G., Brock,H.W. and Dura,J.M. (2003) Identification and characterization of polyhomeotic PREs and TREs. *Dev. Biol.*, 261, 426–442.

[11] Gruzdeva,N., Kyrchanova,O., Parshikov,A., Kullyev,A. and Georgiev,P. (2005) The Mcp element from the bithorax complex contains an insulator that is capable of pairwise interactions and canfacilitate enhancer-promoter communication. *Mol. Cell. Biol.*, 25, 3682–3689.

[12] DeVido,S.K., Kwon,D., Brown,J.L. and Kassis,J.A. (2008) The role of Polycomb-group response elements in regulation of engrailed transcription in *Drosophila*. *Development*, 135, 669–676.

[13] Kozma,G., Bender,W. and Sipos,L. (2008) Replacement of a *Drosophila* Polycomb response element core, and in situ analysis of its DNA motifs. *Mol. Genet. Genomics*, 279, 595–603.

[14] Cunningham,M.D., Brown,J.L. and Kassis,J.A. (2010) Characterization of the polycomb group response elements of the *Drosophila melanogaster* invected Locus. *Mol. Cell. Biol.*, 30, 820–828.

[15] Horard,B., Tatout,C., Poux,S. and Pirrotta,V. (2000) Structure of a polycomb response element and in vitro binding of polycomb group complexes containing GAGA factor. *Mol. Cell. Biol.*, 20, 3187–3197.

[16] Ringrose,L., Rehmsmeier,M., Dura,J.M. and Paro,R. (2003) Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. Dev. Cell, 5, 759–771.

[17] Brown,J.L., Mucci,D., Whiteley,M., Dirksen,M.L. and Kassis,J.A. (1998) The *Drosophila* Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. Mol. *Cell*, 1, 1057–1064.

[18] Matharu,N.K., Hussain,T., Sankaranarayanan,R. and Mishra,R.K. (2010) Vertebrate homologue of *Drosophila* GAGA factor. *J. Mol. Biol.*, 400, 434–437.

[19] S.J. Marygold, P.C. Leyland, R.L. Seal, J.L. Goodman, J.R. Thurmond, V.B. Strelets, R.J. Wilson and the FlyBase Consortium (2013). FlyBase: improvements to the bibliography. Nucleic Acids Res. 41(D1):D751-D757. [FBrf0220350]

[20] Schwartz,Y.B., Kahn,T.G., Nix,D.A., Li,X.Y., Bourgon,R., Biggin,M. and Pirrotta,V. (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster.Nat. Genet., 38,700–705.

[21] Tolhuis,B., de Wit,E., Muijrers,I., Teunissen,H., Talhout,W., van Steensel,B. and van Lohuizen,M. (2006) Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster.Nat. Genet., 38, 694–699

[22] Schuettengruber,B., Ganapathi,M., Leblanc,B., Portoso,M., Jaschek,R., Tolhuis,B., van Lohuizen,M., Tanay,A. and Cavalli,G. (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol.*, 7, e13.

[23] Simon, J. A., & Kingston, R. E. (2009). Mechanisms of Polycomb gene silencing: Knowns and unknowns. Nature Reviews Molecular Cell Biology, 10, 697–708.

[24] Ringrose,L., Rehmsmeier,M., Dura,J.M. and Paro,R. (2003)Genome-wide prediction of Polycomb/Trithorax response elements in *Drosophila melanogaster*. *Dev. Cell*, 5, 759–771.

[25] Sørensen, T. (1957). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons". Kongelige Danske Videnskabernes Selskab 5 (4):1–34.

[26] Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species". *Ecology* 26(3):297302. Doi: 10.2307/1932409. JSTOR 1932409.

[27] Fiedler T., Rehmsmeier M. (2006). jPREdictor: a versatile tool for the prediction of cis-regulatory elements. W546–W550 Nucleic Acids Research, 2006, Vol. 34