Biospectra : ISSN: 0973-7057, Vol. 19(1), March, 2024, pp. 163-170 ***Interdisciplinary Science***

An International Biannual Refereed Journal of Life Sciences

ISSN : 0973-7057

Identification of quasi-specific binding sites of cro-, λ- and gal- repressor proteins within *Escherichia coli* bacterial and *Enterobacteria phage* λ viral genomes

Mahendra Kumar[a]*, Satish Saw[a] , Navin Chandra[a] & Kishore Kumar Gupta[b]

[a]Department of Physics, Vinoba Bhave University, Hazaribag, Jharkhand, India

[b]Department of Zoology, Vinoba Bhave University, Hazaribag, Jharkhand, India

Received : 29th December, 2023 ; Revised : 29th January, 2024

DOI:-https://doi.org/10.5281/zenodo.14558059

Abstract- Binding of Proteins with DNA molecules is one of the fundamental bases of life. A protein may have one or more natural binding site within its cognate genome. There may be some other sites within a genome which have slightly lower binding affinity for the protein than its binding affinity with the natural binding site. These sites are quasi-specific binding sites of the proteins. Very little is known about the presence and role of such quasi-specific binding sites within a genome. In the present work we have predicted large number of quasi-specific sites for Cro-Repressor & λ-repressor within foreign *E. coli* genome. For Gal-repressor we find only one such site within foreign *Bacteriophage* λ genome. While in case of the cognate genomes of the three repressor proteins we found very a smaller number of quasi-specific binding sites for the three proteins. This indicates that quasi-specific sites may be behaving as close competitors for protein's binding hence nature have evolved in such a way that a protein can have only smaller number of quasi-specific sites within its cognate genome. Because large number of such quasi-specific sites of a protein within its cognate genome will increase the competition during target search by the protein and eventually cause interference in natural binding of the protein through slowing down the binding process of the protein to its natural sites. Whereas in case of a foreign genome a protein does not have its natural binding site within the genome hence it may have any number of quasi-specific binding sites within it.

Key words: *E. coli*, *Bacteriophage* λ, Genome, Cro-Repressor, λ-repressor, Gal-repressor, Binding energy difference

INTRODUCTION

Protein-DNA interaction is one of the fundamental bases of life. This interaction happens within the cell where a significant number of DNA-binding proteins exist together with the long chain of the DNA made-up of few thousand to several million numbers of base-pairs. Though the DNA is a very long chain molecule, a protein binds with only specific small section of the DNA called specific binding site of the protein.[1-3] A specific binding site may have upto few hundred base-pairs. The other sections of the DNA to whom protein does not bind/interacts naturally but may have low affinity are non-specific sites for the protein.[4] The non-specific sites may have a role during the protein's binding sites target search process within the genome. Berg *et al.* (1982)[5] had suggested that target search mechanism is enhanced by the proteins through using these non-specific sites of the genome. However, this mechanism is debated heavily. Several other studies point out that target search by the proteins involve several complicated steps.[6,7] Understanding of many of these steps is still not well clear.

*Corresponding author :

Phone : 8210826442,7321084409

E-mail : mahendrakumar12051989@gmail.com

To understand the complications of these steps we need to first understand the interactions going on between protein and their specific binding sites within the cognate genome. Takeda *et al.* (1989)[8, 9] performed the first single base-pair substitution experiments to study how a single base-pair contributes to the total binding energy during the interactions of Cro-repressor and λ-repressor with its specific binding sites. These two DNA binding proteins follow different dynamics during their target search within the *Bacteriophage λ* viral genome.[10] In another similar base-pair substitution experiment Naiya *et al.* (2016)[11] had measured the binding affinity changes of Gal-repressor with its single site mutant operator sites. The base-pair substitutions experiments have reported that each base-pair within the specific binding sites of the proteins do not contribute equally to the total binding energy of the protein-DNA complexes. The single of more than one base-pair within a specific binding site can be mutated to obtain different sequences with similar binding energy. The same idea can also be used for obtaining sequences having slightly less binding energy in comparison to the binding energy of specific sites. Such new sites with slightly less binding energy are called quasi-specific sites.[12] How many such quasi-specific sites are within a genome? What are their possible roles during specific binding site target search of a protein within the genome? Very less is known about these questions. In the present study we have tried to find answer of these important questions. As the experimental data required for the present study are available only in case of Cro-repressor, λ-repressor and Gal-repressor proteins hence we have chosen these proteins as ideal system for our study. Cro-repressor and λ-repressor are two gene-regulatory repressor proteins found inside *Bacteriophage λ* virus.[13] While Gal-repressor is a dimer repressor protein naturally found inside *E. coli* bacteria.[14] The main function of these repressor proteins is to repress the expression of certain genes. Cro-repressor is a 66 amino acids small-polypeptide product of cro gene and λ-repressor is 236 amino acids long polypeptide product of cI gene.[15] Both of these proteins naturally bind to same six different 17-base-pair long operator sites (O_R1, O_R2, O_R3, O_L1, O_L2 and O_L3) in *Bacteriophage λ* genome. On the other hand, Gal-repressor binds to two different 16-base-pair long gal operator sites O_E and O_I in *E. coli.*[14] In the present work we tried to predict the quasi-specific binding sites, of these three repressor proteins, within the genomes of *Escherichia coli* bacteria (*E.coli*) and *Enterobacteria phage λ* (*Bacteriophage λ*) virus. The required *E. coli* genome with accession number NC_000913_1 having 4639675 base-pairs and a *Bacteriophage λ* genome with accession number NC_001416.1 having 48502 base-pairs were downloaded from National Centre for Biotechnology Information (NCBI) website http://www.ncbi.nlm.nih.gov. The quasi-specific binding sites of repressor proteins were predicted by writing FORTRAN programs in which experimental data of single base-pair substitution experiments performed by Takeda *et al.* (1989)[8] and Naiya *et al.* (2016)[11] was used as input data. For this, binding energy differences (BED) of Cro-repressor and λ-repressor proteins (with respect to their binding energy with their wild type operator O_R1) with all possible 17-base-pair long genomic sequences of the two downloaded genomes were calculated. Similarly in case of Gal-repressor the BED for all possible 16-base-pair long genomic sequences of the two genomes was calculated with respect to Gal-repressor's binding energy with its wild type operator O_E. For Cro-repressor and λ-repressor proteins large number of quasi-specific binding sites with BED values less than 5 Kcal/mol was detected within the studied foreign *E. coli* genome, while for Gal-repressor only one quasi-specific site within foreign *Bacteriophage λ* genome was found. In case of their cognate genomes very small number of quasi-specific binding sites for all the three proteins were detected, which is quite important and interesting result. As the binding energy of quasi-specific sites remain very close to that of the binding energy of specific sites hence on the basis of binding affinity, the quasi-specific sites may be very close competitors of specific sites for protein's binding. This may distract the protein form binding to its specific sites or even slow down the binding process of the protein. To prevent form this problem, nature may have evolved the genomes in such a way that a protein has only few quasi-specific binding sites within its cognate genome. Thus, small number of quasi-specific sites is also expected in view of the binding of the protein to its specific site of cognate genome within a reasonable small time period. On the other hand, in case of foreign genome the protein does not have their natural sites hence this problem may not be true due to which a foreign genome may have any possible number of quasi-specific binding sites for a protein. A sharp increase in number of binding sites with a small increase in the value of binding energy difference was also noticed. This sharp increase in number of binding sites with increase in BED is due to increase in possibility of random selection/

mutation/replacement in/of cognate 17-base-pair/16-base-pair sequences to which repressor proteins binds.

MATERIALS & METHODS

Collection of *Escherichia coli* and *Enterobacteria phage λ* genome sequences:

Escherichia coli (*E. coli*) bacteria and *Enterobacteria phage λ* virus (also known as *Bacteriophage λ* virus) both of them have genomic material made-up of DNA. *E. coli* genome has more than 4.6 million base-pairs while *Enterobacteria phage λ* (*Bacteriophage λ*) virus genome is relatively small and made-up of approximately 49 thousand base-pairs. For our calculations an *E. coli* genome with accession number NC_000913_1 having 4639675 base-pairs and a *Bacteriophage λ* genome with accession number NC_001416.1 having 48502 base-pairs were downloaded from National Centre for Biotechnology Information (NCBI) website http://www.ncbi.nlm.nih.gov (Table-1).

Table 1

Sl. No.	Accession number of genome file	Total number of bases
01.	NC_001416.1 (*Enterobacteria phage* λ virus)	48502
02.	NC_000913_1 (*E. coli* genome)	4639675

Cro-repressor, λ-repressor and Gal-repressor:

The main function of repressor proteins is to repress the expression of certain genes. Cro-repressor and λ-repressor are two regulatory repressor proteins found inside *Bacteriophage λ* virus.[13] Cro-repressor is a 66 amino acids small-polypeptide product of cro gene while that of λ-repressor is relatively large, 236 amino acids long polypeptide product of cI gene.[15] Both of these two proteins naturally bind to same six different 17-base-pair long operator sites (O_R1, O_R2, O_R3, O_L1, O_L2 and O_L3) in *Bacteriophage λ* genome. Thus, these two proteins together make a genetic binary switch that controls the expression of cro gene and cI gene by preventing their transcriptions through binding to a series of operator sites in *Bacteriophage λ* genome.[10] On the other hand, Gal-repressor is a dimer repressor protein naturally found inside *E. coli*. This repressor binds to two different 16-base-pair long gal operator sites O_E and O_I in *E. coli.*[14]

Identification of quasi-specific binding sites within *E. coli* and *Bacteriophage λ* genomes:

Cro-repressor and λ-repressor naturally binds to a sequence of six different operators O_R2, O_R3, O_L1, O_L2, O_L3 and O_R1 of *Bacteriophage λ* genome. While Gal-repressor binds to two operator sites O_E and O_I inside *E. coli* genome. These binding sites of the repressor proteins exists naturally within their cognate genomes and are considered as specific binding sites. Through base-pair/base substitution experiments, Takeda *et al.* (1989)[8,9] had measured the binding energy difference (BED) of Cro-repressor and λ-repressor proteins to different chemically-synthesized 17 base-pair long operator sites O_R2, O_R3, O_L1, O_L2, O_L3 and their single-site mutants with respect to their binding energy with the wild type operator O_R1. They found that in comparison to binding energy of O_R1 operator, binding energy of other five operators (i.e. O_R2, O_R3, O_L1, O_L2 and O_L3 operator) varies from -0.6 Kcal/mol to 1.9 Kcal/mol for Cro-repressor and from -0.1 Kcal/mol to 2.8 Kcal/mol for λ-repressor. In a similar kind of another base-pair substitution experiment, Naiya *et al.* (2016)[11] has measured the binding energy difference of Gal-repressor to its 16 base-pair long natural operators O_I with respect to operator O_E and found that BED of natural operator O_I varies by 3.57 Kcal/mol. Binding free energy changes calculated from affinity changes were mostly the additive in nature for multiple mutations. This approximate-additive property of the data paved way for its use in prediction of binding affinity of the repressor proteins to any other nucleic acid sequence(s).[8,9] This idea led us to the work for prediction of specific and quasi-specific binding sites (quasi-specific sites are those sites in genome which have binding affinity less than the affinity of naturally occurring specific binding sites) of these repressor proteins within the *E. coli* and *Bacteriophage λ* genomes. By using the base-pair substitution experiment data of Sarai and Takeda (1989)[9], Chakrabarti *et al.* (2011)[12] had predicted presence of a large number of quasi-specific sites for λ-repressor within *E. coli* genome. In this work we have predicted quasi-specific binding sites inside *E. coli* and *Bacteriophage λ* genomes which have BED value less than 5 Kcal/mol in comparison to binding energy of; 17-base-pair long wild type operator O_R1 (in case of the λ-repressor & Cro-repressor) and 16-base-pair long wild type operator O_E (for Gal-repressor). To calculate the binding energy differences for all possible genomic sequences (17-base-pair long for Cro-repressor and λ–repressor and 16-base-pair long for Gal-repressor) of *E. coli* and *Bacteriophage λ* genomes, one FORTRAN program for each of the three repressor proteins were written. In comparison to the binding energy of wild type natural operator O_R1, binding energy differences of λ-repressor & Cro-repressor with each possible 17-base long sequences of the genomes are calculated. Similarly for Gal-repressor protein the binding energy difference of each possible 16-base-pair long genomic sequences in comparison to binding energy of Gal-repressor with 16-base-pair long wild type operator O_E was calculated. For the calculations of value of binding energy differences for all possible 17-base-pair/16-base-pair long genomic sequences, binding energy data of systematic single base pair substitution experiments was used as input in FORTRAN program. Both strands of each genome were scanned one by one in 5' to 3' direction by sliding a window of 17-base-pairs throughout the whole genome in case of Cro-repressor & λ–repressor and a window of 16-base-pairs for Gal-repressor. The binding of repressor is considered to be very weak with all those sequences having binding energy difference more than 5 Kcal/mol; hence data related to such sequences were ignored in this work.

RESULTS

In present study, a large number of quasi-specific binding sites with BED values less than 5 Kcal/mol were detected in case the studied genome was a foreign genome for the proteins. This result is listed in table-2 where we can see that Cro-repressor has 14512 quasi-specific binding sites (sites with BED value between 2 to 5 Kcal/mol) while λ–repressor has 3801 such quasi-specific binding sites (sites with BED value more than 3 Kcal/mol but not exceeding 5 Kcal/mol) in foreign *E. coli* genome. Further the predicted result of presence of large number of quasi-specific binding sites in case of foreign genomes is mostly similar in nature with a deviation observed in case of Gal-repressor which have only one such site inside foreign *Bacteriophage λ* genome (Table 2). The reason behind this deviation is not much clear. Another quite interesting and very important result to note is that, in case of their cognate genomes, all the proteins have very small number of quasi-specific binding sites. As we can see that Gal-repressor has only 39 such quasi-specific sites (sites having binding energy difference more than natural operator O_1) inside its cognate *E. coli* genome. Similarly, Cro-repressor has 137 quasi-specific binding sites (sites with BED value more than 2 Kcal/mol) while λ–repressor has only 32 such quasi-specific binding sites (sites with BED value more than 3 Kcal/mol) in their cognate *Bacteriophage λ* genome (Table 2). A sharp increase in number of quasi-specific sites with increasing BED-values is to be noted in each case (Figure-1). In comparison to the λ–repressor protein the Cro-repressor has remarkably large number (about four times) of quasi-specific binding sites inside studied foreign *E. coli* genome. Along with the very large number of quasi-specific binding sites of Cro- and λ– repressor proteins, we have also found there are few hundred specific binding sites (sites where repressor binds with the similar strength as it binds with its specific sites in their cognate genomes) of these proteins inside foreign *E. coli* genome.

Table 2- Number of predicted specific and Quasi-specific binding sites within different range of BED value

Table:2a							
Sl. No.	**Genome name**	Number of sites within different range of BED (for Cro-repressor)					
		Below 0 Kcal/mol	0-1 Kcal/mol	1 - 2 Kcal/mol	2 - 3 Kcal/mol	3 - 4 Kcal/mol	4 - 5 Kcal/mol
01.	*E. coli*	29	106	374	1238	3672	9602
02.	*Bacteriophage λ*	4	3	9	10	38	89
Table:2b							
Sl. No.	**Genome name**	Number of sites within different range of BED (for λ-repressor)					
		Below 0 Kcal/mol	0-1 Kcal/mol	1 - 2 Kcal/mol	2 - 3 Kcal/mol	3 - 4 Kcal/mol	4 - 5 Kcal/mol
01.	*E. coli*	0	4	41	285	1047	2754
02.	*Bacteriophage λ*	1	1	2	4	12	20
Table:2c							
Sl. No.	**Genome name**	Number of sites within different range of BED (for Gal-repressor)					
		Below 0 Kcal/mol	0-1 Kcal/mol	1 - 2 Kcal/mol	2 - 3 Kcal/mol	3 - 4 Kcal/mol	4 - 5 Kcal/mol
01.	*E. coli*	0	1	1	1	15	31
02.	*Bacteriophage λ*	0	0	0	0	0	1

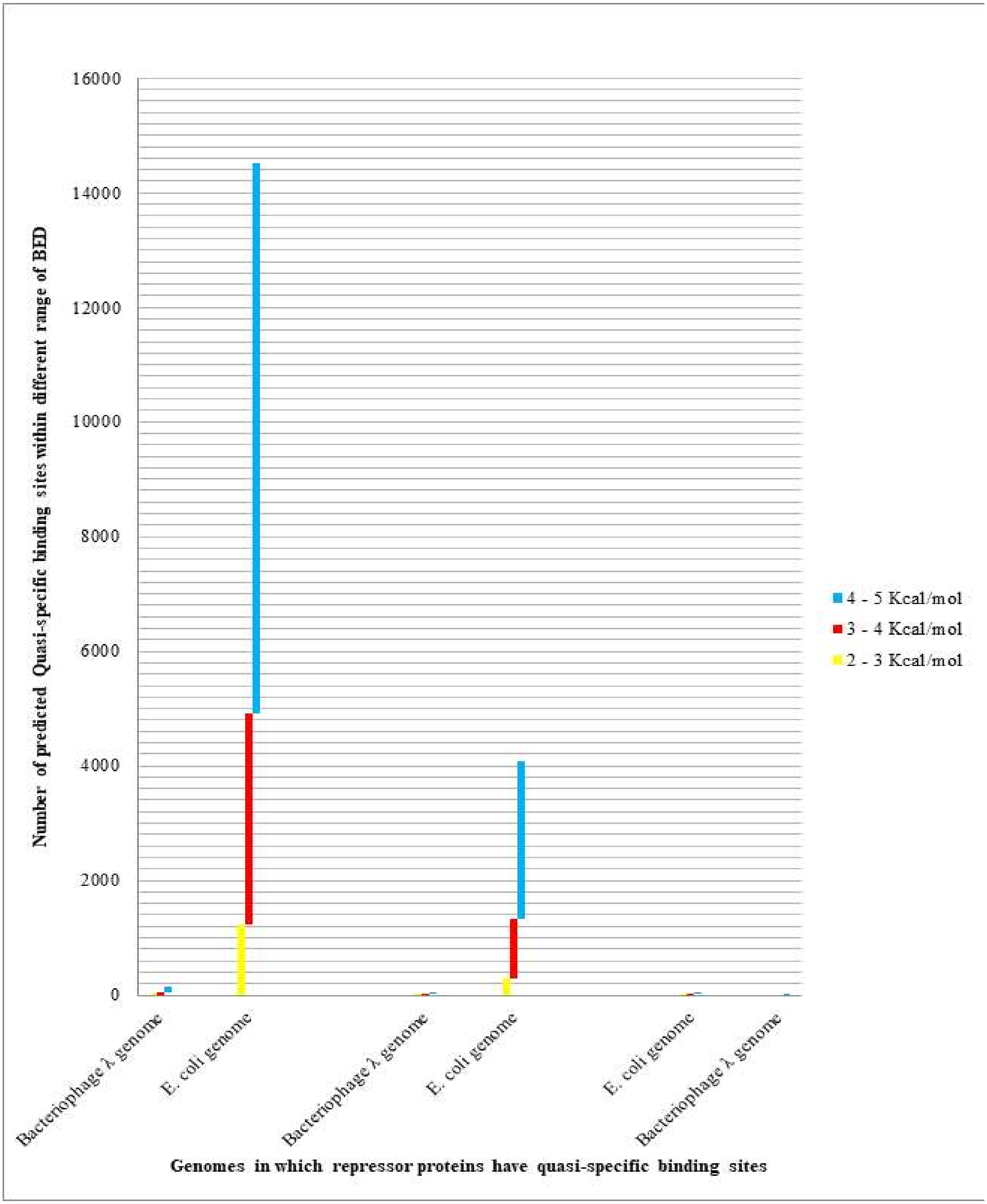

Figure-1: Showing number of quasi-specific binding sites for Cro-repressor, λ–repressor and Gal-repressor proteins predicted inside *Bacteriophage* λ and *E. coli* genomes. This also shows a sharp increase in number of quasi-specific binding sites with small increase in BED value.

DISCUSSION

We noted presence of large number of quasi-specific binding sites along with few hundred specific binding sites of Cro- and λ- repressor proteins inside foreign *E. coli* genome. Whereas this number is very small (only few) in case the studied genome was the cognate genome of the proteins (Table-2). This small number of sites is also expected in view of the binding of the protein to its specific site of cognate genome within a reasonable small time period. However, this may not be true in case of a foreign genome. In a study done by Chakrabarti *et al.* (2011)[12] a large number (more than forty) of such binding sites of λ-repressor protein inside a foreign genome had been predicted.[23] On the other hand, in another study by us (communicated for publishing); large number of quasi-specific binding site of Cro-, λ- and Gal-repressor proteins inside several variants of *SARS-CoV-2* foreign genomes was found earlier. These two studies indicate that a protein have only very few specific and quasi-specific binding sites within its cognate genome but it may have any possible number of specific and quasi-specific binding sites inside a foreign genome. This hypothesis is supported by the results of the present work too, as only few specific as well as quasi-specific binding sites of the three repressor proteins was found within their cognate genomes (Table-2). A possible reason behind the presence of very a smaller number of quasi-specific binding sites of a protein within its cognate genome may be that; since the binding energy of quasi-specific sites remain very close to that of the binding energy of specific sites hence on the basis of binding affinity, the quasi-specific sites may be very close competitors of specific sites for protein binding. This may distract the protein form binding to its specific sites or even slow down the binding process of the protein. If happened so, this may adversely affect the proteins natural function inside its cognate genome. To prevent form this problem, nature may have evolved the genomes in such a way that a protein has only few quasi-specific binding sites within its cognate genome. On the other hand, in case of foreign genome the protein does not have their natural sites hence this problem may not be true due to which a foreign genome may have any possible number of quasi-specific binding sites for a protein. Further we found no specific and only one quasi-specific binding site for Gal-repressor within foreign *Bacteriophage λ* genome, while in case of Cro- and λ- repressor this number is very large in foreign *E. coli* genome (Table-2). Here we should also notice that in comparison to the λ- repressor, the Cro-repressor has relatively large number of such sites within foreign *E. coli* genome (Figure-1). This finding explains a very important mechanism during life cycle of *E. coli*; why for Cro-repressor to bind with O_R1 operator of *E. coli* genome, the concentration of Cro-protein should be high enough, because Cro- will first occupy those sites which have higher affinity to bind. We also know that if λ-repressor protein predominates then *Bacteriophage λ* virus will remain in lysogenic state but if Cro-repressor protein predominates then virus will follow lytic cycle. Considering the important roles of the above discussed two repressor proteins for *Bacteriophage λ* virus, the number of specific and quasi-specific binding sites in the genome may adversely affect the protein binding to essential genomic regions. This also disturbs the replication and translation of the genes resulting in stoppage/disturbance in the *Bacteriophage λ* viral growth inside the host *E. coli* cell. Further experimental study is needed to reach any realistic conclusion. Also, in all the three cases of repressor proteins, we also noticed a sharp increase in number of binding sites with a small increase in the value of binding energy difference (Figure-1). This sharp increase in number of binding sites with increase in BED is due to increase in possibility of random selection/mutation/replacement in/of cognate 17-base-pair/16-base-pair sequences to which repressor proteins binds.

CONCLUSION

On the basis of calculated binding energy differences (BEDs), it was found that repressor proteins have only few specific as well as quasi-specific binding sites within their cognate genomes, whereas in case of the foreign genomes the number of these sites is relatively large. Out of the three repressor proteins studied, Cro-repressor has the highest number of specific and quasi-specific binding sites inside *Bacteriophage λ as well as E. coli* genome. As the quasi-specific binding sites have binding energy very close to that of specific binding sites of proteins, hence they may be very close competitor for proteins bindings. Thus, their presence in large number may distract a protein from binding to its natural specific site within a reasonable time period. This would adversely affect the process of transcription/translation and will eventually slow down the natural process inside cell. Thus, nature has evolved in such a manner that a protein has only very few numbers of quasi-specific binding sites within their cognate genome.

However, in case of a foreign genome binding of the protein is not specific to a site and nature does not care about that hence there may have any possible number of quasi-specific sites. Binding of a protein to a part of genome does depend upon other factors along with the binding energy. Hence further experimental studies are needed to confirm this result. Also, the sharp increase in number of binding sites with increase in BED is due to increase in possibility of random selection/mutation in/of the cognate 17-base-pair/ 16-base-pair sequences to which repressor proteins binds.

REFERENCES

1. **Berg O. G., & von Hippel P. H. 1988**. Selection of DNA binding sites by regulatory proteins. *Trends in Biochemical Sciences,* **13(6):** 207–211. doi:10.1016/ 0968-0004(88)90085-0.

2. **Siggers T., & Gordan R. 2013.** Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Research,* **42(4):** 2099–2111. doi:10.1093/nar/gkt1112.

3. **Tjian R. 1978**. The binding site on SV40 DNA for a T antigen-related protein. *Cell,* **13(1):** 165–179. doi:10. 1016/0092-8674(78)90147-2.

4. **Halford S. E. 2004.** How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Research,* **32(10):** 3040–3052. doi:10.1093/nar/gkh624.

5. **Berg O. G., Winter R. B., & von Hippel P. H. 1982.** How do genome-regulatory proteins locate their DNA target sites? *Trends in Biochemical Sciences*, **7(2):** 52–55. doi:10.1016/0968-0004(82)90075-5.

6. **Stanford N. P. 2000.** One- and three-dimensional pathways for proteins to reach specific DNA sites. *The EMBO Journal,* **19(23):** 6546–6557. doi:10.1093/ emboj/19.23.6546.

7. **Berg O. G., and P. H. von Hippel. 1987.** Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193(4):**723–743. doi.org/ 10.1016/0022-2836(87)90354-8.

8. **Takeda Y., Sarai A., & Rivera V. M. 1989.** Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proceedings of the National Academy of Sciences*, **86(2):**439–443. doi:10.1073/ pnas.86.2.439.

9. **Sarai A., & Takeda Y. 1989.** Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proceedings of the National Academy of Sciences,* **86(17):** 6513-6517. doi:10.1073/pnas.86.17.6513.

10. **Albright R. A., & Matthews B. W. 1998.** How Cro and λ-repressor distinguish between operators: The structural basis underlying a genetic switch. *Proceedings of the National Academy of Sciences,* **95(7):** 3431–3436. doi:10.1073/pnas.95.7.3431.

11. **Naiya G., Raha P., Mondal M. 2016.** Conformational selection underpins recognition of multiple DNA sequences by proteins and consequent functional actions. *Physical Chemistry Chemical Physics,* **18(31):** 21618–21628. doi:10.1039/c6cp0327 8h.

12. **Chakrabarti J, Chandra N, Raha P. 2011.** High-Affinity Quasi-Specific Sites in the Genome: How the DNA-Binding Proteins Cope with Them. *Biophysical Journal.* **101:**1123-1129. doi: 10.1016/j.bpj.2011.07. 041.

13. **Albright RA, Matthews BW. 1998.** How Cro and λ-repressor distinguish between operators: The structural basis underlying a genetic switch. *Proceedings of the National Academy of Sciences.* **95(7):**3431-3436. doi.org/10.1073/pnas.95.7.3431.

14. **Wartell R. M., Adhya S. 1988.** DNA conformational change in Gal repressor-operator complex: involvement of central G-C base pair(s) of dyad symmetry. *Nucleic Acids Research.* **16(24):**11531-11541. doi.org/10.1093/ nar/16.24.11531.

15. **Ohlendorf D. H., Anderson W. F., Lewis M. 1983.** Comparison of the Structure of Cro and λ repressor Proteins from *Bacteriophage* λ. *Journal of molecular biology.* **169(3):**757-769. doi.org/10.1016/S0022-2836(83)80169-7.

***