# Contribution of half-site spacing in bZIP-DNA recognition

Sarkar A.K., Sarkar A., Lahiri A.

*University of Calcutta, Kolkata, India*

albmbg@caluniv.ac.in

The basic leucine zipper (bZIP) proteins form a superfamily of transcription factors that bind to their target DNA sequences as homo- or heterodimers. Although the prototypical bZIPs like GCN4, cJUN and CREB share considerable sequence similarities in their basic and hinge regions, they interact with the AP-1 DNA site in a significantly different manner. While AP-1 is considered a cognate site for the GCN4 and the cJUN proteins, the CREB protein poorly recognizes it. To understand the role of the basic and the hinge regions in recognizing the DNA sites, we have constructed molecular chimeras *in silico* by swapping the basic and the hinge regions of GCN4 with those of cJUN and CREB. We then used molecular dynamics simulations to analyse the interaction of the native GCN4 homodimer and its two chimeric constructs with the AP-1 sequence. Our results showed features which indicated more stable associations between the GCN4 and the GCN4-cJUN constructs with the AP-1 site. MM-GBSA calculations also indicated less favourable free energy of interaction for the GCN4-CREB-AP-1 model compared to those of the GCN4-AP-1 or GCN4-cJUN-AP-1 models. Remarkably, in our simulations, the strongest structural destabilization occurred in the hinge region, indicating a possible role for this region in bZIP-DNA recognition.

*Key words: DNA-protein interaction, MM-GBSA, chimeric bZIP, GCN4-AP-1 complex, basic and hinge region.*

## 1. Introduction

Evolutionarily, bZIPs are ancient proteins as they seem to have appeared at the very early stage of eukaryotic evolution [1]. It was demonstrated that they evolved independently in each major eukaryotic lineage [1]. Although the bZIP-DNA interaction is highly specific in nature, it offers a certain level of flexibility in the selection of the DNA sites. Traditionally, bZIPs were classified into various families on the basis of the names of the binding sites with which they interacted [2]. The bZIPs belonging to the AP-1 family include the GCN4 protein, a yeast transcriptional activator that regulates around 35 genes [3] mainly responsible for regulating amino acid biosynthesis and cJUN, a member of the transcription factor complex in eukaryotes, which regulates cell proliferation in response to the external stimuli [4]. CREB, a bZIP transcription factor that conveys the cAMP mediated responses to the genes that play a role in controlling circadian rhythm, memory, learning, and reproduction [5], was considered to belong to the CREB/ATF family as it was known to interact with the CRE DNA sequence [2].

In spite of the apparent structural simplicity of the bZIP-DNA complexes, the mechanism of the selection of DNA sites by the bZIPs is still an open area for investigation. In an investigation involving plant bZIP proteins, Niu et al. further divided the non-zipper region of the bZIP domain into N-terminal, core basic and hinge regions to better understand the role of specific clusters of amino acids [6]. It was observed from their *in vivo* domain swapping experiments on the EmBP-1 and TGA1a proteins, which bind to G-box and C-box sequences respectively, that neither the core basic region nor the hinge region was sufficient for determining the specificity towards a G-box or C-box DNA sequence. Rather, it was demonstrated that these two regions in combination played an important role in determining the binding specificity to the cognate DNA site [6]. In spite of the fact that the hinge region did not make any direct contact with the DNA, they observed that their amino acid combinations were highly conserved and replacing the hinge region of EmBP-1 with that of TGA1a decreased the G-box binding affinity of EmBP-1.

Surprisingly, there are only a very limited number of reports on the MD simulation study of the bZIP-DNA complexes, in spite of its widespread occurrence, importance, and structural beauty [7-10]. In this study, we have tried to detail in what ways, in a microscopic sense, the interactions of the DNA binding domains of the three bZIPs with the same AP-1 site differ from one another.

To that end, we have performed *in silico* domain swapping to generate two chimeric constructs of GCN4 by replacing the basic and the hinge regions with the corresponding sequences found in the cJUN (named as GCN4-cJUN) and CREB (named as GCN4-CREB) and investigated whether the hinge and the basic regions of the proteins determined the favourable or unfavourable nature of the interaction with the AP-1 site.

## 2. Computational Details

### 2.1. System preparation

The crystal structure of the GCN4 homodimer complexed with the AP-1 site (PDB ID: 1YSA) was subjected to structural manipulations such as removing crystal waters from the crystal structure and renumbering of residue indices. These were carried out using suitable utilities of PyMOL [PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC] and Chimera [11].

The basic and the hinge regions of the GCN4 were identified by multiple sequence alignment and UniProt [http://www.uniprot.org/] annotation (Fig. 1A) and following that, the chimeric models were constructed by swapping the hinge and the basic regions of the GCN4-AP-1 complex with the corresponding sequences from the cJUN and the CREB transcription factors. These *in silico* domain swapping exercises were carried out using suitable utilities of the FoldX suite of programs [12, 13]. Briefly, the crystal structure of the GCN4-AP-1 complex (PDB ID: 1YSA) was first repaired to remove any short contact and then the amino acids of its basic and hinge region were swapped by the corresponding domain of the two bZIP proteins cJUN and CREB respectively using the mutation option of FoldX [12]. Five models were generated in each swapping operation that were then repaired using the FoldX suite [12]. The complex having the lowest energy was considered as a representative model for each chimeric system. In this way we generated three separate sets of coordinates for the bZIP-DNA complexes, one was for the native GCN4-AP-1 complex and the two chimeric models that were generated from it and named as GCN4-cJUN-AP-1 (the complex with the hinge and basic region sequences of cJUN) and the GCN4-CREB-AP-1 (the complex with the hinge and basic region sequences of CREB). These three bZIP-DNA complexes were then manipulated using PyMOL [PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC] to add the ACE and the NHE groups at the N and C termini respectively.

### 2.2. Molecular Dynamics simulation

All the three systems namely, GCN4, GCN4-cJUN and GCN4-CREB, in complex with the AP-1 DNA site, were separately solvated in a truncated octahedral box of TIP3P water such that the minimum distance between the box edge and the solute's surface was 10Å. All molecular dynamics (MD) simulations were carried out using the MD simulation suite AMBER12 [14]. The ff99SB-ildn [15] force field coupled with the parmbsc0 modification [16] was used to model the protein and the DNA parts, respectively.

All the systems were neutralized with the requisite number of $Na^+$ ions. Post neutralization, 500 cycles of steepest descent followed by 1500 cycles of conjugate gradient minimization were carried out by holding the heavy atoms of the solute with 500 kcal/mol.Å$^2$ restraint force. Thereafter, an unrestrained energy minimization was carried out by performing 1000 cycles of steepest descent, followed by 3000 cycles of conjugate gradient steps. After minimization, the systems were equilibrated in two steps: an initial 200 ps NVT simulation was carried out to slowly heat the systems from 0K to 300K restraining the heavy atoms of the solute with 10 kcal/mol.Å$^2$. This was followed by 20 ps of unrestrained NPT simulation to equilibrate the pressure in the systems. Subsequently, production simulations were carried out for 100 ns maintaining the same conditions. During the simulation, the covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm with the default tolerance of 0.00001 Å [17]. Langevin thermostat with random velocity scaling using a collision frequency 1 ps$^{-1}$ was used to control the temperature at 300 K. The particle mesh Ewald (PME) summation method with 10Å real space cut-off was used to calculate the electrostatic interactions [18]. The van der Waals forces were truncated beyond the cutoff of 10Å. Isobaric condition was maintained by turning on the isotropic position scaling and was set at 1 atm. A simple leapfrog integrator was used to propagate the dynamics with a time step of 0.002 ps (2 fs).

### 2.3. Analysis

All analyses were carried out with suitable utilities of AmberTools 15. The backbone RMSD of the proteins was calculated for the backbone atoms C, CA and N, after fitting all the heavy atoms to the corresponding unrestrained minimized structures. Backbone order parameter was calculated according to the methodology proposed by Zhang et al. [19] using the suitable utility of the application s2 (http://spin.ccic.ohio-state.edu/index.php/). The analytical relationship for the estimation of the order parameter ($S^2$) of N-H bond vector of the *i*-th amino acid was taken as

$$S_i^2 = \tanh\left(0.8\sum_k\left(\exp\left(\frac{-r_{i-1,k}^0}{1\text{Å}}\right)+\exp\left(\frac{-r_{i,k}^{\text{H}}}{1\text{Å}}\right)\right)+b,\right.$$

where $r_{i-1,k}^0$ was the distance between the carbonyl oxygen of the ($i$ – 1)-th amino acid to heavy atom $k$ and $r_{i,k}^{\text{H}}$ was the distance between the amide proton H and heavy atom $k$. The parameter $b$ was set to –0.1, which took into account the order parameters of rigid protein. The sum ranged over all heavy atoms $k$ that did not belong to amino acids $i$ and $i$ – 1.

Principal components (PCs) were obtained from the covariance matrix of the Cartesian coordinate data set for the $C_\alpha$ atoms. Eigenvectors and eigenvalues were derived from this matrix, and then the eigenvectors were ranked as PC1 to PC$n$, $n$ is any natural number, according to their eigenvalues in decreasing order. The first PC contained the highest proportion of variance in the data. We used the R package bio3D [20] for PC analysis (PCA).

### 2.4. MM-GBSA Calculations

We have employed the Molecular Mechanics with Generalized Born and Surface Area (MM-GBSA) approximation method to evaluate the relative change in the binding free energy of the protein−DNA complex.

We used the single trajectory method to reduce the noise in the MM-GBSA calculation [21]. In this method, a simulation is carried out with the complex only and the interaction free energy, $\Delta G_{bind}$, will be composed of four terms: the electrostatic interaction energy ($\langle \Delta E_{elec} \rangle$), the van der Waals energy ($\langle \Delta E_{vdW} \rangle$), the solvation energy ($\langle \Delta G_{solvation} \rangle$), and the solute entropic contribution ($T\Delta S$):

$$\Delta G_{bind} = \Delta E_{vdW} + \Delta E_{elec} + \Delta G_{solv} - T\Delta S \qquad (8)$$

In MMGBSA, the $G_{solv}$ calculation is carried out by the Generalized Born solvation model developed by Hawkins et al. 1995 ($GB^{HCT}$) [22] implicitly considering the ionic strength of the monovalent ion to correspond to 100 mM. Entropy was calculated by using the quasi-harmonic approximation [23].

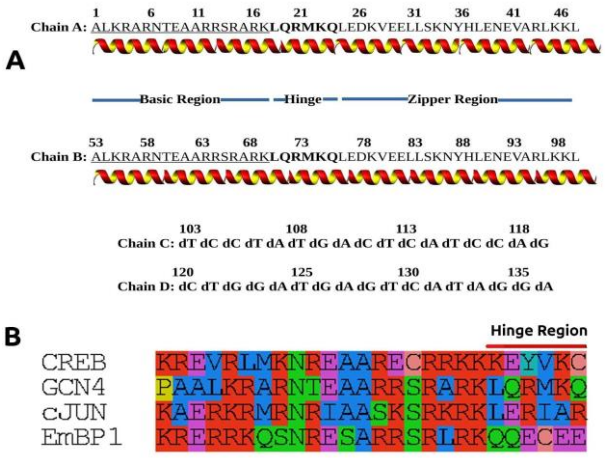


**Fig. 1.** Details about the systems used in this study. **A**. GCN4 and the AP-1 DNA sequences with the residue numbering. In the case of GCN4, the underlined and bold regions are the basic and hinge regions, respectively, which were swapped *in silico* to generate chimeric bZIPs like GCN4-cJUN and GCN4-CREB. **B**. Aligned sequences of the basic region of GCN4, CREB, cJUN and EmBP1. The basic and the hinge regions of GCN4 were substituted in these sequences to generate the GCN4-cJUN and GCN4-CREB models.

## 3. Results and Discussion

### 3.1. Compatibility of the basic region within the AP-1 site: Backbone dynamics

We observed a large change in the RMSD of the GCN4-CREB chimera as compared to the GCN4-cJUN chimera or the native GCN4 protein, in their respective complexes with the AP-1 site (Figures 2A and B). The backbone RMSD values of the GCN4-CREB chimera increased after ~5ns of simulation and attained a maximum value of around 6Å for both the protein chains (Fig. 2A and Fig. 2B, respectively). It suggested that the GCN4-CREB chimeric protein in its complex with AP-1 DNA might have undergone a major structural change as compared to the native GCN4-AP-1 or GCN4-cJUN-AP-1 complexes.

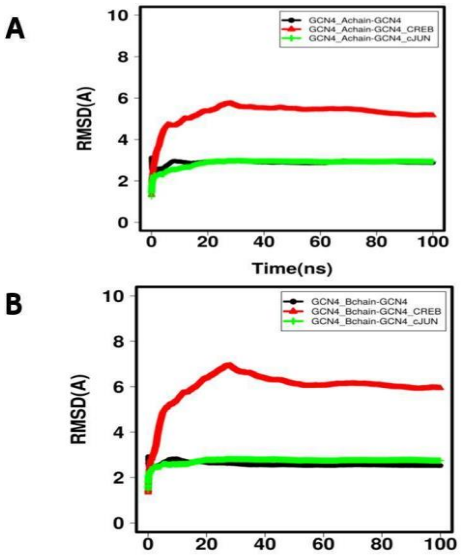


**Fig. 2**. Backbone RMSD of the three bZIPs in the presence of the AP1 sequence. Backbone RMSDs of the GCN4 (black), GCN4-CREB (red) and GCN4-cJUN (green) were measured after fitting the whole bZIP-DNA complexes with their respective reference structure (energy minimized). **A**. RMSD of chain A, **B**. RMSD of chain B.

To understand the possible consequences of the increased flexibility of the amide backbone in the GCN4-CREB chimera, we computed the residue-wise secondary structure formation for all the three models of the protein-DNA complex (Fig. 3). The secondary structure plot indicated that, except for the GCN4-CREB chimera, the rest of the two bZIPs, namely native GCN4 and GCN4-cJUN, maintained a continuous α-helical structure, a characteristic of the bZIP domain. On the other hand, the model of the GCN4-CREB protein showed a loss of helical structure in both the chains. The loss of α-helical structure was mainly restricted to the residues in the hinge regions. It was also noted that, in the case of the GCN4-CREB chimera as well as in the native CREB protein, this hinge region had a relatively higher number of charged residues as compared to the native GCN4, the GCN4-cJUN chimera or the native cJUN (Fig. 1B). Also, the time evolution analysis of residue-wise secondary structure formation indicated that structural transitions were not symmetric in nature, i.e., conversion occurred to different extents in the two helices of the GCN4-CREB homodimer (Fig. 3). While the disruption of helical conformation was restricted within residues ranging from 20–22 in the case of chain A, in the case of chain B, the range was from 72–76, which was relatively large. The loss of the helical nature of the protein backbone in the case of the GCN4-CREB chimera lead to a distorted bZIP-DNA complex, which seemed to be structurally different from the GCN4 or

the GCN4-cJUN complex (Fig. 4) and the structural destabilization may play a role in weakening the bZIP-DNA interaction in the case of the GCN4-CREB chimera.

Based on the scree plot (Fig. 5), we selected 6 PCs that captured ~80 % of the variance in the conformational ensemble. We observed that in the case of the GCN4-CREB model, the PC1 captured ~50 % of the variance as compared to the GCN4 or the GCN4-cJUN chimera. In the latter cases only ~30 % of the variance was captured by the PC1. Next, we calculated the residue-wise contribution to the 6 PCs (Fig. 6). Three distinct residue-wise contributions were observed, namely from the DNA binding region, from the leucine zipper region and from the hinge region, in all the 6 PCs (Fig. 6). In contrast to the DNA binding and the zipper region, the contributions from the hinge region were found to be considerably different between GCN4-CREB and GCN4 and GCN4-cJUN. It indicated that the residues in the hinge region of the GCN4-CREB model had different dynamics as compared to the GCN4 and GCN4-cJUN.

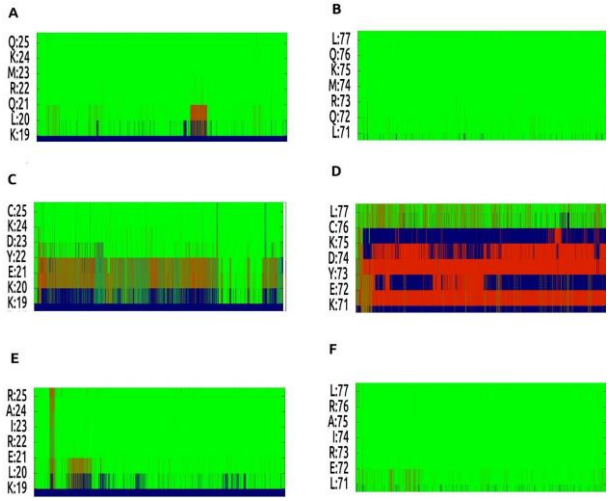


**Fig. 3.** Time evolution of residue-wise secondary structure formation for the residues of the hinge region of chain A and B for GCN4 (**A** and **B**), GCN4-CREB (**C** and **D**), GCN4-cJUN (**E** and **F**). Different secondary structures were colored as: alpha-helix (green), bend (red), and any other secondary structure (blue).

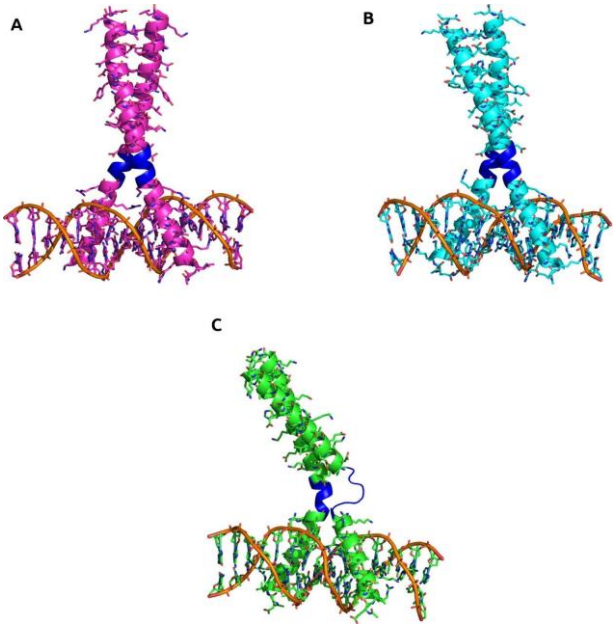


**Fig. 4.** Representative picture of the **A**. GCN4-AP-1, **B**. GCN4-cJUN-AP-1 and **C**. GCN4-CREB-AP-1 complexes. Snapshots are taken from the last frame of the simulation and the hinge region is highlighted in blue color.

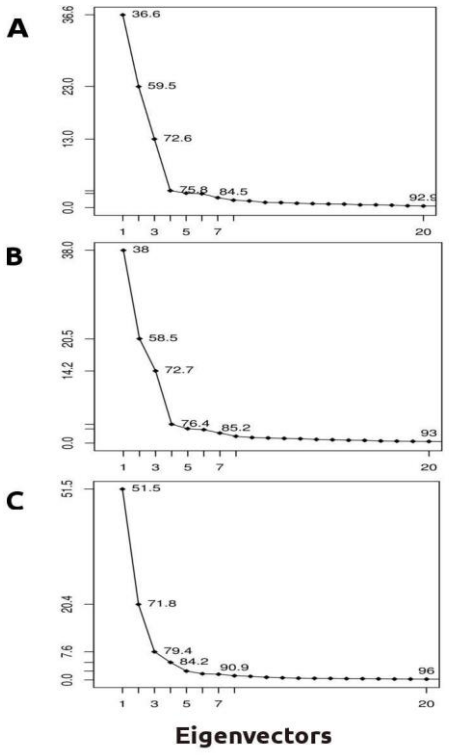


**Fig. 5.** Scree plots for A. GCN4, B. GCN4-cJUN and C. CGN4-CREB in complex with AP-1. The y-axis represents the percentage (%) of variance of the respective ensembles captured by the corresponding eigenvectors.

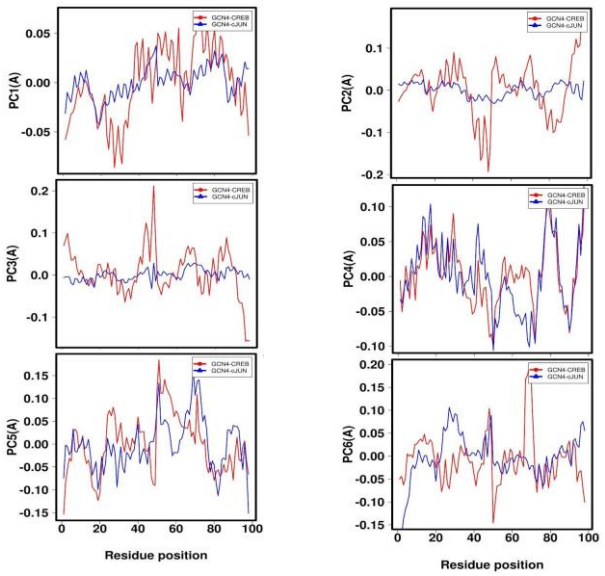


**Fig. 6.** Difference in residue-wise PCA contributions with respect to the GCN4-AP-1 complex: GCN4-cJUN (blue) and CGN4-CREB (red) in complex with AP-1.

### 3.2. Binding energy calculation

To analyze the energetics of the three bZIP-AP-1 complexes, the various components of the interaction free energy ($\Delta G_{binding}$) were evaluated using the entire 100 ns simulation time. The detailed results of the analysis are presented in Table 1. In Fig. 7A we have depicted the entropy and enthalpy terms and in Fig. 7B we have shown the $\Delta G_{binding}$ for the GCN4, GCN4-cJUN and GCN4-CREB in complex with AP-1. In comparison with the GCN4 and GCN4-cJUN, the complex formation between the GCN4-CREB and AP-1 was found to be unfavorable enthalpically and entropically (Fig. 7A). In order to understand the reason for this unfavorable energetics for the GCN4-CREB-AP-1 complex, we considered the individual energy terms listed in Table 1. We found that the bZIP-AP-1 interaction was an enthalpy driven process. The total solvation energy, $\langle \Delta G_{solvation} \rangle$ which was composed of polar ($\langle \Delta G_{GB,electrostatics} \rangle$) and non-polar ($\langle \Delta G_{GB,non-polar} \rangle$) terms, was found to be highly unfavorable for all the three complexes (Table 1). This large positive value of $\langle \Delta G_{solvation} \rangle$ was due to one of its component $\langle \Delta G_{GB,electrostatics} \rangle$ which had a large positive value in all the three complexes. This unfavourable solvation was compensated by the electrostatic interaction energy ($\langle \Delta E_{electrostatic} \rangle$). The molecular mechanics electrostatic term favored the bound state as $\langle \Delta E_{electrostatic} \rangle < 0$ and compensated the $\langle \Delta G_{GB,electrostatics} \rangle$. The sum of the $\langle \Delta E_{electrostatic} \rangle$ and the $\langle \Delta G_{GB,electrostatics} \rangle$ can be thought of as a parameter that represented the total electrostatic energy which was found to favour the formation of the bZIP-AP-1 complexes. Interestingly, this value was relatively favorable (~ –84 Kcal/mol) in GCN4-AP-1 and GCN4-cJUN-AP-1 while it was ~ –45 Kcal/mol in case of GCN4-CREB. Since it seems to have a major contribution to the thermodynamics of the bZIP-AP-1 complex formation, the complex formation for the GCN4-CREB-AP-1 seems to be quite unfavorable.

**Table 1**. Components of the bZIP-AP-1 interaction free energy (kcal/mol)

| Components (kcal/mol) | GCN4-AP-1 | GCN4-cJUN-AP-1 | GCN4-CREB-AP-1 |
|---|---|---|---|
| $\langle \Delta E_{electrostatics} \rangle$ | –9783.43 | –11129.47 | –5875.13 |
| $\langle \Delta E_{vdW} \rangle$ | –107.28 | –125.45 | –103.57 |
| $\langle \Delta G_{GB,\ solvation\_electrostatics} \rangle$ | 9699.00 | 11043.04 | 5829.28 |
| $\langle \Delta G_{solvation\_non\text{-}polar} \rangle$ | –19.12 | –20.90 | –16.95 |
| $\langle \Delta G_{solvation} \rangle$ | 9679.88 | 11022.1 | 5812.33 |
| $\langle \Delta H \rangle$ | −210.83 | −232.82 | −166.37 |
| $T\Delta S$ (at 298 K) | –66.78 | –73.65 | –56.09 |
| $\Delta G_{binding}$ | –144.36 | –159.13 | –110.29 |
| $\Delta\Delta G_{binding}$ (Mutant-WT) | 0 | −14.77 | 34.07 |

## 4. Conclusions

Explicit solvent molecular dynamics simulations of the native GCN4-AP-1 model, the cJUN basic and hinge region substituted GCN4-cJUN-AP-1 model and the CREB basic and hinge region substituted GCN4-CREB-AP-1 model revealed a considerable difference in the backbone dynamics for the GCN4-CREB in comparison with the GCN4 and GCN4-cJUN models in their interaction with the DNA AP-1 site. Secondary structure analysis showed a transition of the peptide chain from its native α-helical structure to turns and bends in the same region. We also explored the principal components of the dynamics of the bZIPs and found that the contribution of the hinge region towards the PC space was considerably different in the GCN4-CREB model in comparison with that of the GCN4 and GCN4-cJUN models.

The overall free energy change due to complex formation as calculated using the MM-GBSA approximation indicated destabilization of the GCN4-CREB-AP-1 complex compared to the other two. We found that due to the relatively less $\langle \Delta E_{electrostatic} \rangle$ contribution, the GCN4-CREB-AP-1 complex was energetically unfavorable. Overall, although the basic region maintained the helical secondary structure during the timespan of this simulation, consideration of detailed interactions of the basic region with the DNA site was sufficient to gain insight into the DNA binding specificity of the bZIPs. The hinge region also seemed to play a role in the stability of binding, but the precise mechanism was not clear from this study.

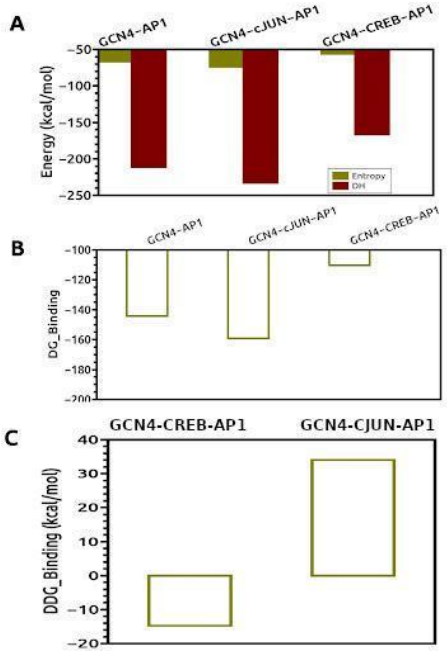


**Fig. 7.** (**A**) Average $\Delta H$ and entropy of binding, (**B**) $\Delta G_{binding}$ and (**C**) relative change in $\Delta G_{binding}$ ($\Delta\Delta G_{binding}$) for the GCN4 and its two chimera GCN4-cJUN and GCN4-CREB. $\Delta G_{binding}$ is calculated using the MM-GBSA method.

## 5. Acknowledgement

This research work was partially supported by the departmental DST-FIST and UGC-DSA programs. The UGC RFSMS and CSIR SRF programs are also acknowledged for providing fellowship to one of the authors (AKS).

## 6. Reference

1. Jindrich K., Degnan B.M. *BMC Evol. Biol*. 2016. V. 16. P. 28. doi: 10.1186/s12862-016-0598-z
2. Miller M. *Curr. Protein Pept. Sci*. 2009. V. 10. P. 244–269.
3. Hope I.A., Struhl K. *EMBO J*. 1987. V. 6. P. 2781−2784.
4. Wisdom R., Johnson, R.S., Moore C. *EMBO J*. 1999. V. 18. No. 1. P. 188–197.
5. Shaywitz A.J., Greenberg M.E. *Annu. Rev. Biochem*. 1999. V. 68. P. 821–861.
6. Niu X., Renshaw-Gegg L., Miller L., Guiltinan M.J. *Plant Mol. Biol*. 1999. V. 41. P. 1–13.
7. Cukier R.I. *J. Phys Chem B*. 2012. V. 116. P. 6071–6086.
8. Robustelli P., Trbovic N., Friesner R.A., Palmer III A.G. *J. Chem. Theory Comput*. 2013. V. 9. P. 5190−5200.
9. Choi Y.H., Yang C.H., Kim H.W., Jung S. *Bull. Korean Chem. Soc*. 1999. V. 20. P. 1319–1322.
10. McHarris D.M., Barr D.A. *J. Chem. Inf. Model*. 2014. V. 54. P. 2869−2875.
11. Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. *J. Comput. Chem*. 2004. V. 25. P. 1605–1612.
12. Schymkowitz J., Borg J., Stricher F., Nys R., Rousseau F., Serrano L. *Nucleic Acids Res*. 2005. V. 33. P. W382–W388.
13. Yong Z.H., Gui M., Chun Peng Z., Yao Sheng C. *Science China* (*Life sciences*) 2011. V. 54. P. 442–449.
14. Case D.A. et al. *AMBER 2015*. San Francisco: University of California, 2015.
15. Lindorff-Larsen K., Piana S., Palmo K., Maragakis P., Klepeis J.L., Dror R.O. Shaw D.E. Proteins. 2010. V. 8. P. 1950–1958.
16. Pérez A., Marchán I., Svozil D., Sponer J., Cheatham T.E., Laughton C.A., Orozco M. *Biophys J*. 2007. V. 92. P. 3817–3829.
17. Ryckaert J.P., Ciccotti G., Berendsen H.J.C. *J. Comput. Phys*. 1977. V. 23. P. 327−341.
18. Darden T., York D., Pedersen L. *J. Chem. Phys*. 1993. V. 98. P. 10089−10092.
19. Zhang F., Bruschweiler R. *J. Am. Chem. Soc*. 2002. V. 124. P. 12654–12655.
20. Grant B.J., Rodrigues A.P.C., ElSawy K.M., McCammon J.A., Caves L.S.D. *Bioinformatics*. 2006. V. 22. P. 2695−2696.
21. Hou T., Wang J., Li Y., Wang W. *J. Chem. Inf. Model*. 2011. V. 51. P. 69–82.
22. Hawkins G.,D., Cramer C.J., Truhlar D.G. *Chem. Phys. Lett*. 1995. V. 246. P. 122–129.
23. Karplus M., Kushick J. *Macromolecules*. 1981. V. 14. P. 325–332.