**FUNCTIONAL RELATIONSHIP BETWEEN AMINO ACID RESIDUES AT N- AND C-TERMINI OF DNA-BINDING REGIONS OF TRANSCRIPTION FACTORS CREB AND AP-1 REVEALED BY ANALYZING THE PAIR CORRELATIONS OF AMINO ACID SUBSTITUTIONS.**

AFONNIKOV D.+, TITOV I.I.

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 10 Lavrentiev Ave., Novosibirsk, 630090, Russia;
e-mail: ada@bionet.nsc.ru;

+Corresponding author

Keywords: transcription factor, DNA-binding, bZIP domain, amino acid substitutions, correlation analysis

Transcription factors of the CREB and AP-1 families bind to the DNA sites CRE and AP-1, respectively, and contain DNA-binding domains of bZIP class [1]. The bZIP domains include the region rich in basic amino acids and a leucine zipper, an amphipathic a -helix, immediately at its
C-terminus. Basic region directly binds DNA and the leucine zipper serves for dimerization of proteins. X-ray structure data demonstrate [2-4] that the bZIP-DNA complexes have a similar spatial structure within this protein family. In this work, we tried to reveal certain peculiarities of amino acid residue interactions in the DNA-binding region of the CREB and AP-1 family proteins using the correlation analysis of amino acid substitutions at the protein position pairs within this region.

To estimate the statistical relation between the residue substitutions we used a linear correlation coefficients approach. The approach estimates the correlation coefficient between the values of a certain physicochemical quantity for the residues observed at the positions of the set of aligned proteins. We analyzed a set of 42 proteins from the CREB and AP-1 transcription factor family. The protein sequences were taken from the SWISS-PROT databank. The isoelectric point values for the amino acids were chosen for our analysis as physicochemical characteristic reflecting the residue charges [5]. Low variable positions (containing not more than two different amino acids in their alignment columns) were excluded from the analysis.

The analysis of the pairwise positional correlations revealed the negative correlations between the isoelectric point values in the pairs formed by positions from the N- and C-termini of the basic region. The highest correlations were observed for the pairs formed by positions 227-229 at the N-terminus and 248-251 at the C-terminus of the domain (alignment positions are numbered according to protein GCN4; Fig. 1).



Figure 1. Amino acid sequence of the basic region of the GCN4 transcription factor bZIP domain. Low-variable positions are underlined.

We suggested that such correlations contribute to the conservation of the total isoelectric point value at the positions in the N- and C-termini of the region in question.

We performed an additional analysis of our protein set to verify this hypothesis (similar approach was used by Lim and Ptitsyn to confirm the conservation of hydrophobic core volume of globins [6]). We calculated the sums of the isoelectric point values in groups of the N- and C-terminal positions of the region (*Q*nand*Q*c, respectively) for each sequence of the set. Different groups of positions were considered: 227-229 and 227-230 at the N-terminus; 248-252, 247-252, and 247-251 at the
C-terminus; totally, six pairs from the groups of N- and C-terminal positions were examined.

The total isoelectric point value at the N- and C- domain termini*Q*+=*Q*n+*Q*cwere calculated for each combination considered. We used the value of sample variance as a measure of*Q*+conservation for the protein set analyzed.

The degree of*Q*+conservation was estimated through a Monte-Carlo test. We used initial DNA-binding regions alignment to generate 50,000 "artificial protein" sets through random amino acid permutations in the alignment columns. Thus, we obtained randomized sets with the amino acid frequencies at the positions that were identical to the frequencies in actual proteins but lacking the correlations between the positions within the region. The value of the sample dispersion D(*Q*+) in the initial set was compared with the corresponding values for the randomized sets. The values*Q*-=*Q*n-*Q*cand the linear correlation coefficient between the*Q*nand*Q*cvalues (*r*(*Q*+,*Q*-)) were analyzed in a similar way.

The value D(*Q*+) in the real set appeared to be essentially lower than its mean estimation over the randomized sets for all the above-mentioned combinations. For example, the D(*Q*+) value for the groups of positions 227-229 and 249-252 was lower than the mean over the randomized sets by almost 2.7 standard deviation values (Table 1). The dispersion value*Q*+was lower than the D(*Q*+) in the real set for only 19 randomized sets of the 50,000 analyzed. These results indicate that the existing correlations between amino acid substitutions in the pairs of N- and C-terminal positions of the DNA-binding region of transcription factors CREB and AP-1 provides the conservativeness of the*Q*+value in the proteins of these families.

Table 1. Comparison of the sample dispersions*D*(*Q*+) and*D*(*Q*-) and the correlation coefficient*r*(*Q*+,*Q*-) for the groups of positions at the N (227-229) and C (249-252) termini of the basic region of the real and randomized sets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Xorig |  x.gif (66 bytes)rand | *s*(Xrand) | N(Xrand>Xorig) |
| D(*Q*+) | 18.07 | 40.53 | 8.16 | 49981 |
| D(*Q*-) | 77.03 | 40.47 | 8.17 | 5 |
| *r*(*Q*+,*Q*-) | -0.62 | 0.00 | 0.16 | 0 |

The corresponding values in the original set (*X*orig), the mean (rand) and root mean square (*s*(*X*rand)) deviations estimated according to the randomized set realizations are listed in the table as well as the number N(Xrand>Xorig) of randomized sets wherein the corresponding value exceeded the values for the original set (absolute value in the case of correlation coefficient).In addition, the dispersion of*Q*-in the real protein set appeared to be essentially higher than its mean over the randomized sets. This fact indicates a considerable variability of the values*Q*in the real proteins CREB and AP-1.

The results described above along with the high absolute value of the correlation coefficient*r*(*Q*+,*Q*-) suggested the negative dependence between the values*Q*nand*Q*cfor the proteins studied. This dependence is apparent in the scatterplot of*Q*nand*Q*cparameters (Fig. 2).



Figure 2. Scatterplot of*Q*nи Qcvalues (determined for the groups of positions 227-229 and 249-252) for proteins of CREB and AP-1 families. The regression line is shown.

We estimated the parameters of this dependence in a linear approximation (*Q*c=*b*+*a· Q*n) by equations given in [7]. The estimations for regression coefficients amounted to*a*= -1.02;*b*=51.15;*a*values fell into the range -1.62<=*a*<= -0.67 with 95% probability. These estimations have confirmed our suggestion that the dependence of*Q*nand*Q*cis*Q*n+*Q*c*const*.

Since the value of isoelectric point characterizes the charge of amino acids, it is reasonable to suggest that such dependence stem from the electrostatic interactions of the residues in the groups of positions considered. This suggestion is supported by the data on mutations in the bZIP domains of proteins CRE-BP1 and GCN4 [8]. In that work, the amino acids from the N-terminal cluster of the protein GCN4 were substituted with the amino acids of the protein CRE-BP1, and vice versa. The free energy of the complexes of original and mutant proteins with DNA sites was measured. These substitutions [8] caused a considerable deviation of the value*Q*+from the mean for the set we have analyzed and, interestingly, an increase in the free energy of the DNA-mutant protein complexes. Thus, it can be concluded that a considerable deviation of*Q*+from its optimal value correlates with the decrease in bZIP-DNA complex stability.

**Acknowledgements**

We are grateful to Yury Kondrakhin for valuable discussions and to Ms Galina Chirikova for assistance in translation.

This work was supported by grants from the Russian Foundation for Basic Research (No.97-04-49740, 97-07-90309, 96-04-50006, 98-04-49479, 98-07-90126); Russian Ministry of Science and Technologies; Russian Human Genome Project; Russian Ministry of High Education.

**References**

1. W.H. Landshultz, P.F. Johnson and S.L. McKnight, "The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins", Science,**240**, 1759-1764 (1988)
2. T.E. Ellenberger, C.J. Brandl, K. Struhl, Harrison S.C. "The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex", Cell,**71**, 1223-1237, (1992).
3. P. Ko nig, T.J. Richmond "The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility", J. Mol. Biol.,**233**, 139-154 (1993).
4. J.N.M. Glover, Harrison S.C. "Crystal structure of the heterodimeric bZIP transcription factor c-Fos - c-Jun bound to DNA", Nature,**373**, 257-261 (1995).
5. White, P. Handler, E.L. Smith, R.L. Hill, I.R. Lehman, "Principles of Biochemistry, vol.1.", McGraw-Hill, Inc. (1978)
6. V.I. Lim and O.B. Ptitsyn "On the constancy of the hydrophobic nucleus volume in molecules of the myoglobins and hemoglobins", Mol. Biol. (USSR),**4**, 372 (1970).
7. Kendall M.G. and Stuart A. The advanced theory of statistics, vol. 2, Inference and relationship, 2nd ed., Charles Griffin & Co Ltd, London.
8. S.J. Metallo, D.N. Paolella and A. Scheparrtz "The role of a basic amino acid cluster in target site selection and non-specific binding of bZIP peptides to DNA", Nucl. Acids Res.,**25**, 2967-2972 (1997).