A possible mode of the specific recognition of nucleic acids by proteins

LI Xuqing[1,3] & LIU Ciquan[2,4]

1. Department of Computer Science, Kunming University of Science and Technology, Kunming 650041, China;
2. Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China;
3. Yunnan Observatory, Chinese Academy of Sciences, Kunming 650011, China;
4. Modern Biological Center, Yunnan University, Kunming 650091, China

Correspondence should be addressed to Liu Ciquan

Abstract **Seven sets of protein target sites, which occur in several gene promoters, have been analyzed. The results suggest that there is a possible mode of specific recognition of double-helical nucleic acids by proteins. This recognition mode is related to a special topological property of double-helical DNA, which is termed base spatial pattern (BSP) of DNA segment. BSP is the spatial topological property determined only by the spatial arrangement of the bases on double-helical DNA segment.**

Keywords: protein, nucleic acids, recognition, base spatial pattern (BSP), target sites.

Protein-DNA interactions play important roles in many cell processes, such as DNA replication, modification, repair and RNA transcription. The structures of many DNA-binding proteins and their DNA complexes have been determined and, on the basis of these structures, a number of highly conserved DNA-binding motifs have been identified. The intensive studies on protein-DNA interactions have mainly focused on how proteins recognize specific DNA sequences[1]. It is now clear that no code will be found that can describe DNA recognition by all DNA-binding proteins. But it is still possible that rules will emerge for members of a single structural family or for a group of families which interact in similar ways with DNA[2—4].

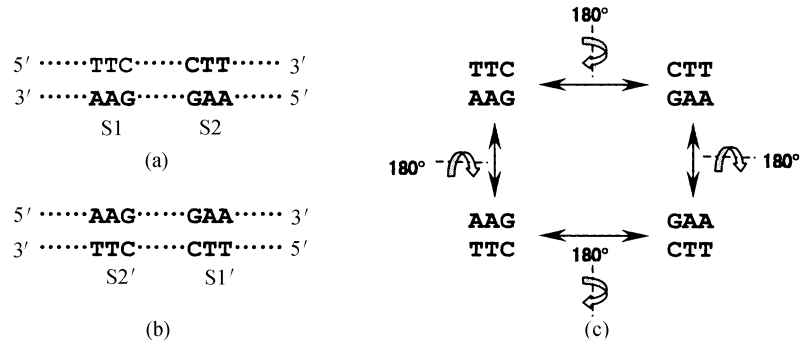

Fig. 1. Schematic diagram for the relationship of four DNA segments S1, S2, S1′ and S2′ and the relationship of their corresponding BSPs.

In this report, we describe a possible mode of specific recognition of nucleic acids by proteins. This recognition mode is related to a special topological property of double-helical DNA, which is termed base spatial pattern (BSP) of DNA segment. We will explain the general features of this recognition mode in terms of the BSP.

1 Method and data

(ⅰ) Base spatial pattern of DNA segment

Definition. For a given segment of double-helical DNA, its corresponding BSP refers to the spatial topological property determined only by the spatial arrangement of the bases on this DNA segment. Also when discussing BSP, the helical features of DNA and the 5′ to 3′ polarity on the two anti-parallel chains of the DNA segment are ignored.

Our reasoning is best explained by an example: In fig. 1(a), there are two DNA segment stretches: S1 and S2. They differ from each other sequentially, but their corresponding BSPs are the same. Rotating the DNA in fig. 1(a) 180° around its transverse axis gives another view (shown in fig. 1(b)). Thus, the DNA in fig. 1(a) and the DNA in fig. 1(b) are identical, they seem different just because they are in different orientation. Based on the traditional linear base sequence analyses, S1=S1′, S2=S2′, S1≠S2 and S1′≠S2′. But from an alternative viewpoint, S1, S2, S1′ and S2′ ave the same BSP; i.e. they present different views of the same BSP.

When searching for its binding site on a DNA sequence, a protein theoretically should be able to approach and identify the site from any direction in three-dimensional space. Rotating a DNA segment in space is equivalent to changing the direction from which this DNA segment is seen. After the rotation, the corresponding BSP itself keeps unchanged, just showing another view of the same BSP. Hence, rotations will be used in the following theorem to decide whether two given DNA segments correspond to the same BSP or not.

Theorem. Two DNA segments with the same number of base pairs have the same BSP, if the BSP of one can be transformed into the BSP of the other by the following transformations of rotation: (a) rotating 180° around its horizontal axis of symmetry; and / or (b) rotating 180° around its vertical axis of symmetry.

Accordingly, the four DNA segments in fig. 1(a), (b) have the same BSP, since one of their corresponding BSPs can be transformed into the others by such rotations (fig. 1(c)).

BSP reflects a special kind of topological property of a double-helical DNA segment. BSP highlights the structural characteristics determined only by the relative spatial arrangement of bases on the DNA segment. From the viewpoint of BSP, some different DNA segments are linked together, i.e. through BSP, the same structural characteristics emerge from some of DNA segments with different base-pair sequences.

(ⅱ) Two sets of DNA target sites for proteins CG and SpGCF1. The two sets of protein target sites occur in the gene *Endo16* promoter of the sea urchin *Sronglocentrotus purpuratus.* This promoter consists of more than 30 regulatory elements dispersed through about 2.3 kb of the upstream sequence. Within this cis-regulation domain, Yuh et al.[5] have mapped target sites for 15 different proteins that bind with high specificity, that is, $\geqslant 10^4$ times their affinity for synthetic double-stranded copolymer of deoxyinosine and deoxycytidine [poly(dI-dC) • poly(dI- dC)]. Yuh et al. also have given the DNA sequence in module A and in basal promoter (Bp) region, and the locations of target sites for some DNA binding proteins[6]. Among those DNA binding proteins, CG and SpGCF1 correspond to more than one DNA target site (fig. 2), while the others correspond to only one DNA target site. We will take the two sets of target sites for CG and SpGCF1 as examples to analyze later. In fig. 2, boxed sequences indicate conserved core elements of the target sites, not the complete target site sequences[6].

(ⅲ) Five sets of DNA target sites for five types of proteins gotten from EMBL Nucleotide Sequence Database.

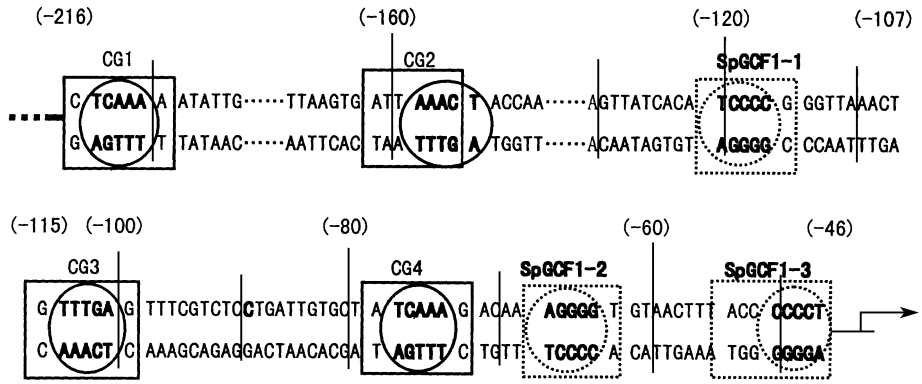

Fig. 2. Partial DNA sequence of module A and the Bp region of the gene *Endo16* promoter of the sea urchin *Sronglocentrotus purpuratus*. Boxed sequences indicate conserved core elements of the target sites with high specificity for transcription factors CG and SpGCF1.

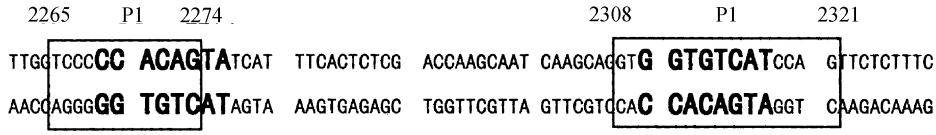

Fig. 3. Two target sites for protein P1 (in boxes). The two sites contain the same BSP (marked by bold characters).

(1) Two sets of DNA target sites for proteins P1 and P18. The two sets of protein-binding sites (figs. 3 and 4) occur in the sequence of the *TZ2 CyIIIa* actin gene regulatory domain of the sea urchin *Sronglocentrotus purpuratus*[7, 8]. The database accession number of the sequence is M64573.

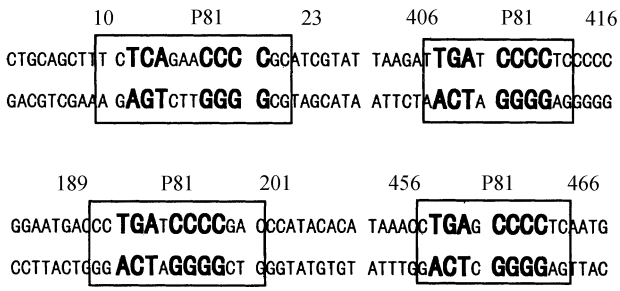

Fig. 4. Four target sites for protein P8I (in boxes). The four sites contain two same sequences (marked by bold characters); this is a special case of containing the same BSP.

(2) Two sets of DNA target sites for proteins LF-A1 and BHLH. These two sets of protein-binding sites (figs. 5 and 6) occur in the sequence of the promoter region and exon Ⅰ of *Bovine* conglutinin gene[9]. The database accession number of the sequence is D25294.

(3) A set of DNA target sites for protein ArgRLF-A1 repressor. This set of protein-binding sites (fig. 7) occurs in the sequence of Streptomyces *clavuligerus* argin ine biosynthesis cluster (argCJBDRGH genes)[10—12]. The database accession numbers of the sequence are Z49111, M83659, Y11134.

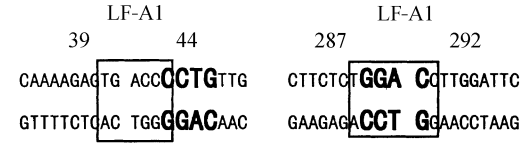

Fig. 5. Two target sites for protein LF-A1 (in boxes). The two sites contain the same BSP (marked by bold characters).

Fig. 6. Three target sites for protein BHLH (in boxes). The three sites contain the same BSP (marked by bold characters).

2 Results and discussion

The nucleotide sequence recognized by a transcription factor is often not a fixed sequence and allows the occurrence of multiple patterns with different levels of degeneracy[13]. In fact, most specific, natural DNA-bind- ing proteins recognize a set of related sequences[14]. What is the common characteristic of those related sequences? Among the different nucleotide sequences bound by the same type of transcription factor, is there, aside from the order of nucleotide sequence, some other common char-

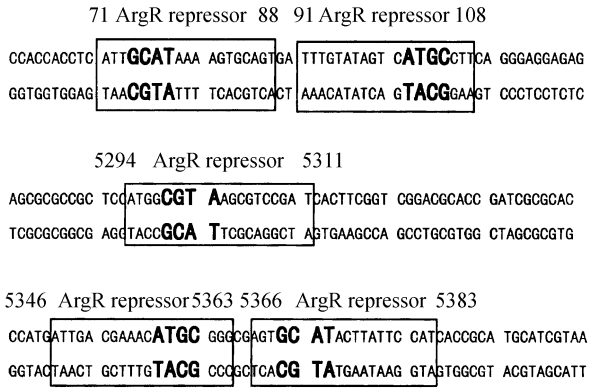

Fig. 7. Five target sites for protein ArgR repressor (in boxes). The five sites contain the same BSP (marked by bold characters).

acteristic recognized by the transcription factor? If so, the common characteristic may vary from type to type of proteins because of the diversity of the structure and characteristic of proteins. Examining, from BSP viewpoint, the seven sets of DNA target sites for the seven types of transcription factors, we found that there is actually a common characteristic related to BSP among the seven sets of DNA target sites, stated differently, for each of the seven sets of DNA target sites, the nucleotide sequence for the same type of transcription factor may be different, but the target sites contain the same type of BSP. We will take the two sets of target sites for transcription factors CG and SpGCF1 as examples to explain the phenomenon.

In fig. 2, the sites labeled CG1, CG2, CG3 and CG4 bind the same type of protein CG. The conserved core elements of the four target sites for CG (boxed by the solid line) contain four 5-basepair DNA segments (circled by the solid line). Based on traditional linear nucleotide sequence analyses, the DNA segments circled by the solid line in CG1, CG3 and CG4 are the same (5′-TCAAA-3′), but are different from the DNA segment circled by the solid line in CG2(5′-AAACT-3′). However, from the viewpoint of BSP, these four DNA segments circled by the solid line correspond to the same BSP.

Examining the conserved core elements of the three target sites for transcription factor SpGCF1 in fig. 2 (boxed by the dashed line) gives similar conclusion. These three target sites also contain three 5-basepair DNA segments (circled by the dashed line). Based on traditional linear nucleotide sequence analyses, the DNA segments circled by the dashed line in SpGCF1-2 and SpGCF1-3 are the same (5′-AGGGG-3′), but they are different from the DNA segment circled by the dashed line in SpGCF1-1 (5′-GGGGA-3′). However, from the viewpoint of BSP, these three DNA segments circled by the dashed line correspond to the same BSP.

The five sets of DNA sequences in figs. 3—7 are the five sets of DNA target sites for the five types of proteins (P1, P8I, LF-A1, BHLH and ArgR repressor), respectively. For each of the five sets of DNA target sites, the nucleotide sequences for the same type of transcription factor may be different, but the target sites correspond to the same type of BSP. The five types of BSPs, which correspond to the five types of proteins, are respectively represented in bold characters in figs. 3—7.

The seven sets of protein-binding sites contain the same type of BSP, respectively. Is this only a coincidence? This maybe implies that the BSP in the seven sets of target sites is a possible mark or one of the combined marks recognized by their corresponding proteins.

How general is the feature possessed by those seven sets of target sites? We have looked up throughout EMSL nucleotide sequence database[15] to search for the data containing more than one target sites for one type of DNA-binding proteins. But there are only a few of such data in the database. Up to June 2, 2000, there were only ten sets of such data in the following database accession numbers:

(1) M65001. Two DNA target sites for proteins SP-1 are available. The two sites contain the same sequence 5′-CCGCCC-3′.

(2)—(5) AF179904, M19353, M24842, X12799. Six DNA target sites for proteins SP-1 are available, the six sites contain the same sequence 5′-CCGCCC-3′ (or its complementary sequence 5′-GGGCGG-3′).

(6) M6457. Two DNA target sites for proteins P6 are available, the two sites contain the same sequence 5′-AGGTAGG-3′ (or its complementary sequence 5′-CCTACCT-3′). Three DNA target sites for proteins P8II are available, the three sites contain the same sequence 5′-CCCTCCCCC-3′. Two DNA target sites for proteins P1 (fig. 3) and DNA target sites for proteins P8I (fig. 4) are available.

(7) D25294. Two DNA target sites for proteins AP-1 are also available, the two sites contain the same sequence 5′-TGAGTCA-3′. Two DNA target sites for proteins LF-A1 (fig. 5) and three DNA target sites for proteins BHLH (fig. 6) are also available.

(8)—(10) Z49111, M83659, Y11134. Five DNA target sites for protein ArgR repressor (fig. 7) are available.

According to the definition of BSP, the above ten sets of protein-binding sites contain the same BSP respectively. There are five sets of DNA target sites, among the above ten sets, containing respectively the same sequence that could be found easily. Containing the same sequence is a special case of containing the same BSP, so we do not give the corresponding figures of the DNA sequence of the five sets of DNA target sites.

Could BSP act as an independent factor to mediate a possible DNA-protein recognition mode? Or is BSP a sub-

factor in some special DNA-protein recognition modes? These questions are waiting for the test and explanation of experimental biologists. Maybe BSP could be used in identifying putative target sites for some DNA binding proteins. We are trying to formalize the feature of BSP to develop related computer program.

There is no ground to exclude any recognition phenomena that have appeared in nature. There are probably other recognition modes that could describe only a limited subset of protein-DNA recognition. The seven sets of binding target sites suggest that there might exist a protein-DNA recognition mode mediated by BSP and this particular recognition mode is probably limited only to a small subset of DNA-binding proteins.

There are various *cis*-regulating elements on DNA regulating regions, such as enhancers, silencers, CAAT boxes and GC boxes. Those *cis*-regulating elements function in a manner relatively independent of their position and orientation with respect to a nearby gene[16]. Is it possible that those *cis*-regulating elements have any connections with BSP?

Faisst and Meyer[17] have collected 156 vertebrate-encoded transcription factors and DNA sequences recognized by those transcription factors. Among those 156 DNA sequences, there are 36 sequences that contain a symmetric DNA sequence consisted of at least four base pairs (i.e. 4, 6, 8,⋯), accounting for 23.1%; and there are 30 that contain a symmetric DNA sequence consisted of at least five base pairs (i.e. 5, 7, 9,⋯), accounting for 19.2%. A symmetric DNA sequence is also a special case of BSP. For those symmetric DNA sequences, does the symmetry DNA sequence itself or the BSP on the symmetric DNA sequence function?

We also discovered that BSP mediated a relation between 4 and 20, called the 4-BSP-20 relation, and developed an algorithm that is based on the description of the mathematical model of equivalent classes and can be used to analyze the BSP of DNA segments. The algorithm has the potential of theoretically predicting target sites for the DNA binding proteins related to BSP (Details are given in another paper).

In this note we have proposed a possible protein-DNA recognition mode mediated by the particular structural characteristics of double-helical DNA (BSP). It is hoped that the proposals set forth here will provide a new clue to the work of theoretically predicting target sites for some DNA binding proteins, and will serve to stimulate experiments which may eventually reveal the mechanisms for protein-nucleic acid recognition.

Acknowledgements We are grateful to Dr. Pekka Mellergard for his many valuable suggestions and help, and for his careful modification of the manuscript. We also thank Profs. Peng Shouli and Xie Guangzhong very much for help. This work was supported by the National Natural Science Foundation of China (Grant No. 39770418).

References

1. Jin, C., Marsdon, I., Chen, X. et al., Dynamic DNA contacts observed in the NMR structure of winged helix protein-DNA complex, J. Mol. Boil., 1999, 289: 683.
2. Choo, Y., Klug, A., Physical basis of a protein-DNA recognition code, Current Opinion in Structural Biology, 1997, 7: 117.
3. Wolberger, C., Vershon, A. K., Lui, B. et al., Crystal structure of a MATa2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions, Cell, 1991, 67: 517.
4. Suzuki, M., Common features in DNA recognition helices of eukaryotic transcription factors, EMBO J., 1993, 12: 3221.
5. Yuh, C. H., Ransick, A., Martinez, P. et al., Complexity and organization of DNA-protein interactions in the 5'-regulatory region of an endoderm-specific marker gene in the sea urchin embryo, Mechanisms of Development, 1994, 47 (2): 165.
6. Yuh, C. H., Bolouri, H., Davidson, E. H., Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene, Science, 1998, 279: 1896.
7. Calzone, F. J., Theze, N., Thiebaud, P. et al., Developmental appearance of factors that bind specifically to *cis*-regulatory sequences of a gene expressed in the sea urchin embryo, Genes. Dev., 1988, 2: 1074.
8. Theze, N., Calzone, F. J., Thiebaud, P. et al., Sequences of the CyII-Ia actin gene regulatory domain bound specifically by sea urchin embryo nuclear proteins, Mol. Reprod. Dev., 1990, 25: 110.
9. Kawasaki, N., Itoh, N., Kawasaki, T., Gene organization and 5′-flanking region sequence of conglutinin: a C-type mammalian lectin containing a collagen-like domain, Biochem. Biophys. Res. Commun., 1994, 198 (2): 597.
10. Ludovice, M., Martin, J. F., Carrachas, P. et al., Characterization of the Streptomyces clavuligerus argC gene encoding N-acetylglutamyl-phosphate reductase: expression in Streptomyces lividans and effect on clavulanic acid production, J. Bacteriol., 1992, 174 (14): 4606.
11. Rodriguez-Garcia, A., Martin, J. F., Liras, P., The argG gene of Streptomyces clavuligerus has low homology to unstable argG from other actinomycetes: effect of amplification on clavulanic acid biosynthesis, Gene, 1995, 167 (1-2): 9.
12. Rodriguez-Garcia, A., Ludovice, M., Martin J. F. et al., Arginine boxes and the argR gene in Streptomyces clavuligerus: evidence for a clear regulation of the arginine pathway, Mol. Microbiol., 1997, 25 (2): 219.
13. Ulyanov, A. V., Stormo, G. D., Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions, Nucleic Acids Research, 1995, 23(8): 1434.
14. Rhodes, D., Schwabe, J. W., Chapman, L. et al., Towards an understanding of protein-DNA recognition, Philos. Trans. R. Soc. Lond. B. Biol. Sci., 1996, 351(1339): 501.
15. Stoesser, G., Sterk, P., Tuli, M. A. et al., The EMBL nucleotide sequence database, Nucleic Acids Research, 1997, 25 (1): 7.
16. Khoury, G., Gruss, P., Enhancers elements., Cell, 1983, 33: 313.
17. Faisst, S., Meyer, S., Compilation of vertebrate-encoded transcription factors, Nucleic Acids Research, 1992, 20(1): 3.

(Received August 7, 2000)