A MODEL FOR THE BINDING OF LAC REPRESSOR TO THE LAC OPERATOR

G. V. GURSKY, V. G. TUMANYAN, A. S. ZASEDATELEV, A. L. ZHUZE,
S. L. GROKHOVSKY, and B. P. GOTTIKH
Institute of Molecular Biology, Academy of Sciences of the U.S.S.R., Moscow 117312, U.S.S.R.

(Received 6 October, 1975)

Abstract. A model is suggested for the lac repressor binding to the lac operator in which the repressor polypeptide chain sequences from Gly 14 to Ala 32 and from Ala 53 to Leu 71 are involved in specific interaction with operator DNA. A correspondence between the protein and DNA sequences is found which explains specificity of the repressor binding to the lac operator. The model can be extended to describe specific binding of other regulatory proteins to DNA.

I. INTRODUCTION

The binding of lac repressor to the lac operator is a good example of a specific and tight interaction between protein and DNA when both protein [1] and DNA sequences [2, 3] have been strictly established. This offers a possibility to determine which repressor amino acid residues and DNA bases are involved in specific interactions. The purpose of this paper is to present a detailed structural model for the lac repressor binding to the lac operator. The model is consistent with a large amount of experimental data available at present and provides strong support for the code suggested by us for description of specific protein-DNA interactions (see accompanying paper [4]). The code requires the presence of amino acid residues coding for AT and GC base pairs in certain places of a stereospecific protein site for recognition of a particular nucleotide sequence on DNA. The code rules, if they exist, allow to determine the protein sequence which exhibits a correspondence with the base pair sequence in the lac operator. Attempts to determine such a correspondence have been undertaken by several previous investigators [5, 6].

II. RESULTS AND DISCUSSION

The ideas presented in this work stem from the stereochemical model that has been advanced for stereospecific sites of regulatory proteins [4, 7]. According to this model stereospecific protein sites contain pairs of antiparallel polypeptide chain segments hydrogen bonded together to form a double polypeptide helix isogeometric to that of DNA. When a protein binds to DNA one polypeptide chain segment (t-chain segment) in a stereospecific site is attached through hydrogen bonds to pyrimidine O2 oxygens and adenine N3 atoms lying in one polynucleotide strand while

Molecular Biology Reports 2 (1976) 427–434.

the other chain segment (g-chain segment) is hydrogen bonded only to guanine bases in the opposite polynucleotide strand. The amide groups serve as specific reaction centres being hydrogen bond acceptors in the g-chain segment and hydrogen bond donors in the t-chain segment. The binding is stereospecific with an $N - C^{\alpha} - C'$ sequence in each polypeptide chain segment co-linear with the $C3' - C5'$ direction in the corresponding polynucleotide strand.

In agreement with the proposed model the control sequences usually exhibit an asymmetric distribution of guanine residues between the two polynucleotide strands as exemplified by the nucleotide sequence determined for the lac operator [2, 3]. Another prominent feature of the lac operator sequence is the presence of a two-fold rotation symmetry [2] which probably dictates a symmetrical attachment of the lac repressor subunits to the operator DNA. The lac repressor contains four identical subunits [8] and appears to bind to the lac operator as a tetramer [9, 10]. To find operator regions serving as attachment sites for particular repressor subunits we used the nucleotide sequence for the entire lac promotor-operator region as determined by Dickson *et al.* [3] and suggested that four attachment sites for the repressor subunits should be all similar. A shorter sequence determined by Gilbert and Maxam and assigned to the lac operator [2, 10] will be subsequently shown to involve the binding sites for two repressor subunits only.

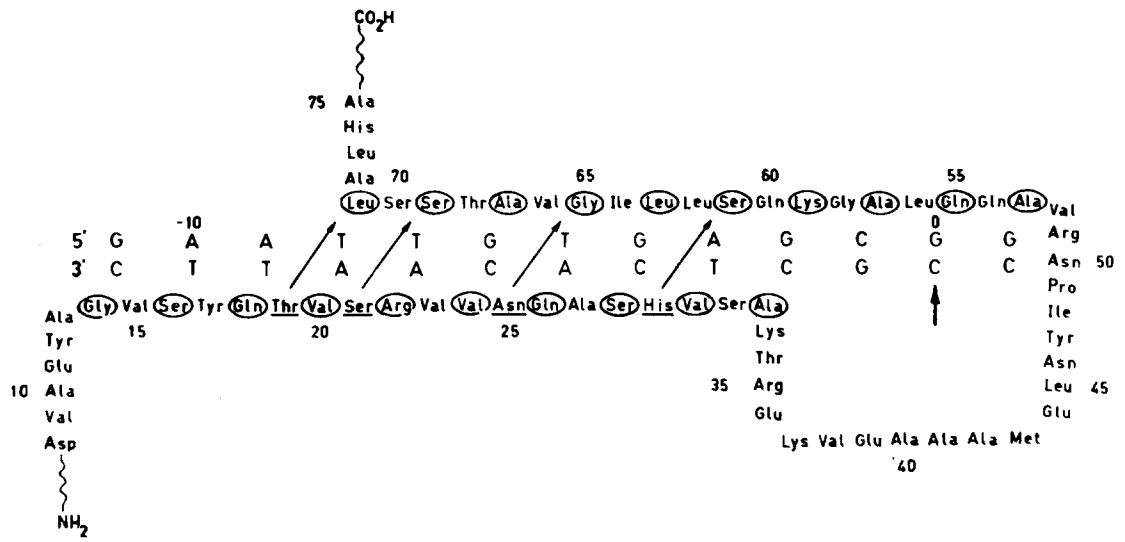
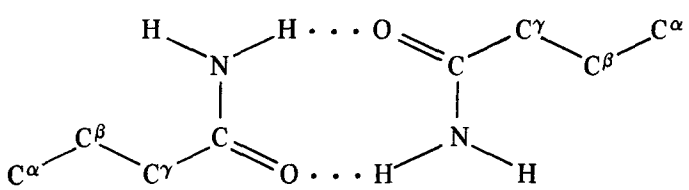


Fig. 1. A diagram illustrating the proposed correspondence between amino acid sequence in the stereospecific repressor site and base pair sequence in the lac operator. The polypeptide chain segment ranging from Gly 14 to Val 30 exists in t-conformation and forms hydrogen bonds with all bases lying in the lower polynucleotide strand of the lac operator starting with cytosine in position – 12 to cytosine in position – 4. The upper polypeptide chain segment from Ala 53 to Leu 71 exists in g-conformation and reacts only with guanine bases in the upper polynucleotide strand. In the t-chain segment all amino acid residues circled are involved in hydrogen bonding with DNA bases. In the g-chain segment each residue can be hydrogen bonded to the corresponding guanine only in the absence of an AT coding residue at the corresponding site of the t-chain segment. As an example, the carbonyl group of Ala 67 can be connected by a hydrogen bond with the 2-amino group of guanine in position – 6 (Val 27 acts as a residue coding for GC). However, such a hydrogen bonding is not possible for Gly 65 even when a GC pair is present in position – 5. Amino acid residues coding for AT are all underlined. The tilting arrows show those residues of the g-chain segment whose hydrogen bonding state is affected by the AT-coding residues. A vertical arrow indicates the position of the two-fold symmetry axis in the operator. Systematic hydrogen bonds (not shown) connect NH and C=O groups of external residues (circled) lying in g- and t-chain segments. As an example, the NH group of Leu 71 is hydrogen bonded to the C=O group of Val 20.

Figure 1 shows one operator region complexed with a stereospecific site of one lac repressor subunit. It can be readily seen from the figure that the 3′ polynucleotide strand of DNA contains an extended region with no guanine whereas the 5′-strand contains several guanine bases required for accurate recognition. It follows therefrom that the t-polypeptide chain segment in the stereospecific repressor site is located closer to the amino-terminus of the repressor polypeptide chain than is the g-chain segment. Since coding amino acid residues are always located in the t-chain segment (see [4]) finding of a correspondence between repressor and operator sequences requires a systematic study for the sequence fitting of all repressor polypeptide chain segments starting from the amino-terminal region. The only polypeptide chain segment found to display such a correspondence is that ranging from Thr 19 to Val 30. In this segment all residues coding for AT (Thr, Ser, Asn and His) are present in absolutely the same places where the lac operator site has AT base pairs and are absent in the positions where GC pairs are present. In addition, in this segment all residues are compatible with the β-structure with no residue negatively charged present. In contrast, the neighbouring chain segment has three negatively charged glutamic acid residues and exhibits no correspondence with the base pair sequence at the central part of the lac operator. For this reason we suggested that this chain segment could not be located in the minor groove of DNA but had to form a loop connecting t- and g-chain segments in the stereospecific repressor site.

The central operator region consists of three consecutive GC pairs two of which (in positions +1 and −1) are related by two-fold symmetry. The two repressor subunits bound to the neighbouring sites at the central operator part are probably symmetrically arranged with their g-chain segments hydrogen bonded to guanine bases lying in the opposite DNA strands. In particular, g-chain segment of the repressor subunit bound on the left operator side interacts with guanines in positions 0 and + 1 (5′ strand) while g-chain segment of the subunit bound on the right operator side interacts with guanine in position −1 (3′ strand). The overlapping regions of the two g-chain segments can be hydrogen bonded together to form a fragment of double polypeptide helix (gg-structure) [4]. This structure may be considered as a right-hand twisted β-sheet bended in the central region to form hydrogen bonds with guanine bases.

Which protein sequence is pertinent for the interaction with the central operator region? From symmetry considerations and inspection of molecular models we found such a sequence to be as shown in Figure 2. In the double-stranded structure formed between the interacting portions of two g-segments the two Gln 54 are related by two-fold symmetry. Their side-chains can interact with each other to form two hydrogen bonds locking the two repressor subunits on their correct positions:



Obviously, these hydrogen bonds can be formed only when the two repressor subunits are bound to their correct binding sites. Inspection of molecular models shows that asparagine and glutamine are the only amino acid residues whose side-chains are capable of interacting in such a way.

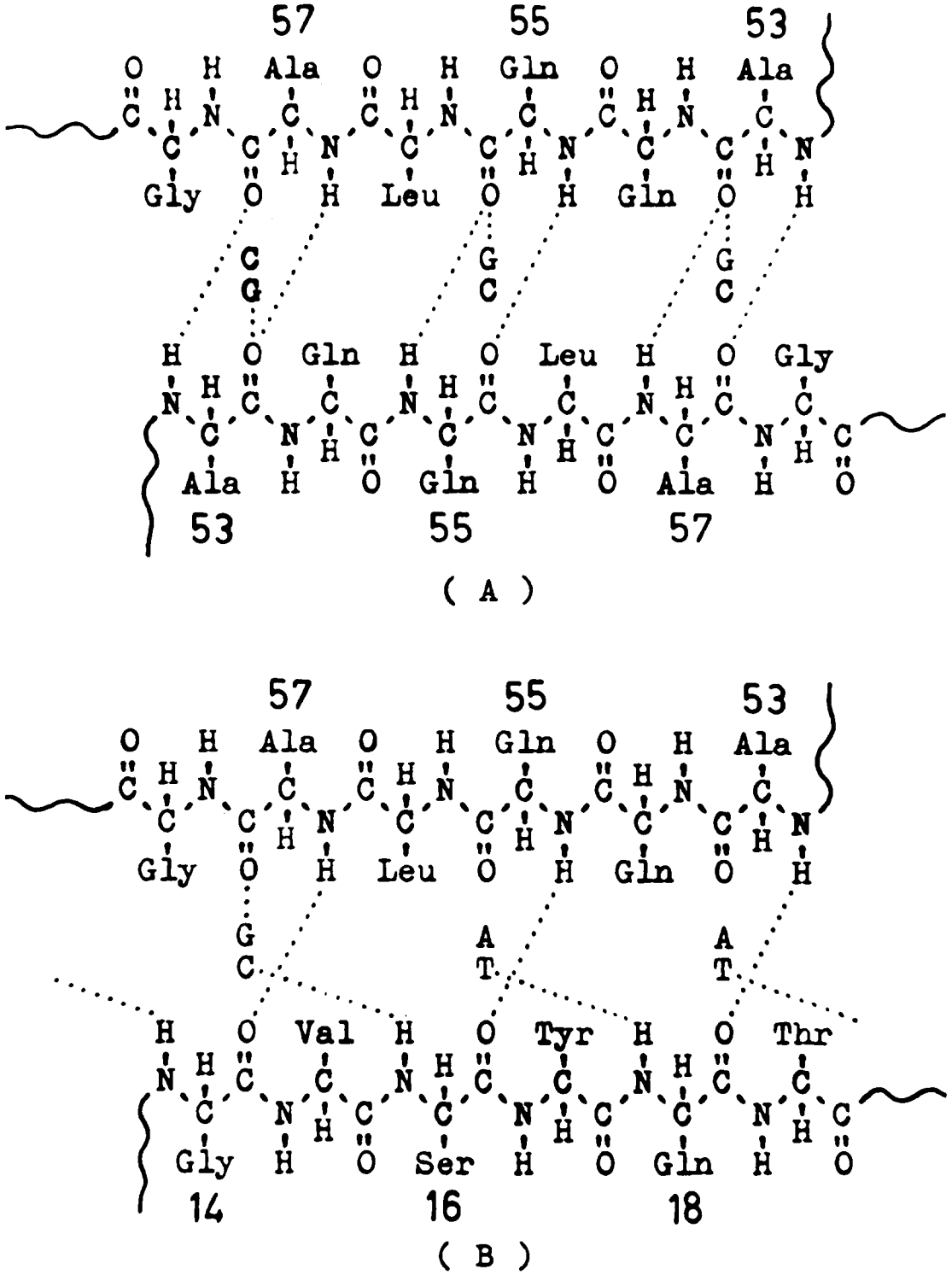


Fig. 2. Contacts between the repressor subunits bound to the central (A) and peripherical (B) parts of the lac operator. The amino acid sequences are presented for the cohesive ends of two repressor subunits bound to the neighbouring sites on the lac operator. The dotted lines show hydrogen bonds formed with the operator DNA and between the two polypeptide chains.

The repressor polypeptide chain sequence ranging from Ala 53 to Leu 71 exists probably in g-conformation and is involved in hydrogen bonding to guanine bases lying in the upper polynucleotide strand in positions −6, −4, −2, 0 and +1 (see Figure 1). The polypeptide chain sequence from Gly 14 to Ala 32 is likely to form hydrogen bonds with all bases in the lower polynucleotide chain segment ranging from cytosine in position −12 to thymine in position −3. We suggest that the stereospecific repressor site involves an antiparallel β-sheet at sequences 20–30 and 61–71 and two cohesive ends at 14–18 and 53–57 representing single-stranded portions of t- and g-chain segments. The repressor polypeptide chain forms turns at sequences 10–13, 31–34, 48–52 and 72–76, which determine the geometry of bonded areas between adjacent repressor subunits. Complex formation with operator DNA results in β-structure deformation as described previously [4] so that the lower polypeptide chain segment exists in t- and the upper one in g-conformation.

The sequences involved in β-turns can be well connected with t- and g-structures. The backbone NH of Ala 32 and Val 52 can participate in hydrogen bond formation with the carbonyl oxygens O2 of thymine in position −3 and cytosine in position −2, respectively. The backbone NH of Thr 34 is probably involved in hydrogen bond formation with the carbonyl oxygen of Ser 31. Another hydrogen bond is formed between the backbone NH of Lys 59 and the carbonyl group of Ala 32. The turns at sequences 10–13 and 48–52 are also stabilized by hydrogen bonds. The formation of a turn at sequence 31–34 does not allow Ser 31 to act as an AT-coding residue, probably because of hydrogen bond formation between its side-chain and the carbonyl group of Arg 35. All other residues coding for AT are present in the stereospecific repressor site in the positions that give the best correspondence with the base sequence in the lac operator. We feel that this correspondence is unlikely to have happened by chance [7]. In principle, this correspondence may be explained in two different ways depending on the mode of action of amino acid residues coding for AT base pairs. The side-chains of these residues may either interact directly with AT base pairs forming hydrogen bonds with thymine oxygens O2 and adenine atoms N3 or act as elements controlling hydrogen bond interactions between the g-chain segment in the stereospecific repressor site and guanine bases (see [4]). In both of these models the stereospecific repressor site is built in terms of a right-hand twisted antiparallel β-sheet involving t- and g-chain segments with the amide groups in the g-chain segment interacting with the guanine 2-amino groups. In both models the β-sheet must have a curvature opposite to that existing in the model of Carter and Kraut [11]. We prefer the second model since it predicts a more symmetrical attachment of polypeptide chain segments to DNA.

Figure 3 shows the proposed arrangement of the lac repressor subunits on the lac operator. Being complexed with operator DNA, lac repressor subunits are related in pairs by two-fold symmetry. In each subunit g- and t-chain segments interact with polynucleotide chain regions extending over 10 bases. The cohesive ends of adjacent repressor subunits bound interact with each other thus forming double-stranded structures of the gg- and gt-type. In the former case single-stranded portions of g-chain segments in the two repressor subunits complexed with the central operator region interact to give a deformed β-sheet (contact A). In the latter case contacts of the B-type are formed between the repressor subunits in the periphery operator parts, with single-stranded portions of t- and g-chain segments in two subunits interacting to form a gt-structure described previously [4]. In both contacts the interacting portions of g- and t-chain segments consist of five amino acid residues whose sequences are shown in Figure 2.

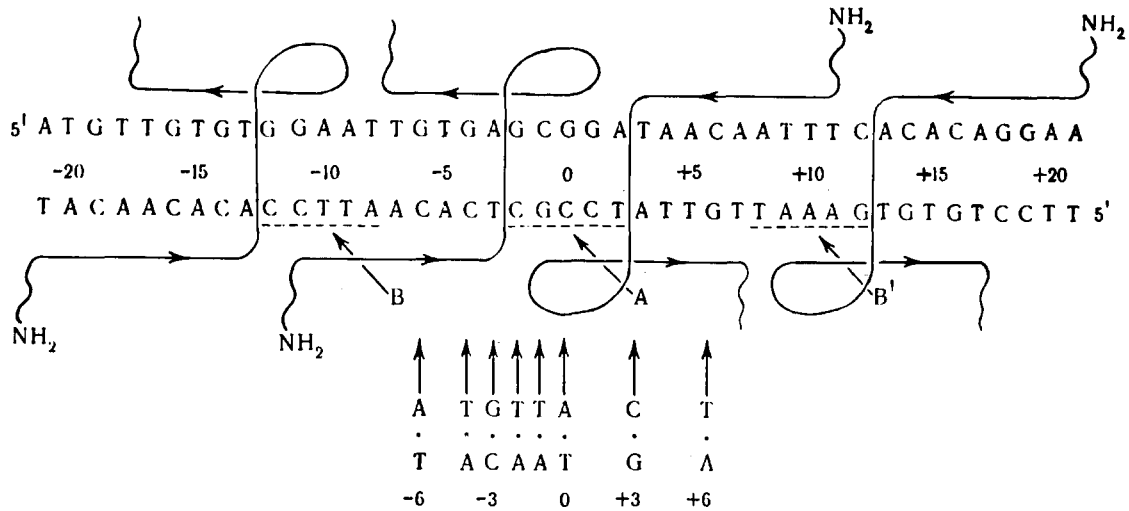


Fig. 3. Diagram illustrating the proposed arrangement of the lac repressor subunits on the lac operator. The stereospecific site of each subunit is represented as a continuous line showing the direction of polypeptide chain progress (See Figure 1). The bottom portion of the figure shows the alterations in the operator base sequence which occurs as a result of various mutations (data from Gilbert *et al.* [13]).

The four attachment sites for lac repressor subunits are all rather similar but not identical. The best correspondence between protein and DNA sequences exists for subunits interacting on the left operator side. The model proposed predicts that those subunits bind much more strongly as compared with subunits interacting on the right operator side. This is in accordance with the fact that effects of operator mutations in disrupting repressor-operator interaction are greater on the left operator side, small on the right [12, 13]. Most mutations are known to be concentrated on the left operator side within the first seven base pairs including the GC pair located on the two-fold symmetry axis [13] (Figure 3). Every mutation involves an AT substitution for GC or vice versa [13] and causes a drastic decrease in the magnitude of the repressor binding constant with operator [12]. We interpret these observations as indicating that one or several specific bonds cannot be formed between the mutant operator and the repressor. If the GC pair is in its proper place in the operator (as determined by the amino acid sequence in the stereospecific repressor site) its substitution by an AT pair following mutation results in a lowered binding constant due to one hydrogen bond formed rather than two. If the AT pair is in its proper position such a substitution of GC for AT is preferable in which guanine is located in the polynucleotide strand complementary to the g-chain segment in the stereospecific repressor site. Otherwise, the repressor may not recognize which of the two chain segments in its stereospecific site stands for the g- or t-chain segment. The AT → GC mutations all meet this requirement. They also weaken the repressor-operator interaction, since a strong hydrogen bond with thymine is substituted by a relatively weak hydrogen bond with cytosine in the mutant operator.

The proposed model shows that the lac operator base sequence does not yet optimally correspond to the amino acid sequence in the stereospecific repressor site. It must, therefore, be expected that such lac operator mutations will be eventually discovered which will exhibit a greater affinity for the lac repressor than does the wild-type operator. Our model predicts that in

an optimal operator AT should be substituted by GC in positions −9, −10, +2, +4, +9 and +10, with AT substitution for GC in position +18 (Figure 3).

The ability of the repressor to recognize the lac operator may also be affected by mutations that alter the primary structure of the repressor itself. These alterations involve replacements of Ser 16 and Ala 53 by Pro and Val, respectively, as well as replacement of Thr 19 by Ala [14]. Each amino acid substitution is sufficient to eliminate repression *in vivo*, probably by altering the operator binding activity. In accord with the proposed model both Ser 16 and Ala 53 occur in the cohesive ends in the stereospecific repressor site and take part in hydrogen bond formation with the operator DNA. The Ala 53 occupies a unique position in the stereospecific site being in the region where the adjacent repressor subunits bound are in contact with each other. From inspection of molecular models it appears unlikely that a side-chain bulkier than that of alanine could be accommodated in this region. Substitution of Thr 19 by Ala seems to be unfavourable since threonine acts as an AT-coding residue and β-former.

The present model is also consistent with the genetic [5, 15, 17] and chemical [17, 18] experiments involving the amino-terminal region of the lac repressor in operator binding. It is also compatible with the chemical experiments showing that the repressor binding to the lac operator and non-operator DNA causes a shielding of adenine atoms N3 from methylation by dimethylsulfate [19, 20]. The repressor also affects the methylation pattern of guanines at position 7 which is exposed in the major groove of DNA. The model presented does not exclude involvement of the major groove in the repressor binding but suggests that specific contacts between the repressor and DNA are mainly formed in the minor groove. It is of interest that the methylation pattern reveals DNA regions where the extent of methylation is increased upon addition of the lac repressor [20]. These occur in that position in the central operator region where the two polypeptide chains of the neighbouring repressor subunits bound make turns to form cohesive ends.

On complex formation with the inductors such as β-galactosides the repressor affinity for the lac operator is lowered [21] with the inductor seemingly causing conformation changes in the loop interconnecting t- and g-chain segments in the stereospecific repressor site. These conformation changes may prevent hydrogen bond formation between cohesive ends of repressor subunits and the central operator region thereby decreasing the strength of repressor binding to DNA. We may, therefore, conclude that the proposed model is consistent with a large amount of experimental data available at present.

The model can be extended to describe specific binding of other regulatory proteins to DNA. We suggest that stereospecific sites of all regulatory proteins are built according to the principle described above for the lac repressor. They all contain t- and g-polypeptide chain segments with coding amino acid residues present in the t-chain segment. In addition, in many cases subunits of regulatory proteins have cohesive ends that possess a flexible structure and are responsible for the cooperative effects in protein binding. It is generally accepted that regulatory proteins must possess a binding surface capable of assuming a helical conformation over a distance of ~ 100 Å. A combination of an ordered structure (a right-hand twisted antiparallel β-sheet) and flexible cohesive ends within the stereospecific protein site provides a basis for possible solutions of the structural and kinetic problems associated with the twisted nature of DNA. Another important factor is a loosely coupled oligomeric structure, allowing protein subunits to act to some extent independently.

In all regulatory sequences determined so far guanine bases are distributed asymmetrically between the two DNA strands, a feature which makes it likely that the code controlling specific binding of regulatory proteins is universal.

ACKNOWLEDGEMENTS

We thank Prof. V. A. Engelhardt for valuable discussions, Drs R. Sh. Beabealaschvilly and A. D. Mirzabekov for useful information and continual interest.

REFERENCES

1. Beyreuther, K., Adler, K., Geisler, N., and Klemm, A., *Proc. Nat. Acad. Sci. U.S.A.* **70**, 3576 (1973).
2. Gilbert, W. and Maxam, A., *Proc. Nat. Acad. Sci. U.S.A.* **70**, 3581 (1973).
3. Dickson, R. C., Abelson, J., Barnes, W. M., and Reznikoff, W. S., *Science* **187**, 27 (1975).
4. Gursky, G. V., Tumanyan, V. G., Zasedatelev, A. S., Zhuze, A. L., Grokhovsky, S. L., and Gottikh, B. P., *Mol. Biol. Reports*, this issue, pp. 413–425.
5. Adler, K., Beyreuther, K., Fannig, E., Geisler, N., Gronneborn, B., Klemm, A., Müller-Hill, B., Pfahl, M., and Schmitz, A., *Nature* **237**, 322 (1972).
6. Patel, D., *Biochemistry* **14**, 1057 (1975).
7. Gursky, G. V., Tumanyan, V. G., Zasedatelev, A. S., Zhuze, A. L., Grokhovsky, S. L., and Gottikh, B. P., *Mol. Biol. U.S.S.R.*, in press.
8. Gilbert, W. and Müller-Hill, B., *Proc. Nat. Acad. Sci. U.S.A.* **56**, 1891 (1966).
9. Steitz, T. A., Richmond, T. J., Wise, D., and Engelman, D., *Proc. Nat. Acad. Sci. U.S.A.* **71**, 593 (1974).
10. Gilbert, W., Maizels, N., and Maxam, A., *Cold Spring Harbor Symp. Quant. Biol.* **38**, 845 (1973).
11. Carter, C. W. and Kraut, J., *Proc. Nat. Acad. Sci. U.S.A.* **71**, 283 (1974).
12. Jobe, A., Sadler, J., and Bourgeois, S., *J. Mol. Biol.* **85**, 231 (1974).
13. Gilbert, W., Gralla, J., Majors, J., and Maxam, A., in: H. Sund and G. Blauer (eds.), *Protein-Ligand Interactions*, Walter de Gruyter & Co., Berlin, p. 193.
14. Weber, K., Platt, T., Ganem, D., and Miller, J. H., *Proc. Nat. Acad. Sci. U.S.A.* **69**, 3624 (1972).
15. Müller-Hill, B., Crapo, L., and Gilbert, W., *Proc. Nat. Acad. Sci. U.S.A.* **59**, 1259 (1968).
16. Pfahl, M., *Genetics* **72**, 393 (1972).
17. Platt, T., Weber, K., Ganem, D., and Miller, J. H., *Proc. Nat. Acad. Sci. U.S.A.* **69**, 897 (1972).
18. Platt, T., Files, J., and Weber, K., *J. Biol. Chem.* **248**, 110 (1973).
19. Mirzabekov, A. D., Kolchinsky, A. M., Gilbert, W., submitted to *Mol. Biol. Reports.*
20. Gilbert, W., Maxam, A., and Mirzabekov, A. D., *Alfred Benzon Symposium IX, Control of Ribosome Synthesis*, in press.
21. Riggs, A. D., Newby, R. F., and Bourgeois, S., *J. Mol. Biol.* **51**, 303 (1970).