A CODE CONTROLLING SPECIFIC BINDING OF REGULATORY PROTEINS TO DNA

A. 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 possible code is suggested that describes a correspondence between amino acid sequences in stereospecific sites of regulatory proteins and nucleotide sequences at the control sites on DNA. Stereospecific sites of regulatory proteins are assumed to contain pairs of antiparallel polypeptide chain segments which form a right-hand twisted antiparallel β-sheet with single-stranded regions at the ends of the β-structure. The binding reaction between regulatory protein and double-helical DNA is accompanied by significant structural alterations at stereospecific sites of the protein and DNA. Half of the hydrogen bonds normally existing in β-structure are broken upon complex formation with DNA and a new set of hydrogen bonds is formed between polypeptide amide groups and DNA base pairs. The code states a correspondence between four amino acid residues at a stereospecific site of the regulatory protein and an AT (GC) base pair at the control site. It predicts that there are six fundamental amino acid residues (serine, threonine, histidine, asparagine, glutamine and cysteine) whose arrangement in the stereospecific site determines the base pair sequence to which a given regulatory protein would bind preferentially.

I. INTRODUCTION

Regulatory proteins are adsorbed on DNA at specific sites 'recognizing' definite nucleotide sequences. Such specific interactions are exemplified by RNA polymerase binding to initial gene sites (promotors) (for a review of the literature see ref. [1, 2]), by interaction of lac and lambda repressors with respective operators [3, 4] and specific binding of some methylases and nucleases [5, 6]. It might be conceived that specific association of regulatory proteins could be accomplished differently. It does, however, seem much more plausible that regulatory proteins all dispose of certain 'rules' for recognizing specific regulatory sequences in DNA. These rules involve a definite correspondence (code) between the sequence of amino acid residues in the stereospecific site of regulatory protein and that of nucleotides at the control site to which a given regulatory protein binds specifically. This is actually the second code underlying the phenomena of life. The first fundamental code is known to account for the relation between the nucleotide

Molecular Biology Reports 2 (1976) 413–425.

sequences in DNA and amino acid sequences in proteins encoded by these nucleotide sequences. This paper deals with the second code. In the accompanying paper [7] we demonstrate which particular DNA bases and protein amino acid residues can participate in specific interaction as exemplified by that of lac repressor and lac operator.

Many authors have attempted to solve the problem of protein-nucleic acid recognition in terms of models based on direct interaction between side chains of amino acid residues and nucleic acid bases [8–12]. The present work is based on the idea that amide groups of the polypeptide chain act as specific reaction centres for protein stereospecific site bonding to DNA bases, whereas side chains are responsible for the formation of a particular configuration of the polypeptide chain backbone but are not specifically bound to bases.

The following general considerations are advanced.

First, the specific adsorption problem is the recognition of nucleotide sequences in DNA. As compared with adsorption of small molecules on DNA its specificity lies not only in the physical dimensions of ligand molecules adsorbed but in the effect of the internal ligand structure on the adsorption equilibrium. It becomes thus of great significance to put forward adequate theoretical models. Many workers deal, therefore, with this problem and in their models the internal ligand structure becomes gradually more involved as the DNA-ligand interaction is found to be more specific. The first case to be studied was that when a ligand covers a definite number of DNA base pairs on binding, but does not react specifically with bases [13–15]. The second case under study was that when a ligand has one [13, 16] or several [17, 18] reaction centres responsible for the specificity of binding interactions.

The main postulate of this work states that regulatory proteins recognize the base sequences in the DNA double helix without unwinding it, with base pairs acting as specific binding centres. It is, of course, presumed that the regulatory protein also contains corresponding reaction sites, specific to AT and GC base pairs. Recent experimental evidence [19, 20] appears to substantiate these considerations. It becomes thus possible to formulate a general principle that will be further referred to as the principle of lattice recognition. This principle states that any regulatory protein can be considered in terms of an equivalent one-dimensional lattice of reaction centres which enables the protein binding properties to be correctly described. In a general case such a lattice consists of four components since AT $\neq$ TA and GC $\neq$ CG on binding. The sequence in which the reaction centres are arranged provides a code for finding the appropriate (complementary) sequence of DNA base pairs. An important property of the model is a strict correspondence between repeating distances of lattices characteristic of regulatory protein and DNA. That is, the distance between successive reaction centres is equal to or multiple of 3.36 Å if DNA is in B conformation.

Second, the code controlling specific protein-DNA interactions might be considered as being universal, or nearly so. The universality of the code and the lattice principle of recognition impose strict stereochemical limitations on the possible structure of the regulatory protein site involved in specific interaction with DNA. The universality of the code requires stereospecific sites of regulatory proteins to be arranged in terms of one and the same structural scheme that is only slightly modified to recognize every particular base sequence. The lattice principle of recognition requires the polypeptide chain involving in specific interaction with DNA bases to be a right hand helix isogeometric to that of DNA. Molecular model building indicate that for such a helix to be formed the polypeptide chain must contain at least two amino acid residues in the

repeating unit. There is another limitation imposed on the structure of stereospecific sites of regulatory proteins by the symmetry of the DNA molecule. These sites are required to possess at least a set of pseudo two-fold axes normal to the helix axis and coinciding with the dyad axes of DNA molecule itself. We, therefore, suppose that α-helix and collagen-like structures are inacceptable as models for stereospecific sites of regulatory proteins.

Third, specific reaction centres for protein binding to regulatory sequences of DNA are amide groups rather than side chains of amino acid residues. The polypeptide chain backbone can readily form helical structures with amide groups acting either as hydrogen bond donors or acceptors depending on their orientation. Side groups of amino acid residues are structurally so different that they are extremely unlikely to be a basis for helical arrangement of reaction centres.

II. STRUCTURE OF STEREOSPECIFIC SITES OF REGULATORY PROTEINS

The structure that we propose for stereospecific sites of regulatory proteins in a complex with DNA is simple. The stereospecific sites consist of two antiparallel polypeptide chain segments hydrogen bonded together to form a double polypeptide helix isogeometric to that of DNA (Figures 1 and 2). Each polypeptide chain contains two amino acid residues in the repeating unit and is hydrogen bonded to the bases lying in one polynucleotide strand. The two polypeptide chain segments are inserted in the minor groove of DNA and form numerous van der Waals contacts with the atomic groups of the phosphate-deoxyribose backbone. It is to be noted that two antiparallel polypeptide chains interacting with two polynucleotide chains meet the main requirements imposed on geometry of protein stereospecific sites by the symmetry of the nucleic acid molecule. This point is well discussed by Carter and Kraut [21]. The structure shown in Figures 1 and 2 was found in the following way. We were initially interested in the helical configurations of a polypeptide chain with all side chains omitted. By analogy to oligopeptide antibiotics such as distamycin [22–27] we presumed that the polypeptide chain was located in the minor groove of DNA and carbonyl O2 atoms of thymine and cytosine (and probably N3 of adenine) acted as acceptors for hydrogen bonds with NH amide groups of the polypeptide chain. The 2-amino group of guanine was considered as a donor for hydrogen bonding to C = O of the amide group. The table of atomic coordinates for DNA indicates [28] that pyrimidine O2 and purine N3 atoms all have very similar spatial coordinates in the DNA duplex. In relation to the specificity of DNA binding interactions the importance of this fact was first noted by Bruskov and Poltev [29] and by Ivanov [11] and Karpeisky (private communication). The spatial position of the donor (guanine 2-amino) group and that of acceptor groups is, however, different. This fact together with difference in the chemistry of interacting groups give rise to two possible types of regular structures for the polypeptide chain backbone. These are readily distinguished by supposing that one DNA strand consists of guanine or thymine bases only to which the polypeptide chain is bonded by a regular system of hydrogen bonds as schematically shown in Figure 1. These two configurations of the polypeptide chain will be referred to as g- and t-structures, respectively. In both structures two of the three successive amide groups are involved in specific hydrogen bonds with DNA bases. The binding is stereospecific in that $N - C^{\alpha} - C'$ sequence in the two chains coincides with the $C_3' \rightarrow C_5'$ direction in polynucleotide chains.

Using standard Pauling-Corey bond-distances and angles [30] we calculated on a computer all

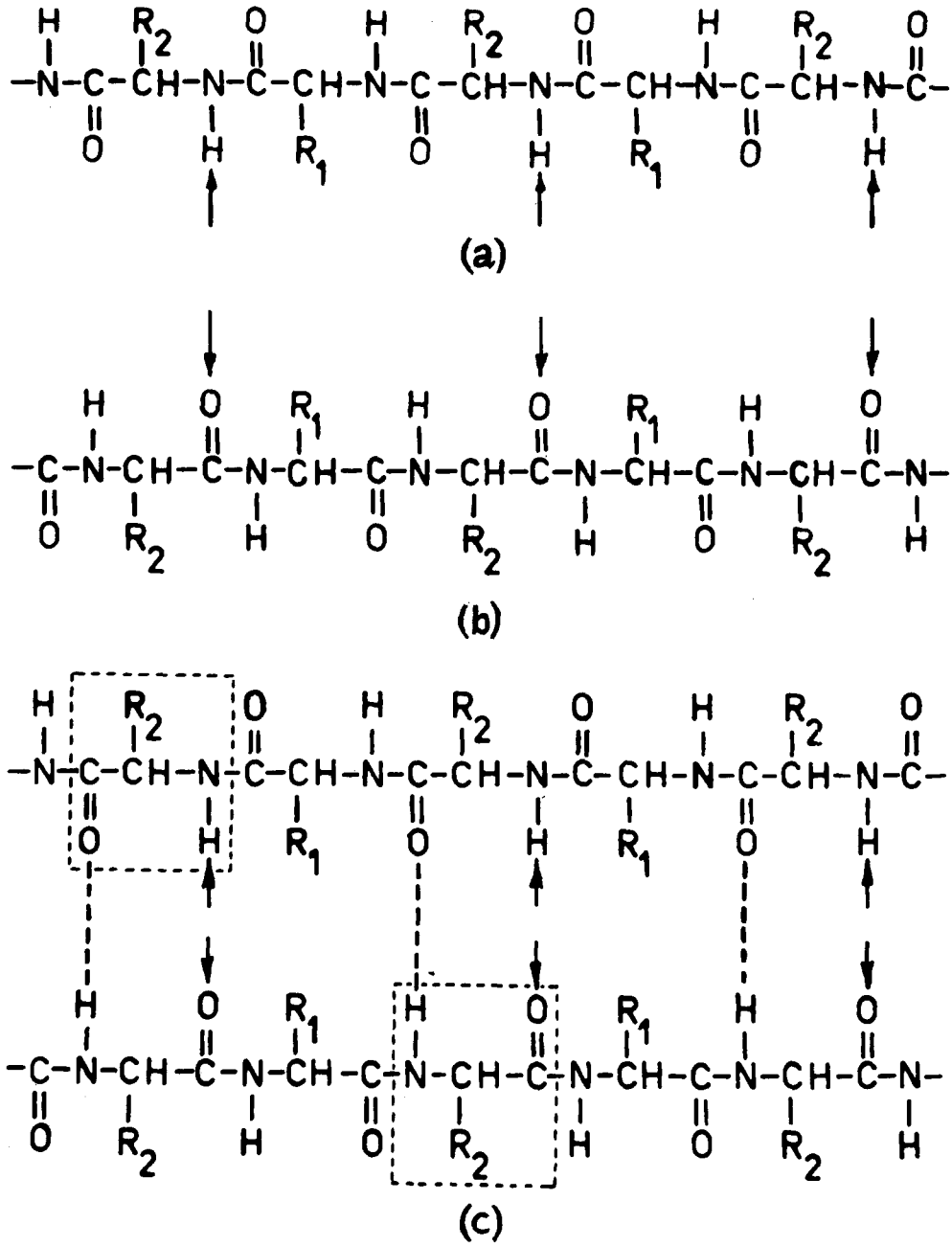

Fig. 1. A diagram illustrating the formation of a double polypeptide helix from the two separate polypeptide chains forming the systematic hydrogen bonds with GC base pairs in poly dG • poly dC. (a) A polypeptide chain segment (t-chain segment) forming the systematic hydrogen bonds which connect the backbone NH with the cytosine oxygens O2 in a poly dG • poly dC duplex. Arrows indicate hydrogen bonds. (b) A polypeptide chain segment (g-chain segment) attached through the systematic hydrogen bonds to the guanine 2-amino groups in a poly dG • poly dC duplex. (c) Two antiparallel polypeptide chain segments shown in (a) and (b) are hydrogen bonded together to form a double polypeptide helix. Boxed are the amino acid residues involved in hydrogen bond formation with a given GC base pair. The dashed lines represent hydrogen bonds between the two polypeptide chains. R_1 and R_2 represent outside- and inward-pointing side-chains, respectively.

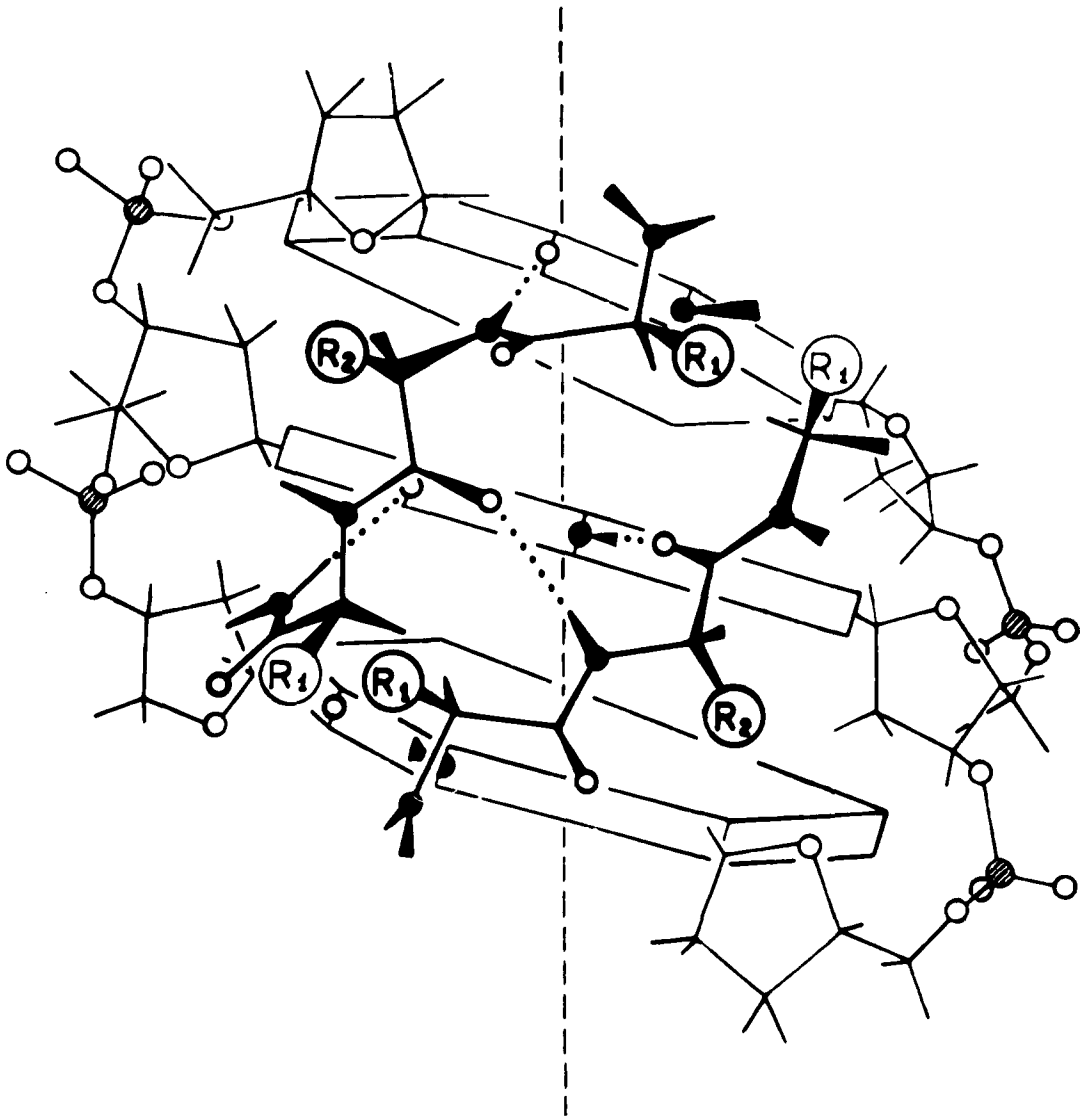

Fig. 2. General structural motif within the stereospecific protein site involved in specific binding interactions with poly dG • poly dC. An asymmetric unit of the proposed complex consists of a GC pair and four amino-acid residues. The structure of the complex is shown as projected on the vertical plane passing through the helix axis and perpendicular to the dyad axis of poly dG • poly dC at the origin of each asymmetric unit. (○) = oxygen; (●) = nitrogen; (⊘) = phosphorus. The dotted lines represent hydrogen bonds.

possible helical polypeptide structures (with two residues in the repeating unit) capable of interacting by a hydrogen bonding mechanism either with guanine 2-amino groups or pyrimidine O_2 atoms. Each structure is specified by four dihedral angles φ_1, ψ_1, φ_2, and ψ_2 which define rotations about $N - C^\alpha$ and $C^\alpha - C'$ bonds.The helix-generating parameters (axial rise and twist angle per asymmetric unit) were calculated as functions of these angles according to a formula derived by Sugita and Miyazawa [31]. Of the several possible helical structures only those were selected in which the dihedral angles φ and ψ were located in the large allowed region of the Ramachandran

diagram [30]. The reason for this is that the polypeptide and polynucleotide chains apparently undergo mutual structural adaptation upon complex formation. This requires some adjustments of polypeptide chain geometry within the allowed region on the (φ, ψ) map so as to fit better the helix-generating parameters of nucleic acid. The permissible range of such adjustments is the greatest one with our choice of allowed region on the (φ, ψ) map. A number of very similar structures were found for t- and g-polypeptide chains, differing in the magnitudes of axial rise and twist angle per repeating unit. Dihedral angles and helix-generating parameters for several such structures are given in Table I. It is to be noted that these structures are similar but not identical to bb- and cd-structures proposed by De Santis *et al.* [32]. These authors aimed at describing all possible helical polypeptide structures with two residues in the repeating unit, axial rise per unit of 3.4 Å and twist angle per unit of 36°. They were not concerned with specific interactions between amino acid residues and DNA bases. The overall number of such structures is very high, but most of them are unacceptable as models for stereospecific sites of regulatory proteins.

TABLE I: Characteristic conformation parameters of selected t- and g-structures

Structure	θ (°)	h (Å)	R_N (Å)	R_O (Å)	φ_1	ψ_1	φ_2	ψ_2
t	25.7	3.63	12.6		−134	39	−52	164
g	25.1	3.63		12.3	−76	128	−151	108
t	32.4	2.80	10.7		−132	36	−55	171
g	31.8	2.80		10.5	−89	137	−178	130
t	32.1	3.20	10.5		−132	37	−54	171
g	31.9	3.20		10.4	−95	142	−171	136
t	40.0	2.81	8.7		−130	39	−64	179
g	39.8	2.80		8.3	−88	142	179	140
t	40.7	3.60	8.0		−138	31	−48	−174
g	41.2	3.60		7.2	−74	136	−173	128
t	45.6	3.83	6.7		−114	37	−56	168
g	45.6	3.80		6.4	−80	140	−179	140
t*	40	3.8	6.3		−150	45	−60	−170
g*	40	3.8		6.2	−110	150	180	145
g'*	40	3.8		6.7	−130	140	170	170

An asymmetric unit of each structure involves two successive amino acid residues. φ_i and ψ_i are the dihedral angles for the i-th residue. The direction of progress along the polypeptide chain is $N \rightarrow C^{\alpha} \rightarrow C'$. The backbone NH and C = O groups of the second residue in each asymmetric unit are hydrogen bonded to GC base pairs, thus giving rise to t- and g-structures, respectively. R_N and R_O are the radial distances from the helix axis for the atoms N and O which can be involved in hydrogen bond formation with GC pairs. h and θ are helix-generating parameters. t*, g*, g'* represent those t-, g- and g'-structures whose conformation parameters were determined from the molecular model of the complex with poly dG • poly dC (see Figure 2). The standard skeletal components (scale 1 Å = 2.5 cm) were used and the dihedral angles were measured with an accuracy of ± 10°.

Our conformational calculations were carried out for another class of helical structures. We consider hydrogen bond formation between polypeptide amide groups and DNA bases as a prerequisite of binding. In addition, we took into account the possibility that the DNA structure could be somewhat affected by interaction with polypeptides. Using standard geometries for base pairs and deoxyribose rings, standard phosphate and peptide groups we searched for the conditions when both polypeptide chain and DNA would have identical helix-generating parameters. Our model building was essentially based on the results of DNA conformation calculations made by one of the present authors (V.G.T.) according to the general procedure outlined previously [33, 34]. Only those DNA conformations were used in our model building study which were shown to be stereochemically and energetically satisfactory.

In g-structures, presented in Table I, both amino acid residues in the repeating unit have dihedral angles close to those occurring in deformed β-structure, whereas in t-structures one of local conformation resembles that of poly-L-proline II [35]. Bearing in mind that this conformation is characteristic of the amino acid residue of the t-chain whose NH group must participate in hydrogen bonding to the carbonyl group of thymine (cytosine) we believe that proline cannot be present either in the g- or the t-chain. Dipeptide maps calculated for residues branched at C^{β} [30, 36] show that most other side-chains are compatible with g- and t-structures.

We were mainly interested in double-stranded polypeptide structures in which two antiparallel polypeptide chains being in g- and t-conformations could be attached together by hydrogen bonds formed between those amide groups of the two chains which did not interact with DNA bases. We found that with dihedral angles φ and ψ lying in the large allowed region of the Ramachandran diagram only gg- and gt-double-stranded structures could be formed while tt-structures were all unacceptable. The structures presented in Table I are combined in pairs each representing a double-stranded polypeptide helix. A structure of gt-type is probably the main element of stereospecific protein sites specifically interacting with DNA. Figure 2 illustrates a portion of a gt-polypeptide double helix wrapped around the minor groove of a poly dG · poly dC duplex. In Table I conformational parameters are presented for the two polypeptide chains involved in the complex. A list of atomic coordinates will be published elsewhere [37]. The gt-structures may be considered as variants of deformed antiparallel β-sheet. It is well-known [38] that antiparallel β-chains twisted in a right-handed sense possess a stable structure common to globular proteins in general and thus, possibly, to regulatory proteins in particular. We suggest that half of the hydrogen bonds normally existing in β-structure are broken upon complex formation between a regulatory protein and DNA, and a new set of hydrogen bonds is formed between polypeptide amide groups and DNA bases. These structural changes permit the two polypeptide chains to be brought closer to each other and ultimately result in the structure described above. By contrast, all interchain hydrogen bonds existing in a β-sheet are preserved when a gg-structure is formed.

Although the proposed gt-structure is the most compact one, our model building study showed that it could not be accommodated in the minor groove of DNA in B conformation due to the formation of unacceptable short Van der Waals contacts between the atomic groups of the t-chain and the phosphate-deoxyribose backbone. This drawback cannot be eliminated by a slight structural modification of the polypeptide double helix. In accord with recent experimental evidence [19, 20] we suggested the DNA configuration to be somewhat affected on binding with regulatory proteins. Short contacts between t-chain and phosphate-deoxyribose backbone can be eliminated by tilting of DNA base pairs so that normals to base pairs planes form an angle of about 15° with

the helix axis. Although a detailed description of the structure can be provided after the completion of calculations of energy minimum conformations it is now clear that base pairs are tilted in the complex in nearly the same way as they are in the A form of DNA. Calculations of the DNA conformation demonstrated the stereochemical possibility for structures with tilting angle of 15° and sugar ring puckering C3′-exo (B form) and C3′-endo (A form) (Tumanyan, V. G., unpublished data). One of these structures with standard C3′-exo deoxyribose ring puckering was used in building of the model shown in Figure 2. The DNA structure has a relatively large axial rise (3.8 Å) and twist angle (40°) per residue while the magnitude of base pair displacement from the helix axis towards the side of minor groove is nearly the same as in the B conformation. A convenient measure of this displacement is the distance from the helix axis to the line connecting C1′ deoxyribose atoms in a base pair. This distance is 2.5 Å, i.e. 0.3 Å greater than that in B DNA. All nonbonded interatomic distances between polypeptide chains and DNA are satisfactory except for the distance between ring oxygen of deoxyribose and C^{β} atom of t-chain. This distance is 2.5 Å which is 0.2 Å shorter than the minimum allowed approach. A more satisfactory model could be constructed if some distortion of bond lengths and angles in the polypeptide chains is allowed.

It is of interest that polynucleotide chains in the proposed model may have C3′-endo furanose ring pucker as in A-type of double-helices. The model can be further adjusted to accommodate the free hydroxyls of ribose rings. This indicates that stereospecific sites of proteins recognizing specific base sequences in double-helical RNA may have a structure similar to that described above. Structures with a considerably greater base pair displacement towards the side of the minor groove are also possible. As it follows from Table I, they should be characterized by lower values for axial rise ($\simeq$ 3 Å) and twist angle ($\simeq$ 30°) per residue. Refinements of the complex structure along these lines are now proceeding.

III. GENERAL FEATURES OF THE CODE CONTROLLING REGULATORY PROTEIN BINDING TO DNA

The suggested model accounts for the main properties of base pair sequences at the control sites on DNA. These sequences all have long stretches along one DNA strand with no guanine whereas guanine is present in the complementary strand in certain places and is necessary for protein-DNA recognition. This can be exemplified by base pair sequences in the lac operator determined by Gilbert and Maxam [39], by a more complete sequence including lac promotor and lac operator [40] as well as by sequences in the right and left lambda operators and promotors [41–44]. GC substitution by AT in essential sites of these sequences markedly decreases the binding constant of the regulatory protein. The stereospecific sites of regulatory proteins are assumed to be complementary to regulatory base sequences with the t-chain segment forming hydrogen bonds with all bases in one DNA strand and the g-chain segment interacting only with guanine bases in another strand. This property permits regulatory proteins to distinguish control sites from the other part of DNA whose guanine residues are distributed more symmetrically between the two polynucleotide strands. The question naturally arises how do regulatory proteins recognize their specific binding sites among all control sites available. To answer this question it seems necessary to understand the effect of amino acid side-chains on the backbone conformations of polypeptide chain segments involved in the stereospecific protein site. In other words, it is essential to elucidate what amino

acid sequences must be present in the stereospecific protein site to recognize a particular base sequence on DNA.

It was already noted that in the minor groove of DNA only guanine can serve as a hydrogen bond donor, whereas thymine, cytosine and adenine may serve as hydrogen bond acceptors. We have, therefore, postulated that the required specificity of protein binding arises from the fact that hydrogen bonding between guanine residues and the g-chain segment is controlled by the amino acid sequence in the stereospecific protein site. Having systematically studied the possible role played by various amino acid residues we conclude that side-chains of certain residues in the t-polypeptide chain segment could form hydrogen bonds with the carbonyl groups of the g-chain segment, thereby breaking or strongly weakening hydrogen bonds between the g-chain segment and guanine bases. Only six such amino acid residues were found: serine, threonine, asparagine, histidine, cysteine and glutamine; two of them forming O – H • • • O hydrogen bonds (serine, threonine), three others N – H • • • O bonds (histidine, asparagine, glutamine) and one S – H • • • O bond (cysteine). In all these residues atomic groups serving as hydrogen bond donors can occupy nearly identical spatial positions. We suggest that all these residues code for AT base pairs while residues which are unable to interact by this mechanism code for GC pairs.

Amino acid residues in the polypeptide double-helix can be also divided into two classes depending on their geometrical positions: i.e., external residues whose side-chains point toward the helix axis and are located on the periphery of the structure, and internal residues whose side-chains point out and are projected on the middle region of the structure (Figures 1 and 2). Their functional role is also different. External residues take part in specific interactions with DNA bases. Internal residues in the t-chain segment participate in coding. If the upper residue R_1 in the t-chain segment (see Figure 2) is an AT-coding residue then its side-chain can be hydrogen bonded to the backbone carbonyl group of the residue R_2 lying in the g-chain segment. Such hydrogen bonding is accompanied by a rotation of the corresponding amide group of the g-chain segment (Figure 3) and results in the weakening or breaking of the hydrogen bond between the above amide group and guanine. In the last case a conformation such as g′ arises (Table I) which is in the allowed region of the Ramachandran diagram, just as are conformations intermediate between g and g′.

Stereospecific protein sites are assumed to contain t- and g-polypeptide chain segments forming a gt-double helical structure with single-stranded regions at the ends of the structure. The single-stranded polypeptide chain regions belonging to stereospecific sites of different protein subunits act probably as cohesive ends. They are responsible for cooperative effects in protein subunit binding, thereby facilitating accurate recognition of the correct protein binding sites. Cohesive ends of neighbouring protein subunits can form gg- and gt-structures only when the protein subunits are bound to their correct positions on DNA (see accompanying paper [7]). The code controlling specific protein binding to DNA is, therefore, to a large extent degenerated in respect to amino acid sequences in stereospecific protein sites. It requires the presence of amino acid 'switchers off' in certain places of the sequence and also requires all amino acid residues in the stereospecific sites to be compatible with an antiparallel β-structure. Such a strong code degeneracy is of biological significance since there are several functional advantages associated with it. Amino acid residues in stereospecific sites of regulatory proteins are likely to perform several functions simultaneously. Besides participating in coding and ensuring specific contacts with DNA bases they also provide for fixing of t- and g-polypeptide chain segments together in a very precise position on the protein surface and determine the steric features of bonded areas in protein-DNA complexes. Clearly, the

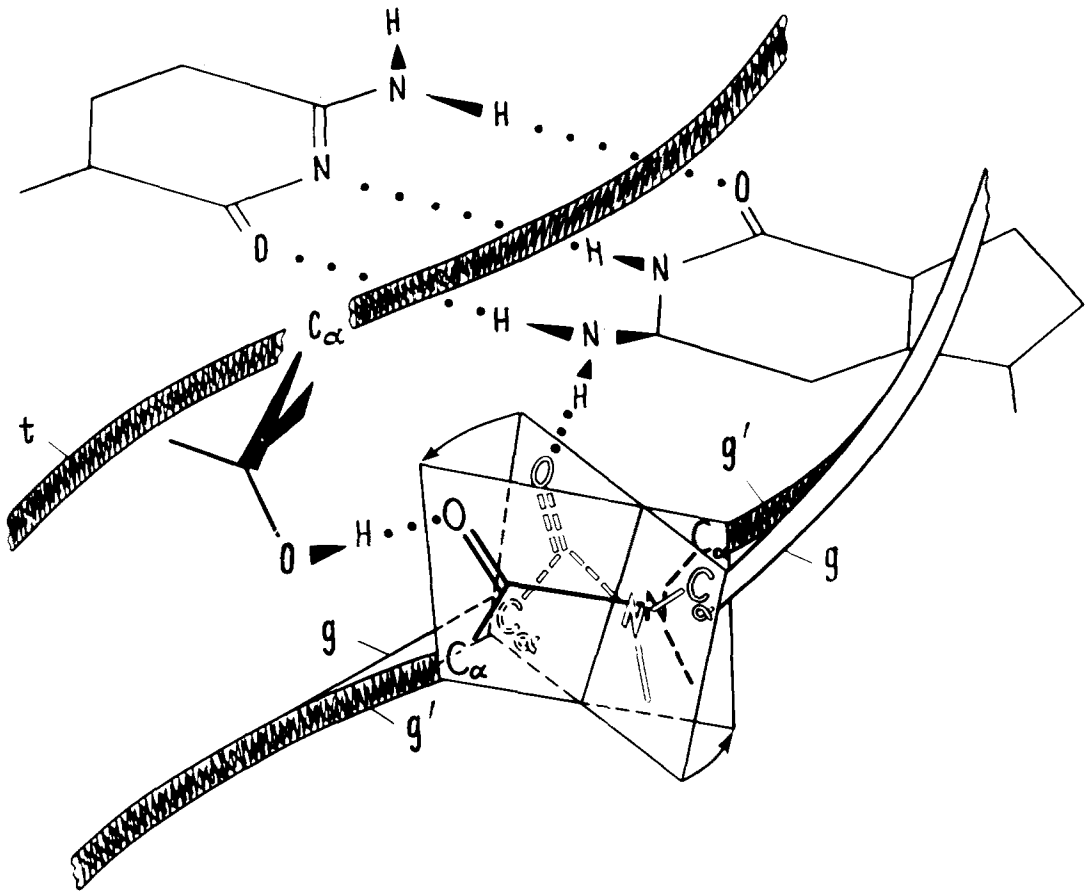

Fig. 3. A diagram illustrating the mode of action of AT-coding amino-acid residues. In the absence of an AT-coding amino acid residue a hydrogen bond is formed between the guanine 2-amino group and the amide group of a g-polypeptide chain segment in the stereospecific protein site. In the presence of an AT-coding residue (serine) this hydrogen bond is weakened or broken, and a new hydrogen bond is formed between the serine side-chain hydroxyl and the amide carbonyl oxygen. This hydrogen bonding is accompanied by a certain rotation of the amide group as a whole around the axis running along the polypeptide chain direction. The two special positions of the amide group in the presence and absence of a hydrogen bond with guanine 2-amine group are shown by continuous and dashed lines, respectively.

high degree of code degeneracy allows one to design proteins with adequate structural and functional properties in a more economical manner. One may, therefore, expect that molecular evolution should favour the maintenance of code degeneracy. Many variations in protein sequence associated with the stereospecific site appears to be not functional because of code degeneracy. This allows the tertiary protein structure to be perfected in the process of molecular evolution leaving the capacity for detecting 'correct' binding sites on DNA to some extent unaffected. Limitations imposed on amino acid sequence by the tertiary constraints are apparently much more rigid than those due to the code existence.

We consider now fundamental restrictions imposed on protein sequence by steric features of bonded areas in protein-DNA complexes. According to the proposed model the stereospecific protein site appears to be a protuberance on the protein molecule that can be wrapped around the minor groove on DNA. Since the minor groove is the region with the greatest density of negative

charge on the DNA surface, the negatively charged residues such as glutamic and aspartic acids should conceivably be not available in stereospecific sites. This holds true in particular for residues with inward-pointing side-chains (R_2) which are in close proximity to phosphate groups in protein-DNA complexes. On the contrary, positively charged side-chains of lysine, arginine and histidine

TABLE II: The code rules[a] and stereochemical limitations[b] imposed on the amino acid sequence in the stereospecific protein site.

Base pair	Type of polypeptide chain segment	Outside-pointing side-chain, R_1	Inward-pointing side-chain, R_2
A T • • • or • • • T A	t	[a] Ser, Thr, Asn, His, Gen, Cys	
		[b] Any residue, except for Pro	
	g	[b] Glu, Asp, Lys, Arg and His are unlikely to occur	
C • • •	t	[a] Gly, Ala, Val, Leu, Phe, Ile, Met, Tyr, Trp	[b] Any residue, except for Pro, Asp, Glu
		[b] Glu, Asp, Lys and Arg are unlikely to occur	
G	g	[b] Any residue, except for Pro	
		[b] Glu, Asp, Lys, Arg and His are unlikely to occur	
G • • C	t g	[a] This situation is unfavourable for accurate recognition and must occur rarely	

[a] The code limitations are imposed on residues with outside-pointing side-chains lying in the t-chain segment. These must be either serine, threonine, asparagine, cysteine, glutamine or histidine for recognition of an AT pair; for recognition of an GC pair the above residues and proline should be absent.

[b] Stereochemical limitations are imposed on both inward and outside-pointing side-chains and concern with the occurrence of charged residues and proline in t- and g-chain segments. A possible role of glutamine residues in recognizing a two-fold symmetry axis in regulatory base sequences is considered in the accompanying paper [7].

and polar side-chain of tyrosine occur preferentially in inward-pointing positions allowing hydrogen bond formation with phosphate groups. Such hydrogen bonding is impossible when these side-chains occur in outside-pointing positions in polypeptide double-helix. In our proposed code each AT and GC base pair is correlated with respective four amino acids two of which are in the t-chain segment and two in the g-segment. Table II summarizes the code rules and stereochemical limitations imposed on these four residues. One must, however, bear in mind a fundamental limitation imposed on the protein sequence in the stereospecific site, that of obligatory formation of an anti-parallel β-structure.

The recognition of control sites by regulatory systems is based on a correspondence between protein and DNA sequences which results in formation of numerous hydrogen bonds between the polypeptide amide groups and DNA bases. When the GC pair is in its proper place in the control site a protein molecule can be attached to it through two hydrogen bonds (one with guanine, the other one with cytosine), with only one such bond formed in the case of an AT pair. When the GC pair is not in its proper place the hydrogen bond with guanine is broken or strongly weakened owing to the presence of an AT-coding amino acid residue in the stereospecific protein site, leaving only one comparatively weak hydrogen bond with a cytosine carbonyl group (the latter already being involved in hydrogen bonding to the guanine 2-amino group in the DNA duplex). The only degeneracy to be allowed for is due to thymine and adenine being recognized as one and the same letter. This degeneracy may, however, be considerably reduced if hydrogen bonding with thymine is much stronger than that with adenine. The stereochemical basis for this is the fact that purine atom N3 lies by 0.5 Å closer to the helix axis than thymine oxygen O2 [28]. Obviously, the code degeneracy can be completely eliminated when a protein forms hydrogen bonds with adenine and/or thymine in the major groove of DNA in addition to the specific contacts in the minor groove. At present, however, it remains unclear to what extent the code is degenerated in various real systems.

ACKNOWLEDGEMENTS

We thank Prof. V. A. Engelhardt for helpful discussions and critical reading of the manuscript and Prof. A. L. Pumpyansky for English translation of the manuscript.

REFERENCES

1. Chamberlen, M. J., *Ann. Rev. Biochem.* **43**, 721 (1974).
2. Von Hippel, P. H. and McGhee, J. D., *Ann. Rev. Biochem.* **41**, 231 (1972).
3. Gilbert, W., Maizels, N., and Maxam, A., *Cold Spring Harbor Symp. Quant. Biol.* **38**, 845 (1973).
4. Maniatis, T., Ptashne, M., and Maurer, R., *Cold Spring Harbor Symp. Quant. Biol.* **38**, 857 (1973).
5. Murray, K. and Old, R.W., *Progr. Nucl. Acid Res. Mol. Biol.* **14**, 117 (1974).
6. Arber, W., *Progr. Nucl. Acid Res. Mol. Biol.* **14**, 1 (1974).
7. 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. 427–434.

8. Adler, K., Beyreuther, K., Fannig, E., Geisler, N., Gronnebern, B., Klemm, A., Muller-Hill, B., Pfahl, M., and Schnitz, A., *Nature* **237**, 322 (1972).
9. Pelc, S. R. and Welton, M. G. E., *Nature* **209**, 868 (1966).
10. Patel, D. J., *Biochemistry* **14**, 1057 (1975).
11. Ivanov, V. I., submitted to *FEBS Letters.*
12. Helene, C., *Nature New Biol.* **234**, 120 (1971).
13. Crothers, D. M., *Biopolymers* **6**, 575 (1968).
14. Zasedatelev, A. S., Gursky, G. V., and Volkenstein, M. V., *Mol. Biol. U.S.S.R.* **5**, 245 (1971).
15. McGhee, J. D. and von Hippel, P. N., *J. Mol. Biol.* **86**, 469 (1974).
16. Gursky, G. V., Zasedatelev, A. S., and Volkenstein, M. V., *Mol. Biol. U.S.S.R.* **6**, 479 (1972).
17. Zasedatelev, A. S., Gursky, G. V., and Volkenstein, M. V., *Studia Biophys.* **40**, 79 (1973).
18. Livschitz, M. A., Gursky, G. V., Zasedatelev, A. S., and Volkenstein, M. V., submitted to *Mol. Biol. U.S.S.R.*
19. Maniatis, T. and Ptashne, M., *Proc. Nat. Acad. Sci. U.S.A.* **70**, 1531 (1973).
20. Wang, J. C., Barkley, M. D., and Bourgeois, S., *Nature* **251**, 247 (1974).
21. Carter, C. W. and Kraut, J., *Proc. Nat. Acad. Sci. U.S.A.* **71**, 283 (1974).
22. Zasedatelev, A. S., Gursky, G. V., Zimmer, Ch., and Thrum, H., *Mol. Biol. Reports* **1**, 334 (1974).
23. Luck, G., Triebel, H., Waring, M., and Zimmer, Ch., *Nucl. Acid Res.* **1**, 503 (1974).
24. Zasedatelev, A. S., Gursky, G. V., Zimmer, Ch., and Thrum, H., submitted to *Nucl. Acid Res.*
25. Zhuze, A. L., Gottikh, B. P., Grokhovsky, S. L., Gursky, G. V., Zasedatelev, A. S., and Tumanyan, V. B., Abstracts 9th Int. Congress of Chemotherapy, M-43 (1975).
26. Kolchinsky, A. M., Mirzabekov, A. D., Zasedatelev, A. S., Gursky, G. V., Grokhovsky, S. L., Zhuze, A. L., and Gottikh, B. P., *Mol. Biol. U.S.S.R.* **9**, 19 (1975).
27. Wartell, R. M., Larson, J. E., and Wells, R. D., *J. Biol. Chem.* **249**, 6719 (1975).
28. Arnott, S. and Hukins, D. W. L., *Biochem. Biophys. Res. Comm.* **47**, 1504 (1972).
29. Bruskov, V. I. and Poltev, V. I., *Dokl. Acad. Sci. U.S.S.R.* **219**, 231 (1974).
30. Ramachandran, G. N. and Sasisekharan, V., *Adv. Protein Chem.* **23**, 284 (1968).
31. Sugita, H. and Miyazawa, T., *Biopolymers* **5**, 673 (1967).
32. De Santis, P., Forni, E., and Rizzo, R., *Biopolymers* **13**, 313 (1974).
33. Tumanyan, V. G. and Esipova, N. G., *Dokl. Acad. Sci. U.S.S.R.* **218**, 1222 (1974).
34. Tumanyan, V. G. and Esipova, N. G., *Biopolymers*, in press.
35. Cowan, P. M. and McGavin, S., *Nature* **176**, 501 (1955).
36. Leach, S. J., Nemethy, G., and Scheraga, H. A., *Biopolymers* **4**, 887 (1966).
37. 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.
38. Ghothia, C., *J. Mol. Biol.* **75**, 295 (1973).
39. Gilbert, W. and Maxam, A., *Proc. Nat. Acad. Sci. U.S.A.* **70**, 3581 (1973).
40. Dickson, R. C., Abelson, J., Barnes, W. M., and Reznikoff, W. S., *Science* **187**, 27 (1975).
41. Maniatis, T., Ptashne, M., Barrel, B. G., and Donelson, J., *Nature* **250**, 394 (1974).
42. Maniatis, T., Ptashne, M., Backman, K., Kleid, D., Flashman, S., and Jeffrey, A., *Proc. Nat. Acad. Sci. U.S.A.* **72**, 1184 (1975).
43. Pirrota, V., *Nature* **254**, 114 (1975).
44. Walz, A. and Pirrota, V., *Nature* **254**, 118 (1975).