Hypothesis

SPECIFIC RECOGNITION OF GUANINE BASES IN PROTEIN–NUCLEIC ACID COMPLEXES

Claude HELENE

Centre de Biophysique Moléculaire, 45045 Orléans Cedex, France

Received 27 December 1976

1. Introduction

The selective association of nucleic acids with proteins is of utmost importance at every step of genetic expression. The most striking examples include the recognition of operators by repressors, of promoters by RNA polymerase and of DNA base sequences by restriction endonucleases. Three main topics have to be considered when dealing with the problem of selective recognition between proteins and nucleic acids:

(i) There should be a structural complementarity between the regions of the two macromolecules which are in contact.

(ii) Direct interactions have to take place between the chemical groups involved in the interacting regions (amino acid side-chains and peptidic group of the protein; bases, sugars and phosphates of the nucleic acid).

(iii) Direct interactions between these chemical groups could be mediated through a third species such as a metal cation [1].

Direct interactions include:

(i) Electrostatic interactions between positively charged amino acid side-chains (Lys, Arg, protonated His) and phosphate groups.

(ii) Stacking interactions between aromatic amino acid side-chains (Trp, Tyr, Phe, His) and nucleic acid basis.

(iii) Hydrogen bonding between several amino acid side-chains or the peptidic group, and phosphate, ribose or bases.

(iv) Hydrophobic interactions between aliphatic amino acid side-chains and nucleic acid bases.

Several studies have been devoted to stacking interactions involving aromatic amino acids and nucleic acid bases [2–4]. It has been demonstrated that peptides containing aromatic residues could recognize single-stranded from double-stranded nucleic acids [3,5]. Thus, a simple tripeptide such as Lys–Trp–Lys is able to bind selectively to unpaired regions in DNA which has been submitted to ultraviolet irradiation [5]. In single-stranded polynucleotides, the stacking interaction of Trp residues is dependent on the base sequence (Boubault, Maurizot and Hélène, to be published). The role of aromatic residues of proteins in their binding to single-stranded nucleic acids has been recently emphasized in the case of gene 5 protein from phage fd [6] and that of gene 32 protein from phage T4 [7a,b]. Proteins could also use aromatic residues to anchor to single-stranded regions of nucleic acids such as tRNAs [8].

Hydrogen-bond formation between amino acid side-chains and nucleic acid bases certainly represents one of the most specific ways in which proteins can recognize base sequences [9–12]. There are several amino acid side-chains which have the capability of forming pairs of hydrogen-bonds with bases. These include the acidic residues Glu and Asp, the amide groups of Gln and Asn, the imidazole ring of His, the guanidinium group of Arg.

Amide side-chains of Gln and Asn have one donor and one acceptor group. They can therefore form two hydrogen-bonds with several bases: adenine [NH_2(6) and N(7)] and guanine [NH_2(2) and N(3)] in the major and minor groove, respectively, of double-stranded nucleic acids (fig.1). All four bases are in single-strands. The same situation prevails for carboxylic acids in their unionized ($-C\overset{=O}{\underset{O-H}{}}$) form. They can give rise to the same type of hydrogen-bonds as amides (fig.1). It should be kept in mind that ionizable

North-Holland Publishing Company – Amsterdam

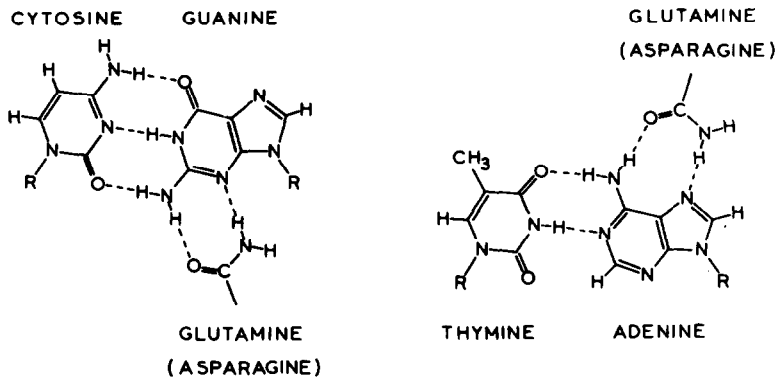

Fig.1

side-chains may have abnormal pK values in proteins and this also holds for protein-nucleic acid complexes. Therefore glutamic and aspartic acids could well exist as neutral species under physiological conditions. The formation of hydrogen-bonded complexes between unionized carboxylic acids and nucleic acid bases has already been demonstrated [9,13].

The imidazole ring of His in its neutral or protonated form could also be considered as a potential bridging group between hydrogen-bonding sites of bases. Such a model has been proposed where neutral His makes a bridge between adenine and thymine in histone-DNA interactions [14].

We now consider specific recognition of guanine in double-stranded and single-stranded nucleic acids. There are two types of side-chains which should be expected to have a more specific behavior. These are the chemical groups which possess only hydrogen-bonding donor (Arg) or hydrogen-bonding acceptor groups (ionized carboxylic acids of glutamyl and aspartyl side-chains).

Only one base has two hydrogen-bond donor groups in a suitable position to form two hydrogen-bonds with ionized Glu and Asp side chains. This base is guanine in single-stranded nucleic acids with NH(1) and NH_2(2) as hydrogen-bond donors (fig.2a).

The side-chain of arginine appears to be of special importance since all its chemical groups are hydrogen-bonding donors (two NH_2 groups and one NH group). In double-stranded nucleic acids, all bases still have hydrogen-bonding possibilities acting either as donor or acceptor. But only one of them, guanine, has two hydrogen-bond acceptor groups (O(6) and N(7)). They are located in the major groove. This should enable this base to form two hydrogen-bonds with an arginine side-chain as shown in fig.3.

The only other possibility for Arg of forming two hydrogen-bonds will be with cytosine in a single-stranded nucleic acid. Cytosine has two hydrogen-bond acceptor groups (N(1) and O(2)) which are involved in pairing with guanine in double-strands but will be available in single-strands (fig.2b).

However the arginine side-chain also bears a posi-

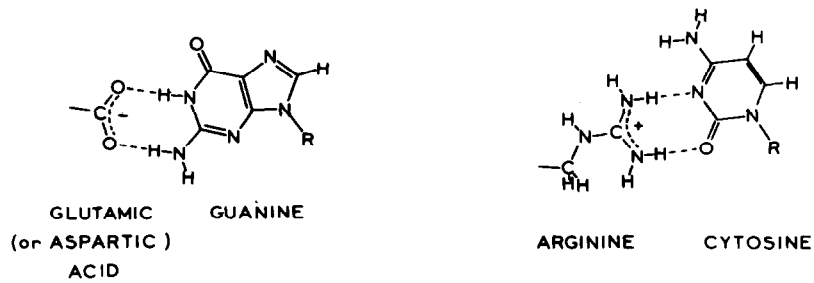

Fig.2

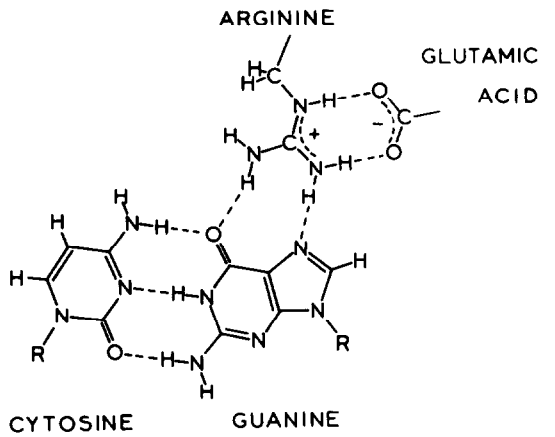

Fig.3

tive charge which makes it a good candidate for electrostatic interaction with a phosphate group of the nucleic acid. This arginine–phosphate interaction has been shown to take place, e.g., in the binding of the inhibitor deoxythymidine-3′,5′-diphosphate to staphylococcal nuclease [15]. The precise nature of this interaction has been characterized by X-ray studies of crystals of methyl guanidinium dihydrogenorthophosphate [16]. This crystal studies provide an excellent model of arginine-phosphate hydrogen-bonding interactions. From these studies one would expect a strong interaction between positively charged arginine side-chains and negatively charged phosphate groups. Thus the arginine side-chain would not be readily available to bind guanine residues.

We would like to propose a model in which this last interaction could take place. The positive charge of the arginine side-chain could be neutralized by a carboxylate anion whose two oxygen atoms could form hydrogen-bonds with two NH groups of Arg (as observed in the case of phosphate anions). We have built a space-filling model of the sequence Arg–Glu. The carboxylic group of Glu and two hydrogen-bond donor groups of Arg (one NH_2, NH) can form the structure shown in fig.1 in which the planar guanidinium group [16] and the carboxylic group are coplanar. The sequence Arg–Asp would be less appropriate since it will not give such a coplanar structure. However it should be noted that the arginine and carboxylic acid side-chains need not be adjacent in a protein molecule. Two such side-chains might interact with each other to form the structure shown in fig.3 if they are brought in close contact by the folding of the polypeptide chain. Nevertheless the Arg–Glu sequence would be quite favorable. It will leave the possibility of forming two hydrogen-bonds between Arg and guanine. It may be proposed therefore that this Arg–Glu sequence should allow the recognition of guanine bases in a double-stranded nucleic acid (of both guanine and cytosine in single-strands). This interaction would be favored if the Arg–Glu sequence was brought in close contact with nucleic acid bases by introducing neighboring basic amino acid residues giving rise to electrostatic interactions with phosphate groups. Thus one would predict that a sequence in which Arg–Glu has a basic neighbor (such as Arg–Glu–Lys or Lys(Arg)–Arg–Glu) would be a good candidate for this specific interaction.

In order to test the proposal made above for a specific role of Arg–Glu sequence in recognizing guanine in double-stranded nucleic acids (guanine and cytosine in single-stranded nucleic acids) we have looked for the presence of such sequences in proteins interacting with nucleic acids. In ribosomal S4 protein the sequence Arg–Glu–Lys appears three times, Arg–Arg–Glu once and Glu–Arg twice. Protein S4 binds to ribosomal 16 S RNA. Ultraviolet irradiation of the complex induces the formation of covalent-bonds between the protein and the nucleic acid [17]. An analysis of the tryptic peptides which are cross-linked to 16 S RNA reveals that two of the three S4 protein regions containing the sequence Arg–Glu–Lys are in close contact with the RNA. (Tryptic digestion gives two cross-linked peptides beginning by the sequence Glu–Lys, tryptic cleavage occurring after the Arg residue). A third cross-linked peptide contains the sequence Glu–Arg. The chemical nature of the covalent-bonds formed between protein S4 and 16 S RNA upon ultraviolet irradiation is not known. Several amino acid side-chains are known to react photochemically with bases [18,19] (Cys, Ser, Thr, Tyr, Lys, Arg). We have recently shown that carboxylic side-chains of Glu and Asp could also participate in the formation of photochemically-induced cross-links (Toulmé and Hélène, to be published).

The sequence of the *lac*-repressor has been determined [20]. Genetic and chemical studies have shown that the N-terminal part of this protein is involved in the selective recognition of the *lac*-operator [21]. The

sequence Arg–Glu–Lys appears in position 35–37 and could be involved in binding selectively to a G–C base-pair such as the single G–C base-pair of the first long symmetrical base sequence of the *lac*-operator ($^{A}_{T}$AATTGT$^{T}_{A}$). It should be noted that the region which contains the Arg–Glu sequence is the most basic sequence of the N-terminal part of the *lac*-repressor ($-Ly\overset{33}{s}$–Thr–Arg–Glu–$Ly\overset{37}{s}$–) which is expected to be strongly bound to the *lac*-operator by electrostatic interactions.

If there is a molecular code for the recognition of nucleic acid base sequences by proteins, the considerations developed above could provide a basis for the selective recognition of guanine bases in double-stranded nucleic acids by Arg–Glu sequences and that of guanine in single-stranded nucleic acids by carboxylic side-chains. Some of the possibilities presented above can be experimentally tested. We are presently synthesizing several amino acid sequences of the *lac*-repressor including the Arg–Glu–Lys sequence. The investigation of base sequence specificity in the binding of these peptides to different nucleic acids should allow us to test the validity of the proposed model.

References

[1] Hélène, C. (1975) Nucleic Acids Res. 2, 961–969.
[2] Dimicoli, J. L. and Hélène, C. (1974) Biochemistry 13, 714–723 and 723–730.
[3] Brun, F., Toulmé, J. J. and Hélène, C. (1975) Biochemistry 14, 558–563.
[4] Gabbay, E. J., Sanford, K., Baxter, C. S. and Kapicak, L. (1973) Biochemistry 12, 4021–4029.
[5] Toulmé, J. J., Charlier, M. and Hélène, C. (1974) Proc. Natl. Acad. Sci. USA 71, 3185–3188.
[6] Anderson, R. A., Nakashima, Y. and Coleman, J. D. (1975) Biochemistry 14, 907–917.
[7](a) Anderson, R. A. and Coleman, J. E. (1975) Biochemistry 14, 5485–5491.
(b) Hélène, C., Toulmé, F., Charlier, M. and Yaniv, M. (1976) Biochem. Biophys. Res. Commun. 71, 91–98.
[8] Hélène, C. (1971) FEBS Lett. 17, 73–77.
[9] Sellini, H., Maurizot, J. C., Dimicoli, J. L. and Hélène, C. (1973) FEBS Lett. 30, 219–224.
[10] Bruskov, V. I. (1975) Mol. Biol. (Moscow) 9, 304–309.
[11] Gursky, G. V., Tumanyan, V. G., Zasedatelev, A. S., Zhuze, A. L., Grokhovsky, S. L. and Gottikh, B. P. (1975) Mol. Biol. (Moscow) 9, 635–651.
[12] Seeman, N. C., Rosenberg, J. M. and Rich, A. (1975) Proc. Natl. Acad. Sci. USA 73, 804–808.
[13] Lancelot, G. (1977) Biophys. Chem (in press) and Biophys. J. (in press).
[14] Gourévitch, M., Puigdomenech, P., Cavé, A., Etienne, G., Méry, J. and Parello, J. (1974) Biochimie 56, 967–985.
[15] Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E., Richardson, D. C., Richardson, J. S. and Yonath, (1971) J. Biol. Chem. 246, 2302–2316.
[16] Cotton, F. A., Day, V. W., Hazen, E. E. and Larsen, S. (1973) J. Am. Chem. Soc. 95, 4834–4840.
[17] Ehresmann, B., Reinbolt, J. and Ebel, J. P. (1975) FEBS Lett. 58, 106–111.
[18] Varghese, A. J. (1976) in: Aging, Carcinogenesis and Radiation Biology (Smith, K. C. ed) 207–223, Plenum Press.
[19] Elad, D. (1976) in: Aging, Carcinogenesis and Radiation Biology (Smith, K. C. ed) 243–260, Plenum Press.
[20] Beyreuther, K., Adler, K., Farming, E., Murray, C., Klemm, A. and Geisler, N. (1975) Europ. J. Biochem. 59, 491–509.
[21] Müller-Hill B. (1975) Prog. Biophys. Mol. Biol. 30, 227–252.