# THE BINARY CODE FOR PROTEIN–NUCLEIC ACID RECOGNITION WITH REPULSIVE GUANINE: APPLICATION TO tRNA CASE

Valery I. IVANOV
*Institute of Molecular Biology, USSR Academy of Sciences, Moscow 117312, U.S.S.R.*

Received 13 May 1975
Revised version received 18 August 1975

## 1. Introduction

Restrictases [1], repressors [2,3], RNA polymerases [4], aminoacyl-tRNA synthetases [5] can recognize specific nucleotide sequences in the double-stranded segments of a nucleic acid molecule. The recognition may be realized by a specific system of hydrogen bonds between the recognizing groups of the protein ('legs') and the polar groups of the base pairs exposed to the helicals grooves of complementary helix [2]. A general consideration of such models from statistico-mechanical point of view is given in the work of Livshitz et al. [6].

Below an actual code is suggested for protein-nucleic acid recognition with the participation of the small, or 'sugar' groove. The main feature of this variant is the formation of an H-bond between the protein 'leg' and any neighbouring base, except guanine, at which, on the contrary, repulsion arises. This scheme explains many data on tRNA recognition by synthetases.

## 2. Stereochemistry of the complex of the 'leg' with the bases of the double helix

Fig.1 shows that the chemically identical groups, exposed to the major, or non-sugar, groove, e.g. $NH_2$ of adenine or cytosine and C=O of guanine or uracyl (thymine), are placed in different way relatively the dyad axis. At the same time the positions of $N_3$ of G or A and $O_2$ of C or U(T) in the sugar groove are practically the same [7].

Thus, the sugar groove side of all the bases, except G, which has an extra 2-$NH_2$ group, looks similar. The

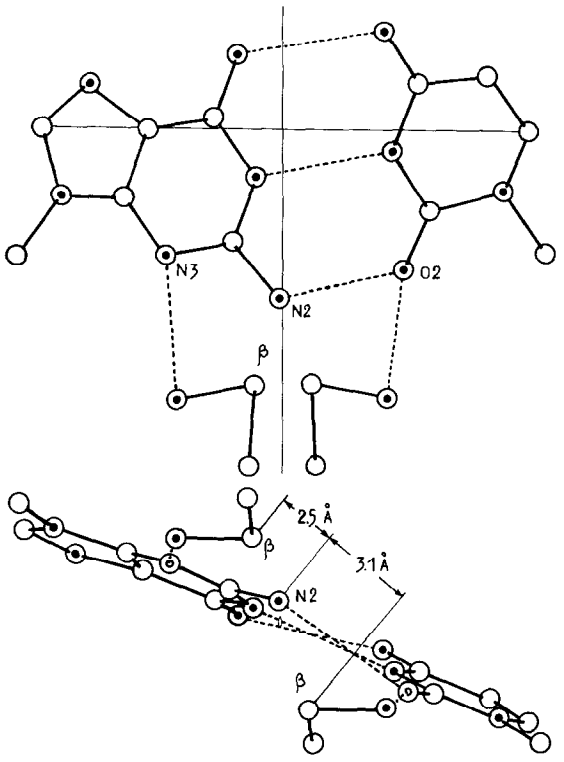


Fig.1. Recognition by the sugar groove of the double helix. Above is the projection along the axis of the helix; below is that along the dyad axis. The H-bonds of the Ser legs of a protein with purine $N_3$ and pyrimidine $O_2$ atoms are shown. If the purine is a guanine, then the H-bond cannot be formed due to a close contact of the $NH_2$ group and the Ser β-carbon atom (see the text). Indicated are the dyad and twist axes.

fact, that antibiotic netropsin binds to DNA, dA·dT, dI·dC, but not to dG·dC [8,9] shows that the 2-$NH_2$ group can indeed play a discriminatory role. It is therefore tempting to construct a scheme, which will demonstrate how such a discrimination can be realised. The main obstacle is that the 2-$NH_2$ group is projected exactly on the dyad axis (fig.1) and therefore this group might be thought to interfere not only with the leg for G but with the other leg for the complementary C as well (fig.1, above). However, if one thinks on the recognition of tRNA by the stems (see further), then no sugar groove mechanism with AU pairs only is probable, because in many cases the tRNA stems consist of GC pairs entirely [10], and the isoacceptor $tRNA^{Met}$ and $tRNA_f^{Met}$ have no identically placed AU pairs at all [10]. Fig.1 (below) shows how the formation of the twist from the bases of a complementary pair, that is quite typical in the models of A-RNAs [11], solves the problem of C recognition. Due to a twist, the 2-$NH_2$ group of G is moved up to the leg, which would bind to the $N_3$ atom, and away from the leg which is intended for C.

Recently Carter and Kraut [12] have proposed a stereochemical model of the complementary RNA–polypeptide complex. It is remarkable that this model can easily be transformed into the one recognizing nucleotide sequences – it is only needed to substitute the inward-pointing $\beta$-$CH_3$ groups of the Carter–Kraut model by $CH_2$-OH or $CH_2SH$ groups of Ser or Cys (fig.1). The atomic co-ordinates given in the study [12] allow calculating the distances between $\beta$-carbon atoms of thus created legs and 2-$NH_2$ groups of G. The contact of the $\beta$-carbon atom of the leg neighbouring G and 2-$NH_2$ group proves to be impossibly tight, 2.5 Å, while the corresponding contact for the cytosine recognizing leg is equal to 3.1, that is close to a sum of Van der Waals radii of N and C (fig.1, below). Thus of all the bases only the G will be repulsive. Since the contact is not due to a group directly involved in H-bond, but is due to the preceding $CH_2$ group, the legs may be built of –$CH_3$ groups as well, providing that some non-specific interactions with the sugar-phosphate backbone are responsible for attraction.

The potential possibility to create a recognition mechanism along the sugar groove of double helical RNA according to the scheme A=U=C≠G is thus demonstrated.

Let us consider now the application of this binary G–non-G code.

### 3. Recognition of tRNA by tRNA synthetase

Recognition of tRNA by the ACC stem [5,13] or DHU stem [14] was repeatedly surmised. No mode of the recognition, however, was suggested.

I shall guess a tRNA synthetase to be a kind of the 'many-legged' protein, whose legs (possibly Ser, Cys or Ala residues) are situated in such a way that they do not encounter the Gs of the cognate tRNAs. Ser and Cys would form the H-bonds depicted in fig.1, with A, C and U.

Direct comparison of the known primary structures of tRNAs in a cloverleaf configuration reveals that in no case the tRNA molecules of different specificities had an identical G-pattern in the double helical moiety. This inference is based upon more than 60 structures collected in the handbook [10]. But some individual stems in the different tRNAs may be identical or similar (in terms of this degenerative G–non-G code). For example, the *E. coli* $tRNA_1^{Arg}$ and $tRNA_{2a}^{Val}$ have the same TΨC and DHU stems and very similar ACC stems with greatly different ANT stems (fig.2). The ANT stems should thus be involved in recognition area. Experimental data on a discriminative role of the 4th nucleotide from 3′-end make this non-paired base be included into the totality of recognizible elements [15,16]. (But its mode of recognition may be different).

Here are three recognition phenomena that I wish to discuss:

*Isoacceptor tRNAs.* They are known to have very different sequences sometimes, but can nevertheless be aminoacylated with the similar kinetic parameters. The examples for $tRNA_{1,2}^{Gln}$ or $tRNA^{Met}$ and $tRNA_f^{Met}$ in fig.2 show that the number of the coincident elements with the discussed degenerative code is great enough for one and the same combination of the synthetase legs can be equally effective in recognition of isoacceptors.

*Mutations in the ACC stem.* Direct confirmation of the repulsive function of G comes from the mutations in $su^+tRNA^{Tyr}$ of *E. coli* [5,17,18]. I shall discuss the double mutants only in which an AU pair was substituted for the GC pair (fig.2). A double mutant

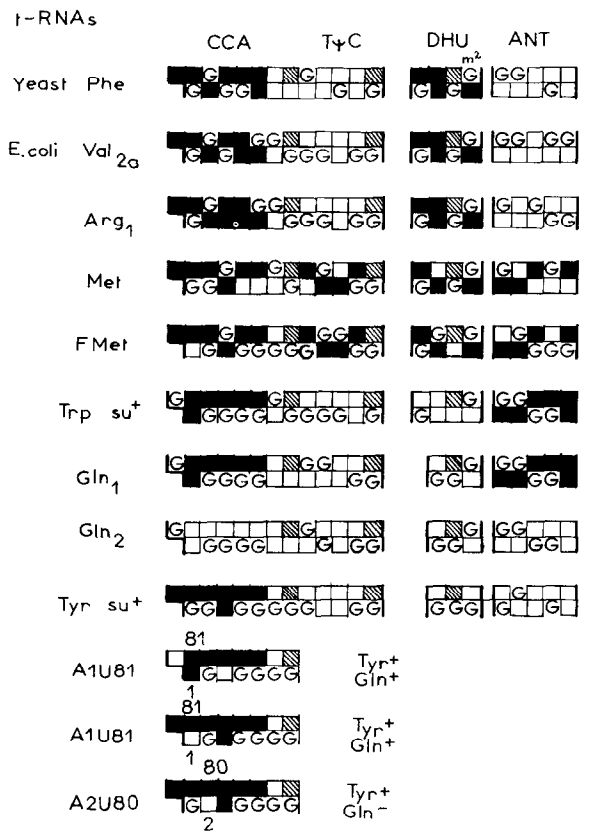


Fig.2. Guanine patterns in the stems of some tRNAs. By the black areas are marked the potential places for the tRNA synthetases' legs. The shaded areas are the positions of the obligatory non-Gs. Below are the double mutants su+3 $tRNA^{Tyr}$ within the ACC stem. To the right of them aminoacylation specificity in vivo is indicated [5,17,18].

A1U81 of su+$tRNA^{Tyr}$ can be aminoacylated in vivo with high efficiency not only by the Tyr-synthetase but by the Gln-synthetase as well [17]. It is evident from fig.2 that the 'mould' for $tRNA^{Gln}$ (Gln-synthetase) becomes complementary to the mutant su+$tRNA^{Tyr}$, which can thus be charged by each of these aminoacids. Another double mutant, A2U80 (fig.2) may not be expected to acquire Gln-specificity, and this is so [5,18].

These mutation data show unambiguously, that it is G that causes repulsion and, that the misacylation in the Tyr–Gln pair is due to a close similarity of the ACC stems of these tRNAs.

Of the known tRNA structures one more only, $tRNA^{Trp}$, has a very similar ACC stem to that of $tRNA^{Gln}$ (fig.2). It is remarkable that in this case too the misacylation of a mutant su+7 $tRNA^{Trp}$ by the Glu-synthetase occurs [19]. But the authors say that in this case such a transformation of specificity is due to a single change C → U in the anticodon, not in a stem. It is possible, that in this particular case, when as it may be seen from fig.2, not only the ACC stems are similar, but the ANT stems are identical, the bases of anticodon itself are used for recognition too. (There are some other data as well on involvement of the anticodon bases in recognition in some cases [20, 21, 22].)

*Erroneous aminoacylation.* The extensive data of Ebel et al. [23–25] on erroneous aminoacylation in the series of the homologous and heterologous tRNA and synthetases supply us with the material to be rationalized on the basis of the G–non-G code. As an example I took the data of the work [23] on aminoacylation of the various tRNAs of *E. coli* and yeast by the Val-synthetase from *E. coli* in the conditions favouring misacylation (20% of dimethyl-sulfoxide). It is natural to suppose, that the less numbers of legs encounters with Gs of a non-cognate tRNA, the greater rate of aminoacylation will be observed. (A discussion on, why it is $V_{max}$ or some 'level' in the studies [23–25], but not $K_m$, that is primarily affected, is given elsewhere [25,26]).

The problem is that the pattern of the legs is not known yet. Let us suppose that the recognizible area in tRNAs consists of (1) a half of turn in the ACC stem beginning from the 4th base from 3′-end (the remaining part of the helix including the TΨC stem looks to an opposite side of space), (2) the DHU stem, which looks to the same side as the area (1) [27,28] and according to [14] is a site of recognition, (3) the ANT stem. Then, taking into account the structures of isoacceptor $tRNA^{Val}_{1,2,2a}$ of *E. coli* and the above assumptions (1)–(3), the 'comb' of legs for the *E. coli* Val-synthetase looks like in fig.3. Superposing this pattern onto

Fig.3. The combination of the legs (circles) for the *E. coli* Val-tRNA synthetase which is used in the present paper for calculating the numbers of the repulsive sites in the misacylation cases. Gs are indicated in the places, occupied by Gs in the isoacceptor $tRNA^{Val}_{1,2a,2b}$ [10]. The TΨC stem and the two last pairs of the ACC stem are supposed non-participating in the recognition of the *E. coli* $tRNA^{Val}$ (see the text).

Table 1
Comparison of the series of misacylation in 20% dimethyl sulfoxide by Val-tRNA synthetase of *E. coli* [23] with those calculated in the present study

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *E. coli* tRNAs | | | | | | | | |
| According to [23] | Val(71[a]) | Met(12) | Ala(9) | Ile(6) | Trp(5) | Arg(4) | | |
| According to the collision number | $Val_{1,2a,2b}$(0/0[b]) | Met(0/1) | $Ala_{1a}$(0/2) | $Ile_1$(0/2) | | $Arg_1$(0/3) | $Leu_1$(0/3) . . . | Trp(1/4) |
| Yeast tRNAs[c] | | | | | | | | |
| According to [23] | Val(83[a]) | Ile(27) | Phe(23) | Ala(23) | Met(17) | Trp(7) | | Tyr(5) |
| According to the collision number | $Val_1$(0/2[b]) | | Phe(0/3) | $Ala_1$(0/3) | Met(0/5)[c] | Trp(0/4) | Ile(0/5) | Tyr(0/6) |

[a] Aminoacylation level in conditional units according to ref. [23].
[b] Collision number with the Gs of tRNAs indicated for the system of legs of fig.3. For the sequences see [10,29] ($Met^c$, mouse myeloma), [30] (Ala, *E. coli*). First figure is collision number for the discriminator (4th base from 3'-end); the second one is the total number of collisions in the two-stranded recognition areas.
[c] The interaction of a leg with C, which is complementary to $M^2G$ in yeast tRNAs is considered as a collision. The cognate synthetase may have a hydrophobic leg in such a place [26].

the schemes of tRNAs [10], presented like on fig.2, and calculating the numbers of collisions, one gets the series of table 1. One may see, that except a $tRNA^{Ile}$ of heterologous system and a $tRNA^{Trp}$ of homologous one, the series coincide with those obtained experimentally [23]. (The case of Ile could possibly be explained by that the real number of legs for the ANT stem is less than the one used).

Detailed consideration of the other series [24,25] as well as a discussion on recognition of the minor nucleosides will be given elsewhere [26].

So, the protein-nucleic acid code suggested here explains the various facts on recognition of tRNA. This scheme predicts that the complementary oligoribonucleotides with the specific sequences can be the competitive inhibitors of tRNA binding to the synthetases and that an artificial tRNA with inosines instead of Gs will fit any synthetase.

*The obligatory non-Gs.* The presence of a universal non-G in the positions marked by the grey squares in fig.2 was checked with approx. 60 structures [10]. Interestingly, while in the TΨC stem it is always C, in the other two stems, ACC or DHU, this can be A, U or C.

These obligatory non-Gs may be thought to be the binding sites for the ribosome, and translation factors. (Incidentally, assembling of the ribosome proteins onto rRNA could also be governed by the G–non-G code).

## Acknowledgements

I thank Drs G. V. Gursky, M. Ya. Karpeisky and V. Sh. Scheinker for valuable criticism.

## References

[1] Arber, W. (1974) Progr. in Nucleic Acid Research and Mol. Biol. 14, 1–37.
[2] Adler, K., Beyreuter, K., Fanning, E., Geisler, N., Gronenborn, B., Klemm, A., Müller-Hill, B., Pfahl, M. and Schmitz, A. (1972) Nature 237, 322–327.
[3] Gilbert, W. and Maxam, A. (1973) Proc. Nat. Acad. Sci. USA 70, 3581–3584.
[4] Chamberlin, M. J. (1974) Annual Rev. of Biochemistry 43, 721–775.
[5] Celis, J., Hooper, U. and Smith, J. (1973) Nature New Biol. 244, 261–264.
[6] Livshitz, M. A., Gursky, G. V., Zassedatelev, A. S. and Volkenstein, M. V., J. Mol. Biol., submitted for publication.
[7] Bruskov, V. I. and Poltev, V. I. (1974) Doklady AN SSSR 219, 231–234.
[8] Krey, A. K., Allison, R. G. and Hahn, F. E. (1973) FEBS Lett. 29, 58–62.
[9] Zassedatelev, A. S., Gursky, G. V., Zimmer, Ch. and Thrum, H. (1974) Mol. Biol. Reports 1, 337–342.
[10] Barrel, B. G., Clark, B. F. C. (1974) Handbook of Nucleic Acid Sequences, Published by Joynson-Bruwers Ltd., Lab. Mol. Biol., Cambridge.
[11] Arnott, S. (1970) Progr. Biophys. Mol. Biol. 21, 256–319.

[12] Carter, Ch. W., Jr, Kraut, J. (1974) Proc. Nat. Acad. Sci. USA 71, 283–287.
[13] Chambers, R. (1971) Progr. Nucl. Acid Res. Mol. Biol. 11, 489–525.
[14] Dudock, B., di Peri, C., Scileppi, K. and Reszelback, R. (1971) Proc. Nat. Acad. Sci. USA 68, 681–684.
[15] Roe, B. and Dudock, B. (1972) Biochem. Biophys. Res. Commun. 49, 399–406.
[16] Crothers, D. M., Seno, T. and Söll, D. G. (1972) Proc. Natl. Acad. Sci. USA 69, 3063–3067.
[17] Ghyssen, A. and Celis, J. E. (1974) J. Mol. Biol. 83, 333–351.
[18] Smith, J. D., Barnett, L., Brenner, S. and Russel, R. L. (1970) J. Mol. Biol. 54, 1–14.
[19] Yaniv, M., Folk, W. R., Berg, R. and Soll, L. (1974) J. Mol. Biol. 86, 245–260.
[20] Kisselev, L. L. and Frolova, L. Yu. (1964) Biokhimia 29, 1177–1189.
[21] Mirzabekov, A. D., Lastit, D., Levina, E. S., Undritzov, I. M. and Bayev, A. A. (1972) Molekularnaya Biologia 6, 87–104.
[22] Roberts, J. W. and Carbon, J. (1974) Nature 250, 412–414.
[23] Giege, R., Kern, D. and Ebel, J. -P. (1972) Biochimie 54, 1245–1255.
[24] Kern, D., Giege, R. and Ebel, J. -P. (1972) Eur. J. Biochem. 31, 148–155.
[25] Ebel, J. -P., Giege, R., Bonnet, J., Karn, D., Befort, N., Bollack, C., Fasiolo, F., Gangloff, J. and Dirheimer, G. (1973) Biochimie 55, 547–557.
[26] Ivanov, V. I., (1975) Molecularnaya Biologia (SSSR), submitted for publication.
[27] Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C. and Klug, A. (1974) Nature 250, 546–551.
[28] Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H. J., Seeman, N. C. and Rich, A. (1974) Science 185, 435–440.
[29] Piper, P. W. (1975) Eur. J. Biochem. 51, 295–304.
[30] Williams, B. J., Nagel, W., Roe, B. and Dudock, B. (1974) Biochem. Biophys. Res. Commun. 60, 1214–1221.