Theoretical analysis of 'addressed' chemical modification of DNA

M.P.Perelroyzen and A.V.Vologodskii*

Novosibirsk Institute of Bioorganic Chemistry, Siberian Division of the Academy of Sciences of USSR, Novosibirsk 630090, and Institute of Molecular Genetics Academy of Sciences of USSR, Moscow 123182, USSR

Received December 8, 1988; Revised and Accepted April 19, 1988

ABSTRACT

Chemical "addressed" modification of DNA involves treatment of single-stranded DNA with oligonucleotides complementary to certain target sequences in this DNA and bearing groupings reactive towards DNA bases. The binding of oligonucleotides can occur both at completely (specific) and incompletely (non-specific) complementary sites. We analyse the modification of a fragment that is flanked by two target sequences complementary to a given oligonucleotide address, contains no more such targets and has some randomly distributed sites for nonspecific binding. Conditions for the maximum ratio between specific and non-specific modification are determined. We find the probability of both target termini being specifically modified without any non-specific modification occurring within the fragment up to a given moment in time. Quantitative analysis is based on the use of known features of the specific and non-specific binding of an oligonucleotide to DNA sites. This analysis shows the possibility of specific cutting of DNA based on addressed modification.

INTRODUCTION

One of the problems in cloning large genomes consists in the specific cleavage of DNA into fragments of 10^4 - 10^5 base pairs. It is difficult to do this by restriction endonucleases because their cleavage sites occur at shorter distances. One of the possible solutions is the so-called "addressed" chemical modification (for review see ref. (1)) which involves the treatment of single-stranded DNA by oligonucleotides complementary to certain target sequences in this DNA and bearing groupings reactive towards DNA bases. After addressed modification (e.g. addressed alkylation) DNA may be chemically cleaved at the modified bases to produce the desired fragments.

However, because of the flexibility of the DNA chain, the oligonucleotide "address" may obviously form not only specific

complexes with the desired targets but also complexes with partial complementarity. The number of partially complementary targets is much greater than the number of completely complementary ones. Therefore it is a legitimate question whether complementary addressed modification can generally be employed to obtain long genome fragments of the kind mentioned above.

The recent publications by Aboul-ela et al. (2) present thermodynamic characteristics of various non-ideal duplexes formed by oligonucleotides. These data make it possible to perform a quantitative theoretical analysis of the selectivity of addressed modification of DNA. Such an analysis has been undertaken in the present study.

THEORETICAL MODEL OF ADDRESSED MODIFICATION

Formulation of the model

Let us consider a fragment of DNA flanked by two target sequences complementary to a given oligonucleotide. Let us assume that this fragment contains no more sequences complementary to the oligonucleotide, and that the length of the fragment is L nucleotide residues. We shall analyze the probability $P(t)$ of selective modification of the ends of the fragment without any non-specific modification of the internal part. This formulation corresponds to the problem of DNA fragmentation at specific modification sites. Let us assume that modification can take place only after binding the nucleotide address to DNA. Specific and non-specific modification proceed independently. Therefore the probability $P(t)$ of specific modification up to time t may be written as

$$P(t) = (p_s(t))^2 \prod_i (1 - p_i(t)), \quad (1)$$

where $p_s(t)$ is the probability of selective modification of one of the target termini up to time t , and $p_i(t)$ are the probabilities of non-selective modification of different sites in the internal part. The product corresponds to all the non-specific binding sites.

The selectivity of modification depends on the ratio of the lifetime of the DNA-oligonucleotide complex τ_o , and the charac-

teristic time τ_m of the chemical reaction of the reactive residue of the oligonucleotide within the complementary complex. Obviously, if

$$\tau_m \ll \tau_o \quad (2)$$

the selectivity of modification will be determined by the ratio of the rate constants of the specific and non-specific complementary binding, because in this situation the first binding act, whether specific or non-specific, will result in modification. It is known (3-5) that the rate constant of the formation of complementary complexes by oligonucleotides does not depend on their length. The limiting stage here is the collision of one of the base residues of the oligonucleotide with the complementary base residue of DNA. Therefore the rate constants of the specific and non-specific binding must not appreciably differ from each other, so that it will not be possible to achieve selective modification in this case.

On the other hand, if

$$\tau_m \gg \tau_o \quad (3)$$

then any act of modification must be preceded by many acts of complex formation and dissociation. In this case the selectivity will depend on the aggregate lifetimes of specific and non-specific complexes. In other words, the rate of modification will be proportional to the equilibrium degrees of occupation of the specific and non-specific sites.

The equilibrium degree of occupation f may be expressed as

$$f = Kc / (1 + Kc), \quad (4)$$

where K is the equilibrium binding constant and c is the equilibrium concentration of free oligonucleotide addresses. As suggested by equation (4), the values of f for specific and non-specific modification may differ appreciably only when

$$Kc \ll 1 \quad (5)$$

In this case

$$f = Kc \quad (6)$$

The condition (5) may always be fulfilled by choosing the appropriate temperature and concentration of addresses.

Hence, the maximum selectivity of modification should be expected under the conditions (3) and (5). The analysis presented below corresponds to these conditions.

If there is an excess of oligonucleotide, the process of modification is described by a kinetic equation of the first order with the rate constant gf , where g is the rate constant characterizing the reagent's reaction ability in the oligonucleotide/DNA complex. The probability of a particular site of a given DNA fragment being modified by time t is

$$p(t) = 1 - \exp(-fgt) = 1 - \exp(1 - Kcgt). \quad (7)$$

For the sake of simplicity we assume that there is a large excess of oligonucleotides, allowing the change of their concentration to be neglected. The equation (1) may then be transformed into

$$P(t) = (1 - \exp(-K_s cgt))^2 \exp(-K_n cgt), \quad (8)$$

where the index s assigns a value to specific binding, and the index n to non-specific binding, and

$$K_n = \sum_i K_i \quad (9)$$

Obviously, the function (8) has a maximum at some t which is the optimum time for the highest yield of selective modification. The maximum value $P_{\max}$ can be found from the condition $dP/dt=0$. Through calculations we obtain

$$P_{\max} = 4r^r / (2+r)^{2+r}, \quad (10)$$

where

$$r = K_n / K_s. \quad (11)$$

It is noteworthy that, as might be expected from general considerations, $P_{\max}$ depends only on the ratio of the sum of constants of non-specific binding to the constant of specific binding.

Statistical analysis of equilibrium binding constants

The thermodynamic constant of the binding of an oligonucleotide address to a given fully complementary site of the target may be calculated as

$$K_s = \beta \prod_j s_j \quad (12)$$

(see, e.g., ref. (6)). Here β is the nucleation constant, and s_j is the stability constant of the j -th pair, which depends on the nature of the pair. Multiplied are s_j values characteristic of all the complementary pairs of the structure formed.

In order to estimate K_n , let us consider the number and the nature of sites in the DNA fragment that are partially complementary to the address. Let us assume that bases are distributed randomly, and that the fraction of each of them is equal to 1/4. Since the mathematical expectation of any of the given sequences occurring within a stretch of length L is proportional to L , the value of K_n should be proportional to the total length of the fragment. This enables us to express the value of K_n for a stretch of arbitrary length L , $K_n(L)$, through the value of $K_n(L_0)$ corresponding to a specified length L_0 :

$$K_n(L) = (L/L_0)K(L_0). \quad (13)$$

For L_0 one can conveniently use the mean distance between two specific binding sites for an address of n pairs ($L_0 = 4^n$). Any particular sequence of length n occurs once, on an average, in the fragment L_0 , which fact simplifies the calculations.

The non-specific binding sites, which make a considerable contribution to K_n , may be divided into four classes.

Class 1 includes sites of length n which carry non paired base residues only at their termini. Consider the case when only one unpaired base is present at one of the termini. This means that here the target carries one of the three bases non-complementary to that of the address. For each of these variants the specific binding constant with a "perfect" address would be s_1 or s_n times greater than the non-specific one. The total contribution to r is

$$3(1/s_1 + 1/s_n) \quad (14)$$

If any number of unpaired bases are present at the termini, their contribution to r is

$$\begin{aligned} r_1 = & 3(1/s_1 + 1/s_n) + 3^2(1/s_1 s_2 + 1/s_1 s_n + 1/s_{n-1} s_n) + \\ & + 3^3(1/s_1 s_2 s_3 + 1/s_1 s_2 s_n + 1/s_1 s_{n-1} s_n + 1/s_{n-2} s_{n-1} s_n) + \dots \end{aligned} \quad (15)$$

Class 2 includes sites carrying non-canonical pairs within the helical region. The presence of even one such pair decreases the binding constant considerably; hence we shall consider only sites having one such pair. Three types of non-canonical pairs may be found at a given position in the helical region. However, unlike the situation in class 1, the destabilizing effects of these three types are different. To estimate them, we shall introduce the product $s_j s_j^*$ instead of the value s_j in eq. (12). The values s_j^* are

$$s_j^* = \exp(-\Delta G_j/RT), \quad (16)$$

where ΔG_j is the change in the free energy of a duplex caused by the substitution of a given non-canonical pair for the j -th canonical pair. The complete contribution to r of the sites having a single non-canonical pair is

$$r_2 = \sum_{j=2}^{n-1} (s_{j1}^* + s_{j2}^* + s_{j3}^*), \quad (17)$$

where the second index of s_{jk}^* corresponds to one of the three possible variants. The values s_{jk} can be estimated from the data reported in (2) (see below).

Class 3 includes sites having one bulged base in the target. The maximum possible contribution of such sites to r may be found as follows. The equilibrium constant of perfect binding (12) must be multiplied by the factor s^* , i.e., the extent of destabilization caused by the bulging base. The number of such sites over the length L_0 is equal, on an average, to $(n-1)$. Hence the contribution of these sites to r is

$$r_3 = (n-1)s^* \quad (18)$$

In the absence of experimental evidence, we assume that the factor s^* is the same for different bulged-out bases. It has been determined (2) for cytosine only.

Class 4 includes sites that cause one base of the address to bulge on complex formation. Compared with class 3, the effective length of the address is in this case reduced by one unit, so that the probability of occurrence of such sites is four times greater than that of ideal sites. The average number of these sites in the DNA fragment under consideration is $4(n-1)$. The binding constant is s_j times smaller (j denotes the bulging base residue). Hence

$$r_1 = 4 \sum_{j=2}^{n-1} s^* / s_j. \quad (19)$$

The values $r_1 - r_4$ make the major contribution to r .

The values of the parameters s_j , which characterize the stability of canonical base pairs, are known over a wide range of external conditions. It is a good approximation for DNA that these parameters depend only on the nature of the pair, and are independent of its neighbours (see, e.g., refs.(7,8)). For 1 M NaCl and temperature 298°K they are (9,10)

$$\begin{aligned} s_{AT} &= 9.2 \\ s_{GC} &= 52.7 \end{aligned} \quad (20)$$

A systematic investigation of the effects of different non-complementary pairs on the stability of oligonucleotide complexes has been performed for the first time by Aboul-ela et al. (2). These workers studied the melting of the complexes $dCA_3XA_3G \cdot dCT_3YT_3G$, where $X,Y = A,C,G,T$. Table 1 of their paper lists the free energies of the formation of various duplexes of this type. That Table can be used to find the ΔG_j values eq. (16). The values averaged for the two orientations of a given pair in the duplex in kcal/mole are presented below:

Base in the address	Base of the target site				$(s_{j1}^* + s_{j2}^* + s_{j3}^*) \cdot 10^3$
	A	T	G	C	
A	4.0	0	3.7	5.0	3.3
T	0	4.0	3.2	4.1	8.3
G	3.7	3.2	3.5	0	9.1
C	5.0	4.1	0	5.3	1.3

The constant s^* can also be found from the data in paper (2), where the free energies of the formation of duplexes $dCA_6G \cdot dCT_6G$ (ΔG_1), $dCA_3CA_3G \cdot dCT_6G$ (ΔG_2), $dCA_6G \cdot dCT_3CT_3G$ (ΔG_3) were calculated. The first of these duplexes forms a regular helix, in the second and third duplexes cytosine is pushed out the helix in one or the other strand.

$$s^* = \exp((\Delta G_2 + \Delta G_3 - 2\Delta G_1)/2RT) \quad (21)$$

For the above-mentioned conditions

$$s^* = 0.004. \quad (22)$$

The thermodynamic constants presented above suggest that the maximum r is reached for addresses entirely consisting of thymidines, and the minimum r for addresses entirely consisting of cytosines. The extreme values of $r(L_0)$ for an address consisting of 7 nucleotide residues are presented below ($L_0 = 16384$):

	r_1	r_2	r_3	r_4	r
dT_7	1.166	0.041	0.024	0.009	1.240
dC_7	0.124	0.006	0.024	0.002	0.156

DISCUSSION

The analysis presented above shows that addressed modification may be achieved under the following conditions:

(i) The life-time of the specific complex must be much smaller than the characteristic time of the chemical reaction τ_m ; to fulfil this condition one may make use of the very steep dependence of τ_o on temperature and ionic strength. For particular conditions and addresses, τ_o can be estimated using the following equation (see (3,11))

$$\tau_o = 1/(2nk_f) \prod_j s_j, \quad (23)$$

where $k_f = 10^6 + 10^7 \text{ sec}^{-1}$ (3,4) is the rate constant of a pair being added to an already existing helical region. For $n=7$ and the above mentioned conditions,

$$\tau_o = 10^{-1} - 10^5 \text{ sec} \quad (24)$$

(ii) The product of the concentration of the address by its equilibrium constant of specific binding must be much smaller than 1.

These two conditions are a necessary pre-requisite for the direct dependence of selectivity on the constants of specific and non-specific binding. The maximum yield of specific binding depends only on r (eq. 11). Fig. 1 shows the dependence $P_{\max}(r)$ calculated according to eq. (10). One can see that $P_{\max} \geq 0.3$ for $r \leq 0.5$. Analysis shows that the yield of any fragment arising due to non specific modification is less than 7%. With the typical $r(L_o)$ values presented above, a region of L may also be found for which $r(L) \leq 0.5$. Hence, for this region at least the yield of selective modification is sufficient for the subsequent selective cleavage.

Figure 1. Dependence of the maximum yield P_{max} of specific modification on the ratio of the sum of constants of non-specific binding to the constant of specific binding.

Figure 2. Time dependence of the probability of specific modification P for different values of r (—). Time dependence of the degree of modification of the specific target p_s (---).

It follows from eq. (8) that for any particular r value there exists an optimum reaction time at which the maximum P is achieved. The dependence $P(t)$ for different r values is presented in Fig. 2. The broken line shows the dependence of the degree of modification of the specific target on the reaction time. It is seen that the maximum P , depending on r , is achieved at a varying degree of modification. The choice of the optimum reaction time is an important condition of achieving the maximum specificity of modification. It follows from the above that terminal non-canonical pairs of the complex make a preponderant contribution to the r values. Therefore, to reduce r , the termini of the address must consist of G and/or C residues. However, when estimating r_1 , we tacitly assumed that the non-canonical pairs are melted, i.e., that they are as good as absent from the complex. In reality these terminal non-complementary pairs may increase the binding constants somewhat at low temperatures, thus leading to an increase of r_1 . A quantitative estimation of the stabilizing role of non-complementary terminal pairs requires a special experimental study.

Furthermore, we assumed the concentration of oligonucleotides c to be much larger than the concentration of specific targets, and so the change in c due to binding was neglected. However, the relationship (10) between $P_{\max}$ and r , and the degree of modification corresponding to the maximum P are not changed if this condition is removed.

In our analysis we did not consider the possible effect of the secondary structure of single-stranded DNA upon the addressed modification. Obviously, the existence of this higher structure should not change our main results because the potential targets involved in the secondary structure will not be subject to modification in this system. Hence, the secondary structure can only increase the real length of the fragments produced, but will not change the yields.

Finally, we did not consider the possibility that the residue of the reagent attached to the address may affect the binding constant. It may happen that the effect will somehow change the quantitative data presented above. Relevant experimental data are not yet available.

* To whom all correspondence should be addressed.

REFERENCES

1. Knorre, D.G. and Vlassov, V.V. (1988) Affinity modification of biopolymers, CRC Press, Boca Raton.
2. Aboul-ela, F., Koh, D. and Tinoco, I., Jr. (1985) Nucl. Acids Res. 13, 4811-4824.
3. Craig, M.E., Crothers, D.M. and Doty, P. (1971) J. Mol. Biol. 62, 383-401.
4. Porschke, D. and Eigen, M. (1971) J. Mol. Biol. 62, 361-381.
5. Porschke, D., Uhlenbeck, O.C. and Martin, F.H. (1973) Biopolymers 12, 1313-1335.
6. Cantor, C.R. and Schimmel, P.R. (1980) Biophysical chemistry, pp. 1192-1194, W.H. Freeman and company, San Francisco.
7. Vologodskii, A.V., Amirikyan, B.R., Lyubchenko, Yu.L. and Frank-Kamenetskii M.D. (1984) J. Biomol. Struct. dyn. 2, 131-148.
8. Wada, A. and Suyama, A. (1986) Prog. Biophys. Mol. Biol. 47, 31-112.
9. Wada, A., Yabuki, S. and Husimi Y. (1980) Crit. Rev. Biochem. 9, 87-144.
10. Kozyavkin, S.A., Mirkin, S.M. and Amirikyan B.R. (1987) J. Biomol. Struct. Dyn. 5, 119-126.
11. Anshelevich, V.V., Vologodskii, A.V., Lukashin, A.V. and Frank-Kamenetskii, M.D. (1984) Biopolymers 23, 39-58.