A Simple Model for Gene Targeting

Tommi Ratilainen, Per Lincoln, and Bengt Nordén
Department of Physical Chemistry, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

ABSTRACT Sequence-specific binding to genomic-size DNA sequences by artificial agents is of major interest for the development of gene-targeting strategies, gene-diagnostic applications, and biotechnical tools. The binding of one such agent, peptide nucleic acid (PNA), to a randomized human genome has been modeled with statistical mass action calculations. With the length of the PNA probe, the average per-base binding constant k_0, and the binding affinity loss of a mismatched base pair as main parameters, the specificity was gauged as a "therapeutic ratio" G = maximum safe $[\mathrm{PNA}]_{\mathrm{tot}}$/minimal efficient $[\mathrm{PNA}]_{\mathrm{tot}}$. This general, though simple, model suggests that, above a certain threshold length of the PNA, the microscopic binding constant k_0 is the primary determinant for optimal discrimination, and that only a narrow range of rather low k_0 values gives a high therapeutic ratio G. For diagnostic purposes, the value of k_0 could readily be modulated by changing the temperature, due to the substantial $\Delta H°$ associated with the binding equilibrium. Applied to gene therapy, our results stress the need for appropriate control of the binding constant and added amount of the gene-targeting agent, to meet the varying conditions (ionic strength, presence of competing DNA-binding molecules) found in the cell.

INTRODUCTION

There are numerous kinds of molecules capable of sequence-specifically recognizing parts of the genetic code carried in the DNA (Nielsen, 1991; Ren and Chaires, 1999). Besides giving valuable insight about the mechanisms of molecular recognition in general, and of nucleic acids in particular, the development and study of various gene-targeting strategies have many other important practical applications. These range from "pure chemical" gene diagnostic applications (Carlsson et al., 1996; Rufer et al., 1998; Wang et al., 1997; Weiler et al., 1997) to various forms of gene therapy based on man-made gene-targeting molecules (Cohen and Hogan, 1994; Crooke, 1995). Also gene-repair strategies based on targeting as a primary step to anchor the correct gene sequence have been proposed (Chan et al., 1999). The most promising gene-targeting agents so far have been those based on various modifications of the nucleic acid molecules themselves. To obtain more favorable properties, e.g., regarding binding strength, sequence specificity and stability in biological liquids, a wide variety of modifications has been explored (Cook, 1998; Uhlmann and Peyman, 1990). The range of modifications include, e.g., exchange of single atoms in the nucleic acid backbone, substituents on the nucleobases, and other alternatives that involve completely newly designed backbones.

Aimed to bind to DNA, and no matter how these molecules look, an important question to be answered is whether the efficiency and safety is high enough. To develop successful gene drugs, i.e., with an optimized sequence specificity in the biological system, one needs to explore some basic binding thermodynamics. It will turn out that it is not only a matter of obtaining as high binding affinity for the target sequence as possible, but in addition, and more importantly, to minimize the drug binding to nontarget DNA, i.e., mismatched target sequences. Moreover, the purpose of this is not only to avoid blocking unwanted genes, but also to keep the drug amount low in the cell to simultaneously avoid non-DNA related toxicity and meet economical requirements.

Trying to optimize the concentration of drug delivered to a cell is therefore a delicate act of balance between repressing the gene (target) of interest while leaving normally working genes unaffected. Here we want, in a simplified manner, to discuss some aspects of this optimization process regarding the molecular recognition.

Peptide nucleic acids (PNA) are DNA mimics with interesting potential for gene-therapeutic, gene-diagnostic, and molecular biology applications (Nielsen, 1999; Nielsen and Egholm, 1999). Compared to other DNA analogs, PNA has several advantageous binding properties, mainly arising from the fact that the negatively charged backbone of DNA is replaced by a neutral pseudopeptide in PNA. The backbone in PNA is composed of *N*-(2-aminoethyl)glycine units, onto which the natural recognizing elements of DNA (the nucleobases) are attached (Fig. 1). Designed to be structurally homomorphous to DNA, a mixed-sequence PNA strand is capable of hybridizing with complementary single-stranded DNA, RNA, or another PNA strand to form highly stable duplexes with high sequence selectivity (Egholm et al., 1993; Kilså Jensen et al., 1997; Nielsen et al., 1991; Tomac et al., 1996). At moderate salt, the thermal stabilities increase in the series DNA–DNA < PNA–DNA < PNA–RNA < PNA–PNA, mainly due to the lack of interstrand electrostatic repulsion in PNA-containing duplexes (Egholm et al., 1993; Tomac et al., 1996).

The hybridization affinity of PNA to DNA has been modeled from a statistical viewpoint using thermodynamic data previously determined for mixed purine–pyrimidine

Received for publication 19 April 2001 and in final form 11 July 2001.

Dr. Ratilainen's present address, and to whom reprint requests should be addressed, is TATAA Biocenter, Chalmers Teknikpark, SE-4121 88 Gothenburg, Sweden. Tel.: +46-70-422-49-52; Fax: +46-31-794-62-68; E-mail: tommi.ratilainen@tataa.com.

0006-3495/01/11/2876/10 $2.00

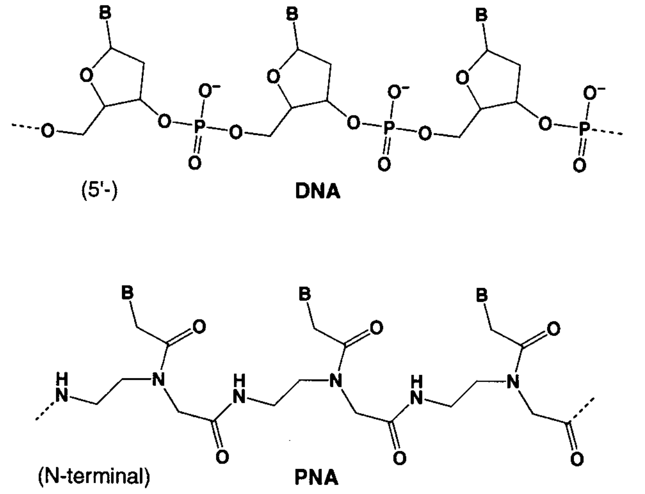

FIGURE 1 Chemical structures of DNA and PNA. The deoxyribose phosphodiester backbone in DNA has been changed to *N*-(2-aminoethyl) glycine in PNA.

sequence PNA–DNA duplexes (Ratilainen et al., 2000). Here, we introduce some basic models for estimating the binding constants of PNA–DNA target sequences in competition with mismatched targets, to a randomized human genome. We show, using a set of simulations, that the models need not to be very complex, but a rather simple approach is sufficient for finding the appropriate drug concentration range to use in a therapeutic or biotechnical application.

METHODS

Statistical modeling

Let the genome (DNA), of a certain length L, have a uniform distribution of the four bases (A, C, G, and T) in a random sequence, with the exception of one target sequence, T, to which a probe can bind with a binding constant of K_0 (M^{-1}). In our example, the probe will be a PNA oligomer of length n bases, in a sequence complementary to T. On average, each base will contribute a factor k_0 ($M^{-1}bp^{-1}$) to the binding constant. The equilibrium between P·T (PNA bound to the target), P (free PNA), and T (free, unblocked target sequences), can then be described by

$$\mathrm{P}+\mathrm{T} \overset{K_0}{\rightleftharpoons} \mathrm{P{\cdot}T}; \quad K_0 = k_0^{\mathrm{n}} = \frac{[\mathrm{P{\cdot}T}]}{[\mathrm{P}][\mathrm{T}]}. \tag{1}$$

Binding model 1: Two-state model (match or not)

In the most simple model of PNA binding to a sequence with m mismatches, we assume that the binding constant can be factorized into sequence- and position-independent binding constants (per base): k_0 for the $(n - m)$ matched basepairs and k_{mis} for m mismatched basepairs,

$$K_{\mathrm{m}} = k_0^{\mathrm{n-m}} k_{\mathrm{mis}}^{\mathrm{m}} = k_0^{\mathrm{n}} f^{\mathrm{m}} = \frac{[\mathrm{P{\cdot}D_m}]}{[\mathrm{P}][\mathrm{D_m}]}, \tag{2}$$

where D_m denotes a nontarget DNA sequence with m mismatches and, for convenience, $f = k_{\mathrm{mis}}/k_0$ defines the "frustration factor," i.e., the penalty of having one mismatch in the sequence.

Assuming the overall binding density of PNA to be low enough that overlap of binding sites could be neglected, the concentration of free sequences with m mismatches, $[D_m]$, is calculated from the mass balance condition,

$$[\mathrm{D_m}] = [\mathrm{D_m}]_{\mathrm{tot}} - [\mathrm{P{\cdot}D_m}], \tag{3}$$

where $[\mathrm{P{\cdot}D_m}]$ denotes the concentration of PNA bound to sequences with exactly m mismatches, and $[D_m]_{\mathrm{tot}}$ denotes the total concentration of such sequences. The latter quantity is given by

$$[\mathrm{D_m}]_{\mathrm{tot}} = P(\mathrm{D_m})L[\mathrm{DNA}]_{\mathrm{tot}}, \tag{4}$$

where $P(D_m)$ is the probability for a consecutive sequence of n bases on the DNA to have exactly m mismatches relative to the PNA, L is the number of possible n-base sequences, which, because $L \gg n$, is equal to the number of bases per DNA molecule, and $[\mathrm{DNA}]_{\mathrm{tot}}$ is the total concentration of DNA molecules. Assuming a random DNA sequence with GC = AT, $P(D_m)$ is given by the binomial coefficient,

$$P(\mathrm{D_m}) = \binom{n}{m}\left(\frac{1}{4}\right)^{\mathrm{n-m}}\left(\frac{3}{4}\right)^{\mathrm{m}}. \tag{5}$$

Note that D_0 denotes a copy of the target that statistically may appear with a certain probability in a random sequence model genome; thus, although $P(\mathrm{T})L \equiv 1$, because, by hypothesis, the target sequence is present in the genome, $P(D_0)L \neq 1$, in general.

The concentration of total (added) PNA can now be readily calculated by summation of terms corresponding to free PNA, PNA targeting the correct sequence as well as being bound to the various mismatched sequences,

$$\begin{aligned}[\mathrm{P}]_{\mathrm{tot}} &= [\mathrm{P}] + [\mathrm{P{\cdot}T}] + \sum [\mathrm{P{\cdot}D_m}] \\ &= [\mathrm{P}] + [\mathrm{P{\cdot}T}] + \sum_{\mathrm{m}=0}^{\mathrm{n}} \frac{k_0^{\mathrm{n}} f^{\mathrm{m}}[\mathrm{P}][\mathrm{D_m}]_{\mathrm{tot}}}{1 + k_0^{\mathrm{n}} f^{\mathrm{m}}[\mathrm{P}]}.\end{aligned} \tag{6}$$

Binding model 2: Mismatches with different costs

One step to refine the model is to categorize the different mismatches into different classes, depending on their impact on the total binding constant of the duplex. Based on previous results for singly mismatched PNA–DNA duplexes (Ratilainen et al., 2000), we have chosen to lump the mismatches into either of two classes: Class 1 with the energetically less costly ones, i.e. G·T, G·A, G·G, and A·A with a frustration factor $f_1 = 10^{-2}$, and Class 2 containing C·T, C·A, T·T, and C·C with a frustration factor $f_2 = 10^{-3}$. The binding affinity (equilibrium constant) of the probe to a target sequence is then given by Eq. 7, with I being the number of mismatches of Class 1 and J the number of mismatches of Class 2,

$$K_{\mathrm{IJ}} = k_0^{\mathrm{n}} f_1^{\mathrm{I}} f_2^{\mathrm{J}}. \tag{7}$$

The calculation of the concentrations $[D_{\mathrm{IJ}}]$ is done analogously with Model 1, i.e., at the assumption that there be no effect of interactions with the neighboring bases in the sequence. Thus, the PNA sequence can be split into four parts, each containing only one type of base, and the probability of having i mismatches of Class 1 and j mismatches of Class 2 in, for example, the C (cytosine) part can be computed by

$$P(\mathrm{C_{ij}}) = \left(\frac{1}{4}\right)^{\mathrm{w-i-j}}\left(\frac{CX_1}{4}\right)^{\mathrm{i}}\left(\frac{CX_2}{4}\right)^{\mathrm{j}}\binom{w}{i}\binom{w-i}{j}, \tag{8}$$

where w is the number of C bases in the PNA and CX_x is the number of basepairs in mismatch Class x, which contains at least one C base (with our mismatch categories, $CX_1 = 0$ and $CX_2 = 3$). The three other parts, with only A, G, and T bases, give analogous expressions.

The probability of a PNA sequence with totally I Class 1 mismatches and J Class 2 mismatches, $P(D_{IJ})$, is given by all combinations of the product $P(A_{ij})P(C_{ij})P(G_{ij})P(T_{ij})$ fulfilling the conditions $\Sigma\, i = (i_a + i_c + i_g + i_t) = I$ and $\Sigma\, j = (j_a + j_c + j_g + j_t) = J$:

$$P(D_{IJ}) = \sum_{\Sigma i=I,\ \Sigma j=J} P(A_{i_a j_a})P(C_{i_c j_c})P(G_{i_g j_g})P(T_{i_t j_t}). \quad (9)$$

The concentrations can be expressed analogously with Model 1, according to

$$[P]_{tot} = [P] + [P{\cdot}T] + \sum [P{\cdot}D_{IJ}]$$

$$= [P] + [P{\cdot}T] + \sum_{I,J=0}^{I+J=n} \frac{K_{IJ}[P][D_{IJ}]_{tot}}{1 + K_{IJ}[P]}, \quad (10)$$

where

$$[D_{IJ}]_{tot} = P(D_{IJ})L[DNA]_{tot} \quad (11)$$

Choice of parameters

The choice of parameters is quite critical for the outcome of the simulations. As described above, estimations of the various binding constants are taken from a set of experimental thermodynamic data for forming duplex complexes with DNA (Ratilainen et al., 2000). We use a length-averaged (6–20 bp) binding constant, $k_0 = 8.6\ M^{-1}bp^{-1}$ at 37°C, of matched PNA–DNA duplexes (Ratilainen et al., 2000), with the different mismatches categorized in two classes as described above. Furthermore, as default, we use a hypothetical randomized (with %GC = %AT) genome with human DNA length ($L = 3 \times 10^9$ bases).

To obtain a value relevant for in vivo gene-targeting conditions, we have used the local concentration of DNA in the nucleus of a eukaryotic cell; on average 0.17 g/mL (Daban, 2000). This converts to ~0.1 nM using $M_w \approx 650\ g{\cdot}mol^{-1}bp^{-1}$ and $L = 3 \times 10^9$ bases.

Output from the calculations

With the free PNA concentration [P] as independent variable, concentrations of blocked target, [P·T] and the sum of the concentration of PNA bound to mismatched sequences $\Sigma[P{\cdot}D_{IJ}]$ were calculated and plotted versus total PNA concentration (Eq. 10). To make the results of the simulation easier to visualize, the dominating terms in the summation, i.e., $[P{\cdot}D_{IJ}]$ for those mismatched sequences that were significantly (>50%) blocked, were plotted separately.

RESULTS AND DISCUSSION

Figure 2 shows a schematic graphical presentation of the results in this work. Logarithmic concentration profiles of free PNA, PNA bound to target, [PNA·T], as well as the sum of PNA bound to mismatched sequences, $\Sigma[P{\cdot}D_m]$ (Model 1) and $\Sigma[P{\cdot}D_{IJ}]$ (Model 2), respectively, and a stepwise curve representing the sum of PNA bound to heavily blocked mismatched sequences ("critical mismatches" for which $K_{IJ}[P] > 1$) are all plotted as a function of total PNA concentration, $[PNA]_{tot}$. These are indicated, together with the limit for binding to a mismatch to become "dangerous" (*left*), and the "therapeutic range" (*top*). By therapeutic range, we define the span from the point where

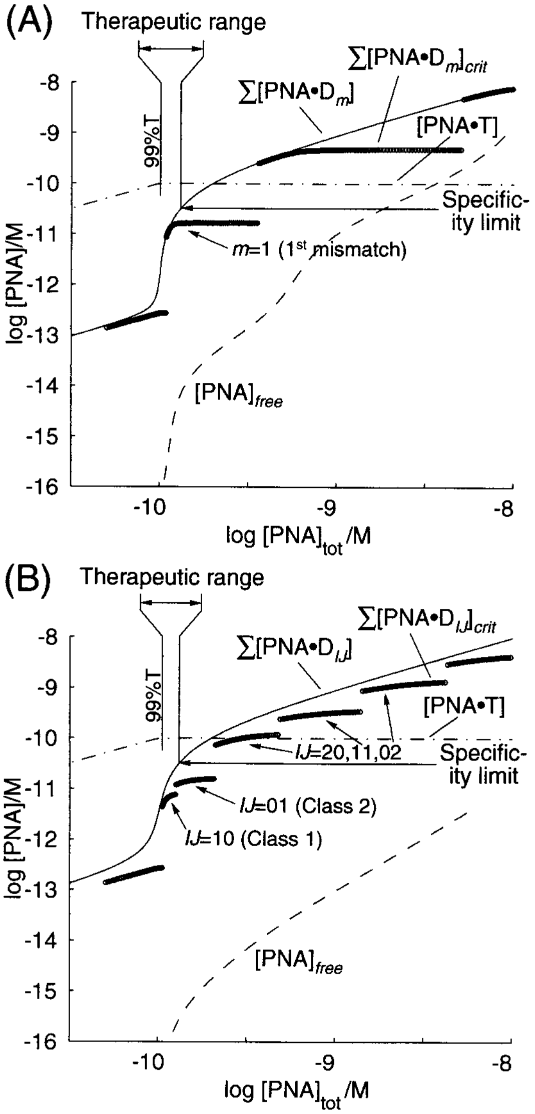

FIGURE 2 Graphical representation of the principle and some parameter definitions used in the simulations for the two different binding models. (*A*) Model 1 (one class of mismatches): The graphs show profiles of [PNA·T] (– · –), $[PNA]_{free}$ (– – –), $\Sigma[P{\cdot}D_m]$ (——), and $\Sigma[P{\cdot}D_m]_{crit}$ (∘∘∘∘∘∘). The therapeutic range is defined starting from the concentration at which 99% targeting occurs until the integrated concentration of mismatched complexes reach the specificity level (here chosen to be $10^{-0.5}$ times lower than the concentration of targeted sequence). (*B*) Model 2 (two classes of mismatches): The graphs show profiles of [PNA·T] (– · –), $[PNA]_{free}$ (– – –), $\Sigma[P{\cdot}D_{IJ}]$ (——), and $\Sigma[P{\cdot}D_{IJ}]_{crit}$ (∘∘∘∘∘∘). The binding curves appear somewhat more complex, and several details are changed. The singly ($I + J = 1$, indicated as 10 and 01) mismatched sequences now appear as two branches and the doubly ($I + J = 2$, i.e., 20, 11, and 02) as three (marked with *arrows* in graph). Also, the curve for the free PNA concentration appears more smoothed with different mismatches (Model 2) than with only one type (Model 1). The therapeutic range, however, does not change significantly.

binding to target is considered complete (99%), to the point where the collective (integrated) concentration of PNA-bound mismatched sequences grows above a certain level (*left*, typically a factor $10^{0.5} \approx 3.2$ lower than the covered target concentration). The vertical lines in the graphs show these two limits.

In the simplest case with only one type of mismatch (Model 1), a rather clear picture of the binding is obtained (*panel A*) even when considering the individual mismatches (*open circles*). For two different types of mismatches (Model 2; *panel B*) the binding curves become more complex, and several details are changed, especially regarding the individual mismatches (*open circles*). For appropriate cases, we will discuss the underlying values of the individual mismatch concentrations. The concentration of these will asymptotically approach a limiting concentration; in the figures, each curve is truncated at the point where binding to the subsequent mismatch starts to dominate.

Finding an optimal length of the PNA (mixed sequences)

Using the previously determined (Ratilainen et al., 2000) average binding constant per basepair PNA–DNA, $k_0 = (8.6 \pm 1)\ \mathrm{M}^{-1}$ (at 37°C), and the mismatch costs described in the Methods section, we performed a series of calculations with mixed-sequence PNA–DNA duplexes of varying lengths, from $n = 18$ to $n = 20$ (Fig. 3). Panel *A* ($n = 18$) represents a situation with conditions such that we do not achieve sequence-specific binding. The vertical line pointing at the integrated mismatch curve, where it corresponds to our criterion of $10^{0.5}$ times lower concentration than the blocked target, appears before the shorter line (marked 99%T) showing where the target is sufficiently blocked. Moving on to longer PNA probes changes the order of these limits and thus represents specific targeting, although the therapeutic range is very narrow in both cases (*panels B–C*). That further increasing *n* has little effect on the collective concentration of mismatched sequences is also notable.

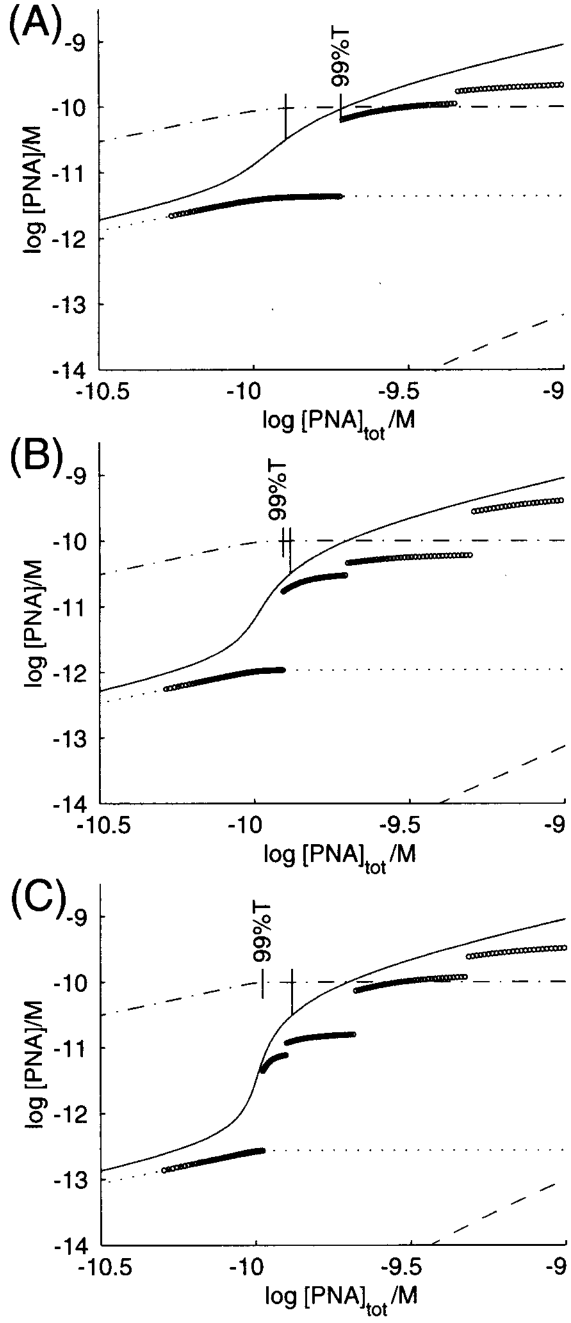

FIGURE 3 Profiles of [PNA·T] (– · –), $[\mathrm{PNA}]_{\mathrm{free}}$ (– – –), $\Sigma[\mathrm{P{\cdot}D_{IJ}}]$ (——), $\Sigma[\mathrm{P{\cdot}D_{IJ}}]_{\mathrm{crit}}$ (○○○○○○) simulated for three mixed PNA–DNA sequences of varying length ($N = 18$, 19, and 20). All three panels are simulated using $k_0 = 8.6\ \mathrm{M}^{-1}\mathrm{bp}^{-1}$, at a DNA concentration of 0.1 nM. Sequences: (*A*) $A_4C_5G_5T_4$, (*B*) $A_5C_5G_5T_4$, and (*C*) $A_5C_5G_5T_5$.

Different mismatches give different behavior

In Fig. 4, binding curves for four different 20-mer sequences are shown, each representing one pure sequence of the four nucleobases, respectively. Because our categorization of mismatches is such that all three C-containing mismatches belong to Class 1 (less costly, frustration factor $f_1 = 10^{-2}$) and all the G-containing mismatches are of Class 2 ($f_2 = 10^{-3}$), whereas the homo-T- and homo-A-oligomers represent the mixing of the two mismatch classes (33% and 67% Class 2, respectively), these four panels show the meaning of different frustration factors.

We see that panel *A* with the C_{20} sequence and $f_1 = 10^{-2}$ giving a reasonably specific binding for a range from $10^{-10.00}$ to $10^{-9.88}$ M, is quite similar to the other extreme, panel *D* corresponding to a G_{20} with $f_2 = 10^{-3}$, even though the therapeutic range is narrowed down to $\approx 10^{-9.95}$ to $10^{-9.88}$ M. Investigating the individual mismatch branches, the same pattern emerges. However, looking at the mixed cases, panels *B* and *C*, gives a completely different picture

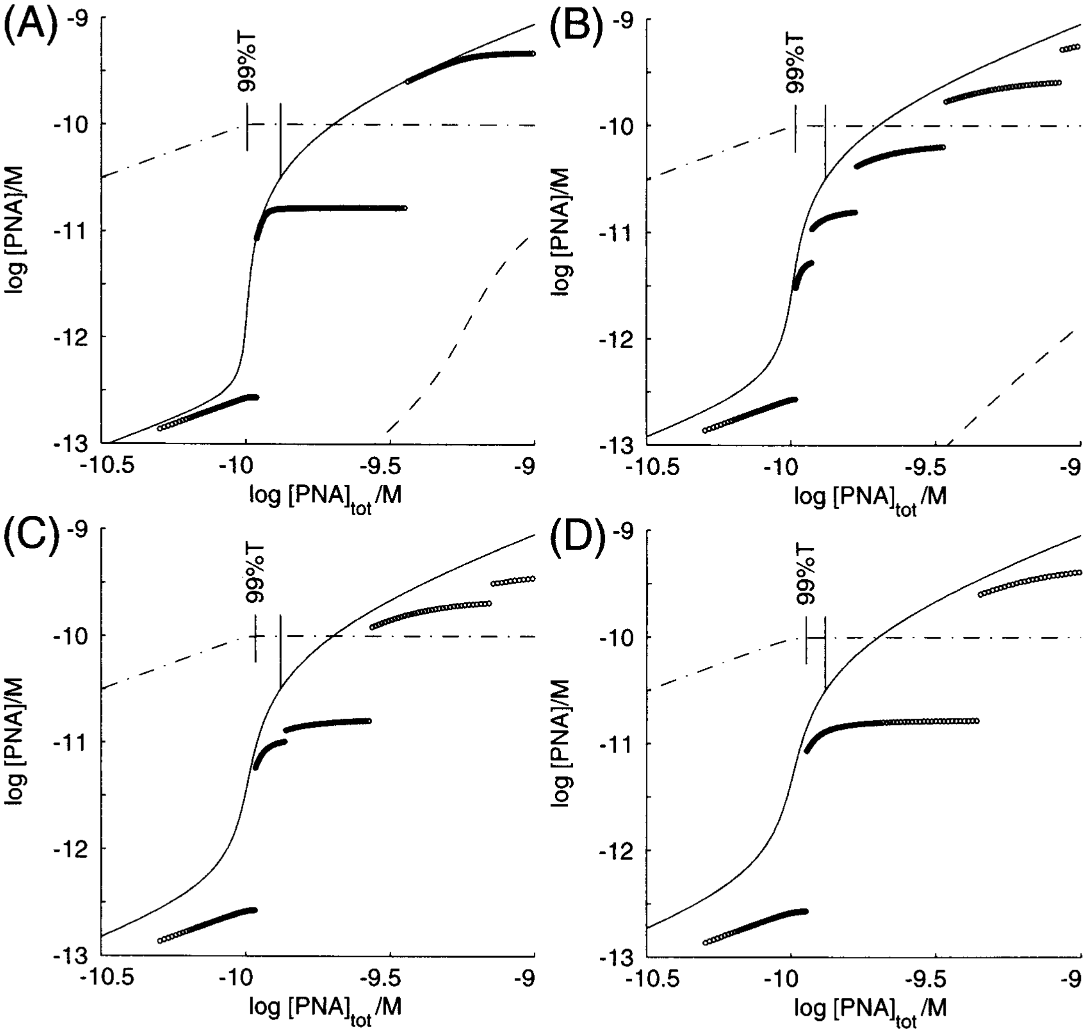

FIGURE 4 Profiles of [PNA·T] (–·–), $[PNA]_{free}$ (– – –), $\Sigma[P{\cdot}D_{IJ}]$ (——), $\Sigma[P{\cdot}D_{IJ}]_{crit}$ (○○○○○○) simulated for 20-mer PNA–DNA sequences containing only one type of base, at a DNA concentration of 0.1 nM, and using $k_0 = 8.6\ M^{-1}bp^{-1}$. (*A*) (C_{20}) represent only Class 1 mismatches with a frustration factor of 10^{-2}, whereas panel (*D*) is Class 2 only (G_{20}) with frustration factor 10^{-3}. (*B*) (T_{20}) and (*C*) (A_{20}) illustrate mixing of the two Classes, i.e., 33% and 67% Class 2, respectively.

of the individual mismatches. The two variants of singly mismatched species appear in different shapes, and especially the first of the doubly mismatched sequences in panel *B* (appearing around $[PNA]_{tot} = 10^{-9.78}$ M) may thus have to be considered with special caution. The overall mismatch concentration, though, gives binding curves and therapeutic ranges that fall linearly between the two extreme cases, *A* and *D*. What needs to be remembered is that examples of our simulations are based on rather arbitrary values of frustrations factors corresponding to the two classes of mismatches, and that the free-energy penalties are previously determined values taken from studies of single mismatches in 9- and 12-mer PNA–DNA duplexes formed with single-stranded DNA at room temperature and low salt (Ratilainen et al., 2000).

Choice of binding constant

Using mixed-sequence PNA–DNA duplexes, we previously determined a length-averaged contribution of $\partial\Delta G°/\partial n = -(6.5 \pm 0.3)$ kJ·mol^{-1}bp^{-1}, corresponding to a binding constant for one basepair (k_0) of $\approx$14 M^{-1} (at 25°C) (Ratilainen et al., 2000). Moreover, resolving the thermodynamics of PNA–DNA binding into enthalpic and entropic terms yielded $\partial\Delta H°/\partial n = -(30.0 \pm 2.5)$ kJ·mol^{-1}bp^{-1} and $\partial T\Delta S°/\partial n = -(23.5 \pm 2.3)$ kJ·mol^{-1}bp^{-1}, respectively (Ratilainen et al., 2000). Using these values, it is possible to calculate the binding constant at any temperature from $-RT \ln k_0(T) = \Delta H° - T\Delta S°$, if the enthalpy and entropy changes upon hybridization are considered independent of temperature. Converting the observed value of the average binding

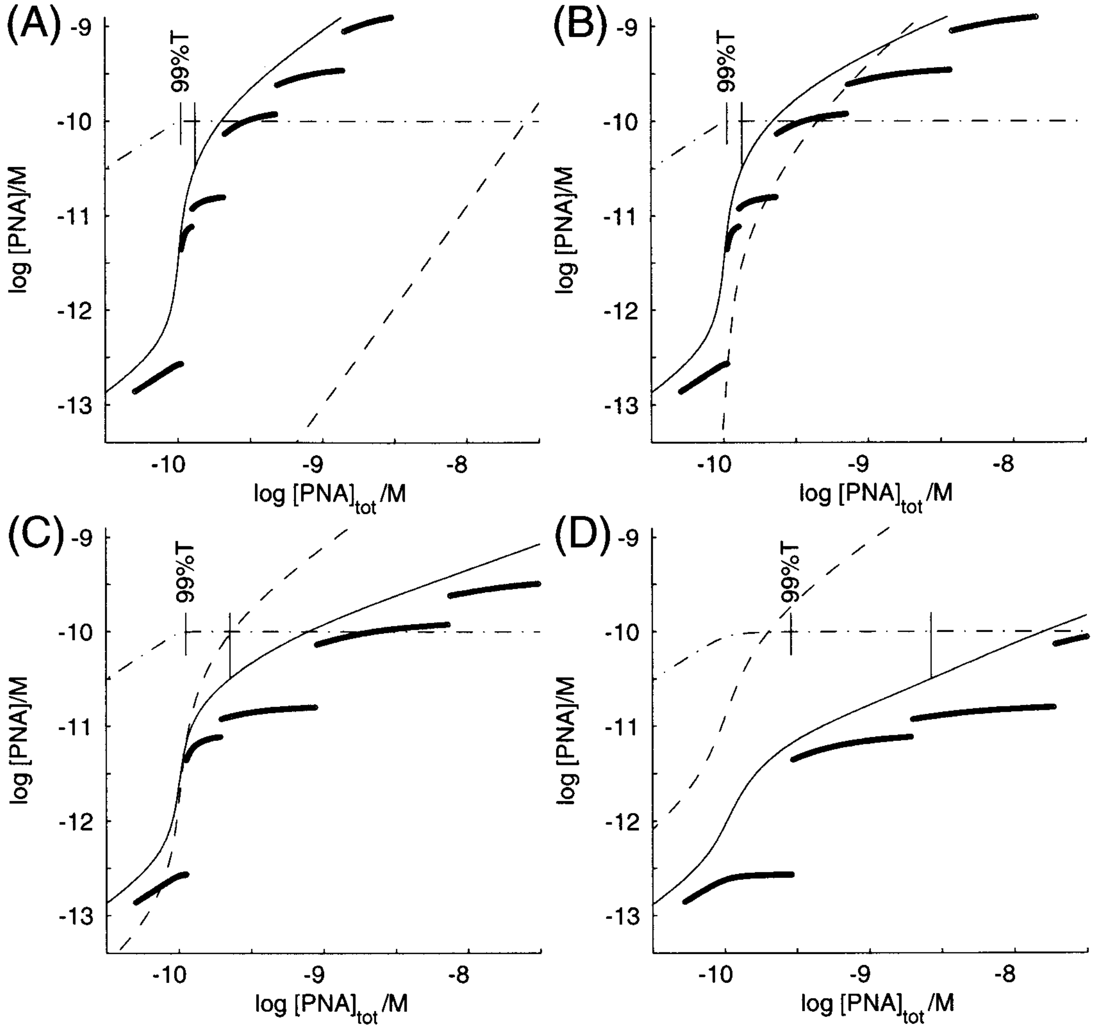

FIGURE 5 Profiles of [PNA·T] (– · –), $[PNA]_{free}$ (– – –), $\Sigma[P{\cdot}D_{IJ}]$ (———), $\Sigma[P{\cdot}D_{IJ}]_{crit}$ (○○○○○○) simulated for four identical 20-mer PNA–DNA sequences ($A_5C_5G_5T_5$) at different temperatures (37, 50, 55, and 60°C) resulting in different values of the binding constant k_0. (*A–D*) represent k_0 = 8.6, 5.4, 4.5, and 3.9 $M^{-1}bp^{-1}$, respectively. All four panels are simulated at a DNA concentration of 0.1 nM.

constant for formation of one basepair (k_0) yields ≈8.6 M^{-1} at physiological 37°C. This value corresponds to conditions with low ionic strength (10–100 mM NaCl), which implies that the PNA–DNA value outperforms the binding constant for DNA–DNA duplex formation by a factor of 3, and may serve as a guideline for estimating duplex stability for perfectly matched duplexes. At physiological salt conditions (140 mM NaCl) the difference is expected to be even lower, because the electrostatic repulsion between two DNA backbones is reduced substantially in presence of salt (Tomac et al., 1996). Consequently, in our simulations, we have in addition used lower values of k_0.

To keep the model calculation as simple as possible, emphasizing the effect of correct versus incorrect basepairing, we shall use one and the same k_0 value for AT and GC basepairs. For explicit cases of target calculations, of course, different k_0 values can be used and easily incorporated into the model.

In Fig. 5, we performed a series of simulations for the mixed-sequence PNA 20-mer, letting the binding constant correspond to different temperatures according to thermodynamic properties of PNA–DNA duplexes determined previously (Ratilainen et al., 2000). In panel *A*, the human physiological case (37°C) is shown, whereas panels *B–D* demonstrate how delicately the balance is shifted between 50 and 60°C for this system. This might be very useful in diagnostic applications where temperature often is used to tune or select the detection of the correct sequence relative to incorrect sequences (Carlsson et al., 1996). In a diagnostic application, and from a detection-limit standpoint, one may also have to consider which species gives rise to signal, e.g., what is labeled. For example, if the free PNA contributes to the detected signal, the case in panel *C* might imply an upper limit to the temperature used, because the free PNA concentration is of the same magnitude as the sum of mismatches. Still, though, they are both much below the target sequence concentration.

Figure 5 serves an additional purpose of showing the critical (lowest) k_0 value that still discriminates the target sequence. For our mixed 20-mer, this corresponds to $k_{0,\mathrm{crit}} \approx 4\ \mathrm{M^{-1}bp^{-1}}$, and hence panels *C* ($k_0 = 4.6\ \mathrm{M^{-1}bp^{-1}}$) and *D* ($k_0 = 3.9\ \mathrm{M^{-1}bp^{-1}}$) represent situations on either side of this limit. In panel *D*, the curve of the free PNA concentration (*dashed*), crosses the curve of target binding (*dash-dotted*) at a total PNA concentration around $10^{-9.7}$ M, i.e., before reaching the 99% targeting limit. Thus, any further addition of PNA only increases correct targeting marginally at the expense of significantly decreasing drug economy.

The same series also shows how the therapeutic range changes quite dramatically as k_0 is varied in the vicinity of the critical value. The point where the target is blocked "enough" moves to even higher values as the binding strength decreases (going from $k_0 = 5.4\ \mathrm{M^{-1}bp^{-1}}$ to $k_0 = 3.9\ \mathrm{M^{-1}bp^{-1}}$), though, accompanied by a relatively larger change of the lower specificity limit in the same direction thus making the therapeutic range larger at the expense of having to add unrealistic amounts of PNA.

At higher values of k_0 (Fig. 5 *A*), we see, as in the other simulations (Figs. 3 and 4), that the most striking feature is the very narrow range of useful concentrations of drug (PNA), thus setting a severe implication on the usefulness of this class of sequence-specific DNA binders. Just adding one extra base onto the probe can alter the concentration profile(s) of the mismatched species (that needs to be accounted for) considerably. Therefore, the optimal length will be greatly dependent on such factors as temperature and other physical conditions, e.g., salt. In the gene-targeting application, the total concentration of DNA and temperature is given, but we must keep in mind that the local structure of DNA may vary widely, and the accessible (effective) concentration may be significantly decreased. Another factor that may attenuate the binding scenario is the effective PNA (drug) concentration. As for single-stranded DNA (Holbrook et al., 1999), PNA sequences (particularly partially self-complementary ones) may form back-folded or partially double-stranded conformations, thereby lowering the effective PNA concentration and thus competing with the DNA targeting.

In more biotechnological or analytical applications, we may ourselves choose optimal conditions with respect to temperature and salt within certain practical limits. Also, in diagnostic contexts, the analyses are seldom made directly on the whole (human) genome, but on shorter stretches containing the gene of interest.

General features of simulations

Even though the following discussion can be applied to more sophisticated models, it is sufficient to consider the simplest model to illustrate the main important factors that govern the binding of PNA to DNA. Going back to the key equation when calculating the total concentration of bound PNA for the simplest case of binding, i.e., with only one type of mismatches present (Model 1):

$$\sum[\mathrm{P{\cdot}D_m}] = L[\mathrm{DNA}]_{\mathrm{tot}}\sum\binom{n}{m}\left(\frac{1}{4}\right)^{\mathrm{n-m}}\left(\frac{3}{4}\right)^{\mathrm{m}}\frac{k_0^{\mathrm{n}}f^{\mathrm{m}}[\mathrm{P}]}{1+k_0^{\mathrm{n}}f^{\mathrm{m}}[\mathrm{P}]}$$

$$= L[\mathrm{DNA}]_{\mathrm{tot}}\sum P(\mathrm{D_m})R. \qquad (12)$$

In the summation, the first part $P(\mathrm{D_m}) = \binom{\mathrm{n}}{\mathrm{m}}(1/4)^{\mathrm{n-m}}(3/4)^{\mathrm{m}}$, the probability of occurrence for any sequence of *n* bases in the genome to have exactly *m* mismatches, increases monotonically with *m*. This increase can be expressed by the ratio of probabilities for sequences with *m* versus $m - 1$ mismatches, given by $P(\mathrm{D_m})/P(\mathrm{D_{m-1}}) = 3((n + 1)/m - 1)$. The second factor $R = k_0^{\mathrm{n}}f^{\mathrm{m}}[\mathrm{P}]/(1 + k_0^{\mathrm{n}}f^{\mathrm{m}}[\mathrm{P}])$, the fraction of PNA binding to such a sequence, decreases monotonically with *m*. However, as long as conditions are such that $k_0^{\mathrm{n}}f^{\mathrm{m}}[\mathrm{P}] > 1$, *R* decreases very slowly with *m*, and the terms grow with increasing *m* until $k_0^{\mathrm{n}}f^{\mathrm{m}}[\mathrm{P}] \approx 1$. We shall call this "breaking point" *m* value m_{bp}. The behavior of the terms after this point depends on the size of the effective frustration factor *f* (likely to decrease as the length of sequence or the number of mismatches increases). If $1/f \gg 3((n + 1)/m_{\mathrm{bp}} - 1)$, the subsequent terms in the summation will rapidly decrease, and the concentration of PNA bound to nontarget sequences can be approximated by the maximum term in the summation with $m = m_{\mathrm{bp}}$. However, with a larger (less discriminating) frustration factor, subsequent terms will decrease slowly, or even increase, in case the total concentration of PNA bound has to include also partially covered mismatched sequences with $k_0^{\mathrm{n}}f^{\mathrm{m}}[\mathrm{P}] < 1$ (nonspecific binding).

Therapeutic range

The total amount of PNA added to the solution (cell) will, of course, always be the sum of free PNA and PNA bound to the correct target and to all possible mismatched sequences. The species distributions will, however, depend strongly on the binding situation, i.e., that part of the binding isotherm with which we are interested. Starting with the expression for total PNA concentration according to Model 1 (Eq. 5), and using the frustration factor for an average mismatch, $f = k_0/k_{\mathrm{m}}$, we have

$$[\mathrm{PNA}]_{\mathrm{tot}} = [\mathrm{P}] + [\mathrm{DNA}]_{\mathrm{tot}}\left(\frac{k_0^{\mathrm{n}}[\mathrm{P}]}{1+k_0^{\mathrm{n}}[\mathrm{P}]} + L\sum_{\mathrm{m=1}}^{\mathrm{n}}P(\mathrm{D_m})R\right). \qquad (13)$$

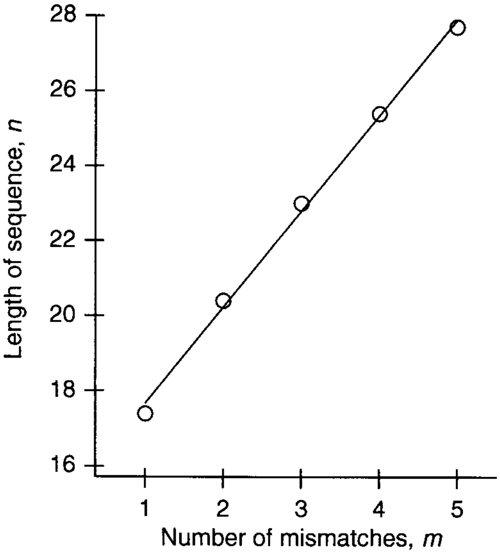

FIGURE 6 Lengths of PNA sequences, n, corresponding to the critical number of mismatches, m_{crit}, one has to consider while designing a specific binding scenario for a genome with $L = 3 \times 10^9$ bases (see text). Line represents linear fit $n = 15 + 2.6m_{\mathrm{crit}}$. The gamma function was used to calculate m_{crit} from the condition $P(\mathrm{D}_{m_{\mathrm{crit}}-1}) = 0.1$ (see Eq. 5), to be able to compute $\binom{n}{m}$ even for noninteger values of n and m.

Consider now the free PNA concentration [P] in two limiting situations. Case A, defined by the minimal free PNA concentration [P] that will satisfactorily block the target, i.e., when $k_0^{\mathrm{n}}[\mathrm{P}] = [\mathrm{P{\cdot}T}]/[\mathrm{T}] = \alpha$, where α, the ratio of blocked to unblocked target, is introduced as an efficiency parameter. Typically, we may want α to be on the order of 10^2–10^3. Because, by hypothesis, [P] is small enough that $L \sum P(\mathrm{D_m})R$ can be neglected, Eq. 13 then reduces to

$$[\mathrm{PNA}]_{\mathrm{tot,A}} \approx \frac{\alpha}{k_0^{\mathrm{n}}} + [\mathrm{DNA}]_{\mathrm{tot}}. \tag{14}$$

The other limit, case B, is the maximum [P] that can be tolerated before the onset of binding to mismatched sequences that have significant probabilities of occurring in the genome. Let m_{crit} be the smallest value of m that gives $P(\mathrm{D_m})L \geq 0.1$, i.e., the minimal number of mismatches necessary for a sequence to occur with a significant (here 10%) probability. A graph of n versus m_{crit}, shown in Fig. 6 for a genome of human size, reveals an approximately linear relationship in which the length n of the PNA has to be increased by between 2 and 3 bases to increase m_{crit} by one. For $m = m_{\mathrm{crit}}$, we require that $k_0^{\mathrm{n}} f^{\mathrm{m}}[\mathrm{P}] = [\mathrm{P{\cdot}D_m}]/[\mathrm{D_m}] < \beta$, where β is the tolerance for nontarget blockage ($\beta = 0.1$ means that ~10% of a sequence with m_{crit} mismatches will be blocked). Because only a small fraction of the mismatched sequences $\mathrm{D_m}$ are blocked, the expression within parenthesis in Eq. 13 is still close to unity and thus leads to the approximation,

$$[\mathrm{PNA}]_{\mathrm{tot,B}} \approx \frac{\beta}{k_0^{\mathrm{n}} f^{\mathrm{m}}} + [\mathrm{DNA}]_{\mathrm{tot}}. \tag{15}$$

To more easily assess what parameter to change to tune the binding, we shall consider the "therapeutic ratio," G, defined by the limiting cases A and B,

$$G = \frac{\text{highest safe } [\mathrm{PNA}]_{\mathrm{tot}}}{\text{lowest efficient } [\mathrm{PNA}]_{\mathrm{tot}}} = \frac{[\mathrm{PNA}]_{\mathrm{tot,B}}}{[\mathrm{PNA}]_{\mathrm{tot,A}}}$$

$$\approx \frac{\beta/m + k_0^{\mathrm{n}}[\mathrm{DNA}]_{\mathrm{tot}}}{\alpha + k_0^{\mathrm{n}}[\mathrm{DNA}]_{\mathrm{tot}}}. \tag{16}$$

There are several implications of Eq. 16: 1) If $m_{\mathrm{crit}} = 0$, no discrimination is possible, thus the minimal length n of the PNA must be >19 bases (see Fig. 6). 2) The therapeutic ratio G will be very close to unity unless $\beta/f^{\mathrm{m}} > k_0^{\mathrm{n}}[\mathrm{DNA}]_{\mathrm{tot}}$, and $\beta/f^{\mathrm{m}} > \alpha$. Because the DNA concentration is not variable in this context, k_0 may not be excessively large. 3) If k_0 is too large in comparison to the effective frustration factor, leading to $1/f < k_0^{(2-3)}$, an increase in the length n of the PNA will, in fact, lead to a decrease in G. 4) In contrast, because $\alpha = k_0^{\mathrm{n}}[\mathrm{P}]$, a too-small k_0 will result in the free PNA concentration [P] being too large, which would be uneconomical and of potential danger with regard to side effects from non-nucleic acid interactions.

A narrow therapeutic range has previously been observed in vitro. Administration of PNA complementary to different targets in mRNA was found to sequence-specifically inhibit translation only within narrow limits (Bonham et al., 1995; Gambacorti-Passerini et al., 1996). The usable range of PNA concentrations is thus quite restrained if one wants to achieve total binding to the target and, at the same time, avoid extensive binding to nontarget sites, again stressing the need for detailed understanding of the basic binding properties of PNA and other sequence-specific DNA-binding agents.

Improving the model

The model we have presented here is quite simplistic from several aspects. Again, it is easily expanded to include additional determinants of the binding strength of PNA to DNA, e.g., relative distribution of GC and AT basepairs (GC%) and fractional pyrimidine content on the PNA strand (Giesen et al., 1998). Expanding on such parametric factors, one could incorporate influence of nearest-neighbor interactions as has been determined for DNA–DNA hybridization (Peyret et al., 1999). However, because the PNA–DNA basepair is asymmetric, a much larger experimental data set is required to extract nearest-neighbor parameters even for matched basepairs, as pointed out previously (Giesen et al., 1998).

For DNA–DNA hybridization, it has been shown that the nearest neighbors in the sequence also influence mismatch stability significantly, and the first complete study of thermodynamic nearest-neighbor parameters was reported by Peyret et al. (1999). Stability of PNA–DNA basepairs is likewise expected to be highly context dependent, which one should take into account when designing PNA oligomers for targeting DNA or RNA sequences in both diagnostic and therapeutic applications and also when choosing targets (Kilså Jensen et al., 1997; Ratilainen et al., 2000). A first step to improving the model, besides a more sophisticated categorization of mismatches, would be inclusion of the composition of neighboring basepairs. We previously found, for PNA–DNA duplexes with single mismatches that, for example, flanking GC pairs, in general, seem to stabilize the mismatch relative to flanking AT basepairs, a behavior similar to what has been observed for DNA–DNA (Peyret et al., 1999) and RNA–RNA duplexes (Zhu and Wartell, 1997).

Relevance to gene targeting

As pointed out, the concrete output data of the simulations should be treated with caution when applied to gene targeting. The results should mainly serve the purpose of identifying critical features and how the present approach could be applied once relevant thermodynamic data is available. Thus, the results of our simulations particularly refer to PNA, forming duplexes with single-stranded DNA. They would thus apply, in principle, to antisense applications once correction for the smaller mRNA genome and lower concentration is made. The approach also directly applies to minor-groove binding gene-targeting agents, such as those invented by Dervan and co-workers (Herman et al., 1999). However, for triplex-forming PNA gene-targeting strategies, particular care has to be taken because of the kinetically governed slow processes of strand invasion (Wittung et al., 1996, 1997).

CONCLUSIONS

The statistical modeling of the hybridization properties of a model ligand (PNA) to a model genome target (single-stranded DNA) has been used to get insight into the basic determinants of binding specificity. The main parameters in this treatment have been the length of the PNA probe, the average per-base binding constant k_0, and the binding affinity loss, given by the frustration factor $f = k_{\mathrm{mis}}/k_0$, of a mismatched basepair. The specificity has been measured as the range of safety of the total drug (PNA) concentration or dose, defined as a therapeutic ratio G = maximum safe $[\mathrm{PNA}]_{\mathrm{tot}}$/minimal efficient $[\mathrm{PNA}]_{\mathrm{tot}}$.

We find that, by far, the most important parameter with respect to G is the average per-base binding constant k_0, and that only a rather narrow range of values gives rise to high binding selectivity. Thus, the in vitro value determined previously (Ratilainen et al., 2000) is too high to permit any significant selectivity. For diagnostic purposes, the value of k_0 is readily modulated due to the substantial $\Delta H°$ associated with the equilibrium, and we find that a temperature of 55–60°C should be optimal for selective binding of PNA to a single-stranded DNA of genomic proportions. For gene-therapeutic purposes, however, the higher ionic strength in the cell and the presence of other, competing DNA-binding molecules (Boffa et al., 1996) including histone proteins in chromatin will reduce the binding, which could be handled by choosing a higher effective PNA concentration. Appropriate control of the binding constant of the gene-targeting agent is thus required to optimize in vivo selectivity.

Our model is quite general, and, though simple, gives insight about sequence-specific binding. Thus, it can be applied to a variety of situations of molecules binding sequence-specifically to a target, ranging from gene-targeting at various levels (antigene, antisense) to biotechnical applications such as gene-sequencing and gene-diagnostics.

We thank Mr. Andreas Hall for writing the program code.

Financial support from the Swedish Cancer Foundation (B.N.), the European Union Biomed 2 Program (BMH4-CT96-0848 and BMH4-CT96-1081, B.N.). is gratefully acknowledged.

REFERENCES

Boffa, L. C., P. L. Morris, E. M. Carpaneto, M. Louissaint, and V. G. Allfrey. 1996. Invasion of the CAG triplet repeats by a complementary peptide nucleic acid inhibits transcription of the androgen receptor and TATA-binding protein genes and correlates with refolding of an active nucleosome containing a unique AR gene sequence. *J. Biol. Chem.* 271:13228–13233.

Bonham, M. A., S. Brown, A. L. Boyd, P. H. Brown, D. A. Bruckenstein, J. C. Hanvey, S. A. Thomson, A. Pipe, F. Hassman, J. E. Bisi, B. C. Froehler, M. D. Matteuci, R. W. Wagner, S. A. Noble, and L. E. Babiss. 1995. An assessment of the antisense properties of RNase H-competent and steric-blocking oligomers. *Nucleic Acids Res.* 23:1197–1203.

Carlsson, C., M. Jonsson, B. Nordén, M. T. Dulay, R. N. Zare, J. Noolandi, P. E. Nielsen, L.-Z. Tsui, and J. Zielenski. 1996. Screening for genetic mutations. *Nature.* 380:207.

Chan, P., M. Lin, A. F. Faruqi, J. Powell, M. M. Seidman, and P. M. Glazer. 1999. Targeted correction of an episomal gene in mammalian cells by a short DNA fragment tethered to a triplex-forming oligonucleotide. *J. Biol. Chem.* 274:11541–11548.

Cohen, J. S., and M. E. Hogan. 1994. The new genetic medicines. *Sci. Am.* 271:50–55.

Cook, P. D. 1998. Antisense medicinal chemistry. *In* Antisense Research and Application. S. T. Crooke, editor. Springer, Berlin. 51–101.

Crooke, S. T. 1995. Oligonucleotide therapeutics. *In* Burger's Medicinal Chemistry and Drug Discovery. M. E. Wolff, editor. Wiley and Sons, New York. 863–900.

Daban, J.-R. 2000. Physical constraints in the condensation of eukaryotic chromosomes: local concentration of DNA versus linear packing ratio in higher order chromatin structures. *Biochemistry.* 39:3861–3866.

Egholm, M., O. Buchardt, L. Christensen, C. Behrens, S. M. Freier, D. A. Driver, R. H. Berg, S. K. Kim, B. Nordén, and P. E. Nielsen. 1993. PNA hybridizes to complementary oligonucleotides obeying the Watson–Crick hydrogen-bonding rules. *Nature.* 365:566–568.

Gambacorti-Passerini, C., L. Mologni, C. Bertazzoli, P. le Coutre, E. Marchesi, F. Grignani, and P. E. Nielsen. 1996. In vitro transcription and translation inhibition by anti-promyelocytic leukemia (PML)/retinoic acid receptor α and anti-PML peptide nucleic acid. *Blood.* 88: 1411–1417.

Giesen, U., W. Kleider, C. Berding, A. Geiger, H. Ørum, and P. E. Nielsen. 1998. A formula for thermal stability (T-m) prediction of PNA/DNA duplexes. *Nucleic Acids Res.* 26:5004–5006.

Herman, D. M., J. M. Turner, E. E. Baird, and P. B. Dervan. 1999. Cycle polyamide motif for recognition of the minor groove of DNA. *J. Am. Chem. Soc.* 121:1121–1129.

Holbrook, J. A., M. W. Capp, R. M. Saecker, and M. T. Record, Jr. 1999. Enthalpy and heat capacity changes for formation of an oligomeric DNA duplex: interpretation in terms of coupled processes of formation and association of single-stranded helices. *Biochemistry.* 38:8409–8422.

Kilså Jensen, K., H. Ørum, P. E. Nielsen, and B. Nordén. 1997. Kinetics for hybridization of peptide nucleic acids (PNA) with DNA and RNA studied with the BIAcore technique. *Biochemistry.* 36:5072–5077.

Nielsen, P. E. 1991. Sequence selective DNA recognition by synthetic ligands. *Bioconjugate Chem.* 2:1–12.

Nielsen, P. E. 1999. Peptide nucleic acids as therapeutic agents. *Curr. Opin. Struct. Biol.* 9:353–357.

Nielsen, P. E., and M. Egholm. 1999. An introduction to peptide nucleic acids. *In* Peptide Nucleic Acids (PNA). Protocols and Applications. P. E. Nielsen, and M. Egholm, editors. Horizon Scientific, Norfolk, VA. 1–19.

Nielsen, P. E., M. Egholm, R. H. Berg, and O. Buchardt. 1991. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science.* 254:1497–1500.

Peyret, N., P. A. Seneviratne, H. T. Allawi, and J. SantaLucia. 1999. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A · A, C · C, G · G, and T · T mismatches. *Biochemistry.* 38:3468–3477.

Ratilainen, T., A. Holmén, E. Tuite, P. E. Nielsen, and B. Nordén. 2000. Thermodynamics of sequence-specific binding of PNA to DNA. *Biochemistry.* 39:7781–7791.

Ren, J., and J. B. Chaires. 1999. Sequence and structural selectivity of nucleic acid binding ligands. *Biochemistry.* 38:16067–16075.

Rufer, N., W. Dragowska, G. Thornbury, E. Roosnek, and P. M. Lansdorp. 1998. Telomere length dynamics in human lymphocyte subpopulations measured by flow cytometry. *Nature Biotech.* 16:743–747.

Tomac, S., M. Sarkar, T. Ratilainen, P. Wittung, P. E. Nielsen, B. Nordén, and A. Gräslund. 1996. Ionic effects on the stability and conformation of peptide nucleic acid (PNA) complexes. *J. Am. Chem. Soc.* 118:5544–5552.

Uhlmann, E., and A. Peyman. 1990. Antisense oligonucleotides: a new therapeutic principle. *Chem. Rev.* 90:543–584.

Wang, J., G. Rivas, X. Cai, M. Chicharro, C. Parrado, N. Dontha, A. Begleiter, M. Mowat, E. Palecek, and P. E. Nielsen. 1997. Detection of point mutation in the p53 gene using a peptide nucleic acid biosensor. *Anal. Chim. Acta.* 344:111–118.

Weiler, J., H. Gausepohl, N. Hauser, O. N. Jensen, and J. D. Hoheisel. 1997. Hybridisation based DNA screening on peptide nucleic acid (PNA) oligomer arrays. *Nucleic Acids Res.* 25:2792–2799.

Wittung, P., P. Nielsen, and B. Nordén. 1996. Direct observation of strand invasion by peptide nucleic acid (PNA) into double-stranded DNA. *J. Am. Chem. Soc.* 118:7049–7054.

Wittung, P., P. Nielsen, and B. Nordén. 1997. Extended DNA-recognition repertoire of peptide nucleic acid (PNA): PNA-dsDNA triplex formed with cytosine-rich homopyrimidine PNA. *Biochemistry.* 36:7973–7979.

Zhu, J., and R. M. Wartell. 1997. The relative stabilities of base pair stacking interactions and single mismatches in long RNA measured by temperature gradient gel electrophoresis. *Biochemistry.* 36:15326–15335.