Design of Sequence-Specific DNA-Binding Ligands

Peter E. Nielsen*

Abstract: Double-stranded DNA can be viewed as a multifunctional, modular receptor that can be read sequence-selectively in a digital way (base pair per base pair) by a complementary, similarly modular ligand. This principle has been exploited in several approaches to design sequence-specific DNA-binding ligands, such as triplex-forming oligonucleotides, peptide nucleic acids and minor groove binding polyamides.

Keywords: DNA recognition · triplexes · molecular recognition · peptide nucleic acids

Introduction

Molecular recognition is a fundamental principle in biology, exploited at various levels. Examples are enzyme–substrate, antigen–antibody, hormone–receptor and drug–DNA as well as protein–DNA and protein–protein interactions. In the effort to understand the physicochemical principles of molecular recognition per se and further to exploit such knowledge in the design of specific ligands and receptors, a discipline of supramolecular and host–guest chemistry is growing fast. In medicinal chemistry the concept of rationally designed ligands for medicinally relevant receptors is also attracting much attention and could provide the basis for development of novel potent therapeutic drugs.

In the border area between molecular biology and chemistry, a field concerned with DNA recognition is thriving. The field was sparked by studies of the huge variety of natural ligands in the form of gene regulatory or processing proteins or potent low molecular weight anticancer drugs that bind sequence selectively to double-stranded DNA. Inspired by nature, chemists have now for more than 10 years developed principles with the ultimate goal of being able to design and synthesize compounds to order that recognize and bind to any desired sequence in double-stranded DNA.[1–3]

[*] P. E. Nielsen
Center for Biomolecular Recognition
Department for Biochemistry and Genetics
Biochemistry Laboratory B, The Panum Institute
Blegdamsvej 3c, DK-2200 N Copenhagen (Denmark)
Fax: Int. code +(31)39-6042
e-mail: penpanum@biobase.dk

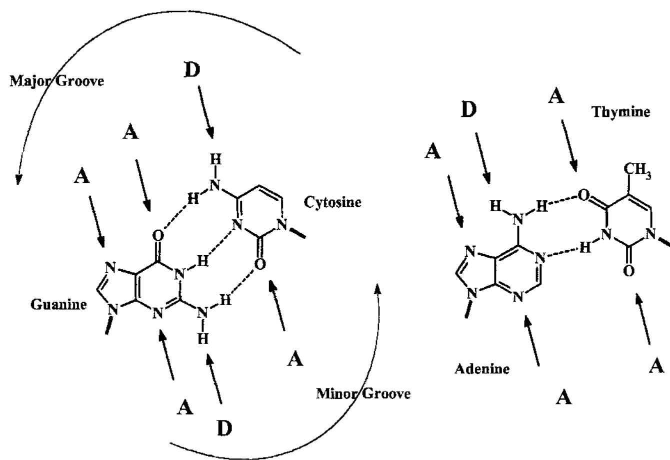
One could easily argue that DNA is "the dream of a receptor" for designing ligands, since X-ray crystal structures as well as NMR solution structures of numerous DNA molecules have been described, and even the structures of a multitude of complexes between natural DNA binding ligands—both proteins and low molecular weight drugs—and their DNA targets have been solved. Furthermore, as discussed below, the receptor is modular due to the base pair sequence, and thus a general (and modular) solution for designing a ligand for a DNA receptor of any sequence should be attainable. Interestingly, nature does not seem to have exploited such a modular principle in protein–DNA recognition (possibly with the exception of at least some Zn-finger proteins), but have in most cases evolved a "shape"-complementary "analogue readout" of the DNA helix major groove. In fact it may be that nature has avoided a "digital" (base pair by base pair) readout of DNA for exactly the opposite reason, namely, in order to evolve proteins that would recognize not only one single DNA sequence, but rather a family of related DNA sequence with differentiated affinity.

Discussion

DNA-recognizing ligands: Although the Watson–Crick "face" of the nuclobases used for interstrand recognition obviously is not accessible from the outside of the DNA double helix, ample nucleobase recognition features are still available from both the minor and especially from the major groove (Figure 1, top). In fact, each of the four base pairs can be unambiguously identified from the major groove solely by the pattern and position of hydrogen-bonding acceptor and donor sites. In the minor groove it is possible to distinguish G–C, C–G and A/T–T/A in this way, but not A–T and T–A base pairs (Figure 1, bottom). Consequently, both the minor and especially the major groove of double-stranded DNA have been targeted as sequence-specific "DNA receptors". The inspiration in these approaches has come from natural ligands that bind with sequence preferences in either groove and have been known for several decades.[1–3] Most interestingly, recent developments have shown that even the Watson–Crick "face" of the nucleobases can be made available for recognition by reagents, such as PNA (see below) that invade the DNA double helix.

Triple helix forming oligonucleotides: Close to 40 years ago, it was discovered that polynucleotides can form triple helices in

 © VCH Verlagsgesellschaft mbH, D-69451 Weinheim, 1997 0947-6539/97/0304-0505 $ 17.50+.50/0



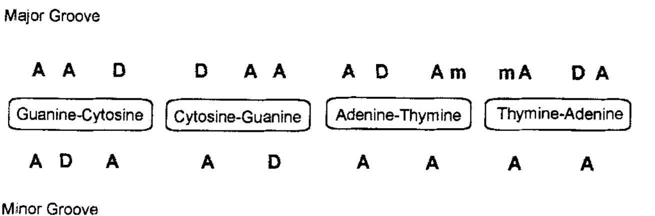


Figure 1. Base pair recognition. Top: Hydrogen-bonding donor (D) and acceptor (A) patterns of DNA base pair in the major and minor grooves of a double helix. Bottom: Schematic representation of the four Watson–Crick base pairs (m: hydrophobic site (methyl group of thymine)).

which a third strand can bind in the major groove of a normal Watson–Crick duplex by T·A or C^{+}·G Hoogsteen base pairings (Figure 2).[4] This principle was adapted by Moser and Dervan who showed that, under appropriate conditions, even short homopyrimidine oligonucleotides (15-mers) can bind sequence specifically to their "Hoogsteen-complementary" double-stranded DNA targets.[5] Simultaneously, Hélène et al. also discovered triple-helical binding of short oligonucleotides.[6] Although subsequent intensive studies and developments have demonstrated the potential of this principle for sequence-specific DNA recognition, for example, for developing specific antigene drugs[7] or reagents for genomic mapping,[2] general sequence recognition has not yet been achieved. Recognition is by and large still restricted to homopurine DNA targets despite the discovery of alternate triplex motifs and the development of a few new "nucleobases" and "strand-switching" strategies.[8] Basically, efficient recognition of thymine and cytosine in a triplex context is still lacking, and most likely will require the design of new recognition units.

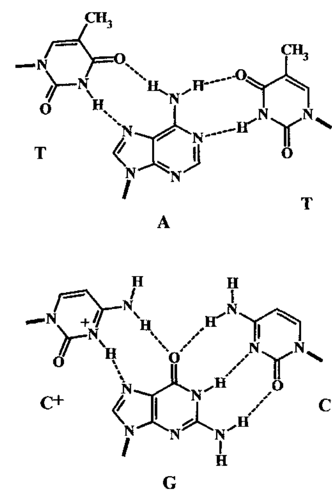


Figure 2. Recognition of A·T and G–C base pairs by thymine and protonated cytosine Hoogsteen base pairing.

Duplex-invading PNA: PNA (peptide nucleic acid), a DNA mimic having an uncharged, achiral pseudopeptide backbone,[9] was constructed in an attempt to devise a new and more efficient strategy to attack this problem. It was the hope that such oligomers by virtue of their lack of negative charge (or even engineered positive charge) would bind more efficiently to duplex DNA. Because PNA is based on amide chemistry,[10] it should also be easier to incorporate and explore the properties of new recognition entities to primarily recognize cytosine and thymine, but ultimately also to replace thymine and cytosine (for adenine and guanine recognition). However, quite surprisingly homopyrimidine PNAs turned out to bind to their targets in double-stranded DNA by a new mechanism, duplex invasion, in which a strand-displacement complex composed of an internal PNA·DNA-PNA (Hoogsteen·Watson–Crick) triplex and a single DNA strand is formed[11, 12] (Figure 3). Although this binding mode in itself by no means solves the issue of C and T recognition, it profoundly redirects the problem, and presents two new strategies. At face value, the simplest solution would be to construct a PNA oligomer with sufficient binding strength to form a stable strand-displacement complex without the need of

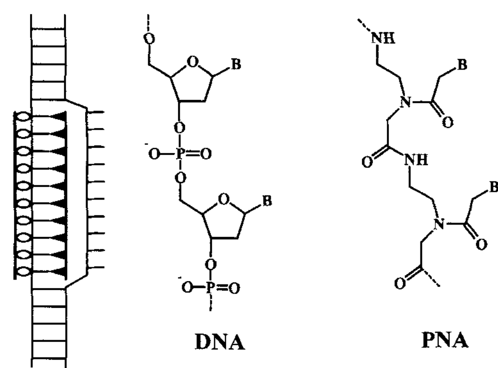
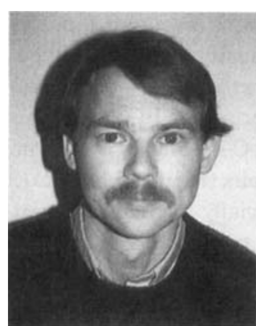


Figure 3. Left: Schematic drawing of a PNA_2·DNA triplex strand-displacement complex. Watson–Crick base pairs are indicated by triangles, while Hoogsteen base pairs are depicted by ellipses. Middle and right: Chemical structures of DNA and PNA.



Editorial Board Member:[*] *Peter E. Nielsen was born in Copenhagen, Denmark, in 1951. He received his Ph.D. in Chemistry in 1980 from the University of Copenhagen and has, apart from a postdoctoral stay at U. C. Berkelyey, worked at the Department of Biochemistry, The Panum Institute, University of Copenhagen, where he is now professor and director of the Center for Biomolecular Recognition. His research interests are molecular recognition in general and DNA recognition in particular. He has developed the uranyl photofootprinting technique and is co-inventor of PNA.*

[*] Members of the Editorial Board will be introduced to the readers with their first manuscript.

© VCH Verlagsgesellschaft mbH, D-69451 Weinheim, 1997
0947-6539/97/0304-0506 $ 17.50+.50/0

a stabilizing Hoogsteen triplex strand. In fact this solution could be within reach, since a recent study showed that homopurine PNAs which do not form triplexes, but do form especially stable PNA–DNA duplexes, do indeed invade a duplex target.[13] Alternatively, the solution could be to construct new recognition units for T and C as with DNA triplexes. However, in the PNA triplex case the obstacles seem less severe since recognition also takes place through the Watson–Crick PNA strand. Thus, by making a "bis-PNA" having a normal ACGT Watson–Crick recognizing first strand and a second strand composed of thymines and cytosines (or pseudoisocytosines[14]) for recognizing A and G plus two novel "nucleobases" that do not need to provide specific T and C recognition, but need only contribute triplex stabilization, one could devise a general sequence recognition principle based on triplex strand displacement. Finally, a "double duplex" strand invasion regime can be imagined with the invading oligomers forming duplexes with both of the DNA strands (Figure 4). This, of course, requires the development of new nucleobases that prevent hybridization of the two "self-complementary" invading oligomers without interfering with the DNA hybridization properties of the oligomer.[15]

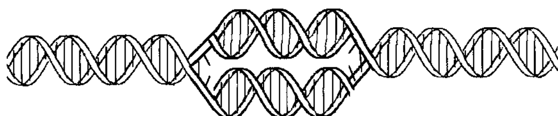


Figure 4. Schematic drawing of a double-duplex invasion complex.

Minor groove binding polyamides: Several natural (netropsin, distamycin) and synthetic (DAPI, Hoechst 33258) ligands that bind specifically in the minor groove of $(A/T)_n$ ($n \geq 4$) sites in double-stranded DNA have been known for many years. Most researchers now agree that this binding is primarily governed by the significantly narrowed minor groove of high electronegative potential of these A/T DNA regions, in which the positively charged and curved minor groove binders snugly fit.[16, 17]

Based on a hydrogen-bonding recognition model for netropsin and distamycin (which may not be correct) it was originally proposed that the introduction of hydrogen-bond acceptor units—such as imidazoles or thiopyrroles—in the minor groove binding molecules could accomplish recognition of G–C base pairs (but without distinguishing between G–C and C–G). This could take place through interaction with the 2-amino group of guanine, which is placed at the floor of the minor groove. However, the resulting "lexitropsins" turned out to still preferentially bind to A/T regions, although intervening G–C base pairs were to some extent tolerated.[18]

The design of sequence-selective minor groove binders took a dramatic new turn when Pelton and Wemmer showed in NMR structural studies that two distamycins, under some circumstances, could bind side by side in an antiparallel fashion in the minor groove[19] (a binding mode that in passing is not unlike that found for mithramycin binding to GC-rich regions[20]).

Dervan and Wemmer quickly picked up this side-by-side binding mode as they realized that it might be possible to develop a principle by which to read out the hydrogen-bonding information of the minor groove such that each unit of the ligand reads only half a base pair (or in essence a single nucleobase)[21, 22] (Figure 5). However, as mentioned earlier, a discrimination between A–T and T–A base pairs is difficult (or even impossible) and has not yet been accomplished. Nonetheless, using such polyamides, Dervan and co-workers have recently been able to target a range of A/T G A/T sequences as well as G-tracts with high specificity.[23] It remains to be established what the detailed specificity of these reagents are, and the generality of the approach in terms of the types, number and sizes of sequences that can be successfully targeted is not fully elucidated.

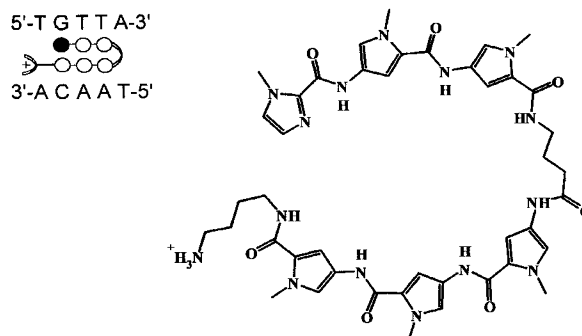


Figure 5. Left: Schematic representation of the recognition of the DNA helix from the minor groove by a hairpin polyamide consisting of pyrrole (o) and imidazole (•) units, where a pyrrole recognizes half of an A–T base pair or the cytosine of a C–G base pair, whereas an imidazole recognizes the guanine of a G–C base pair. Right: Chemical structure of the hairpin polyamide.

Proteins and peptides: Apart from purely chemical approaches to DNA recognition, Zn-finger proteins have successfully been used in a biological approach that exploits the phage display combinatorial library system to select ligands that bind specific DNA sequences. Thereby a Zn-finger sequence/DNA sequence recognition code can be approached.[24]

Although nature, in the form of the cyclic depsipeptide antibiotic echinomycin that recognizes G–C base pair steps, has utilized a small peptide to obtain some degree of sequence specificity, it has not yet been possible to successfully exploit peptides in synthetic DNA-binding ligands. This is probably due to the great conformational freedom of small peptides. Thus one could hope that by restricting this freedom, for example, by cyclization, DNA-recognizing peptides could be developed. Recent results with "distamycin-peptide" chimeras that recognize AT structures with intervening GC base pairs[25] are encouraging in this respect, and could lead the way to DNA-binding ligands on the basis of peptide–DNA recognition.

Closing Remarks

It is my hope that this brief account has demonstrated that efforts towards the design of sequence-specific DNA-binding ligands are quite successful, and that there is good reason to believe that the final goal of being able to target any desired sequence in double-stranded DNA will be reached within the not too distant future. Whether this ability will enable the development of potent gene specific therapeutic drugs is of course yet another story.

The results in this field also emphasize that structural information is not sufficient to understand a system at a level that allows de novo design. A quantitative evaluation of the various molecular interactions as well as a much better description and understanding of the dynamics (including thermodynamics) of the receptor (the DNA in this case) and the ligand both prior and subsequent to the binding event is also required. The dynamics of moving molecules instead of just static molecules must be dealt with, and in doing so solvent (water) displacement and rearrangement must also be taken into account. Thus I believe that this field would benefit very much if the DNA binding of the variety of already existing ligands was understood from a (thermo)dynamic perspective.

Received: November 29, 1996 [C530]

[1] P. E. Nielsen, *Bioconjugate Chem.* **1991**, *2*, 1.
[2] P. B. Dervan, *Nature*, **1992**, *359*, 87.
[3] C. Hélène, *Curr. Opinion Biotechnol.* **1993**, *4*, 29.
[4] G. Felsenfeld, D. R. Davies, A. Rich, *J. Am. Chem. Soc*, **1957**.
[5] H. Moser, P. B. Dervan, *Science*, **1987**, *238*, 645.
[6] T. Le Doan, L. Perrouault, D. Praseuth, N. Habhoub, N. Decout, J. L. Thuong, N. T. Lhomme, C. Hélène, *Nucleic Acids Res.* **1987**, *15*, 7749.
[7] L. J. Maher, *Cancer Invest.*, **1996**, *14*, 66.
[8] K. N. Ganesh, V. A. Kumar, D. A. Barawkar in *Perspectives in Supramolecular Chemistry* (Ed.: A. D. Hamilton), Wiley, New York, **1996**, Vol. 3, 263.
[9] P. E. Nielsen, M. Egholm, R. H. Berg, O. Buchardt, *Science* **1991**, *254*, 1457.
[10] B. Hyrup, P. E. Nielsen, *Bioorg. Med. Chem.* **1996**, *4*, 5.
[11] D. Y. Cherny, B. P. Belotserkovskii, M. D. Frank-Kamenetskii, M. Egholm, O. Buchardt, R. H. Berg, P. E. Nielsen, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 1667.
[12] P. E. Nielsen, M. Egholm, R. H. Berg,. O. Buchardt, *J. Mol. Recognit.* **1994**, *7*, 165.
[13] P. E. Nielsen, L. Christensen, *J. Am. Chem. Soc.* **1996**, *118*, 2287.
[14] M. Egholm, L. Christensen, K. Dueholm, O. Buchardt, J. Coull, P. E. Nielsen, *Nucleic Acids Res.* **1995**, *23*, 217.
[15] J. Woo, R. B. Meyer, Jr., H. B. Gamper, *Nucleic Acids Research* **1996**, *24*, 2470.
[16] P. E. Pjura, K. Grzeskowiak, R. E. Dickerson, *J. Mol. Biol.*, **1987**, *197*, 257.
[17] A. A. Wood, C. M. Nunn, A. Czarny, D. W. Boykin, S. Neidle, *Nucleic Acids Res.* **1995**, *23*, 3678.
[18] J. W. Lown, *J. Mol. Recognit.* **1994**, *7*, 79.
[19] J. G. Pelton, D. E. Wemmer, *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 5723.
[20] M. Sastry, D. J. Patel, *Biochemistry*, **1993**, *32*, 6588.
[21] W. S. Wade, M. Mrksich, P. B. Dervan, *J. Am. Chem. Soc.* **1992**, *114*, 8783.
[22] M. E. Parks, E. E. Baird, P. B. Dervan, *J. Am. Chem. Soc.* **1996**, *118*, 6147.
[23] S. E. Swalley, E. E. Baird, P. B. Dervan, *J. Am. Chem. Soc.* **1996**, *118*, 8198.
[24] Y. Choo, A. Klug, *Curr. Opinion Biotechnol.* **1995**, *6*, 431.
[25] V. A. Nikolaev, S. L. Grokhovsky, A. N. Surovaya, T. A. Leinsoo, N. Yu Sidorova, A. S. Zasedatelev, A. L. Zhuze, G. A. Strahan, R. H. Shafer, G. V. Gursky, *J. Biomol. Struct. Dynamics* **1996**, *14*, 31.

© VCH Verlagsgesellschaft mbH, D-69451 Weinheim, 1997
0947-6539/97/0304-0508 $ 17.50+.50/0