From transcription factors to designed sequence-specific DNA-binding peptides

M. Eugenio Vázquez†, Ana M. Caamaño and J. L. Mascareñas*

Departamento de Química Orgánica y Unidad Asociada al CSIC, Universidad de Santiago de Compostela, Santiago de Compostela, 15782, Spain. E-mail: qojoselm@usc.es; Fax: (+34) 981-595-012

Received 7th March 2003
First published as an Advance Article on the web 3rd July 2003

Transcription factors are DNA-binding proteins responsible for initiating the transcription of particular genes upon interacting with specific DNA sequences located at their promoter or enhancer regions. The DNA recognition process, which is extremely selective, is mediated by non-covalent interactions between appropriately arranged structural motifs of the protein and exposed surfaces of the DNA bases and backbone. The great variability in DNA recognition by transcription factors has hampered the characterization of an amino acid–base step recognition code, making it very difficult to design non-natural peptides that can mimic the DNA-binding properties of these naturally occurring counterparts. However, in recent years, several transcription factor-based miniature proteins capable of tight interaction with specific DNA sites have been successfully constructed, most of them using bottom-up synthetic approaches.

1 Introduction

Cellular behavior relies to a great extent on the controlled expression of proteins, a process that is mainly regulated at the

† Current address, Massachusetts Institute of Technology, Department of Chemistry, Cambridge, USA.; E-mail: eugeniov@mit.edu

M. Eugenio Vázquez was born in 1973, he studied Chemistry at the University of Santiago de Compostela, and after obtaining his degree (Hon) in 1996, he worked for his PhD with Dr. José Luis Mascareñas in the design and preparation of new DNA-binding peptides. During his predoctoral studies he spent six months in Cambridge in the laboratories of Professor Alan Ferscht and three months in the University of Barcelona in the group of Professor Modesto Orozco. After receiving his PhD in 2001 he was awarded with the Human Frontier Science Program Fellowship, and joined the group of Professor Barbara Imperiali at the Massachusetts Institute of Technology as a postdoctoral fellow, where his current work is centered on the design of caged and fluorescent peptides as tools for understanding underlying molecular mechanisms of complex biological processes.

Ana M. Caamaño was born in 1974, she studied Chemistry at the University of Santiago de Compostela, obtaining her degree (Hon) in 1997. She then joined the laboratory of Professor José Luis Mascareñas for PhD studies on the design and preparation of switchable DNA-binding peptides. During her predoctoral studies she spent four months in the University of Rochester, in the laboratories of Professor Benjamin Miller. She received her PhD degree in March 2002. Currently she is working for the Galquimia chemical company.

José L. Mascareñas was born in 1961 in Allariz (Spain). After completing his early education in Ourense he moved to the University of Santiago where he completed a BSc Degree in Chemistry in 1984. He completed his PhD in the Department of Organic Chemistry of this University in 1988 in the laboratory of Professors A. Mouriño and L. Castedo. He was a post-doctoral fellow of the Spanish Ministry of Education and Science at Stanford University from January 1989 to October 1990, working under the supervision of Professor Paul Wender. On returning to the University of Santiago he took up an assistant professor position from 1991 to 1993, when he become Permanent Professor. In 1992 and 1995 he made two four-month stays as visiting researcher in the Department of Chemistry of Harvard University, in the group of Professor Greg Verdine. His current research interests are split between a synthetic program focused at development of efficient cycloaddition and cyclization approaches to medium-sized carbocycles and a bioorganic program aimed at the design and synthesis of new DNA-binding peptides.

M. Eugenio Vázquez

Ana M. Caamaño

José L. Mascareñas

 DOI: 10.1039/b206274g

transcription stage, where the DNA is copied into a messenger RNA. The initiation of this process is highly dependent on the interactions of certain proteins, which are called *transcription factors*, with specific DNA sequences located at promoter or enhancer regions of the genes. The formation of these complexes orchestrates the assembly of the RNA Polymerase II machinery, which is ultimately responsible for triggering the expression of those particular genes.[1]

One of the more remarkable aspects of the above process is the tremendous DNA-binding selectivity exhibited by most transcription factors, as they are capable of selecting the correct binding sequence in the genome out of the vast number of potential alternative sites. Understanding the molecular and physical basis of this selectivity, as well as its implications on the control of gene expression, is a fundamental problem of modern chemical biology. In addition to biological methods like knockout organisms or expression profiles, the use of chemical approaches based on probing simplified versions of naturally occurring transcription factors can be of great value.

These miniature versions of the natural transcription factors may also have important future applications in gene-based medicine. The growing amount of information on the human genome, in conjunction with our increased understanding of molecular mechanisms of many major diseases, will unveil a lot of new genetic targets that can be exploited for the control of illnesses. Therefore it will be very important to obtain molecules that can be delivered to selective sites in the genome and effectively discriminate between closely related DNA sequences.[2]

Unfortunately the design of high-affinity DNA binding peptides consisting of simplified versions of naturally occurring transcription factors is not an easy task, and indeed the progress in this area has been relatively slow. However, in recent years, several transcription factor-based miniature proteins capable of tight interaction with specific DNA sites have been prepared. In this review we summarize the basic principles of transcription factor–DNA interactions and their use for the rational assembly of minimized sequence-specific DNA-binding peptides, focusing mainly on those constructed using synthetic methods. This review will not provide in-depth coverage of closely related areas of research dealing with DNA binding by small drugs or nucleic acids; those interested are referred to other interesting reviews.[3,4]

2 Strategies for selective DNA recognition

The classical approach to DNA targeting by artificial agents is based on the use of small molecules. There are a few small molecules capable of specific DNA interaction, however most of them have relatively low affinity and specificity and therefore their use as therapeutic agents in medicine is problematic owing to unavoidable secondary toxicities.[5]

DNA-binding molecules are usually classified according to the type of DNA-binding strategy (Fig. 1): *a) Intercalating agents*: are a large family of compounds with considerable diversity in terms of structure, ranging from simple aromatic heterocyclic systems such as acridines to oligopeptide bis-intercalators. Their mode of DNA recognition is based on intercalation between base pairs, a mechanism that does not allow high DNA affinities nor selectivities. *b) Alkylating agents*, molecules that upon recognition form covalent bonds with the DNA bases. *c) Minor-groove binders*, which include natural products such as Netropsin and Distamycin A, as well as a large number of artificial molecules like Berenil, Hoechst 33258 or Pentamidine. These compounds are characterized by a concave shape, that favors insertion in the narrow minor groove of sequences rich in A-T and T-A base pairs.[3,6]

In the last decade there has been a significant progress in the small molecule DNA-binding arena, particularly after the finding by Wemmer *et al.* that Distamycin A can interact with the minor groove of DNA as an antiparallel dimer.[7] On this basis, Dervan and coworkers have designed and synthesized a variety of synthetic *hairpin polyamides* made of *N*-methylpyrrole and *N*-methylimidazole units, which bind sequence specifically to the minor groove of DNA as side-by-side stacked antiparallel dimers (Fig. 2).[8] It has been shown that by using simple "pairing rules" it is possible to target in a predictable way the minor groove of a variety of sequences of 4 to 7 base pairs in length. Importantly, some of these compounds have been shown to be able to regulate gene expression at the transcriptional level.[9]

However, a major problem in using these molecules for interfering with the DNA-binding interaction of natural transcription factors, derives from the fact that the latter occurs mainly through the major groove of DNA, whereas oligoamides interact through the minor groove. It has also been difficult to extend the recognition capabilities to relatively long DNA sequences, which is very important for precise sequence targeting within a complete genome.

A good way of targeting specific DNA sequences through the major groove consists of using *triple-helix forming molecules*.[4,10] The classical examples are short oligonucleotides (10–20 bp) that bind to the major groove of oligopyrimidine·oligopurine sequences in double-stranded DNA by establishing Hoogsteen contacts with the oligopurine strand. The fact that the target sequence must contain consecutive purines on the same strand limits considerably the repertoire of potential target sites. Additionally, problems of chemical

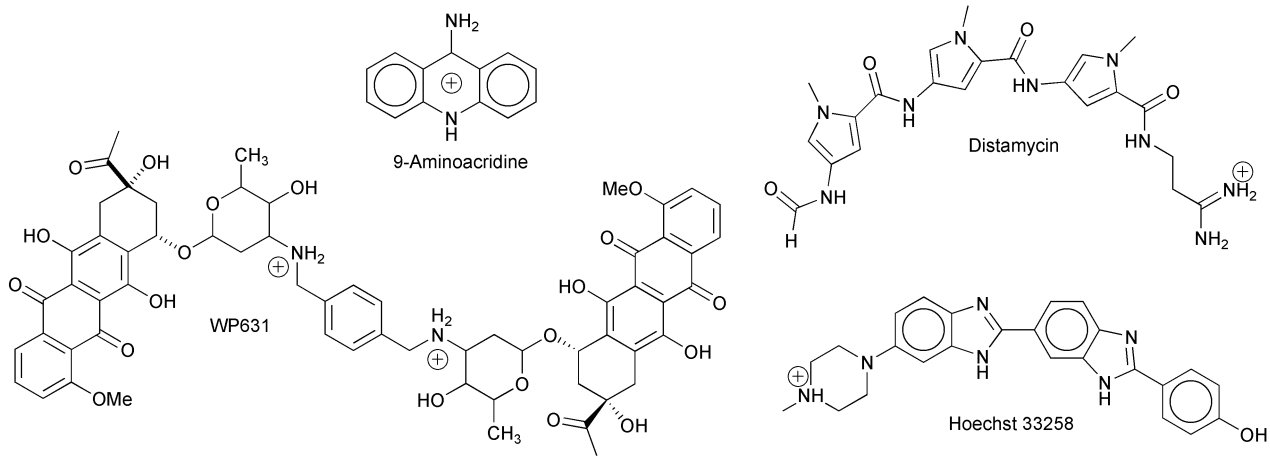

Fig. 1 Some minor groove and intercalating agents. Note the cationic character of Distamycin and Hoechst 33258 and the aromatic nature of the intercalating agents 9-Aminoacridine and WP631, this latter compound being a bisintercalating anthracycline antibiotic.

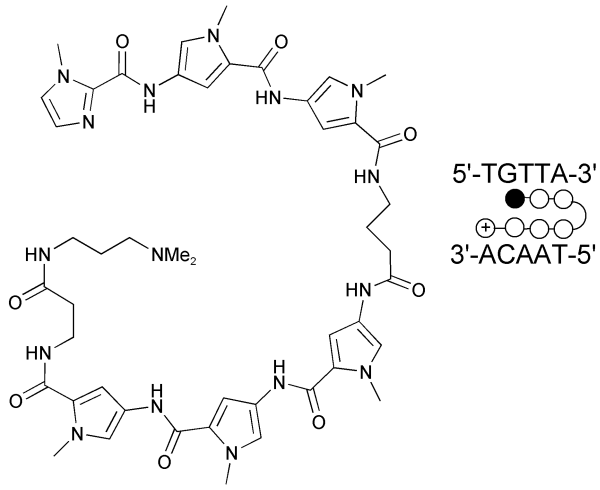

Fig. 2 A Dervan's hairpin polyamide which selectively recognizes a 5′-TGTTA-3′sequence and schematic drawing of the recognition process; Pyrrole (○) and imidazole (●).

instability and poor membrane permeability have made difficult their application as chemotherapeutic agents. These problems are being addressed by the development of other triple-helix forming molecules like PNAs, nucleic acid analogs made of a peptidic instead of a sugar backbone.

3 DNA recognition by transcription factors

In contrast to small natural or synthetic molecules, which interact with DNA mainly through the minor groove, naturally occurring transcription factors bind to specific DNA sequences by contacting primarily to the major groove. In order to facilitate the understanding of the basic characteristics of this recognition process we will first summarize the basic aspects of the double stranded DNA structure.

3.1 DNA structure and its implications on sequence recognition

The DNA forms a double helix in which both chains are associated by hydrogen bonds between complementary base pairs. The three dimensional structure of DNA is a consequence of the sugar geometry and the hydrophobic nature of the bases that tend to minimize their contact with water. The DNA double helix adopts different conformations depending on the conditions (pH, ionic strength, solvent *etc.*), but the most relevant conformation under physiological conditions is the B-form. This is the conformation that is recognized by most transcription factors, and it is therefore pertinent to make some observations about its structure:[11]

a) The B form of DNA consists of a right-handed double helix of polydeoxynucleotides with an approximate diameter of 20 Å (Figure 3). *b)* The bases are almost perpendicular to the helix axis and each base is bound with its complementary base on the opposite chain, forming a base pair (bp). *c)* There are about 10 bp per helix turn, the distance between consecutive bp (axial rise) is 3.4 Å, and the rotation per residue is 36°. *d)* The B form of DNA contains two grooves of different size, each one with very different geometric attributes. The *major groove* is wide and relatively shallow, whereas the *minor groove* is narrow. *e)* The width of the minor groove is much more variable in regions with consecutive A/T base pairs than in G/C rich tracts, but in general A-T rich regions are narrower. Although the structure of B-DNA is fairly regular and uniform, there are local variations in structure and flexibility depending on the base sequence.

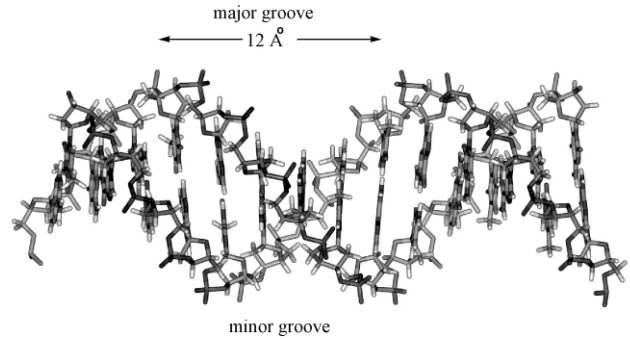

Fig. 3 B-form of double stranded DNA.

In most cases the binding of transcription factors to DNA does not affect the interaction between the DNA base pairs because the recognition process takes place through the functional groups exposed by the bases in the grooves, although it can cause local conformational alterations in the duplex. Inspection of an ideal B-DNA structure shows that the functional variability in the major groove is higher than in the minor groove. In fact, in the minor groove the base pairs A-T and T-A are degenerate in terms of hydrogen bond capabilities, while in the major groove all four possible base pair

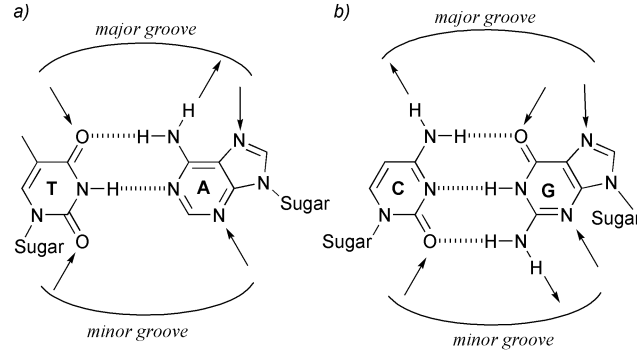

Fig. 4 Hydrogen bonding patterns of the exposed functional groups in (a) A-T, and (b) C-G base pairs. The arrows indicate the hydrogen bond donor or acceptor characteristics.

combinations present a different recognition pattern (Fig. 4). These facts, together with the small size of the minor groove, are consistent with the experimental observation that proteins prefer to interact with DNA through the major groove.

3.2 General thermodynamic and kinetic considerations of protein–DNA interactions

There are a number of reviews on the thermodynamics of protein–DNA complexes,[12,13] so here we only comment on fundamental aspects of the process that must be considered when designing new DNA-binding peptides. As in any other chemical process, the association of a transcription factor with its target DNA depends on the Gibbs free energy change ($\Delta G°_{bind}$) between the protein–DNA complex and the free protein and DNA states. $\Delta G°_{bind}$ determines whether the association is going to take place, and what will be the affinity of the protein for its target DNA, which is measured by the equilibrium association constant K_a ($\Delta G°_{bind} = -RT \ln K_a$).

Specific protein–DNA binding conveys the formation of a kinetically and thermodynamically stable molecular association with a well-defined geometry and stoichiometry (consequence of the number of available binding sites), whereas nonspecific binding is considered as a random association that can take place at any point along the DNA chain. Any protein–DNA interaction with an affinity in the range $K_a \approx 10^8 - 10^{11}\ M^{-1}$ is generally considered specific while weaker interactions, with

binding constants in the range 10^3–10^5 M^{-1}, are usually nonspecific.

In order to rationalize the different contributions to the free energy of binding, it is convenient to decompose ΔG_{bind} into its enthalpic ($\Delta H°_{bind}$) and entropic ($-T\Delta S°_{bind}$) terms: $\Delta G°_{bind} = \Delta H°_{bind} + (-T\Delta S°_{bind})$. In many cases enthalpic and entropic terms have opposing effects on the free energy of complexation, thus for some proteins a favorable enthalpy variation ($\Delta H°_{bind} < 0$) drives an unfavorable entropy change ($-T\Delta S°_{bind} > 0$) and in some other cases it is the favorable change in the entropic term ($-T\Delta S°_{bind} < 0$) that compensates for an unfavorable change in enthalpy ($\Delta H°_{bind} > 0$).[12] Hydrogen bonds, Van der Waals contacts or electrostatic interactions between the protein and the DNA usually contribute favorably to the enthalpic term, whereas desolvation of polar groups and deviation from ideal structural parameters in the complex are a source of unfavorable $\Delta H°_{bind}$. Entropy changes reflect the contribution of different processes associated with the complex formation, hence the release of water from non-polar surfaces and the redistribution of ions are favorable to complex formation, whereas loss of translational, rotational and vibrational degrees of freedom, as well as folding of peptidic chains during the formation of the protein–DNA complex are entropically disfavored processes.

From a kinetic point of view, a simple hypothesis for a general mechanism for specific DNA recognition by transcription factors involves two main steps (Fig. 5); in the first step

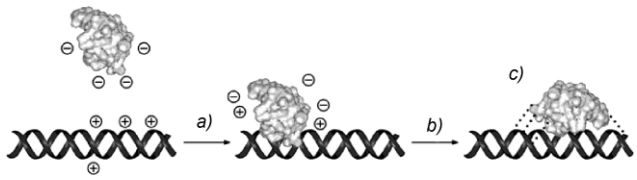

Fig. 5 (a) Non-specific binding to DNA and reorganization of the counterion atmospheres, (b) and (c) sliding of the transcription factor along the DNA chain and formation of specific interactions once the transcription factor finds its target sequence.

nonspecific long-range electrostatic interactions bring together the protein and the DNA. The formation of this initial complex is followed by sliding of the protein along the DNA in a one-dimensional diffusion process, that leads to an accelerated rate of target sequence location. Once the protein has found its target sequence, specific interactions are established and the high-affinity sequence-specific complex is thus formed, in many cases with conformational readjustments of the DNA and/or the protein.[14]

3.3 Structural motifs in sequence-specific protein–DNA interactions

Until recently, the underlying structural factors that determine the recognition process between specific DNA sequences and proteins (particularly transcription factors) were poorly understood, but during the last few years our knowledge has expanded tremendously owing to the exponential increase in the number of X-ray and NMR structures of protein–DNA complexes that have been solved.[15] As a consequence we now have a very good overall picture of the architecture of DNA-binding proteins and how they bind to DNA.

Although transcription factors recognize DNA using a variety of folds, in many cases they share similar structural recognition motifs, which facilitates a classification into families.[16,17] In many of these families the most relevant contacts to DNA occur in the major groove from amino acids of α-helical regions, known as *recognition helices*, embedded in those motifs.

HTH and homeodomain families. Most proteins of the HTH family belong to the prokaryotic kingdom and are characterized by the use of a conserved bihelical DNA-binding motif (helix-turn-helix, HTH), being quite dissimilar in structure outside this region.[18] This motif, composed of approximately 22 amino acids, consists of two α-helices connected by a tight bend. The second helix, referred to as the recognition helix, inserts in the major groove of DNA and makes several contacts with the bases and the phosphate backbone (Fig. 6a).

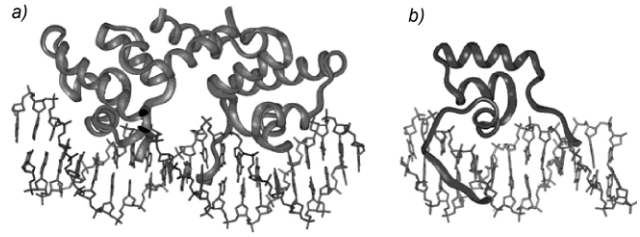

Fig. 6 (a) A view of the structure of the DNA complex of the λ repressor HTH protein. (b) The HTH recognition region of Hin recombinase. Note the arm inserted into the adjacent minor groove

The first helix is not embedded in the groove but in some cases also makes additional contacts to the phosphate backbone. Although the angle between the two helices is fairly conserved, there are small variations in the orientation of the recognition helix in the groove. It should be remarked that isolated HTH motifs are not capable of DNA recognition; this process requires the whole protein and in most cases its homo- or heterodimerization.

Homeodomain transcription factors are considered to be the eukaryotic equivalent of HTH proteins. An important difference between bacterial HTH and eukaryotic homeodomain proteins is that the latter can bind to the target DNA sequences as monomers.[19] To some extent this is possible because in addition to the major groove recognition *via* the HTH motif they establish accessory interactions in flanking positions of the minor groove by means or C- or N-terminal arms (Fig. 6b).

The zinc finger family. Zinc fingers are among the most widespread DNA recognition motifs used by regulatory DNA-binding proteins.[20] The DNA binding domain of the more general class of these proteins (Cys_2-His_2) is about 30 amino acids long and contains the sequence C–X_{4-5}–C–X_{12}–H–X_{3-5}–H (C = cystein, H = histidine, X = any other amino acid). Upon coordination of Zn^{2+} to the two Cys and His residues, the motif folds into a compact unit consisting of an α-helix packed against a β-hairpin (ββα-domain). Sequence-specific DNA recognition is achieved by presentation of the α-helix into the major groove of the double helix, where it comes into contact with a 3–4 base pair-long site, but it must be noted that the recognition process requires several modules joined by short linkers (Fig. 7c). In addition to the above zinc-finger group, there are other families of Zn^{2+}-containing transcription factors, the most prominent members being those belonging to the nuclear hormone receptor family, which bind DNA as non-covalent dimers.

BZIP and bHLH families. The basic region-leucine zipper (bZIP) motif is employed for DNA recognition by a wide number of eukaryotic transcription factors involved in the control of cellular growth.[21] From a structural point of view it is probably among the simplest of all DNA binding motives as it consists of dimers of uninterrupted α-helices of about 60 residues. There are two different subdomains in each helix, a C-terminal leucine-rich area (LZIP, Fig. 8a), which mediates the dimerization through a parallel coiled-coil, and the *basic region* (BR), a domain of about 20 amino acids located at the N-

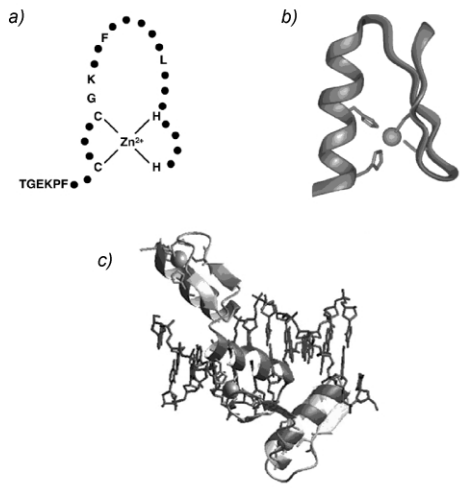

Fig. 7 (a) Schematic representation of the structure of a Cys_2-His_2 zinc finger monomeric recognition unit showing conserved residues. (b) Ribbon diagram of the same unit. (c) Structure of the Zif268 transcription factor–DNA complex. The zinc atoms are shown as spheres.

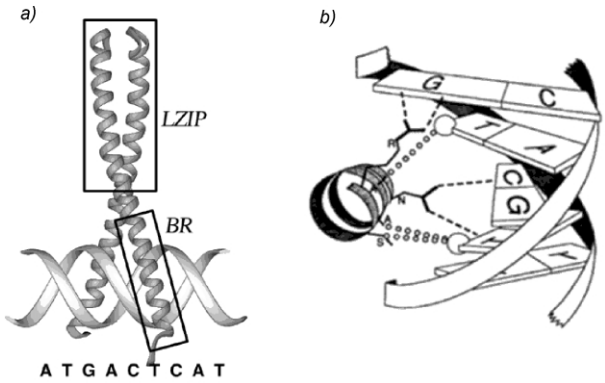

Fig. 8 X-ray structure of the GCN4 DNA-binding domain bound to the AP-1 site showing the C-terminal leucine zipper (*LZIP*) and the basic region (*BR*). (b) Some of the specific DNA contacts made by the basic region with the DNA major groove (N235, A238, A239, S242, R243).

terminus of the leucine region. This region, which is rich in basic amino acids, is inserted in the major groove of DNA and therefore is responsible for most of the direct contacts to the DNA bases and phosphates. The leucine zipper and basic regions are connected through a spacer that is 6 amino acids in length.

A very important DNA-binding signature of the bZIP family is seen in solution, in the absence of their target DNA sequence, where the basic region is poorly structured and only adopts the characteristic α-helical conformation upon specific DNA binding. From a thermodynamic point of view, the interaction of bZIP proteins with DNA has a strongly unfavorable entropic term ($-T\Delta S$) arising from the loss of degrees of freedom associated with the folding of the random coil basic region into the α-helix. This high entropic cost determines that monomers or isolated basic regions cannot bind by themselves to their cognate DNA sequences with sufficient affinity. It is the enthalpy gain from the simultaneous interaction of two chains as homo- or heterodimers that provides the energy to compensate for the unfavorable entropic term, thereby permitting complexation with the target DNA.

Although it could be thought that dimerization precedes DNA binding, it has recently been demonstrated that the preferred kinetic pathway for DNA recognition consists of an initial low affinity interaction of a monomer followed by dimerization on the DNA (Fig. 9). This pathway ensures a rapid assembly of the dimer into the cognate recognition site and avoids kinetic trapping at non-specific sequences.[22]

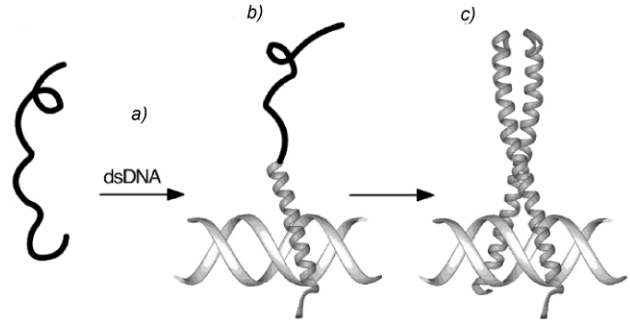

Fig. 9 (a) DNA-binding is coupled to folding of the basic region. (b) A bZIP monomer binds to DNA with low affinity, but the initial complex recruits the second peptidic chain through the leucine zipper to form the final high affinity complex (c).

The basic region-helix-loop-helix (bHLH) proteins share with the bZIP proteins a similar mode of DNA binding with the only salient difference lying in the dimerization region, which is composed of two helices separated by a loop. Both bZIP and bHLH proteins have many members that can form both homo- or heterodimers, a feature that expands the repertory of DNA sequences that the proteins can recognize. The relative simplicity of the DNA recognition mode of these transcription factors has led to their use as main reference framework for the design of new DNA-binding peptides (see later in the article).

Other DNA-binding motifs. While many transcription factors can be classified into the general categories presented above, some others combine features of different families. This is the case for *Skn-1*, a transcription factor that shares elements of the bZIP and homeodomain families; the DNA-binding helix is homologous to the bZIP basic region and indeed recognizes the same DNA half-site as GCN4 but, unlike bZIP proteins, Skn-1 binds as a monomer and it does not have the leucine zipper region. In order to stabilize the interaction, Skn-1 makes use of a homeodomain-like structure that even has an extended arm that makes contacts with the DNA minor groove (Fig. 10a).[23]

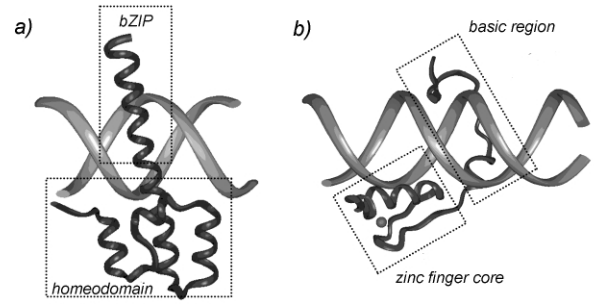

Fig. 10 (a) Structure of the Skn-1 transcription factor binding domain complexed to DNA; the regions that are structurally homologous to those of other families of transcription factors are remarked, (b) Structure of the GAGA factor–DNA complex.

Another interesting example is the *GAGA factor*, a 519 residue-long transcription factor containing a single Cys_2-His_2 zinc finger module which exhibits high DNA affinity. This affinity arises because in addition to this unit, which binds a typical GAG triad, there are a number of additional contacts made by two N-terminal highly basic segments termed BR1 and BR2. BR2 forms a helix that interacts in the major groove while BR1 wraps around the DNA in the minor groove (Fig. 10b).[24] These examples illustrate how nature combines multiple and/or mixed recognition motifs to achieve the desired DNA recognition.

In addition to the above motifs, there are transcription factors and other DNA-binding proteins that do not rely on α-helices

for specific DNA recognition, although they are less common. For instance, the ribbon-helix-helix proteins, exemplified by the *MetJ* and *arc repressor*, form dimers that insert antiparallel β-sheets into the major groove of DNA with the side chains on the face of the β-sheet contacting the base pairs. Usually these proteins bind cooperatively to two or more adjacent DNA binding sites. Biologically very important transcription factors, such as *P-53*, *NF-kB* or *NFAT* proteins bind DNA using immunoglobulin-like folds. While a general feature of these proteins is DNA recognition through loops, there is a great deal of variation in the ways they make base contacts.

3.4 Common features of DNA recognition by transcription factors

As can be deduced from the above discussion, transcription factors use a large variety of architectural motifs for achieving DNA recognition, being thereby extremely difficult to extract general rules to explain the selectivity of the binding process. Many groups have already remarked that there is no simple code that links specific secondary peptide structures and amino acids with specific sequences.[25] Even in the case of zinc fingers, where all the members of the family that have different DNA recognition sequences share the same structure, it has been extremely difficult to decipher a general relationship between amino acid residues in the recognition helix and the corresponding DNA target.

Yet, despite the lack of simple rules governing sequence recognition, it is possible to deduce some general principles, which can be of great help when considering the design of transcription factor-based DNA-binding peptides. Proteins recognize a particular DNA sequence by having a surface that is chemically complementary to the exposed functional groups of the base pairs, but in most protein–DNA complexes there are also a large number of contacts with the deoxyribose-phosphate backbone. Usually, the crucial interactions take place on the major groove of the DNA, as it is there where each base pair can be uniquely distinguished, and in most of the cases, the specific contacts are made by side chains of an α-helix, called recognition helix, which inserts in the groove.

There are basically four major types of direct interactions between proteins and nucleic acids:[26] *a) Salt bridges and hydrogen bonds* between the DNA phosphodiester backbone and amino acid residues with basic side chains (Lys, Arg or His). These contacts do not usually confer specificity to the binding but increase the thermodynamic stability of the complex and help to anchor the recognition domain in a correct orientation. *b) Hydrogen bonds* between the sugars or bases in the DNA and polar side chains in the proteins, which are critical interactions from the specificity point of view. Analysis of the available X-ray structures of protein–DNA complexes show that Arg, Lys, Ser and Thr are the most common amino acids which participate in this type of hydrogen bonding. Curiously, acidic residues such as Asp or Glu are scarcely used, probably because of their unfavorable electrostatic interaction with the DNA backbone. Especially important are bidentate hydrogen bonds formed by a single side-chain with a base or base pair, as these provide an inexpensive way of increasing the bond energy per amino acid–base pair while conferring a higher degree of specificity for a given sequence (Fig. 11). *c) Non-polar contacts* between the DNA base pairs and non-polar amino acid-side chains. Although, because of their lack of directionality requirements these are thought to play a smaller role in specificity than hydrogen bonding, they are now recognized as an important component in protein–DNA binding. For instance, hydrophobic interactions between protein side chains and the methyl group of thymine have been observed to play a key role in sequence specificity in a number of cases. *d*) *Water-mediated hydrogen bonds* are relatively common. Most of these are

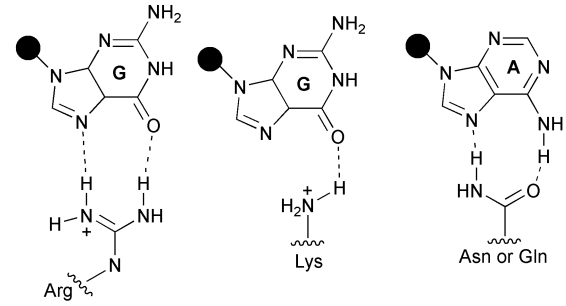

Fig. 11 Schematic diagrams showing some commonly observed hydrogen bonding interactions between bases and amino acid side chains.

established between polar or charged amino acids such as Arg, Lys, Asn, Gln, and even negatively charged residues Glu and Asp, with the DNA backbone. It is believed that most of the water-mediated contacts function as space fillers.

In several families of transcription factors there are important direct contacts between amino acid residues and the edges of the bases in the minor groove – interactions that are not essential for sequence selectivity but provide an extra binding energy. These interactions are important for attaining tight affinities.

What about the function of the rest of the protein which does not participate in the direct contacts with the DNA? It is considered that in most cases it serves as a structural scaffold to preorganize, stabilize and deliver the recognition elements, in particular the recognition α-helix, in an appropriate orientation, although it can also establish additional contacts to the phosphate backbone. In many cases, these DNA non-contacting protein regions mediate homo or hetero oligomerizations with other transcription factors or with other elements of the transcriptional machinery. The ability to multimerize is very important not only by allowing tight DNA binding from proteins that, on their own, show low affinity, but also because it permits the recognition of long sequences, which warrants site-selective binding within genomes as large as the human one. Furthermore, this ability to multimerize with diverse partners drastically expands the possibilities for recognizing diverse sets of DNA sequences (combinatorial gene regulation) from a few protein partners.[27]

Another important aspect that influences the specificity of protein–DNA interactions is derived from different local DNA propensities of certain sequences to adopt unusual or distorted conformations (*indirect readout*). The sequence-dependent deformability of duplex DNA provides site-selectivity by virtue of the predisposition of some nucleic-acid sequences to adopt a particular structure required for binding to a protein at a lower free energy cost than other sequences.

4 Design of DNA-binding proteins

Given that the solutions to the problem of how an organism evolves a protein to recognize specific DNA sequences are many and varied, the *de novo* design of non-natural proteins, particularly miniature derivatives, capable of reproducing the DNA recognition properties of transcription factors is not trivial. Although in this review we will focus on the approaches used for designing relatively small DNA-binding peptides (next section), the reader should be aware that several chimeric DNA-binding proteins constructed by combining different naturally occurring recognition motifs have been successfully prepared.

Unquestionably, greater progress has been made in the area of zinc finger proteins, particularly in the Cys_2-His_2 class. The modularity of both structure and function of this recognition framework has offered excellent opportunities for reprogramming the site selectivity of designed analogs.[28,29,30] Several stitched three-finger proteins derived from Zif268 have been constructed and shown to bind designated 9 base pair sites with

subnanomolar affinities. However it is very difficult to predict whether the designed mutants will have the expected DNA selectivity as in many cases the mutations affect the positioning and orientation of the ββα framework.

Computer modelling has also been successfully applied to obtain structure-based chimeric DNA-binding proteins by combining DNA-binding domains of different transcription factors. For instance, Pabo *et al.* have used this methodology to obtain a fused conjugate of a zinc finger (*ZIF268*) and a homedomain (*Oct-1*) using the crystal structures of their DNA complexes as the starting point for the designing process.[31]

5 Design and synthesis of DNA-binding peptides

The construction of artificial proteins with non-natural DNA-binding specificities is of high interest, however, an even greater chemical challenge consists of using the underlying principles of DNA recognition by transcription factors to design minimized peptides that maintain the DNA affinity and specificity characteristics of the natural counterparts.[32] It would be even better if these molecules could be approached using the tools of organic synthesis because this might allow to introduce non-natural elements into the peptidic framework.

Obtaining tailored DNA-binding molecules would be important not only from a fundamental point of view, by allowing us to test our knowledge about the molecular recognition principles involved in the formation of specific protein–DNA complexes, but also from a practical perspective as potential biomedical gene targeting agents.

Ideally, when trying to design a DNA-binding peptide one should not only consider the structural references of the DNA-transcription factor complexes but also the thermodynamic factors influencing the DNA recognition event. In practice, however, the experimental and even theoretical determination of many energetic contributions (*i.e.* redistribution of ions, desolvation of polar groups *etc*.) is very difficult, and therefore different approaches will require trial and error tests of the various designs based on structural data.

A key initial issue that must be taken into account before embarking on the preparation of DNA-binding peptides deals with the downsizing limits of natural transcription factors. How far can one truncate a DNA-binding transcription factor without losing most of the DNA-binding. As might be anticipated, removal of C- or N-terminal residues of naturally occurring transcription factors reduces their recognition capability drastically. Thus, truncation of *λ-Cro*, a HTH protein which binds DNA as a non-covalent dimer, to a monomeric peptide containing only the helix-turn-helix motif or to a peptide containing just the recognition helix, resulted in complete loss of sequence-specific binding.[33] In the case of monomeric DNA-binding homeodomain transcription factors, such as *antennapedia homeodomain*, a 60 amino acid synthetic peptide that retains a stabilized HTH motif and the minor groove-binding arm has similar activity, whereas truncation of the homeodomain to a peptide containing only the HTH motif drastically reduces the DNA affinity.[29] Isolated zinc finger modules of the Cis_2-His_2 class of transcription factors are unable to bind to their cognate sequence with enough affinity, so at least two modules are needed to obtain affinities in the nanomolar range.

It has been shown that a 61 residue minimized version of the GAGA transcription factor previously discussed, containing only the zinc finger recognition unit and the N-terminal basic regions is able to achieve the DNA recognition with a dissociation constant of 5 nM.[34] However, removal of 27 residues from the basic region leads to suppression of the DNA affinity. Monomers of bZIP proteins or their isolated basic regions cannot bind to their cognate DNA sequence with high affinity. The enthalpic gain of the interaction does not compensate for the considerable loss of entropy of the process, as the binding of these regions is accompanied by folding to an α-helix.

Therefore, it seems clear that truncation of DNA-binding domains of naturally occurring transcription factors in general leads to suppression of their DNA affinity. Restoring the DNA affinity of the isolated DNA-reading sequences of these proteins requires the implementation of innovative chemical strategies, some of which will be discussed below. As will be shown, most of the work carried out to obtain functionally active, miniature versions of transcription factors has been based on the mode of recognition of the bZIP family of proteins.[35]

5.1 Artificial dimerization of bZIP basic domains

Since the early discovery of the structural basis of DNA recognition by HTH, bZIP or zinc finger proteins, it was recognized that isolated recognition helices – when presented as monomeric reading heads – are unable to achieve the binding process, with the rest of the protein being necessary to obtain the required affinities. In 1990 the group of Peter Kim, in ground-breaking work, demonstrated that removal of the leucine zipper region of the bZIP protein GCN4, and dimerization of the remaining basic domains by means of a covalent disulfide bond produces peptides capable of binding to the cognate sequence of the natural protein with nanomolar affinities (Fig. 12).[36]

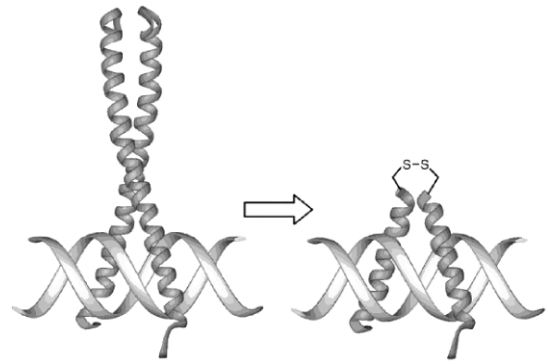

Fig. 12 Schematic diagram for the formation of a specific peptide–DNA complex by substitution of the leucine zipper dimerization motif by a disulfur bridge.

However, in contrast to the natural transcription factor that binds to its target DNA at room temperature with high affinity, the disulfur dimers only bind at low temperatures (4 °C). This seems to point out that the leucine zipper is performing an additional role to being a mere dimerization element.

The group of Goddard III demonstrated that this sulfur–sulfur dimerization strategy can be extended to make heterodimers or even trimers which recognize the predicted composite DNA sequences.[37] For instance the group synthesized a C to N conjugate by coupling appropriately modified basic regions of v-Jun bZIP proteins. The resulting dimer is capable of specific recognition of sites rearranged with respect to those targeted by the natural protein.

After Kim's discovery, several groups demonstrated that the basic regions can be dimerized using other type of covalent or non-covalent connectors. Among the most interesting examples are those reported by the group of Morii, in which both basic regions are dimerized through a non-covalent adamantane–cyclodextrin inclusion complex.[38] The required hydrocarbon molecules can be readily introduced at the C or N-terminal position of fully deprotected peptides by taking advantage of the nucleophilic window provided by a cystein sulfur. Reaction of the C-terminal cysteine with mono-6-deoxy-6-iodo-8-cyclodextrin or *N*-(bromoacetyl)-1-adamantanemethyl-amine in a

slightly basic aqueous buffer yields each of the required peptides (Fig. 13). As in the case of the disulfur dimer, each of the basic regions by itself is not capable of high-affinity DNA recognition, whereas the complex between the β-cyclodextrin and its guest compound efficiently generates a dimer that specifically binds DNA with almost native affinity, although low temperatures (4 °C) are required.

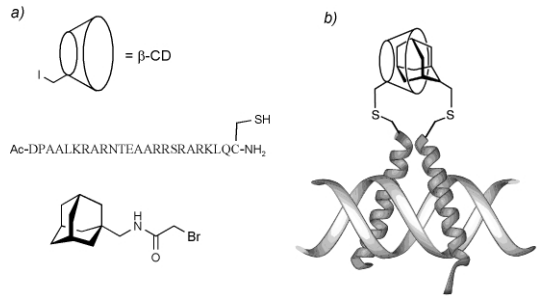

Fig. 13 Formation of non-covalent cyclodextrin–adamantane complexes for dimerization of the GCN4 basic region and specific DNA recognition. (a) Coupling partners. (b) Model of DNA binding by the synthetic dimer.

This strategy has been used by the same group to control the formation of heterodimers of basic regions corresponding to two different members of the bZIP family of transcription factors, *GCN4* and the enhancer binding protein (*C/EBP*). These DNA-binding proteins recognize palindromic sequences with half-sites of 5′-ATGAC-3′ and 5′-ATTGC-3′ respectively, and thus an adamantyl-cyclodextrin heterodimer of the basic regions of both proteins interacts with high affinity with the composite nonpalindromic DNA sequence 5′-ATGACG-CAAT-3′.[39]

Another particularly interesting example of the application of artificial dimerization strategies to obtain DNA-binding peptides is derived from the work of the group of A. Schepartz. This group coupled several activated terpyridine units with a cysteine residue placed at the C-terminus of the *GCN4* basic regions. Dimerization of the resulting hybrids by addition of Fe(II) provided several iron complexes with slightly different geometries (Figure 14). One of the three peptide dimers tested was capable of recognizing the expected DNA target site (CRE: 5′-ATGAcgTCAT-3′) with high affinity, and also showed surprising selectivity, as it did not bind the closely related sequence AP1 (5′-ATGAcTCAT-3′).[40]

In a variation of the above dimerization strategy, the group of Mascareñas has shown that the introduction of rigid photoresponsive azobenzene groups as dimerization units allows control of the DNA binding affinity of the resulting dimers. The covalent dimer was synthesized by coupling a C-terminal cysteine with appropriate azo-bromoacetyl derivatives (Fig. 15). While the *cis* derivative binds at low nanomolar affinity, and even better than the homologous sulfur–sulfur dimer, the

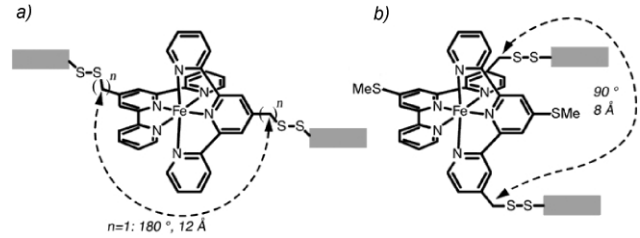

Fig. 14 Schematic representation of the Fe(II) complexes used as dimerization domains showing the different orientations of the attachment points. (a) 180° complexes, $n = 1$ gives the more efficient binder. (b) 90° inactive complex. The grey rectangle represents the basic region of the bZIP protein.

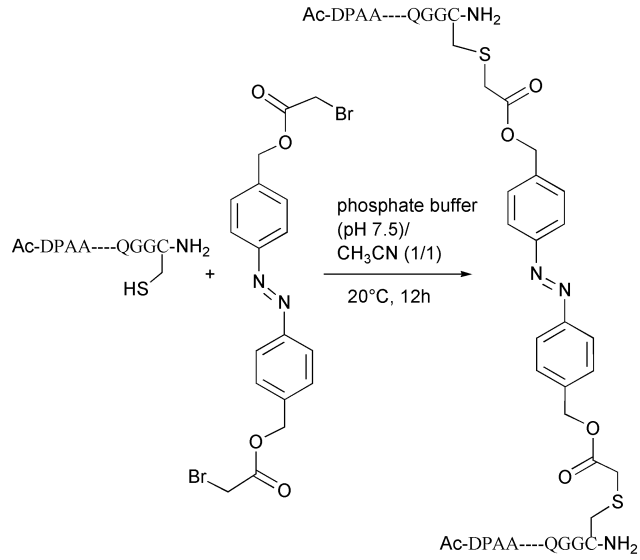

Fig. 15 Synthesis of photomodulable DNA-binding peptides. Coupling of the free peptide with trans-azo bromoacetyl derivative to form the trans-peptide dimer.

trans isomer binds with an approximately 60-fold decrease in affinity to the same sequence (Fig. 16).[41]

This concept of externally modulated DNA-binding peptides might find important applications both in biology and medicine as their activity could be regulated in time and space. Indeed, the activity of many transcription factors is also controlled by external cellular signals, and the DNA affinity of some of them is even controlled through conformational changes.

5.2 Conjugation of small molecules to "DNA-reading" modules

As commented above, isolated monomeric DNA-contacting motifs of most DNA-binding proteins are incapable of tight,

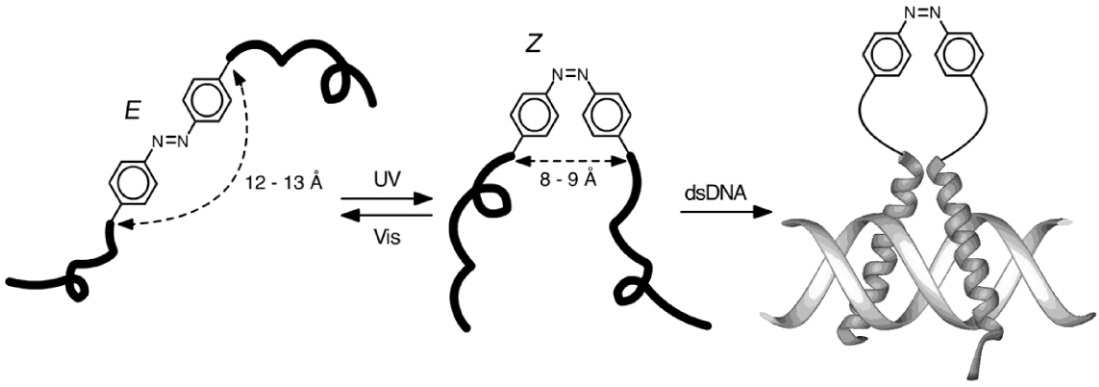

Fig. 16 Upon irradiation, the *E*-azobenzene undergoes a conformational switch to the *Z* isomer, a change that shortens the distance between the two basic regions allowing for high affinity binding.

specific DNA-binding. Recently, several groups have investigated whether combining these isolated elements with other DNA binders could provide conjugates that exhibit tighter affinity and selectivity. Thus, a 52 amino acid-long helix-turn-helix (HTH) portion of a naturally occurring DNA-binding protein (Hin recombinase) was linked to an intercalating cyanine dye, and the DNA-binding properties of the hybrid were investigated by fluorescence spectroscopy (the fluorescence of the dye increases upon intercalation). The synthesis of the labeled and unlabeled hybrids was carried out using standard Fmoc solid-phase peptide synthesis methods. The labeled peptides were obtained by coupling a carboxylic acid derivative of the cyanine dye to a selectively deprotected lysine.[42]

Thermodynamic analysis of the interaction indicated that the conjugate binds DNA almost 100 times better than the untethered peptide, and with a specificity similar to that of the native protein. Apparently the intercalating dye is playing the role of augmenting the nonspecific binding affinity. It is interesting to note that the presence of the cyanine dye not only provides increased affinity and favorable fluorescence properties, but can also be used as a photocleaving agent. This dye-tethering strategy has also been successfully applied to increase the affinity of a 29 amino acid-long zinc finger moiety of the native *glucocorticoid receptor protein* (GR).[43] In this case the hybrid interacts with the native glucocorticoid response element with a dissociation constant of roughly 25 nM.

A fairly similar strategy to increase the otherwise poor DNA-binding affinity of isolated major groove contacting α-helices has been investigated by Barton and coworkers. The authors conjugated the recognition α-helix of the phage 434 repressor protein (HTH family) to the synthetic metallic complexes $[Rh(phi)_2(bpy)]^{3+}$ and $[Rh(phi)_2(phen)]^{3+}$, which intercalate into the DNA major groove and therefore anchor the recognition helix directly into the major groove of the DNA. The peptidic metal complexes were synthesized by coordination of a precoupled N-terminal phenanthroline to $[Rh(phi)_2(DMF)_2]^{3+}$ (Fig. 17). The metal–peptide complexes were shown to be stable to the standard Fmoc-SPPS peptide cleavage/deprotection conditions.[44]

The resulting hybrid molecule recognizes preferentially the sequences ACAA and ACGA with affinities in the range of 50 nM. Although the hybrid constructs show a higher affinity than either of the components by themselves, the site selectivity is modest, and the authors conclude that future designs based on this type of strategy will require the preorganization of the peptide secondary structure to be maximized.

Although the above bivalence-binding approach somewhat follows the polyvalence strategy used by nature for increasing the affinity of bimolecular recognition processes including transcription factor–DNA interactions, a better imitation would be accomplished if both components of the bivalent conjugate were sequence specific in their own right. A successful realization of this idea was recently reported by the group of Mascareñas, who showed that structure-based, rational conjugation of a minor-groove binding Distamycin analog with the basic region of the bZIP transcription factor GCN4, generates a construct with remarkable DNA-binding properties.[45] The synthesis of these conjugates was carried out by coupling appropriately elaborated tripyrroles with a chemoselectively deprotected glutamic acid side chain of the peptide, while the latter is still attached to the solid-phase resin (Fig. 18).

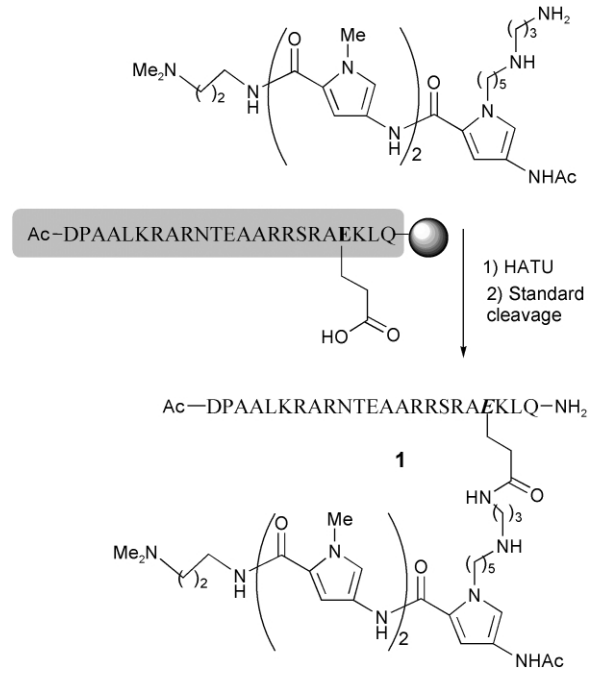

Fig. 18 Synthesis of tripyrrole–peptide conjugate **1**.

CD spectroscopy and gel-shift electrophoresis studies indicate that the hybrid **1** exhibits low nanomolar affinity for the composite site 5′-TCATAAAA-3′, an affinity considerably better than that of any of its components for their respective subsites (Fig. 19).

The bipartite major/minor groove binding mode of this designed peptide somewhat mimics the recognition strategy used by several natural transcription factors, such as homeodomains, with an α-helix inserted into the major groove, and another recognition element (usually the N or C terminal end of the protein) inserted into the minor groove of adjacent sequences.

Arriving at a successful design was not straightforward and required several modifications of the initial design. Hence a linear conjugate **2**, which incorporates a 5-aminovaleric-ornithine-5-aminovaleric linker between the tripyrrole unit and

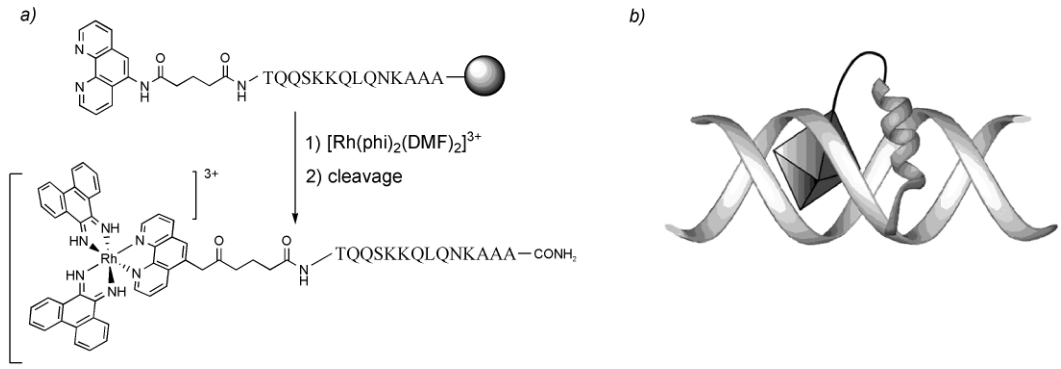

Fig. 17 (a) Synthesis of peptide–$[Rh(phi)_2(DMF)_2]^{3+}$ chimeras and (b) hypothetical model of their binding to DNA through simultaneous interaction of the peptide and the metal complex (represented by an octahedron).

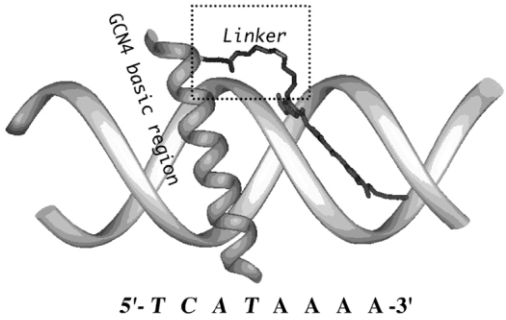

Fig. 19 (a) Hypothetical model of the specific interaction between the GCN4 basic region–tripyrrole hybrid **1** and DNA. The recognition sequence is the result of the basic region half-site DNA recognition sequence (TCAT) plus the sequence preference of the Distamycin unit (AAAA).

the peptide C-terminus (Fig. 20),[46] failed to bind the desired sequence with the required specificity. On the other hand, a derivative similar to **1** but bearing a linker shorter by two methylenes (**3**, Fig. 20) also failed to bind specifically. These

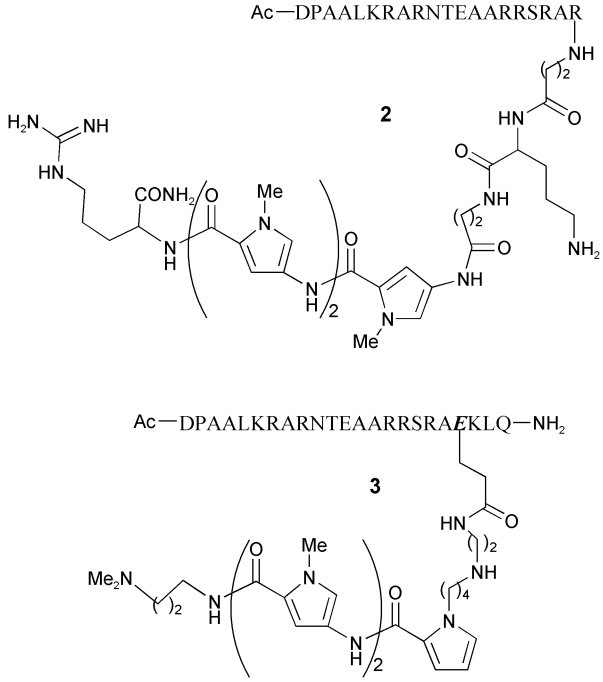

Fig. 20 Tripyrrole–peptide conjugates that failed to bind to the designated DNA sites.

failures illustrate the difficulty in choosing an appropriate linker that does not introduce strain in the docking of the DNA recognition portions into their respective sites or pay by itself a high entropic and/or enthalpic penalty upon binding.

5.3 Thermodynamic stabilization of the DNA-binding motif

Since major groove-binding of the peptidic regions in the native bZIP transcription factors is accompanied by a folding transition from a highly unstructured peptide to an α-helix, it is intriguing to know whether preorganization of the helical secondary structure might elicit DNA-binding of isolated monomeric regions by hypothetically decreasing the entropic cost of the binding process.

Several strategies for stabilizing α-helices have been described, but most of them require the presence of particular amino acids at specific positions of the peptidic chain. Therefore one must be very cautious when applying these strategies to naturally occurring DNA-binding fragments because the required mutations or additions could affect key contacts for specific recognition of the DNA sequence.

The design approaches based on this strategy can be grouped in two types: (a) stabilization of the α-helical secondary structure, and (b) incorporation of key residues required for DNA binding on well-established scaffolds containing preorganized α-helices.

a) Stabilization of the α-helix conformation. One of the most commonly used strategies to increase the helical content of a given peptide is the introduction of conformational constraints such as lactam bridges between appropriate residues. Taylor *et al*. have synthesized a constrained GCN4 basic region analog incorporating two Lys^{i}–Asp^{i+4} side chain lactam bridges.[47] Circular dichroism studies revealed that introduction of these two macrocyclic bridges generates a greater helical content in the resulting peptide than in the natural domain, also inducing higher resistance to thermal denaturation, although the authors did not report DNA binding results.

Later, the same authors reported the synthesis of an N-terminal fluorescence-labeled peptide containing 25 amino acids of the basic region of GCN4 and four additional C-terminal amino acids, three Ala and one Asp. The aspartic acid residue was intramolecularly bridged through a lactam bond to a Lys four residues away towards the N-terminus (Fig. 21).[48]

Fig. 21 Schematic representation of the peptide based on GCN4 basic region showing the modifications introduced: Lactam bridge at the C-terminus, and fluorescent moiety at the N-terminus.

The authors found that there is an increase in the affinity of the modified peptide compared to that of the natural basic region (K_D values of 3.9 ± 0.5 μM for natural basic region and 0.65 ± 0.09 μM for the modified peptide, which corresponds roughly to stabilization of the complex by 1.1 kcal mol^{-1}). This affinity increase correlates with the greater helical content of the lactam-bridged peptide (about 64%) over the unmodified basic region peptide (over 41%).

On the basis of these results it seems reasonable to anticipate that a further increase of the helical propensity, by introducing other constrains or by enriching the alanine content of the peptide could lead to tighter DNA binders. An interesting study by Shin *et al.* on the influence of replacing multiple residues of the basic region of GCN4 by alanines reveals that the strategy is very effective to increase the α-helical content of the unbound domain.[49] These studies also showed that the loss of important polar and non-polar interactions as a result of the mutation of relevant amino acids is compensated by the decrease in entropic cost of the binding process as a consequence of the higher preorganization.

b) Residue grafting strategies. Schepartz *et al*. have studied an alternative strategy for stabilizing the α-helical monomeric basic region of GCN4. The approach consists of mutating solvent-exposed residues of certain protein scaffolds that contain stabilized α-helical folds by other residues needed for specific DNA recognition. In particular, they made use of a small folded structure, the avian pancreatic polypeptide (aPP), which consists of a short α-helix stabilized through hydrophobic interactions with a type II polyproline helix (Figure 22).

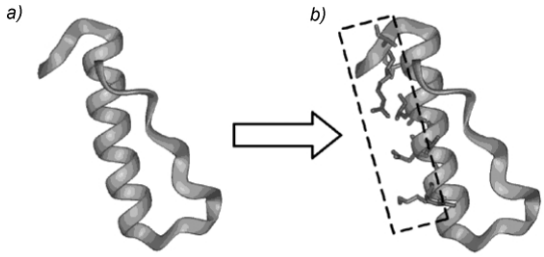

Fig. 22 Residue grafting using the aPP scaffold and introducing the DNA-interacting residues of GCN4 basic region. (a) Structure of the aPP fold (b) the mutated residues responsible for DNA recognition are located at one side of the α-helix of the aPP fold (boxed area).

Since the aPP structure is maintained only by the residues packed in the interior of the α-helix against the polyproline, it was possible the substitution of key residues in the exposed face with those considered critical in the interaction of the transcription factor GCN4 with DNA.

Using this strategy they were able to construct a 42 amino acid peptide that exhibited extremely tight DNA affinity and specificity: association constants of 1.5 nM at 4 °C.[50] Remarkably, the use of evolution techniques to optimize the N-terminal sequences that facilitate peptide folding increased the affinity constant to 0.5 nM at room temperature, therefore exhibiting the same range of affinities as natural transcription factors but without the need to dimerize.

In a related approach, Makino *et al.* used a small compact domain of the F-actin bundling protein villin as a tertiary structure scaffold to graft the residues used by GCN4 to recognize a specific DNA sequence.[51] The presence of the small folded domain seems to increase the thermal stability of the complex between the constructed protein and its target DNA compared with natural GCN4–DNA complex, however the resulting monomeric conjugates were unable to achieve the specific DNA binding with sufficient affinity.

The residue grafting strategy has very interesting possibilities for designing a variety of peptides capable of recognizing distinct DNA sequences, but much remains to be done until this approach can be routinely applied to the synthesis of DNA binding peptides. It is still very difficult to predict the possibilities for success of a particular design and to identify the reasons why certain structure-based designed peptides do not exhibit the expected DNA-binding capabilities.

6 Conclusion and future prospects

As we have seen, several relatively small peptides showing interesting DNA-binding properties have been successfully designed using as reference the DNA-binding mode of natural transcription factors. However, the progress in this area has been relatively slow compared to the developments in the structural elucidation of protein–DNA complexes. Despite the availability of a large body of structural data, we are a long way from fully understanding the molecular and biophysical basis underlying the selective interactions of transcription factors and other proteins with DNA, and subsequently from using those fundamentals for the *de novo* design of DNA-binding peptides with tailored specificities.

Future progress in the area may further combine rational designs with combinatorial-selection methods, so that a larger number of hits can be obtained. The discovery of new small stable folded structures, which can be used as scaffolds to introduce DNA-binding functionalities, can be of great help for engineering new functional derivatives. In order to obtain systems that exhibit higher specificities it would be interesting to design peptides which are not excellent binders by themselves, but interact with DNA efficiently after non-covalent homo- or heterodimerization, similarly to what occurs with naturally occurring transcription factors. The development of peptides equipped with appropriate sensitive functionalities so that their activity can be externally controllable or measurable (*i.e.* by the introduction of fluorescent probes in the sequence) will also be relevant. There is no doubt that the tools of organic synthesis, which allow the introduction of non-natural elements into peptides, will play an important role in future developments in this area. Additionally, as the theoretical framework of protein and peptide structure and the computational methods available are improved we will see examples of *in silico* successful designs.

Although one could doubt the potential medical utility of these designed peptides owing to membrane permeability and stability problems of peptides *in vivo*, recent progress in peptide delivery systems, and the prospect of developing peptidomimetic analogs of the active derivatives, seem to warrant the therapeutic potential of the approach.

Acknowledgements

We thank the Ministry of Science and Technology and the ERDF (SAF2001–3120) and the Galician Government (PGIDIT02BTF20901PR) for supporting our work in this area. M. E. V. thanks the Human Frontier Science Program for a postdoctoral grant (LT00001/2002-C).

References

1 For an updated revision of the protein expression process and control of gene expression see: G. Orphanides and D. Reinberg, *Cell*, 2002, **108**, 439.See also: B. Lemon and R. Tjian, *Genes Dev*., 2000, **14**, 2551.
2 P. Dervan, *Bioorg. Med. Chem.*, 2001, **9**, 2215.
3 L. H. Hurley, *Nature Rev. Cancer*, 2002, **2**, 188.
4 J. B. Opalinska and A. M. Gewirtz, *Nature Rev. Drug Discovery*, 2002, **1**, 503.
5 J. B. Chaires, *Curr. Opin. Struct. Biol*., 1998, **8**, 314.
6 S. Neidle, *Nat. Prod. Rep.*, 2001, **18**, 291.
7 J. G. Pelton and D. E. Wemmer, *Proc. Nat. Acad. Sci. USA*, 1989, **86**, 5723.
8 P. Dervan and R. W. Buerli, *Curr. Opin. Chem. Biol*., 1999, **3**, 688.
9 R. E. Bremen, E. Baird and P. B. Dervan, *Chem. Biol.*, 1998, **5**, 119.
10 M. Faria and C. J. Giovannangeli, *Gene Med.*, 2001, **3**, 299.
11 R. Lavery, C. Zardecki and J. Westbrook, *Oxford Handbook of Nucleic Acid Structure*. S. Neidle, Ed. Oxford Science Pub. Oxford, 1999.
12 L. Jen-Jacobson, L. E. Engler and L. A. Jacobson, *Structure*, 2000, **8**, 1015.
13 M. Oda and H. Nakamura, *Genes to Cells*, 2000, **5**, 319.
14 V. A. Bloomfield, D. M. Crothers and I. Tinoco, Jr., *Nucleic Acids. Structures, Properties, and Functions*. University Science Books Sausalito, CA, 2000.
15 At the end of 2002 more than 280 structures of protein–DNA complexes had been deposited in the protein Data Bank.
16 N. M. Luscombe, S. E. Austin, H. M. Berman and J. M. Thornton, *Genome Biol*., 2000, **1**, 1.
17 C. W. Garvie and C. Wolberger, *Mol. Cells*, 2001, **8**, 937.
18 R. Wintjens and M. Rooman, *J. Mol. Biol*., 1996, **262**, 294.
19 S. Khorasanizadeh and F. Rastinejad, *Curr. Biol*., 1999, **9**, R456.
20 A. Klug and J. W. R. Schwabe, *FASEB J*., 1995, **9**, 597.
21 H. C. Hurst, *Protein Profile*, 1995, **2**, 101.
22 J. J. Kohler and A. Schepartz, *Biorg. Med. Chem.*, 2001, **9**, 2435.
23 P. B. Rupert, G. W. Daughdrill, B. Bowerman and B. W. Matthews, *Nat. Struct. Biol.*, 1998, **5**, 484.
24 J. G. Omichinski, P. V. Pedone, G. Felsenfeld, A. M. Gronenborn and G. M. Clore, *Nat. Struct. Biol*., 1997, **4**, 122.
25 C. O. Pabo and L. Nekludova, *J. Mol. Biol.*, 2000, **301**, 597.
26 N. M. Luscombe and R. A. Laskowski, *Nucleic Acids Res.*, 2001, **29**, 2860; C. L. Larson and G. L. Verdine, *Bioorganic Chemistry: Nucleic acids* Ed. S. M. Hetch, Oxford University Press, NY, 1999.

27 N. Jones, *Cell*, 1990, **61**, 9.
28 M. Moore, A. Klug and Y. Choo, *Proc. Nat. Acad. Sci. USA*, 2001, **98**, 1437 and references therein.
29 C. O. Pabo, E. Peisach and R. A. Grant, *Annu. Rev. Biochem.*, 2001, **70**, 313.
30 M. Nagaoka and Y. Sugiura, *J. Inorg. Biochem.*, 2000, **82**, 57.
31 J. L. Pomerantz, P. A. Sharp and C. O. Pabo, *Science*, 1995, **267**, 93.
32 There is a general interest in downsizing protein size while maintaining native function: W. F. DeGrado and T. R. Sosnick, *Proc. Nat. Acad. Sci. USA*, 1996, **93**, 5680.
33 I. Ghosh, S. Yao and J. Chmielewski, in *Comprehensive Natural Products Chemistry*; Eds: D. Barton and K. Nakanishi, Pergamon, Chap. 7.13, 1999.
34 P. V. Pedone, R. Ghirlando, G. M. Clore, A. M. Gronenborn, G. Felsenfeld and J. G. Omichinski, *Proc. Nat. Acad. Sci. USA*, 1996, **93**, 2822.
35 S. Sato, M. Hagihara, K. Sugimoto and T. Morii, *Chem Eur. J.*, 2002, **8**, 5067.
36 R. V. Talanian, C. J. McKnight and P. S. Kim, *Science*, 1990, **249**, 769.
37 C. Park, J. L. Campbell and W. A. Goddard III, *Proc. Nat. Acad. Sci. USA*, 1993, **90**, 4892.
38 M. Ueno, A. Murakami, K. Makino and T. Morii, *J. Am. Chem. Soc.*, 1993, **115**, 12575.
39 M. Uenu, M. Sawada, K. Makino and T. Morii, *J. Am. Chem. Soc.*, 1994, **116**, 11137.
40 B. Cuenoud and A. Schepartz, *Science*, 1993, **259**, 510.
41 A. M. Caamaño, M. E. Vázquez, J. Martínez-Costas, L. Castedo and J. L. Mascareñas, *Angew. Chem. Int. Ed.*, 2000, **39**, 3104.
42 M. Thompson and N. W. Woodbury, *Biophys. J.*, 2001, **81**, 1793.
43 M. Thompson and N. W. Woodbury, *Biochemistry*, 2000, **39**, 4327.
44 N. Sardesai and J. K. Barton, *J. Biol. Inorg. Chem.*, 1997, **2**, 762.
45 M. E. Vázquez, A. M. Caamaño, J. Martínez-Costas, L. Castedo and J. L. Mascareñas, *Angew. Chem. Int. Ed.*, 2001, **40**, 4723.
46 E. Vazquez, A. M. Caamaño, L. Castedo, D. Gramberg and J. L. Mascareñas, *Tetrahedron Lett.*, 1999, **40**, 3625.
47 M. Zhang, B. Wu, J. Baum and J. W. Taylor, *J. Peptide Res.*, 2000, **55**, 398.
48 M. Zhang, B. Wu, H. Zhao and J. W. Taylor, *J. Peptide Sci.*, 2002, **8**, 125.
49 G. H. Bird, A. R. Lajmi and J. A. Shin, *Biopolymers*, 2002, **65**, 10.
50 N. J. Zondlo and A. Schepartz, *J. Am. Chem. Soc.*, 1999, **121**, 6938.
51 T. Morii, S. Sato, M. Hagihara, Y. Mori, K. Imoto and K. Makino, *Biochemistry*, 2002, **41**, 2177.