Mini-review

Expanding the specificity of DNA targeting by harnessing cooperative assembly

Rocco Morettia,b, Aseem Z. Ansari,a,b,*

a Department of Biochemistry, The Genome Center of Wisconsin, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706, United States
b The Genome Center of Wisconsin, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706, United States

Received 3 January 2008; accepted 18 February 2008
Available online 4 March 2008

Abstract

The precise control of developmental and regulatory processes in the cell requires accurate recognition of specific DNA sites. For genomes as large as that of humans, single-molecule-DNA binders have difficulties accurately and specifically recognizing the intended targets. Natural transcription factors overcome these difficulties by forming non-covalent complexes on the DNA with other transcription factors. These cooperative complexes overcome the difficulties of single-molecule transcription factors, allowing specific, combinatorial control of a range of transcriptional targets. Artificial transcription factors have been designed to take advantage of this technique of cooperative assembly, facilitating future studies in whole genome targeting. In contrast to a simple model of component independence, cooperative complexes as a whole often display slightly altered DNA specificity from what would be expected from the analysis of their separate components. The true sequence specificity of cooperative complexes, and thus their presumed in vivo targets, have to be experimentally probed. A number of techniques, such as the cognate site identity array, now allow for rapid, high-throughput determination of the specificity of cooperative complexes.

Abbreviations: Hox, homeobox transcription factor; Exd, extradenticle; TALE, three amino acid loop extension homeodomain; ATF, artificial transcription factor; bp, base pairs.

Keywords: Protein–DNA dimerizers; DNA targeting; Artificial transcription factors (ATF); Cooperative assembly; Coupled equilibria

One of the key steps in the regulation of cellular processes for both eukaryotes and prokaryotes is the control of transcription. In order to accurately produce the particular proteins needed in response to external signals and internal cellular conditions, RNA polymerase must specifically recognize a short segment of DNA, the promoter, for transcription initiation. Not only must this location be recognized in preference to others, it also must only be recognized at the appropriate time in response to appropriate signals. Regulatory circuits can wreak havoc on the proper functioning of the cell if expressed at inopportune times [1]. This is especially daunting in the eukaryotic genome, where the polymerase must distinguish between tens of thousands of other transcriptional start sites in gigabases of non-relevant DNA. In this task, polymerase is assisted by a multitude of gene-specific transcription factors. These transcription factors not only recruit the polymerase to the appropriate transcriptional start site for transcriptional activation, but can alternatively act as transcriptional repressors, either by reducing the activity of polymerase directly or by bringing in DNA and histone modifying enzymes or other DNA occluding proteins [2]. Transcription factors can perform this selective regulation of particular genes in a sea of many others because they recognize and bind selectively to specific DNA sequences.

Reliant as the cell is on this critical dance between transcription factors and polymerase, it is not surprising that a number of diseases, including developmental defects, cancer, and diabetes, can arise from malfunctioning transcription factors. For example, 25% of all diagnosed pediatric pre-B cell leukemias result from a chromosomal translocation in which the transcriptional activation domain of E2a is fused to the...
Fig. 1. (a) Zinc finger proteins (Zif268 pictured [83]) bind DNA using repeated motifs of approximately 30 amino acids. (b) To a first approximation, each of these fingers can be thought to bind a separate three base pair sequence. Using a variety of techniques, a large number of fingers have been developed, each recognizing one of the 64 triplets. (c) These individual fingers can be combined to make DNA-binding proteins with an arbitrary recognition sequence, including sites much larger than the 9–12 base pairs recognized by native zinc finger proteins. (d) Pyrrole–imidazole polyamides also use a modular binding technique [14], with each ring recognizing one nucleotide. (e) Using the appropriate combination of monomers, either in hairpin or in side-by-side format a desired DNA sequence can be targeted (W = A or T). Extending the size of the polyamide results in compounds with enlarged DNA-binding sites.
DNA-binding domain of Pbx1 [3], and 70% of all solid tumors are correlated with mutations in the DNA-binding domain of the tumor suppressor p53 [4]. Mutations in DNA sites recognized by the liver-specific transcription factor Hnf4 can result in reduction in insulin production and the onset of diabetes [5].

1. Targeting genes with artificial transcription factors

One of the promising avenues of research in treating these diseases, and in the control of transcription in general, is the use of artificial transcription factors (ATFs). These compounds are designed to site-specifically recognize defined locations in the genome and execute a defined regulatory program, substituting for missing transcription factors or ameliorating the effects of aberrant ones.

The ability to target any desired DNA sequence with artificial compounds was buoyed by the recognition that the zinc finger class of DBD has a modular design, with each “finger” of 3–4 finger domain recognizing a three base pair subset of the 9–12 base pair recognition site [6,7]. All 16 million dodecamer sites could theoretically be recognized by various combinations of just 64 fingers, each recognizing a unique three base pair sequence (Fig. 1a–c). Much work has been done to expand the zinc finger code to recognize all 64 triplets, with the majority of the code being filled [7–9]. These engineered zinc fingers have been expressed as fusions with a number of different regulation domains, resulting in ATFs with designed DNA recognition [10–12]. However, certain sequences are, as-of-yet, still untargetable. Although mostly modular, the binding sequences are not completely separable — fingers show base contacts outside of their own triple, influencing neighboring bases [6,9,13].

Sequence programmable synthetic molecules have also been developed. In the late 1970s, a number of groups were investigating the minor-groove binders netropsin and distamycin. Natively these poly-pyrrole compounds prefer to bind A/T stretches [14–17]. However, structural studies indicated that changing a pyrrole to an imidazole would result in a preference for Gs [17–20]. As these pyrrole-imidazole polyamides can bind side-by-side in the minor groove, each strand of the DNA can be targeted (Fig. 1d–e) [14,15,21]. The discovery that hydroxypyrrole can target thymidines specifically resulted in a polyamide—DNA-binding code, from which one can design a molecule to target any specified DNA sequence [22]. Further research has discovered other monomers that can be applied in polyamide building, each with their own slight differences in sequence preference [23–29].

2. DNA targeting specificity using cooperative binding

Although zinc fingers and polyamides can be designed to bind desired DNA sequences, a recognition sequence of 16 base pairs, on average, is required to be able to uniquely distinguish a site in the three gigabase human genome. There are examples of both zinc fingers and polyamides designed to bind to 13–18 base pairs [8,30,31]. However, the number of contacts, and thus the total affinity, for a 13–18 base pair binding site is so great that one or two missed contacts does not substantially change the affinity. The compounds bind these extended sites with great affinity, but show lowered specificity and also bind a large number of sub-optimal sites as well [32]. For example, a six-finger zinc finger ATF needed to match just 13 out of the 18 base pairs targeted in order to show appreciable levels of binding [32,33].

Most metazoan transcription factors, however, target just a 6–10 base pair site and accommodate a high level of degeneracy [34,35]. Despite being able to bind over most of the genome, transcription factors work primarily in a small subset of those sites. The DNA at these “enhancers” consists of a number of clustered transcription factor binding sites (Fig. 2a). Cooperative interactions between transcription factors at the enhancer enhance specificity, as gene expression is only significantly induced when all of the binding partners occupy the enhancer [34,36–38]. The energy of non-covalent protein—protein interactions between the transcription factors stabilizes the multimeric complex on these sites. These interactions are relatively weak (high micromolar to low millimolar), and only form stably when the partners are assembled on adjacent DNA sites. In the example of the interferon-β promoter, seven different transcription factors must bind cooperatively to the enhancer to induce high levels of interferon-β transcription [39,40].

The use of cooperative assembly in transcriptional regulation offers several advantages. The precision of DNA targeting is increased, as the complex can only form where all binding sites are present. By having individual transcription factors be involved in multiple complexes, different genes can be combinatorially controlled by varying which partners are present. For example, by expressing different binding partners in different tissues, the same transcription factor can be directed to different, tissue specific genes in the different cell types, based on the differences in DNA-binding preference of the partners. If multiple partners, each with their own set of variants, assemble together, a large number transcriptional programs can be executed using a minimal number of transcription factors [41]. Cooperative assembly facilitates rapid regulatory switching that is exquisitely responsive to combinations of cellular signals, while buffering genes from being expressed randomly due to a low degree of “noise” [34,42,43].

3. Targeting via cooperative assembly

This concept of cooperative assembly has been applied to the design of a new class of ATFs. In the Hox–TALE system of developmental regulators, the different Hox proteins recognize the almost same DNA-binding site, yet show markedly different effects on animal development [44–46]. It is thought that cooperative binding with TALE transcription factors mediates some of these differences [47]. The primary contact between the Hox and TALE proteins is a short, 4–6 amino acid peptide (Fig. 2b) [48]. Although the extent of the peptide’s contribution to complex formation in the native system is debated [47,49,50], conjugating the peptide to a DNA-binding polyamide recapitulates the Hox protein sufficiently to recruit Exd, a TALE protein, to the DNA to form a ternary complex (Fig. 2c) [51,52]. This does not, however, imply that other
van der Waals interactions or even allosteric modulations of DNA geometry do not play a significant role in the native structure [49,53,107].

The stability of this complex shows only minor effects due to binding site separation. The polyamide binding site can be moved towards, away, or even be inverted with regards to the Exd binding site, and still result in measurable Exd recruitment [53]. Because the polyamide binds the minor groove and Exd binds the major groove, the sites can even be overlapped slightly. The only requirement for successful binding is to have a peptide—polyamide linker with sufficient length to allow the interaction. Having a “too long” linker only imposes a minor entropic penalty, resulting in a minor decrease in complex affinity at elevated temperature [52]. These experiments demonstrate the applicability of ATFs towards cooperative transcriptional systems, as Exd binds only very weakly, if at all, to the DNA in the absence of the peptide-displaying ATF [51–53,107].

4. Modularity of artificial transcription factors

These modular ATFs follow a similar design to their natural counterparts. Natural transcription factors tend to be multi-domain proteins, with the DNA recognition and effector functionalities (e.g. an activator or a repressor domain) on different, separable domains [2,34]. Some of the earliest successes in artificial transcription factor design were simple “domain-swap” experiments, where the DNA-binding domain (DBD) of one transcription factor was tethered to an effector domain from a second transcription factor (Fig. 3a) [2,34,54]. If the domains are appropriately chosen (i.e. not all binding and regulation domains are suitably modular), this hybrid transcription factor causes the transcriptional results (activation or repression) of the second factor at the DNA locations recognized by the first factor. Neither the DNA-binding domain nor the effector domain is affected by the exchange.

The binding and effector domains need not be proteins [1,2]. Many other DNA-binding molecules developed have the potential to work as DNA-binding domains, a large number of which already have clinical applications (Fig. 3b) [55,56]. DNA intercalators are planar, aromatic compounds, which bind to DNA by inserting themselves between two base pairs, stretching and unwinding the DNA slightly [57]. The aromatic portion of the molecule that inserts itself into the DNA is then equivalent to an additional “rung” on the DNA ladder. Many intercalators are usually not “pure” intercalators, being decorated with additional moieties which add to DNA affinity and specificity by binding to the grooves [57]. These additional moieties can add to the sequence specificity of the compound. Even the simple intercalators show some sequence preferences, due to the effect local sequence has on DNA structure and deformability [57]. A well-known example is ethidium, which, although more sequence-neutral than most other DNA binders, has a slight preference in binding CpG and GpC base steps [58]. Groove binders, in contrast,
bind not between base pairs but alongside them, forming hydrophobic and hydrogen-bonding contacts with the walls and floors of the groove [57,59]. These small molecule-DNA binders have been shown to bind in a cooperative fashion, in that binding of one molecule increases the affinity for further molecules to bind to neighboring sites [59]. For example, the intercalators such as ethidium and actinomycin D show cooperative effect in binding DNA [60,61]. Echinomycin has been shown to bind cooperatively to two appositioned sites, but only in the appropriate DNA context [62]. Minor-groove binders like netropsin, distamycin, and Hoechst 33258 can also show cooperative binding effects in the correct environments [63,64]. Most of the time these cooperative effects in small molecule-DNA binding are not a result of direct molecule–molecule interactions, but are rather mediated by larger scale DNA structural changes [65].

In addition to zinc fingers and polyamides, there have been a number of efforts to create modular synthetic systems for targeted DNA recognition [66]. Like zinc fingers, the ideal DNA binder would be modular, with each module specifying the identity of a base pair or small group of base pairs. One of the earliest efforts in modular DNA binding was the use of triplex forming oligonucleotides (TFOs), or triplex forming peptide nucleic acids (PNAs) [67–70]. TFOs bind to duplex DNA using Hoogsteen, rather than Watson–Crick, base pairing arrangements. While affording a modest amount of sequence targeting, TFOs can only bind purine or pyrimidine stretches due to requirement imposed by the Hoogsteen pairing [71]. Mixed or alternating stretches of purines and pyrimidines are either unrecognized or poorly recognized by TFOs.

5. Ternary complex formation

Small molecules, polyamides, PNAs, and TFOs have only a limited binding site, and, like polydactyl zinc fingers, are unable to escape the limitation of single-molecule binding. From a simplified theoretical perspective, any one of these DNA-
binding compounds should be able to function as a DNA-binding domain in a cooperative artificial transcription factor. For ATFs with flexible linkers, the linker and effector domains have little effect on the DNA-binding specificity function of the ATF, so any DNA-binding molecule should function adequately in cooperative complex formation [52,53].

Using a model of coupled binding equilibrium (Fig. 4a), one can obtain an equation for the binding of the complex to DNA for situations where the ATF–DNA-binding domain is changed. Using the relationship between binding constant and free energy change, the population of each binding state (TF–DNA, ATF–DNA, TF–DNA–ATF, etc.) can be found. This analysis is made easier by assuming that the affinity of individual interactions is not affected by the state of the others (no allosterism), and by the use of free concentration constants, rather than total concentrations. This assumption is valid in situations like quantitative EMSA, where the DNA concentration is much lower than either of the transcription factors and the TF–ATF interaction is weak. Using these assumptions equations can be derived which model EMSA results, like that in Fig. 4b, which looks at the fraction of DNA bound with TF.

Using rough parameters for the Hox–TALE system, we can examine the effect of varying the DNA-binding affinity on complex formation (Fig. 4c). A bifunctional compound with even a modest DNA-binding affinity ($K_D = 0.1$ mM) is able to promote a 10-fold increase in binding of TF to DNA. Using coupled equilibria means that a number of DNA-binding compounds, including ones with low affinity, could potentially be used to promote complex formation.

6. Other compounds as DNA-binding domains

Given this theoretical background, we have tested the hypothesis that other small molecule binders can also function as DNA-binding domains in ATFs. Several compounds were made, each containing the Hox peptide conjugated to one of several small molecule-DNA-binding domains (bisbenzimidazole, 9-acridinecarboxylic acid, anthraquinone-2-carboxylic acid, and daunorubicin) using peptide and oxime conjugation chemistries. On their own, these DNA-binding moieties show affinities for DNA in the high nanomolar to micromolar range (3 nM for bisbenzimidazole compound [57,72], 3 μM for acridine derivatives [73], ~10 μM for anthraquinone conjugates [74], and 1.3 μM for daunorubicin [57]). The DBD–peptide conjugates were tested for Exd recruitment in EMSAs, under similar conditions that were used to detect Exd recruitment by the polyamide–peptide conjugates.

Preliminary results of these experiments indicate that there are additional factors at work in the artificial Hox–TALE system than simple additivity of binding energy. It is possible that the detection of the bound complex is limited by kinetic, rather than thermodynamic considerations. If the dissociation rate of the complex is sufficiently fast, the complex may dissociate faster than the resolution time of the assay technique.

Another possibility is that there are additional allosteric effects occurring. It has been shown previously that the binding of one partner in a complex can distort the DNA slightly, allowing for better binding of the second partner in the complex [75,76,107]. In the published Hox–TALE–DNA crystal structures, the DNA deviates from B-form [48,77,78]. Such structural perturbations may assist in stable Exd binding to DNA. Polyamides without the Hox interaction peptide, when presented in a suitable orientation, can promote Exd binding even in the absence of any predicted polyamide–Exd interaction [107]. Presumably, this recruitment is the result of Exd responding to the allosteric modification of the DNA by the polyamide. Alternatively, the polyamide may effect the dynamics of the DNA structure, by limiting or directing the conformational flexure of the DNA so as to promote Exd binding.

7. DNA-binding preference for cooperative complexes

These non-additive effects can subtly perturb the binding of a complex away from the simple combination of binding at half sites. To ensure that the ATFs target only the desired
genes in the cell, it would be preferred to examine the sequence specificity and affinity of the cooperative complex. In this way, the length and sequence of the binding site for the cooperative complex as a whole can be confirmed.

A number of techniques have been developed to measure the DNA-binding preferences of compounds (Table 1). The classic techniques are low throughput, measuring binding affinity to only one or just a few sequences. More recently, several techniques have been developed which are able to pull out the highest affinity binding sequences.

Other techniques use direct binding to double-stranded DNA on a microarray, allowing for the determination of binding to thousands of sequences at once [53,79–82]. Perhaps the most versatile of these techniques is the cognate site identifier (CSI) array (Fig. 5a). The CSI techniques use maskless array synthesizer technology to form over one million single stranded DNA features on a single slide [53]. Self-hybridization of these features then forms double-stranded hairpins. Fluorescently labeled compound is then bound to the array, and detected with standard microarray scanning techniques. The fluorescence intensity thus obtained is a measure of sequence affinity. By using a set of probes which represent each sequence in a given N-mer sequence, a complete binding profile can be obtained (Fig. 5b). Once calibration using a secondary technique is performed for a small subset of sequences, the CSI results for all sequences can even be converted to absolute affinities (Fig. 5c) [53,82].

CSI provides a method of comprehensive determination of the binding preferences of a DNA-binding molecule or DNA-binding complex. The direct readout of CSI is a relative binding complex. The direct readout of CSI is a relative binding preference, necessitating the use of an additional secondary binding assay to convert the relative binding affinities to absolute affinities. CSI also suffers from the current limitations in maskless array synthesis. Maskless arrays are currently limited to 0.4–2 million features per chip, putting a 10–12 base pair upper limit on the number of effective positions of full degeneracy. The binding preference of very large complexes can be inferred through the randomization of smaller to medium sized subcomponents within the larger sequence. Another challenge is that other proteins in the cell, such as histones, DNA methylases, and other, unknown, binding partners may skew the localization of complexes on DNA.

Despite such limitations, the DNA-binding preferences of complete complexes, as distinct from the preferences of individual components, can be successfully examined with CSI. For example, CSI has been applied in determining the binding specificity of a cooperative polyamide–Exd complex [53]. Two different polyamide–peptide conjugates were assayed by CSI, differing by a single methylene in the linker. This small change allowed for a noticeable change in binding profile. In contrast to what was expected, multiple recognition sequences were observed. Both compounds showed the ability to bind to two different separation differences in the half site, but for the longer linker binding was also observed when the polyamide half site was inverted with respect to the Exd half site (Fig. 5b) [53]. Although, given the length of the linker, this binding orientation can be rationalized post hoc, productive

<p>| Table 1: Methods for assaying DNA affinity and sequence specificity |</p>
<table>
<thead>
<tr>
<th>Technique</th>
<th>Comments</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gel electrophoresis (gel shift)</td>
<td>Only works for compounds which diffuse DNA.</td>
<td>[85–86]</td>
</tr>
<tr>
<td>Nitrocellulose filter binding</td>
<td>Ligand must cause labeled DNA to stick to filter (proteins only).</td>
<td>[87]</td>
</tr>
<tr>
<td>Competition dialysis</td>
<td>DNAs are dialyzed against the same free ligand solution. The amount of bound ligand is related to the affinity.</td>
<td>[88]</td>
</tr>
<tr>
<td>Isothermal titration calorimetry</td>
<td>Binding energy of ligand–DNA interaction measured in heat.</td>
<td>[89]</td>
</tr>
<tr>
<td>Nuclear magnetic resonance (NMR)</td>
<td>NMR signals change from the unbound to ligand bound forms.</td>
<td>[90]</td>
</tr>
<tr>
<td>Circular dichroism</td>
<td>Optical rotation of light varies based on ligand binding state.</td>
<td>[91]</td>
</tr>
<tr>
<td>Nuclease protection (footprint)</td>
<td>Bound ligand protects DNA from exogenous cleavage agent.</td>
<td>[92–94]</td>
</tr>
<tr>
<td>Fluorescence anisotropy</td>
<td>A change in mass changes the tumbling, and thus fluorescence polarization of a labeled molecule.</td>
<td>[95]</td>
</tr>
<tr>
<td>Systematic evolution of ligands by exponential enrichment (SELEX)</td>
<td>The highest affinity ligands are pulled from a library by increasing binding stringency.</td>
<td>[98]</td>
</tr>
<tr>
<td>Cyclic amplification and selection of targets (CAST)</td>
<td>Same as SELEX.</td>
<td>[99,100]</td>
</tr>
<tr>
<td>Restriction endonuclease protection selection and amplification (RESPA)</td>
<td>Non-bound DNA sequences are degraded, leaving only those with highest affinity.</td>
<td>[101]</td>
</tr>
<tr>
<td>Surface plasmon resonance (SPR)</td>
<td>Binding to anchored DNA is detected by a change in surface optical properties.</td>
<td>[102]</td>
</tr>
<tr>
<td>Chromatin immunoprecipitation on microarray (ChIP-chip)</td>
<td>Protein bound to genomic DNA. Bound sequences are immunoprecipitated.</td>
<td>[103,105]</td>
</tr>
<tr>
<td>DNA immunoprecipitation with microarray detection (DIP-chip)</td>
<td>Like ChIP-chip, but with purified DNA sequences.</td>
<td>[106]</td>
</tr>
<tr>
<td>Duplex DNA microarray</td>
<td>DNA immobilized on a chip, spotted or synthesized by polymerase on the chip.</td>
<td>[104]</td>
</tr>
<tr>
<td>Protein binding microarray (PBM)</td>
<td>Double-stranded DNA spotted or synthesized by polymerase on the chip.</td>
<td>[107]</td>
</tr>
<tr>
<td>Cognate site identifier (CSI) array</td>
<td>Hairpin DNA chemically synthesized on the chip.</td>
<td>[53,82]</td>
</tr>
</tbody>
</table>
binding in this arrangement of sites was not predicted prior to performing the CSI analysis. Broad scale binding site analysis techniques such as CSI are thus invaluable in examining cooperative complexes, as they are able to evaluate binding for the complete range of sites, and are able to point out sequence recognition features which are not initially apparent from the binding sites of separated components.

8. Conclusion

Artificial transcription factors have the potential to influence many of the transcription-related processes in the cell, including those involved with disease. Key in this transcriptional regulation is the ability of the ATF to site-specifically recognize the intended DNA targets in a genomic context. This DNA recognition is facilitated by the availability of many different DNA-binding molecules, each with their own sequence specificity. This specificity can be assayed by a number of methods, not only to determine how well they bind their intended targets, but also to determine how well they discriminate against mismatch sites. Ultimately, however, a single-molecule-DNA binder is unlikely to find use for targeting sites in the human genome, due to the high likelihood of mismatch site recognition. The use of cooperative assembly of macromolecular complexes affords a powerful way of achieving site-specific targeting in the context of genomes.

Acknowledgments

We gratefully acknowledge the support of the NIH (GM069420), March of Dimes, W.M. Keck foundations, and Shaw Scholar and Vilas Associate awards to AZA. RM was supported by an NIH-MBTG pre-doctoral fellowship.

References


