Published in final edited form as:

Biochemistry. 2016 November 29; 55(47): 6467–6474. doi:10.1021/acs.biochem.6b00757.

Thermodynamic additivity for impacts of base-pair substitutions on association of the Egr-1 zinc-finger protein with DNA

Abhijnan Chattopadhyay, Levani Zandarashvili[†], Ross H. Luu, and Junji Iwahara^{*}

Department of Biochemistry & Molecular Biology, Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX 77555, USA

Abstract

The transcription factor Egr-1 specifically binds as a monomer to its 9-bp target DNA sequence, GCGTGGGCG, via three zinc fingers and plays important roles in the brain and cardiovascular systems. Using fluorescence-based competitive binding assays, we systematically analyzed the impacts of all possible single nucleotide substitutions in the target DNA sequence and determined the change in binding free energy for each. Then, we measured the changes in binding free energy for sequences with multiple substitutions and compared them with the sum of the changes in binding free energy for each constituent single substitution. For the DNA variants with 2 or 3 nucleotide substitutions in the target sequence, we found excellent agreement between the measured and predicted changes in binding free energy. Interestingly, however, we found that this thermodynamic additivity broke down with a larger number of substitutions. For DNA sequences with 4 or more substitutions, the measured changes in binding free energy were significantly larger than predicted. Based on these results, we analyzed the occurrences of high-affinity sequences in the genome and found that the genome contains millions of high-affinity sequences that might functionally sequester Egr-1.

Graphical abstract

^{*} To whom correspondence should be addressed. Contact Information: [Address] 301 University Blvd, Medical Research Building 5.104C, Galveston, Texas 77555-1068, USA j.iwahara@utmb.edu [Phone] 409-747-1403 [Fax] 409-772-6334.

[†] Present affiliation: Department of Biochemistry and Biophysics, University of Pennsylvania

Supporting Information Available

Data showing that the covalent attachment of TAMRA to the 3'-terminus of the DNA duplex does not influence the binding of Egr-1 to its recognition sequence (Figure S1); comparison of uncertainties in dissociation constants estimated from 4 replicates and those from fitting for single datasets (Table S1); and high-affinity Egr-1-binding sequences and their total numbers in the human genome (Table S2). The supporting information is available free of charge on the ACS Publication website.

Introduction

Gene regulation at a transcriptional level requires the association of transcription factors with particular DNA sites in the cis-regulatory elements of the genome. In many cases, the transcription factors locate their target sites due to strong affinity for specific DNA sequences, which are typically shorter than 10 bp for eukaryotic transcription factors.¹ Sequence specificity in DNA binding is one of the most important properties of transcription factors and has thus been studied by means of biochemistry, biophysics, and structural biology.²

To represent the sequence specificity in DNA association of proteins, researchers often use consensus sequences, which are typically obtained through the identification and alignment of high-affinity DNA sequences. DNA foot-printing, gel-shift, and systematic-evolution-of-ligands-by-exponential-enrichment (SELEX) methods are traditionally popular for biochemically identifying such sequences.³ Some recent studies have employed high-throughput methods such as SELEX-seq⁴⁻⁶ and protein-binding microarray⁷⁻⁹, which identify numerous DNA sequences of high affinity for a particular protein. Alignment of the identified high-affinity sequences and statistical analysis of base type at each position yield the consensus sequences that represent the sequence specificity.

Although it is common, the alignment-based information is not sufficient for our quantitative understanding of sequence specificity. Ideally, the binding free energy should be given as a function of nucleotide sequence.^{10,11} This requires measuring the dissociation constants K_d for the protein-DNA complexes of various nucleotide sequences. In practice, however, obtaining such data for all possible sequences is difficult, especially when using conventional quantitative methods. If thermodynamic additivity¹² is applicable, K_d data for a relatively small subset of the possible sequences would be sufficient to estimate the binding free energy for the rest. In fact, the thermodynamic additivity of the impacts of base-pair substitutions was examined and confirmed for some sequence-specific DNA-binding proteins, including bacteriophage λ repressor, Cro repressor, and the eukaryotic transcription factors Max and SYCRP1.¹³⁻¹⁷ However, the thermodynamic additivity does not necessarily hold true for other proteins. For example, non-additive behavior of binding free energy differences was reported for variant operators of the *E. coli lac* repressor and Mnt repressor.^{18,19}

Our goal in this study is to examine the thermodynamic additivity for the zinc-finger protein Egr-1 (also known as Zif268 or NGFI-A), which recognizes a 9-bp sequence, GCGTGGGCG, via three zinc fingers and binds to DNA as a monomer. The Egr-1 DNA-binding domain has been used as a scaffold for zinc-finger technology, which has produced artificial transcription factors and DNA-modifying enzymes that target desired DNA sequences.²⁰⁻²² Although this technology has been successful, its off-target effects (e.g., DNA cleavage by a zinc-finger nuclease at undesired sites) remain concerning and limit its applicability.^{23,24} The energetics of the sequence specificity should be better understood to address this concern. Furthermore, naturally abundant sequences that differ from but resemble the consensus sequences can also exhibit high affinity and may preclude the

transcription factor from reaching the target sites.^{25,26} Thus, it is important to analyze the binding free energy as a function of DNA sequence for this protein.

In the current work, for the interaction between the zinc-finger DNA-binding domain of Egr-1 and its target DNA, we systematically measure the changes in binding free energy upon nucleotide substitutions. For each variant with multiple substitutions, we compare the measured change in binding free energy with the sum of the changes in binding free energy for the single substitutions involved. Some previous microarray-based studies examined additivity for the Egr-1–DNA interactions but did so only for limited positions of base-pair substitutions.^{27,28} Our current study extensively examined the thermodynamic additivity for energetic impacts of base-pair substitutions on the DNA association of Egr-1 in solution at equilibrium. Based on the results, we also assess potential high-affinity Egr-1-binding sites in the human genome.

Materials and Methods

Protein and DNA

The Egr-1 zinc-finger protein (human Egr-1 residues 335–423) was expressed in the *E. coli* strain BL21 (DE3) and purified as previously described.^{29–31} The purified protein was quantified using the BCA protein assay kit (Pierce Biotechnology; Rockford, IL). Individual strands of TAMRA-labeled probe DNA containing the Egr-1 recognition sequence were purchased from Integrated DNA Technologies, Inc. (Coralville, IA) and purified via anion-exchange chromatography using a Mono-Q column together with a ÄKTA Purifier system (GE Healthcare; Chicago, IL). The complimentary strands were mixed and annealed from ~85 °C to room temperature over ~1–2 hours. The resultant TAMRA-labeled duplex was purified again via Mono-Q anion-exchange chromatography. Unlabeled duplexes of the target DNA variants were also purchased from Integrated DNA Technologies, Inc.

Fluorescence-based competitive binding assays

The affinities of the variant DNA duplexes for the Egr-1 zinc-finger protein were measured with competitive binding assays using fluorescence anisotropy as a function of unlabeled competitor DNA. In each assay, we mixed Egr-1 (50 nM) with two different DNA duplexes (Figure 1A). One of them was the 12-bp DNA duplex (10 nM) that includes the Egr-1 recognition sequence and a fluorescent probe, tetramethylrhodamine (TAMRA), which is attached to the 3'-terminus. The other duplex was 12-bp competitor DNA of a different sequence with a central 9-bp region homologous to the Egr-1 recognition sequence. Anisotropy of TAMRA fluorescence from the probe DNA was measured at 20 °C as a function of unlabeled competitor DNA using an ISS PC-1 spectrofluorometer equipped with light polarizers and a temperature controller. To avoid strand exchange between the probe and the competitor DNA, we chose different flanking sequences outside of the 9-bp core region for these duplexes. The strand exchange is unlikely because the sequence-based prediction³² of hybridization free energies for the original and mismatched duplexes suggests that the mismatched duplexes from the potential strand exchange would be more unstable by > 5 kcal/mol than the original duplexes. The excitation and emission wavelengths were 533 nm and 580 nm, respectively. This assay used solutions of 10 nM 3'-

TAMRA-labeled 12-bp DNA containing the Egr-1 recognition sequence, 50 nM protein, and various concentrations (0-100,000 nM) of unlabeled competitor DNA. The buffer conditions were 10 mM Tris•HCl (pH 7.5), 150 mM KCl, and 0.2 μM ZnCl₂. Fluorescence anisotropy was measured for each solution as a function of the concentration of competitor DNA. To calculate the K_d from the fluorescence anisotropy data, we used the following equation for analysis of the competition assay data:³³

$$\sigma = \frac{CK_{d,p} + K_{d,p}K_{d,c} - PK_{d,p} + 2PK_{d,c} - K_{d,p}\sqrt{(C + K_{d,c} - P)^2 + 4PK_{d,c}}}{2\{CK_{d,p} - (K_{d,p} - K_{d,c})(K_{d,p} + P)\}}$$

[1],

$$A_{obs} = (1 - \rho) A_{free} + \rho A_{bound} \quad [2],$$

where ρ is the fraction of the probe DNA bound to the protein; $K_{d,c}$ and $K_{d,p}$ are the dissociation constants for the competitor and probe DNA duplexes, respectively; A_{obs} is the observed anisotropy; A_{bound} and A_{free} are those of protein-bound DNA and free DNA, respectively; and P , D , and C are the total concentrations of the protein, probe DNA, and competitor DNA, respectively. When the $K_{d,c}$ or $K_{d,p}$ value is known, the other dissociation constant can be determined from the competition assay data via nonlinear least-squares fitting. When $P \ll C$ is satisfied, Eq. 2 becomes equivalent to a popular form based on the binding polynomial:³⁴

$$\rho = \frac{P / K_{d,p}}{1 + C / K_{d,c} + P / K_{d,p}} \quad [3]$$

The fitting calculations were performed with Mathematica software (Wolfram; Champaign, IL). For the competitor DNA duplexes of the Egr-1 recognition sequence and two single-substitution variants (G1→C and G6→A), the K_d constants were measured 4 times. Although it would be ideal to measure the K_d several times for the other competitor DNAs as well, this was not practical, as there were 60 different DNA duplexes and the duration of a single experiment was approximately one day for each DNA. Therefore, for the other DNA duplexes, the uncertainties in the measured K_d values were estimated from single datasets using a confidence level³⁵ of 95% in the fitting calculations. This uncertainty estimation was justified by the fact that, for the Egr-1 recognition sequence and the two single-substitution variants, the estimated uncertainties from individual fitting calculations were comparable to two times the standard deviation of 4 replicates of the K_d data (see Table S1 in the Supporting Information).

Analysis of high-affinity sites for Egr-1 in the human genome

Analysis of high-affinity sites for Egr-1 in the human genome was conducted using MATLAB software together with the Bioinformatics Toolbox (MathWorks, Inc.; Natick, MA). First, all possible 9-bp sequences were generated. For each of them, the difference in the binding free energy, ΔG , for Egr-1 was predicted from the ΔG data for single substitutions, and the number of base-pair matches with the Egr-1 recognition sequence was counted. High-affinity sequences for Egr-1 were identified as those that exhibit $\Delta G < 1.3$ kcal/mol and ≥ 6 base-pair matches with the Egr-1 recognition sequence. The total number of each of these high-affinity Egr-1-binding sequences in the human genome was counted using the GenBank GRCh38.p7 assembly.

Results

We examined the thermodynamic additivity for the impacts of base-pair substitutions on the binding free energy for the specific DNA interactions of the Egr-1 DNA-binding domain (human Egr-1 residues 335-423). This domain consists of three zinc fingers and binds to DNA as a monomer. The same protein construct was also used in our previous NMR, X-ray, and fluorescence studies.^{25,29-31,36-41} To analyze the thermodynamic additivity, we investigated changes in binding free energy for 59 different DNA variants of the Egr-1 recognition sequence.

Competitive binding assays

Using fluorescence anisotropy-based competitive binding assays, we determined the dissociation constants K_d of the Egr-1 zinc-finger protein at physiological ionic strength for various 12-bp DNA duplexes, in which the central 9-bp region is homologous to the Egr-1 recognition sequence. Utilizing the TAMRA-labeled 12-bp DNA containing the Egr-1 recognition sequence, unlabeled competitor DNA and protein, this assay monitored the fraction of the fluorescent probe DNA bound to the protein as a function of the concentration of the competitor DNA. As shown in the Supporting Information, the covalent attachment of TAMRA to the 3'-terminus of the DNA duplex does not affect the binding of Egr-1 to its recognition sequence (Figure S1). This competitive binding assay directly assessed the relative affinity of variant DNA with respect to that of the target DNA. This method is more precise than direct measurements using protein titration assays when the dissociation constant (K_d) is close to or smaller than the probe concentration.³³ To determine the K_d value, anisotropy of TAMRA fluorescence from the probe DNA (10 nM) containing the Egr-1 recognition sequence was monitored as a function of the concentration of the unlabeled competitor DNA in the presence of 50 nM Egr-1. The fluorescence anisotropy values of the protein-bound and free states of the probe DNA were 0.22 and 0.16, respectively. When the concentration of competitor DNA was relatively low, the majority of the probe DNA was in the protein-bound state. An increase in competitor DNA caused Egr-1 to transfer from the probe to the competitor DNA, thereby decreasing the population of the protein-bound state of the probe DNA (Figure 1B).

G for all possible single substitutions in the target DNA sequence

First, we systematically measured the dissociation constants of all single-substitution variants of the Egr-1 recognition sequence. Figure 1B shows the binding isotherm data for all possible substitutions at each position within the Egr-1 recognition sequence. Table 1 shows the values of the measured K_d constants for these DNA duplexes. The observed K_d constants strongly depended on both the type and the position of the substitution. For example, the G-to-T substitution at the 3rd position of the Egr-1 recognition sequence rendered a ~900-fold reduction in affinity, whereas the G-to-T substitution at the 9th position left the affinity virtually unaffected; however, the G-to-A substitution at the same position caused a ~5-fold reduction in affinity. From the K_d data shown in Table 1, we calculated the changes in binding free energy (ΔG) for individual substitutions using the following equation:

$$\Delta\Delta G = \Delta G_{var} - \Delta G_{ori} = RT \ln (K_{d,var} / K_{d,ori}), \quad [4]$$

where the parameters for the variant and original recognition sequences are annotated with 'var' and 'ori', respectively; R is the gas constant; and T is the temperature. Using the error propagation method,³⁵ uncertainties in ΔG values were estimated to be

$RT \sqrt{(\sigma_{var}/K_{d,var})^2 + (\sigma_{ori}/K_{d,ori})^2}$, where σ_{var} and σ_{ori} are uncertainties in $K_{d,var}$ and $K_{d,ori}$ respectively. Figure 2 shows the ΔG data for individual single base-pair substitutions. The measured ΔG values ranged from -0.1 kcal/mol to 4.0 kcal/mol. This range is comparable to those observed in similar studies of other proteins.¹³⁻¹⁷

Excellent thermodynamic additivity observed for DNA variants with 2 or 3 substitutions

Using the same competitive binding assays, we measured the dissociation constants for the target DNA variants with two or more base-pair substitutions. The changes in binding free energy upon substitution, ΔG , were determined for these variants as well. For each variant DNA duplex with multiple substitutions, the measured ΔG was compared with the values predicted as the sum of the ΔG s for each constituent substitution. In this manner, we tested the thermodynamic additivity for the impacts of base-pair substitutions on the binding of Egr-1 to DNA.

Because it is impractical to measure all possible variants, we randomly selected 10 double-substitution variant sequences (No. 1-10 in Figure 3) for which the substitution sites are widely distributed throughout the Egr-1 recognition sequence. Due to the practical difficulty in measuring $\Delta G > 4.5$ kcal/mol, we chose the sequences for which ΔG was expected to be smaller than 4.5 kcal/mol. We randomly chose 5 triple-substitution variant sequences (No. 11-15 in Figure 3) as well, and designed 5 more triple-substitution variant sequences (No. 16-20) by adding an extra substitution to the 5 double-substitution variant sequences (No. 6-10). For these double- and triple-substitution variant DNA duplexes, we measured the affinities of the Egr-1 zinc-finger protein, from which the ΔG values were determined. Figure 3 shows the correlation between the measured and predicted ΔG values for Egr-1–DNA binding. For all 20 variants with 2 or 3 substitutions in the Egr-1 recognition sequence, the measured and predicted ΔG values were in excellent agreement. This result clearly

indicates that these substitutions in the Egr-1 recognition sequences produce thermodynamically additive effects. Perhaps surprisingly, even variants with two adjacent substitutions showed good agreement between the measured and predicted G values, although one may not expect thermodynamic independence of two such substitutions. The root-mean-square difference (RMSD) between the measured and predicted G values for the data shown in Figure 3 was 0.18 kcal/mol. For the 5 triple-substitution variants based on the double-substitution variants (the sequences No. 15-20 in Figure 3), the energetic impact of the third substitution, $G_{obs}(\text{triple substitutions}) - G_{obs}(\text{double substitutions})$, agreed well with the G expected from the single-substitution data (RMSD, 0.15 kcal/mol). These data clearly indicate that the effects of 2 or 3 base-pair substitutions within the Egr-1 recognition sequence are additive in terms of binding free energy and can therefore be predicted from the data on the constituent single base-pair substitutions.

Breakdown of the thermodynamic additivity with 4 or more substitutions

Interestingly, however, we found that the thermodynamic additivity broke down for variants containing 4 or more substitutions. We measured the affinities for 10 quadruple-substitution variants (the sequences No. 21-30 in Figure 4) and 2 quintuple-substitution variants (the sequences No. 31 and 32). All sequences of the quadruple-substitution variants were designed by adding an extra base-pair substitution to those of the 10 triple-substitution variants (the sequences No. 11-20). Figure 4A shows the correlation between the observed and expected G values for these variants with 4 or 5 base-pair substitutions within the Egr-1 recognition sequence. The G values measured for these variants were systematically and substantially larger than the value predicted from the G values for single substitutions. For example, the G for a quadruple-substitution variant of GGGTTGGAT (the sequence No. 25 in Figure 4) is predicted to be 1.0 kcal/mol, however, the actual G was as large as 2.8 kcal/mol. In other words, the actual affinities of these variants are far weaker than those predicted from the G data for single substitutions. The RMSD between the measured and predicted G values for the data shown in Figure 4A was 1.38 kcal/mol. This non-additive effect seems to occur when there are 4 or more substitutions. As shown in Figure 4B, the observed impact of the fourth substitution, $G_{obs}(\text{quadruple substitution}) - G_{obs}(\text{triple substitution})$, was significantly larger than expected (RMSD, 1.28 kcal/mol). Though the substitutions of the same type at the same positions exhibited thermodynamic additivity for the double- and triple-substitution variants, the fourth substitutions made the affinity far weaker than expected from the G data for single substitutions. This effect considerably reduces the affinities of sequences that are only weakly similar to the recognition sequence and should therefore enhance the sequence specificity of Egr-1.

Diversity of high-affinity sequences

Based on the findings described above, we identified high-affinity Egr-1-binding sequences among all possible 9-bp nucleotide sequences. Our criteria for “high affinity” were 1) $G < 1.3$ kcal/mol and 2) there are 6 or more base-pair matches with the Egr-1 recognition sequence (i.e., 3 or less substitutions). The second criterion is based on the finding that more than 3 base-pair substitutions cause far weaker affinities than expected, as described in the previous subsection. The first criterion, $G < 1.3$ kcal/mol, means that the K_d for the

sequence differs from that for the recognition sequence by a factor < 10 . This factor represents a relatively high affinity compared to completely nonspecific sequences because the ratio of $K_d(\text{nonspecific}) / K_d(\text{specific})$ is $> 3,000$ for Egr-1 under the current conditions.^{29,38} Using these criteria, a total of 348 different 9-bp sequences (out of 262,144 possible sequences) were identified as high-affinity Egr-1-binding sequences. These sequences and their predicted ΔG are shown in the Supporting Information (Table S2). Importantly, the majority of these high-affinity sequences differ from the Egr-1 recognition sequence by at least 2 nucleotide positions (i.e., quasi-specific sequences^{25,42}), as summarized in Table 2.

Abundance of high-affinity sites in the genome

To gain insight into the extent to which these high-affinity sequences affect Egr-1 in the nucleus, we determined the number of each high-affinity sequence in the human genome using NCBI GenBank data for chromosomes 1-22 and X. The results are shown in the Supporting Information and are summarized in Table 2. Our current data show that the genomic DNA contains over 6 million sites of the high-affinity sequence for Egr-1. This is consistent with our previous estimate from our kinetic study on Egr-1 using naked genomic DNA.²⁵ Even if 85% of the high-affinity sites are buried in nucleosomes, $\sim 10^6$ high-affinity sites should remain to be accessible for Egr-1. It is estimated that the maximum number of the Egr-1 molecules is $\sim 10^4$ copies per nucleus when induced.²⁵ Because Egr-1 regulates $\sim 10^2$ genes^{43,44} and each cis-regulatory element involve only a small number of Egr-1 binding sites, the total number of functional target sites for Egr-1 is likely $\sim 10^2$ - 10^3 . Therefore, the number of the non-functional high-affinity sites in the genome is far greater than the numbers of Egr-1 molecules and their functional targets.

Discussion

Comparison with others' data on Egr-1

Our current study provides comprehensive data regarding the impacts of base-pair substitutions on the DNA-binding affinity of Egr-1. For a limited number of nucleotide positions, other research groups also reported quantitative investigations of changes in the binding affinity of Egr-1 (Zif268/NGFI-A) upon base-pair substitutions in DNA. Our ΔG data on the impacts of single substitutions on the Egr-1–DNA association are at least qualitatively consistent with the affinity data for some DNA variants described by Hamilton *et al.*⁴⁵ and Elrod-Erickson and Pabo⁴⁶. However, our data in Figure 2 differ considerably from the ΔG data of Mikles *et al.* for the same substitutions.⁴⁷ Although the cause of this discrepancy is not clear, it should also be mentioned that the affinity of Egr-1 measured by Mikles *et al.* (e.g., $K_d = 310$ nM for the Egr-1 recognition sequence)⁴⁷ was far weaker than those observed by our and other groups ($K_d < 10$ nM for the same sequence) under similar conditions.^{29,45,46,48-51} A previous microarray-based study on the central 3-bp region initially claimed interdependent effects of base-pair substitutions.²⁸ However, a subsequent study reanalyzed the same data and validated additivity for the substitutions.²⁷ For DNA variants with 2 or 3 substitutions, our solution-based data support validity of the thermodynamic additivity for the entire region of the Egr-1 recognition sequence (Figure 3).

Consideration on breakdown of thermodynamic additivity

For DNA variants with 4 or more substitutions, however, our data clearly demonstrate breakdown of the thermodynamics additivity; the measured changes in binding free energy were significantly larger than those predicted (Figure 4). This could be partially due to sequence dependence of DNA shape and deformability.^{2,52} A large number of substitutions might strongly influence the energy terms for DNA deformation and invalidate the additivity that assumes independence of the impact of each substitution. In fact, the target DNA bound to Egr-1 is moderately deformed from the canonical B form: the pitch of the DNA in the complex is 11.2 bp per turn (10.5 bp per turn for the B form), and the major groove is deeper by 1.6 Å.^{53,54} Another possible explanation is that 4 or more nucleotide substitutions might adversely impact domain-domain packing and weaken positive cooperativity among the three zinc fingers within an Egr-1 molecule bound to DNA.^{55,56}

High-affinity sites as natural decoys

Our current study demonstrates that the human genome contains millions of sites with high-affinity sequences (Table 2). As discussed recently,²⁶ although these high-affinity sites could potentially work as natural decoys that may functionally sequester Egr-1, such functional sequestration could be greatly diminished if other proteins block these sites. For example, methyl-CpG-binding domain (MBD) proteins such as Mbd1, Mbd2, and MeCP2 might play such a role.⁵⁷ Because the Egr-1 recognition sequence contains two CpGs, a majority of the high-affinity Egr-1 binding sites are likely methylated. Thus, there exists a possibility that CpG methylation in high-affinity Egr-1 binding sites can cause binding of MBD proteins and thereby make these sites inaccessible for Egr-1 *in vivo*. This would potentially reduce the risk for Egr-1 to be trapped at nonfunctional high-affinity sites, because functionally important Egr-1 bindings sites are located in the CpG islands (CGIs) near transcription start sites,^{58,59} where DNA methylation is generally rare.⁵⁷ This possibility seems to be high particularly in neurons, where MeCP2 protein expression levels are very high (10⁷ copies per nucleus).⁶⁰ Because CGIs comprise less than 1% of the total genome,⁵⁷ the vast majority of the high-affinity sequences should be located outside the CGIs. Considering these statistics, it is reasonable to speculate that the functional sites within the CGI are unmethylated, whereas the nonfunctional high-affinity Egr-1 sites are methylated. Although our current study suggests the presence of millions of natural Egr-1 decoys in the genome, functional sequestration in these decoys might not be very strong due to the blocking by other proteins. Further studies are necessary to examine this possibility.

Concluding remarks

Our extensive analysis of the changes in binding free energy upon nucleotide substitutions has provided important insight into the Egr-1–target DNA association. For the double and triple-substitution variants, the observed G values were in excellent agreement with the sum of G for individual single substitutions, indicating that thermodynamic additivity is valid for 2 or 3 substitutions. However, with a larger number of substitutions, the thermodynamic additivity broke down, and the observed affinities were substantially weaker than the predicted affinities. This effect should enhance sequence specificity in DNA-binding of Egr-1. These findings were implemented in our analysis of the high-affinity

Egr-1-binding sequences in the genome. Our current study shows that the human genome contains millions of sites with high-affinity Egr-1-binding sequences. Further studies are required to elucidate the role of these very abundant high-affinity sites.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Alexandre Esadze, Catherine Kemme, and Dan Nguyen for useful discussion.

Funding Source Statement:

This work was supported by Grant R01-GM107590 from the National Institutes of Health (to J.I.).

Abbreviations

bp	base pair
CGI	CpG island
MBD	methyl-CpG-binding domain
RMSD	root-mean-square difference
SELEX	systematic evolution of ligands by exponential enrichment
TAMRA	tetramethylrhodamine

References

1. Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* 2009; 25:434–440. [PubMed: 19815308]
2. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci.* 2014; 39:381–399. [PubMed: 25129887]
3. Moss, T. DNA-protein interactions: principles and protocols. 2nd. Humana Press; New Jersey: 2001.
4. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpaa MJ, Bonke M, Palin K, Talukder S, Hughes TR, Luscombe NM, Ukkonen E, Taipale J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* 2010; 20:861–873. [PubMed: 20378718]
5. Riley TR, Slattery M, Abe N, Rastogi C, Liu D, Mann RS, Bussemaker HJ. SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol Biol.* 2014; 1196:255–278. [PubMed: 25151169]
6. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, Mann RS. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell.* 2011; 147:1270–1282. [PubMed: 22153072]
7. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulky ML. Diversity and complexity in DNA recognition by transcription factors. *Science.* 2009; 324:1720–1723. [PubMed: 19443739]

8. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc.* 2009; 4:393–411. [PubMed: 19265799]
9. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol.* 2006; 24:1429–1435. [PubMed: 16998473]
10. Gerland U, Moroz JD, Hwa T. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A.* 2002; 99:12015–12020. [PubMed: 12218191]
11. Slutsky M, Mirny LA. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys J.* 2004; 87:4021–4035. [PubMed: 15465864]
12. Dill KA. Additivity principles in biochemistry. *J Biol Chem.* 1997; 272:701–704. [PubMed: 8995351]
13. Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. *Science.* 2007; 315:233–237. [PubMed: 17218526]
14. Omagari K, Yoshimura H, Suzuki T, Takano M, Ohmori M, Sarai A. DeltaG-based prediction and experimental confirmation of SYCRP1-binding sites on the *Synechocystis* genome. *FEBS J.* 2008; 275:4786–4795. [PubMed: 18761670]
15. Sarai A, Takeda Y. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc Natl Acad Sci U S A.* 1989; 86:6513–6517. [PubMed: 2771938]
16. Takeda Y, Sarai A, Rivera VM. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc Natl Acad Sci U S A.* 1989; 86:439–443. [PubMed: 2911590]
17. Zhao Y, Granas D, Stormo GD. Inferring binding energies from selected binding sites. *PLoS Comput Biol.* 2009; 5:e1000590. [PubMed: 19997485]
18. Frank DE, Saecker RM, Bond JP, Capp MW, Tsodikov OV, Melcher SE, Levandoski MM, Record MT Jr. Thermodynamics of the interactions of lac repressor with variants of the symmetric lac operator: effects of converting a consensus site to a non-specific site. *J Mol Biol.* 1997; 267:1186–1206. [PubMed: 9150406]
19. Man TK, Stormo GD. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* 2001; 29:2471–2478. [PubMed: 11410653]
20. Durai S, Mani M, Kandavelou K, Wu J, Porteus MH, Chandrasegaran S. Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res.* 2005; 33:5978–5990. [PubMed: 16251401]
21. Segal DJ, Barbas CF 3rd. Custom DNA-binding proteins come of age: polydactyl zinc-finger proteins. *Curr Opin Biotechnol.* 2001; 12:632–637. [PubMed: 11849947]
22. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet.* 2010; 11:636–646. [PubMed: 20717154]
23. Isalan M. Zinc-finger nucleases: how to play two good hands. *Nat Methods.* 2012; 9:32–34.
24. Pattanayak V, Ramirez CL, Joung JK, Liu DR. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat Methods.* 2011; 8:765–770. [PubMed: 21822273]
25. Kemme CA, Esadze A, Iwahara J. Influence of quasi-specific sites on kinetics of target DNA search by a sequence-specific DNA-binding protein. *Biochemistry.* 2015; 54:6684–6691. [PubMed: 26502071]
26. Kemme CA, Nguyen D, Chattopadhyay A, Iwahara J. Regulation of transcription factors via natural decoys in genomic DNA. *Transcription.* 2016; 7:115–120. [PubMed: 27384377]
27. Benos PV, Bulyk ML, Stormo GD. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* 2002; 30:4442–4451. [PubMed: 12384591]
28. Bulyk ML, Johnson PL, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 2002; 30:1255–1261. [PubMed: 11861919]
29. Esadze A, Iwahara J. Stopped-flow fluorescence kinetic study of protein sliding and intersegment transfer in the target DNA search process. *J Mol Biol.* 2014; 426:230–244. [PubMed: 24076422]

30. Takayama Y, Sahu D, Iwahara J. NMR studies of translocation of the Zif268 protein between its target DNA Sites. *Biochemistry*. 2010; 49:7998–8005. [PubMed: 20718505]
31. Zandarashvili L, Vuzman D, Esadze A, Takayama Y, Sahu D, Levy Y, Iwahara J. Asymmetrical roles of zinc fingers in dynamic DNA-scanning process by the inducible transcription factor Egr-1. *Proc Natl Acad Sci U S A*. 2012; 109:E1724–1732. [PubMed: 22675124]
32. SantaLucia J Jr, Hicks D. The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct*. 2004; 33:415–440. [PubMed: 15139820]
33. Zandarashvili L, Nguyen D, Anderson KM, White MA, Gorenstein DG, Iwahara J. Entropic Enhancement of Protein-DNA Affinity by Oxygen-to-Sulfur Substitution in DNA Phosphate. *Biophys J*. 2015; 109:1026–1037. [PubMed: 26331260]
34. Dill, KA.; Bromberg, S. *Molecular Driving Forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science; New York: 2011. Multi-site & cooperative ligand binding; p. 559-584.
35. Bevington, PR.; Robinson, DK. *Data reduction and error analysis for the physical sciences*. 3rd. McGraw-Hill; Boston: 2003.
36. Chen C, Esadze A, Zandarashvili L, Nguyen D, Pettitt BM, Iwahara J. Dynamic Equilibria of Short-Range Electrostatic Interactions at Molecular Interfaces of Protein-DNA Complexes. *J Phys Chem Lett*. 2015; 6:2733–2737. [PubMed: 26207171]
37. Esadze A, Chen C, Zandarashvili L, Roy S, Pettitt BM, Iwahara J. Changes in conformational dynamics of basic side chains upon protein-DNA association. *Nucleic Acids Res*. 2016
38. Esadze A, Kemme CA, Kolomeisky AB, Iwahara J. Positive and negative impacts of nonspecific sites during target location by a sequence-specific DNA-binding protein: origin of the optimal search at physiological ionic strength. *Nucleic Acids Res*. 2014; 42:7039–7046. [PubMed: 24838572]
39. Zandarashvili L, Esadze A, Kemme CA, Chattopadhyay A, Nguyen D, Iwahara J. Residence Times of Molecular Complexes in Solution from NMR Data of Intermolecular Hydrogen-Bond Scalar Coupling. *J Phys Chem Lett*. 2016; 7:820–824. [PubMed: 26881297]
40. Zandarashvili L, Esadze A, Vuzman D, Kemme CA, Levy Y, Iwahara J. Balancing between affinity and speed in target DNA search by zinc-finger proteins via modulation of dynamic conformational ensemble. *Proc Natl Acad Sci USA*. 2015; 112:E5142–E5149. [PubMed: 26324943]
41. Zandarashvili L, White MA, Esadze A, Iwahara J. Structural impact of complete CpG methylation within target DNA on specific complex formation of the inducible transcription factor Egr-1. *FEBS Lett*. 2015; 589:1748–1753. [PubMed: 25999311]
42. Chakrabarti J, Chandra N, Raha P, Roy S. High-affinity quasi-specific sites in the genome: how the DNA-binding proteins cope with them. *Biophys J*. 2011; 101:1123–1129. [PubMed: 21889449]
43. Baumgartel K, Tweedie-Cullen RY, Grossmann J, Gehrig P, Livingstone-Zatchej M, Mansuy IM. Changes in the proteome after neuronal zif268 overexpression. *J Proteome Res*. 2009; 8:3298–3316. [PubMed: 19374395]
44. Fu M, Zhu X, Zhang J, Liang J, Lin Y, Zhao L, Ehrenguber MU, Chen YE. Egr-1 target genes in human endothelial cells identified by microarray analysis. *Gene*. 2003; 315:33–41. [PubMed: 14557062]
45. Hamilton TB, Borel F, Romaniuk PJ. Comparison of the DNA binding characteristics of the related zinc finger proteins WT1 and EGR1. *Biochemistry*. 1998; 37:2051–2058. [PubMed: 9485332]
46. Elrod-Erickson M, Pabo CO. Binding studies with mutants of Zif268. Contribution of individual side chains to binding affinity and specificity in the Zif268 zinc finger-DNA complex. *J Biol Chem*. 1999; 274:19281–19285. [PubMed: 10383437]
47. Mikles DC, Schuchardt BJ, Bhat V, McDonald CB, Farooq A. Role of promoter DNA sequence variations on the binding of EGR1 transcription factor. *Arch Biochem Biophys*. 2014; 549:1–11. [PubMed: 24657079]
48. Morisaki T, Imanishi M, Futaki S, Sugiura Y. Rapid transcriptional activity in vivo and slow DNA binding in vitro by an artificial multi-zinc finger protein. *Biochemistry*. 2008; 47:10171–10177. [PubMed: 18754679]

49. Nalefski EA, Nebelitsky E, Lloyd JA, Gullans SR. Single-molecule detection of transcription factor binding to DNA in real time: specificity, equilibrium, and kinetic parameters. *Biochemistry*. 2006; 45:13794–13806. [PubMed: 17105198]
50. Swirloff AH, Milbrandt J. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol Cell Biol*. 1995; 15:2275–2287. [PubMed: 7891721]
51. Yang WP, Wu H, Barbas CF 3rd. Surface plasmon resonance based kinetic studies of zinc finger-DNA interactions. *J Immunol Methods*. 1995; 183:175–182. [PubMed: 7602135]
52. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*. 2010; 79:233–269. [PubMed: 20334529]
53. Elrod-Erickson M, Rould MA, Nekludova L, Pabo CO. Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions. *Structure*. 1996; 4:1171–1180. [PubMed: 8939742]
54. Pavletich NP, Pabo CO. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*. 1991; 252:809–817. [PubMed: 2028256]
55. Imanishi M, Nakamura A, Morisaki T, Futaki S. Positive and negative cooperativity of modularly assembled zinc fingers. *Biochem Biophys Res Commun*. 2009; 387:440–443. [PubMed: 19615972]
56. Lee J, Kim JS, Seok C. Cooperativity and specificity of Cys2His2 zinc finger protein-DNA interactions: a molecular dynamics simulation study. *J Phys Chem B*. 2010; 114:7662–7671. [PubMed: 20469897]
57. Klose RJ, Bird AP. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci*. 2006; 31:89–97. [PubMed: 16403636]
58. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
59. Kubosaki A, Tomaru Y, Tagami M, Arner E, Miura H, Suzuki T, Suzuki M, Suzuki H, Hayashizaki Y. Genome-wide investigation of in vivo EGR-1 binding sites in monocytic differentiation. *Genome Biol*. 2009; 10:R41. [PubMed: 19374776]
60. Skene PJ, Illingworth RS, Webb S, Kerr AR, James KD, Turner DJ, Andrews R, Bird AP. Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol Cell*. 2010; 37:457–468. [PubMed: 20188665]

Figure 1.

Competitive binding assays for the Egr-1 zinc-finger protein. **(A)** Probe and competitor DNA duplexes. The probe DNA was 12-bp DNA with the fluorescent probe TAMRA attached to the 3'-terminus. As shown in the Supporting Information (Figure S1), the covalent attachment of TAMRA to the 3'-terminus of the DNA duplex does not affect the protein-DNA interaction. Competitors were 12-bp DNA. The box represents 9 bp homologous to the target sequence. Different flanking sequences were used to avoid strand exchange between the probe and competitor DNA duplexes. Egr-1 possesses three zinc-finger domains and binds to the target DNA as a monomer. **(B)** Competitive binding assay data for the single-substitution variants. For each panel, the population of the protein-bound probe DNA is shown as a function of the concentration of competitor DNA. For each position, data for 4 different bases are shown.

Figure 2. Changes in the binding free energy ($\Delta\Delta G$) for Egr-1 upon single base-pair substitutions in the target DNA.

Figure 3.

Thermodynamic additivity for impact of double and triple-substitutions in the target DNA. The sequences of the 10 double-substitution and 10 triple-substitution variant DNA duplexes are shown, where the base-pair substitutions are indicated in red or orange. For 5 triple-substitution variants (No. 16-20), the sequences were designed by adding an extra substitution (indicated in orange) to the double-substitution variants (No. 6-10). The vertical axis represents experimentally measured ΔG , and the horizontal axis represents those predicted as the sum of ΔG values for single nucleotide substitutions. Data points of double and triple-substitution variants are shown with circles and squares, respectively. Uncertainties in observed ΔG values are smaller than the size of the symbols.

Figure 4.

Breakdown of the thermodynamic additivity with 4 or more base-pair substitutions. **(A)** Correlation between the observed and predicted G data for the quadruple and quintuple-substitution variants of the target DNA. The quadruple-substitution variant DNA duplexes were designed by adding an extra substitution (indicated in orange) to the triple-substitution variant DNA (the sequences No. 11-20 in Figure 3). The axes are the same as those in Figure 3. **(B)** Correlation between the expected and predicted impacts of the 4th substitution on the binding free energy. The vertical axis represents the difference between the observed G values for the related triple- and quadruple-substitution variants, whereas the horizontal axis represents the G value of the 4th substitution expected from the data of Figure 2.

Table 1

Dissociation constants K_d for the complexes of the Egr-1 zinc-finger protein with 12-bp DNA duplexes with a single base substitution in the Egr-1 consensus sequence. The values shown are in nM.

	Nucleotide substitution			
Position	A	T	G	C
G1	11 ± 2	56 ± 4	$[6.3 \pm 0.4]^a)$	60 ± 9
C2	8.2 ± 1.0	5.2 ± 0.6	8.4 ± 1.0	$[6.3 \pm 0.4]^a)$
G3	$(1.4 \pm 0.1) \times 10^3$	$(6.0 \pm 1.5) \times 10^3$	$[6.3 \pm 0.4]^a)$	$(1.5 \pm 0.2) \times 10^3$
T4	28 ± 4	$[6.3 \pm 0.4]^a)$	13 ± 2	57 ± 8
G5	9.8 ± 1.3	10 ± 1	$[6.3 \pm 0.4]^a)$	26 ± 4
G6	$(4.0 \pm 0.2) \times 10^2$	14 ± 1	$[6.3 \pm 0.4]^a)$	19 ± 2
G7	$(3.5 \pm 0.5) \times 10^2$	$(2.4 \pm 0.3) \times 10^2$	$[6.3 \pm 0.4]^a)$	$(2.1 \pm 0.3) \times 10^2$
C8	15 ± 2	10 ± 1	26 ± 3	$[6.3 \pm 0.4]^a)$
G9	27 ± 4	7 ± 1	$[6.3 \pm 0.4]^a)$	24 ± 3

^{a)} Values in square brackets are for the competitor DNA with no substitutions in the Egr-1 consensus sequence (i.e., 12-bp DNA TGCGTGGGCGAT).

Table 2

Total number of 9-bp sequences that are predicted to exhibit high affinity for Egr-1 and their occurrences in the human genome.

# of high-affinity sequences ^{a)}	Occurrences in the genome	
9-bp match	1	3.7×10^3
8-bp match	19	113.8×10^3
7-bp match	105	$1,300.9 \times 10^3$
6-bp match	223	$5,227.3 \times 10^3$
Total:	348	6.6 million

^{a)} Sequences for which $G < 1.3$ kcal/mol with reference to the binding free energy for the Egr-1 recognition sequence. Actual sequences, together with their G and total occurrences, are individually listed in the Supporting Information.