Published online August 2, 2004

4026–4032 *Nucleic Acids Research, 2004, Vol. 32, No. 13*
doi:10.1093/nar/gkh729

Quantitative modeling of DNA–protein interactions: effects of amino acid substitutions on binding specificity of the Mnt repressor

Tsz-Kwong Man, Joshua SungWoo Yang and Gary D. Stormo*

Department of Genetics, Washington University, 660 S. Euclid, Box 8232, St Louis, MO 63110, USA

Received May 27, 2004; Revised and Accepted July 10, 2004

ABSTRACT

Understanding DNA–protein recognition quantitatively is essential to developing computational algorithms for accurate transcriptional binding site prediction. Using a quantitative, multiple fluorescence, relative affinity (QuMFRA) assay, we determine the binding specificity of 11 different position 6 variants of the Mnt repressor for operators containing all 16 possible dinucleotides at operator positions 16 and 17. We show that the wild-type and all variant proteins interact with the two positions in a non-independent manner, but that a simple independent model provides a close approximation to the true binding affinities. The wild-type His at amino acid 6 is the only protein to prefer the AC sequence of the wild-type operator, whereas most of the variant proteins prefer TA. H6R is unique in having a strong preference for C at position 16. A comparison of the quantitative binding data for all of the protein variants with a model for recognition of the early growth response (EGR) zinc finger family suggests that interactions of Mnt with positions 16 and 17 are similar to interactions of EGR with positions 1 and 2, respectively. This information leads to an augmented model for the interaction of Mnt with its operator.

INTRODUCTION

Discovering the interactions that control gene expression in a cell remains one of the important challenges in molecular biology. Given a genome sequence it is now routine to identify computationally the probable set of transcription factors (TFs) because they fall into well-known protein families (1). There are also a variety of computational methods that can be used to identify likely transcription factor binding sites (TFBSs). These typically employ some method to search for significantly conserved sequence patterns in sets of genes that are likely to be co-regulated, or in 'phylogenetic footprinting' methods that identify conserved sequences for specific promoter regions among multiple species, or both (2). However, none of these approaches solves the problem of identifying the regulon for each TF, the set of genes that are directly regulated through binding at TFBSs for each specific TF. That requires knowledge about the connections between the TFs and the TFBSs, i.e. which TF binds to which TFBS.

There are some experimental approaches to making such connections. For example, yeast one-hybrid assays can identify TFs that bind to a specific TFBS (3). ChIP-chip experiments are a good, high-throughput method to identify the set of promoters that are bound by a specific TF, and then identifying common motifs among those promoters can reveal the TF–TFBS connections (4). One might also use SELEX methods with purified TFs (5) or apply them to promoter arrays to identify those that are bound, and again apply pattern discovery methods to find the TFBSs (6). While these methods are relatively efficient, they can still fail and are still fairly laborious and expensive. It would be much more efficient if there were a computational method to connect TFs to TFBSs.

There is a long history of attempts to define a recognition code for protein–DNA interactions (7). If such a code could be found it would allow for the prediction of a TFBS pattern given any TF sequence and some knowledge of its structure, which is available for each of the TF families (1). Or, given a TFBS, it would be possible to identify the TF in that genome that is most likely to be the one that binds to it. However, a variety of studies have shown that nature does not employ a simple, deterministic code (7–10). But more complex codes have been proposed and shown to be reasonable approximations, at least for some TF families (7,11–14). Currently, the recognition models are limited by the available data, especially the lack of quantitative binding data for a sufficient collection of binding sites. An efficient method to collect quantitative binding data can help enormously in the quest to develop good recognition models. At this point it is not even clear how complicated such models need to be. Most DNA binding site models assume that each position contributes independently to the interaction (15). While clear non-independent interactions have been observed for several proteins (16,17), it has also been shown for those examples that independent models can provide good approximations to the true binding data (18). Whether this is generally true for most TF protein families remains an open question. It is also

*To whom correspondence should be addressed. Tel: +1 314 747 5534; Fax: +1 314 362 7855; Email: stormo@genetics.wustl.edu

Present address:
Tsz-Kwong Man, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA

not known if independence is a valid model for the protein side of the interaction, and whether the contribution of each amino acid contributes independently to the binding affinity. Some evidence suggests that such a model may be valid, at least in some cases (19), but again there is a lack of sufficient quantitative data.

In this paper, we determine the quantitative specificity of several variants of the Mnt repressor protein for a collection of 16 different binding sites, the complete set of dinucleotide pairs at two operator positions. Mnt is a member of the ribbon–helix–helix family of DNA binding proteins (20). The histidine at position 6 (H6) of the Mnt repressor has been shown to interact with positions 16 and 17 of the *mnt* operator (O_{mnt}) DNA (21–23). Using native Mnt repressor protein and *mnt* operator variants containing all possible changes at positions 16 and 17, we previously showed that the interactions of H6 of native Mnt repressor protein and these two base positions of the O_{mnt} are non-independent (16) although an additive model can be obtained that provides a good approximation to the true binding affinities (18). We have also shown previously that replacement of H6 with other amino acids changes the preferred operator sequence, primarily at position 17 and also at positions 16–19, consistent with amino acid 6 being a 'master residue' that effects the binding at several operator positions (22,23). In this paper, we quantify the effects of several position 6 variants for all possible binding site combinations at operator positions 16 and 17.

MATERIALS AND METHODS

Materials and reagents

Chemicals and reagents were purchased from either Sigma Chemical Co. (St Louis, MO) or Fisher Scientific (Pittsburgh, PA) unless specified. *Taq* DNA polymerase and 1× reaction buffer used in PCR were from Promega (Madison, WI). Fluorophore-labeled SK-1 oligo and unlabeled mnt wild-type and mutant oligos were purchased from Integrated DNA technologies (Coralville, IA). The sequences of these oligos were listed in (16). Mnt proteins used in this study were constructed and purified as described previously (23).

Fluorescent labeling of DNA

Double-stranded fluorophore-labeled wild-type and mutant mnt operator (O_{mnt}) DNAs were synthesized by PCR. The reaction contained 200 nM of wild-type or mutant O_{mnt} oligo, 1× reaction buffer, 200 nM dNTPs, 500 nM of fluorophore-labeled (FAM, HEX, TAMRA and ROX) SK primer, 500 nM of KS primer, 4 mM $MgCl_2$, and 1 U of *Taq* DNA polymerase in 100 µl reaction. The PCR was cycled 30 times at 94°C for 1 min, 62°C for 1 min and 72°C for 1 min.

QuMFRA assay

The QuMFRA assays were performed similar to those described in (16). Mnt proteins were incubated with the fluorescently-labeled wild-type O_{mnt} and mutant O_{mnt} PCR DNAs in 1× binding buffer at room temperature for 1 h. The bound and unbound fractions were separated by a 10% TBE polyacrylamide gel at 100 V for 1 h. The gel was then scanned by Typhoon Variable Scanner (Molecular Dynamics, Sunnyvale, CA) using excitation laser at 532 nm and various output voltages and emission filters for different fluorophores (600 V and 536 nm for FAM, 475 V and 550 nm for HEX, 475 V and 580 nm for TAMRA, and 475 V and 610 nm for ROX). The four fluorophores were detected by two scans using a built-in splitter. The fluorescence intensities of the labeled DNAs were deconvoluted by the method described in (16) and the resultant intensities provide the ratio of DNA amounts in the bound and unbound bands. The relative equilibrium binding constants (K_{ref}) were calculated as the bound to unbound ratio of each mutant O_{mnt} DNA divided by the ratio of the wild-type. The K_{ref} of the wild-type O_{mnt} is equal to 1 by definition.

Sequence logos

The sequence logos were created using a modified version of the Makelogo program (24) in which the bases are not sorted according to their frequency but rather as determined by the user. The letter 'M' is added to indicate the amount of mutual information between the two positions, with half of the total placed above each position.

RESULTS AND DISCUSSION

Using 11 His-tagged variants of the Mnt protein described by Silbaq *et al*. (23), we determined the binding affinity, relative to the wild-type sequence, for all 15 possible variants at operator positions 16 and 17 using the QuMFRA assay described in (16). Briefly, for a specific Mnt protein, each lane of an electrophoretic mobility shift assay (EMSA) gel contains the wild-type operator and three variants, each labeled with a different fluorescent dye. De-convolution of the fluorescence intensities at various wavelengths, for both the bound and unbound bands of the gel, is sufficient to determine the relative affinity of the four operators. Therefore, for a single protein, five EMSA lanes are sufficient to determine the relative affinity to all 16 operator sequences. However, we repeated the measurements at least 3 times for each protein–operator combination to also determine the variability of the measurements. Supplementary Table S1 contains the relative affinity measurements and standard deviations for all of the 12 proteins, including the native Mnt reported in (16), binding to each of the 16 operators. All of the other proteins contain His-tags on the C-terminus, as described in (23). 'H6' is the wild-type protein with the His-tag. Each of the other variant proteins has the amino acid substitution listed following the H6 designation; e.g. H6A has alanine substituted for histidine at position 6. H6S, reported in Silbaq *et al*. (23), was not recovered in sufficient quantity to be used in these analyses, but all of the other variants were included.

Table 1 summarizes all of the relative affinity data as 'specific binding constants', $K_s(b)$, which is the affinity for the particular operator, b, divided by the average value of all operators:

$$K_s(b) \equiv K_{ref}(b) \times \frac{N}{\sum_b K_{ref}(b)}, \qquad \mathbf{1}$$

where N is the total number of operators considered, 16 in this case. $K_s(b)$ is a measure of the specificity of the protein for some binding sites compared to others. For a completely non-specific protein, $K_s(b) = 1$ for every sequence; for a protein that

Table 1. Specific binding constants for Mnt variants

DNA positions		$K_s(b)$ of Mnt proteins											
16	17	Mnt	H6	H6A	H6G	H6I	H6L	H6M	H6N	H6Q	H6R	H6T	H6V
A	A	0.37	0.27	0.77	0.68	7.12	0.93	0.48	1.81	0.76	0.18	2.00	3.25
	C	5.31	4.03	0.23	0.17	0.14	0.13	0.09	1.10	0.14	0.13	0.49	0.33
	G	0.16	0.53	0.44	0.44	0.46	0.95	0.65	1.04	0.61	0.28	0.32	2.42
	T	0.39	0.32	0.22	0.22	0.07	0.27	0.20	1.02	0.22	0.09	0.22	0.13
C	A	1.70	1.58	3.64	1.91	1.49	0.83	0.56	1.83	1.34	4.50	1.97	2.33
	C	3.13	2.77	2.17	0.58	0.15	0.32	0.26	0.76	0.73	1.86	0.53	0.57
	G	0.58	0.33	0.71	0.40	0.23	0.14	0.33	0.55	0.76	1.81	0.72	0.57
	T	0.48	0.43	0.42	0.78	0.07	0.19	0.19	0.74	1.83	2.17	1.04	0.19
G	A	0.16	0.30	0.33	0.40	0.57	2.08	0.53	0.46	0.34	0.53	0.61	0.24
	C	0.69	1.52	0.08	0.10	0.07	0.25	0.20	0.25	0.19	0.10	0.16	0.06
	G	0.16	0.14	0.26	0.33	0.45	1.06	0.61	0.73	0.38	0.15	0.73	0.95
	T	0.21	0.25	0.08	0.34	0.06	0.45	0.17	0.34	0.12	0.08	0.91	0.11
T	A	0.32	0.78	5.44	6.29	4.18	6.52	9.41	1.97	4.86	3.23	3.18	3.63
	C	2.02	2.41	0.25	0.41	0.12	0.19	0.36	0.41	0.64	0.23	0.33	0.13
	G	0.16	0.15	0.71	2.25	0.75	1.31	1.62	2.30	2.46	0.59	2.35	0.99
	T	0.16	0.19	0.22	0.68	0.08	0.38	0.34	0.69	0.61	0.08	0.44	0.10
	I_{spec}	0.99	0.85	1.06	1.05	1.55	0.99	1.56	0.26	0.72	1.03	0.48	0.88
	MI	0.09	0.07	0.13	0.05	0.06	0.09	0.13	0.07	0.13	0.10	0.13	0.11
	CC	0.95	0.95	0.94	0.99	0.99	0.97	0.98	0.85	0.91	0.95	0.90	0.95

binds exclusively to one sequence, $K_s(b) = 16$ for that sequence and 0 for all others. $K_s(b)$ divided by N is the fraction of sequence b that would be bound to the protein in an experiment with DNA excess and all of the operators in equal concentration. We can treat those fractions as probabilities to calculate the informational specificity (25,26), over all 16 operators, for each protein variant. The informational specificity is

$$I_{spec} = 1/N\Sigma_b K_s(b)\log_2 K_s(b). \quad \mathbf{2}$$

The non-specific protein, described above, would have $I_{spec} = 0$, and a protein with complete specificity for one sequence would have $I_{spec} = 4$. In general, I_{spec} will be less than the 'information content' of proteins because it is measured over a complete set of binding sites, not just those that have sufficient affinity to be selected *in vivo* or *in vitro* (25). Table 1 also shows the I_{spec} for each protein. The wild-type protein has about 1 bit total, similar to many of the variants. The His-tagged wild-type protein has slightly decreased specificity, consistent with SELEX results (23). A few proteins have higher specificity, especially H6L and H6M. A few proteins have less, especially H6N which is quite non-specific in its binding compared to the others.

From the specific binding constants, one can also determine the best additive model, the probabilities for each base at each position that provides the best fit to the data while assuming the positions contribute independently (18). Those probabilities are provided for each protein in Supplementary Table S2. For the wild-type proteins, these independent base probabilities are very similar to those determined experimentally using a different assay (25). We can then determine how different the additive model is from the true binding affinities of each dinucleotide variant. A convenient measure of that is the mutual information between the two positions, which is the amount of the total information that is not captured by the best independent model (27). The mutual information between positions 16 and 17 is

$$\mathrm{MI} = \Sigma_{bb} P(b_{16}b_{17})\log_2 P(b_{16}b_{17})/P(b_{16})P(b_{17}), \quad \mathbf{3}$$

where $P(b_{16}b_{17})$ is the true probability distribution for all 16 dinucleotides at positions 16 and 17, and $P(b_{16})$ and $P(b_{17})$ are the best independent base probability models for each position. MI is also included in Table 1. As with the wild-type Mnt (16), all of the protein variants show non-independence between the positions. Also, as with Mnt (18), the amount of mutual information is generally a small fraction, usually <10%, of the total information. The proteins that are less specific show the greatest fraction of the total information that is non-independent, as was seen with several zinc-finger proteins (17,18). This close approximation between the additive model and the true data can also be seen by calculating the correlation coefficient between the true probability values for each dinucleotide with those predicted from the independent model. These correlation coefficients are also shown in Table 1 (CC) and are all above 0.9 except for the least specific protein, H6N. In fact, most of the CC values are above 0.95. These results confirm the conclusions of Benos *et al*. (18) that even when protein–DNA interactions are not completely additive, such simple additive models can still be quite good approximations to the true binding probabilities.

Using the independent base probabilities at each position, we can also calculate the informational specificity for each position separately. A sequence logo is a convenient method for displaying the specificity of a DNA binding protein (24). Figure 1 shows the standard logo for each protein variant using the independent base probabilities, and it also includes the amount of mutual information which is displayed as the letter M over each column in the logo (half of the total mutual information is plotted over each position). A similar plot has been used previously to show the mutual information in conserved RNA structures, which are often quite large compared to the information content of the individual positions (28). These plots help to make several points.

(i) MI is small compared to I_{spec} except for the most non-specific proteins. This is quite different from RNA binding proteins where the interaction may depend more on the structure than the sequence and therefore have higher mutual information than information content (28).

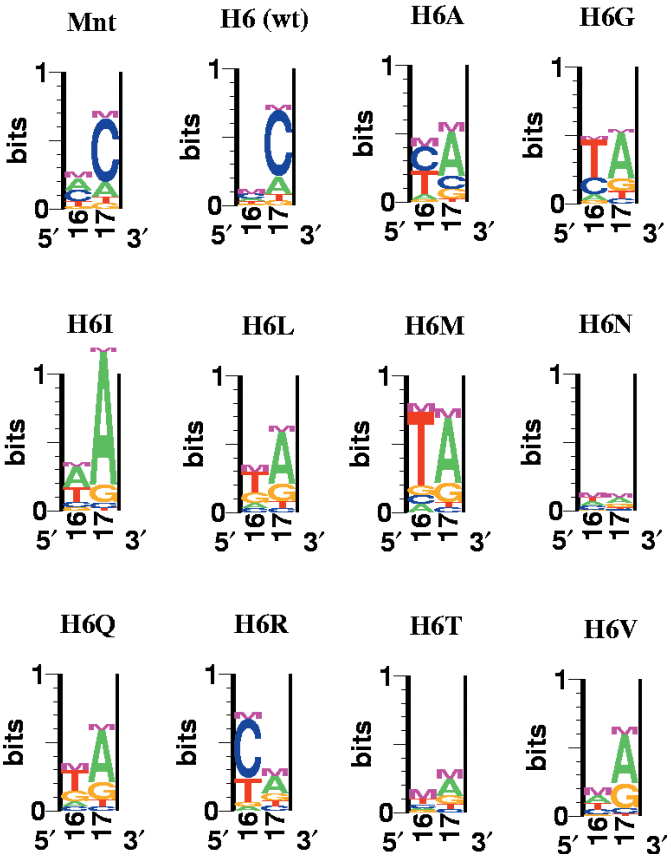

Figure 1. Logo representations (24) for the binding specificity of each Mnt variant. 'Mnt' refers to the native protein; all of the other proteins have a His-tag at their C-terminus. 'H6' is the wild-type protein with the His-tag and each amino acid replacement is indicated by its name. The letter 'M' in the logos refers to the mutual information between positions 16 and 17 for the protein, with half of the total value placed at each position (see text).

(ii) Every protein without H6 prefers A at position 17, instead of the wild-type C.

(iii) Most proteins prefer T at position 16, H6M quite strongly. H6A prefers C and T almost equally, and H6I prefers A and T almost equally.

(iv) Most of the proteins still interact primarily with position 17, although most show increased interaction with position 16 compared to the wild-type protein. H6Q and H6R both have stronger preferences at position 16. H6Q is the equivalent to the N-terminus of the Arc protein, which also prefers the sequence TA at the equivalent positions (22,29). The result with H6R is a special case, as also seen in the SELEX data (23), in that the primary interaction is now with position 16 where a C is preferred. This is consistent with our previous hypothesis (23) that the spacing between the protein at amino acid 6 and the DNA at position 17 is too small to accommodate an arginine, but at the adjacent base pair there is a suitable geometry for a strong interaction.

In general, the quantitative results we obtained in this study are consistent with the binding sites selected *in vitro* using the same set of proteins (23), although there are some significant differences. These can be attributed to at least two sources. In the SELEX study the entire binding site region was randomized; while the inner section of the binding site, positions 7–15, was highly conserved for all proteins, the outer positions, 3–6 and 16–19, were all variable and differed depending on the protein used in the selection. In our study, where all of the positions are constant except 16 and 17, the context is not optimal for some proteins, and that may alter the preference for these positions. For example, in the SELEX data, H6N was highly specific for TA at positions 16 and 17, but it also had a preference for TA at positions 18 and 19 where our construct contains CT at those positions. Furthermore, those H6N SELEX binding sites all had an overlapping site that may have contributed to the affinity, and the results were compiled from a fairly small sample. However, even in that case, we get the same preferred sequence of TA; the specificity is just much lower than we would have predicted based on the SELEX results. In nearly every case, we get the same preferred sequence in the two experiments, an exception being H6I where SELEX gave a strong preference for TA, and we now find a weak preference for AA over TA.

Based on the binding probabilities for all dinucleotides, we can determine the similarity in the specificity between different amino acids. Table 2 shows the correlation coefficient between all pairs of amino acid probability vectors (i.e. between the columns of Table 1). As one would expect based on the logos in Figure 1, the wild-type protein has quite different specificity than every other protein. It is negatively correlated with all of them except for H6A and H6R which have low positive correlations of 0.10 (or even somewhat lower with the native Mnt). On the other hand, all of the other amino acids have positive correlations with each other, consistent with the fact that they all bind to TA better than average [i.e. K_s(TA) > 1], and they tend to have other preferences in common too. If one clusters all of the amino acids with correlations above 0.9 (single link clustering), one group emerges containing A, G, L, M and Q. None of the other amino acids are linked to one another. If the threshold for clustering is reduced to 0.8, then the group expands to include R, T and N, and I and V form another group, but the wild-type H remains alone. In fact, all amino acids except H merge into one cluster at a cutoff of 0.73, but H does not until the cutoff is reduced to 0.1. Figure 2 is a diagram of the clusters, with the stronger correlations (CC > 0.9) indicated by solid lines and the weaker ones (0.8 < CC < 0.9) indicated by dotted lines. These groups overlap but are not identical to those obtained with the SELEX data (23). In particular, H was not so distinctive in the SELEX experiments and grouped with A and T, where R was the most distinctive with its strong preference for C at position 16. In these quantitative studies, C is still preferred at 16, but the overall probability distribution is more similar to the other amino acids than in the SELEX data. Consistent with the SELEX results, it is seen that amino acids with quite different physical properties can have very similar preferences for binding sites, indicating that a specific base pair can be selected for different reasons. For example, while it is not surprising that L and M should have very similar specificities, it is unexpected for G and Q to have such strong similarities to each other and to L and M. This indicates that amino acids with distinct characteristics can select the same base pairs, presumably through interactions with different features. We suppose that G, L and M probably selected T through hydrophobic

Table 2. Correlations between binding probabilities for Mnt variants

	Correlation coefficients of Mnt proteins											
	Mnt	H6	H6A	H6G	H6I	H6L	H6M	H6N	H6Q	H6R	H6T	H6V
Mnt	1.00	0.96	0.06	−0.14	−0.18	−0.26	−0.18	−0.03	−0.20	0.08	−0.21	−0.18
H6		1.00	0.10	−0.10	−0.19	−0.22	−0.13	−0.08	−0.17	0.10	−0.22	−0.19
H6A			1.00	0.91	0.44	0.75	0.79	0.58	0.78	0.83	0.75	0.68
H6G				1.00	0.47	0.91	0.95	0.61	0.93	0.67	0.84	0.66
H6I					1.00	0.50	0.46	0.61	0.42	0.19	0.68	0.82
H6L						1.00	0.96	0.48	0.83	0.40	0.74	0.63
H6M							1.00	0.50	0.90	0.44	0.75	0.60
H6N								1.00	0.65	0.42	0.83	0.73
H6Q									1.00	0.60	0.86	0.58
H6R										1.00	0.57	0.52
H6T											1.00	0.73
H6V												1.00

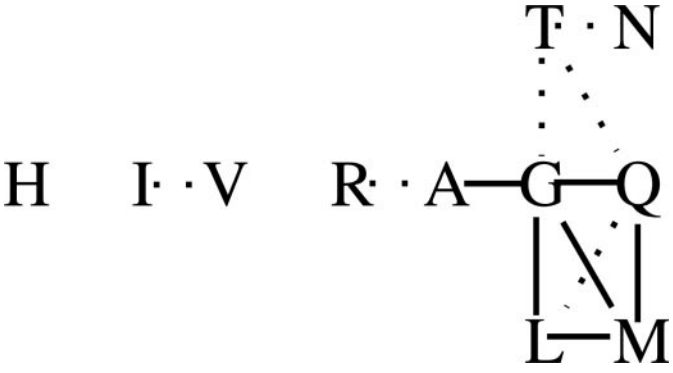

Figure 2. Similarities of the specificities for the amino acid variants. Amino acids with correlations greater than 0.9 are linked by a solid line. Those with correlations between 0.8 and 0.9 are linked by a dotted line.

interactions with the T-methyl group, whereas Q can form hydrogen bonds with the T–A base pair.

The term 'recognition code' refers to modeling the relationship between the amino acid sequence of a TF and the base sequence of a DNA binding site. With a perfect recognition code one could quantitatively predict the binding specificity of a TF from its amino acid sequence. However, all attempts to uncover such a model have fallen well short of the goal. From SELEX and phage-display selections for proteins of the early growth response (EGR) family of zinc-finger transcription factors, we developed a quantitative recognition code that showed moderate success at predicting quantitative specificity for a test set of proteins (12). This method took advantage of the known pattern of interactions, which are that the amino acids at positions −1, 2, 3 and 6 of the α-helix of the zinc-finger protein, interact with positions 3, 4, 2 and 1, respectively, of the binding site. For each of those amino acid: base pair interactions (−1:3; 2:4; 3:2; 6:1) we obtained a 4 × 20 table of predicted binding energies, from which one can determine the relative affinity of each base pair for each amino acid, according the model for zinc-finger–DNA interactions (12). Each of the four amino-acid:base-pair interactions had a distinct relationship, as is expected from the distinct geometries of their interactions. We can compare each of those tables with the relative affinities we obtained for different variants of the Mnt protein, using the best estimates of the independent base contributions (Supplementary Table S2). Comparing the results of our study, using the independent models for each protein, we find that there is a strong correlation, 0.74, between position 17 (using the complementary strand as the basis for comparison) and position 2 of the binding sites for the zinc-finger proteins. Based on our previous SELEX results, we noticed that qualitatively the interaction of Mnt with operator position 17 was more similar to the interaction of zinc-finger proteins with position 2 of their binding sites than with either positions 1 or 3 (23). We now see that quantitatively, over the set of 11 amino acids in our set of Mnt variants, the correlation between *mnt* operator position 17 with zinc-finger binding site position 2 is much higher than for positions 1, 3 or 4, which are all between −0.22 and 0.28 (Supplementary Table S3). However, we cannot simply use position 2 from the zinc finger sites as a substitute for Mnt position 17; a correlation of 0.74 is not nearly as high as we would like and, furthermore, the correlation varies considerably for different amino acids. If we calculate correlations for each amino acid across the probability distributions of all four bases, there are several with very high correlations, 0.89 to 0.99 for A, G, H, I, L, V, a few with more moderate correlations, 0.67 to 0.78 for M, R, T, and two with negative correlation, N, Q. All of the amino acids with weak and moderate correlations show higher correlations between *mnt* position 16 and zinc-finger position 1, where H6R has a very high correlation of 0.94. Overall the correlation is only moderate, 0.38 (Supplementary Table S3), but several amino acids show high correlations between the *mnt* position 16 and zinc-finger binding site position 1, including H6 which has a correlation of 0.98. In zinc fingers His is not often used in protein position 6, which interacts with base position 1, and it does not show much specificity for different bases, only a slight preference for A and C [on the complementary strand; (12)]. This is exactly the case for the *mnt* operator position 16, where there is a weak preference for A and C [Figure 1; (25)]. Therefore, we believe that, in addition to the contacts between Mnt and its operator in the model of Raumann *et al*. (22), there may be an interaction between the H6 from the other strand of the β-ribbon with position 16. This would make the Mnt interaction with its operator more similar to the Arc interaction with its operator, where the homologous Gln amino acids on both β-strands interact with both positions 16 and 17 of the operator (22,29). It also helps to explain how variants of H6 dramatically effect the specificity of both positions 16 and 17

(22,23). How the protein variants interact with those positions then also influences how the outer section of the operator interacts with the protein, so that the change in specificity extends to include positions 18 and 19 as well (23).

CONCLUSIONS

Using the QuMFRA assay (16) we have obtained the quantitative specificity for a set of 11 variants of the Mnt repressor protein for each of the 16 different operator sequences. By examining all possible base combinations at two adjacent positions we have shown that each protein interacts with the two positions in a non-independent manner, but a simple independent model provides a close approximation to the true binding probabilities. In searching for recognition codes for various protein families, it is important to know whether independent models will suffice as more complex models are possible but require many more parameters to be estimated. For Mnt, positions 16 and 17 were much more highly correlated than any other pair of positions in the data from a SELEX experiment (25), so the fact that they can be fit fairly well with an independent model indicates that such a model should work for the entire protein.

Assuming that independent contributions of the positions works well in general, an efficient strategy to determine the quantitative specificity of a protein for its entire set of high affinity binding sites emerges. The protein can be used in a SELEX experiment to obtain a collection of binding sites, all with reasonably high affinity. There should be only a few rounds of selection so that the sites still maintain considerable variability. Each site should be similar enough to the consensus that they are easily aligned, but an average of two to four differences from the consensus will ensure a large sampling of the high affinity sequence set. Then each selected site can be assayed using QuMFRA to determine their quantitative relative affinities, and if desired one site can be used to determine its absolute affinity thereby providing the absolute affinity to every site in the collection. Assays on at least 50 sites could easily be determined in a single gel, and the quantitative binding data could be used to calculate a quantitative specificity model for the protein. Measuring the fit of the model back to the quantitative model will also give an indication of whether the independence assumption is reasonably accurate (15,18). The whole SELEX experiment could then be repeated for variants of the protein in order to develop a quantitative recognition code for the protein family. One important unknown is how well the independence assumption works for the protein side of the interaction. So far most tests have assayed whether changes in the positions of the DNA sites are approximately independent, but equally important for developing general recognition models is whether changes in the amino acids of the protein contribute independently to the binding. It is unlikely to be exactly true, but as with the DNA side, a good approximation to independence would allow for simple models to provide reasonably accurate predictions of specificity for other proteins. This would, in turn, help to uncover the regulatory networks in genome sequences by suggesting which TFs are most likely to interact with which regulatory motifs, currently a major challenge in genome annotation and cellular modeling.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Frauzi Silbaq for the purified His-tagged Mnt proteins used in this study and to Xing Xu for help with the figures. This work is supported by the National Institutes of Health through grant GM28755.

REFERENCES

1. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein–DNA complexes. *Genome Biol.*, **1**, REVIEWS001.
2. Stormo,G.D. and Tan,K. (2002) Mining genome databases to identify and understand new gene regulatory systems. *Curr. Opin. Microbiol.*, **5**, 149–153.
3. Alexander,M.K., Bourns,B.D. and Zakian,V.A. (2001) One-hybrid systems for detecting protein–DNA interactions. *Methods Mol. Biol.*, **177**, 241–59.
4. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E., Volkert,T.L., Wilson,C.J., Bell,S.P. and Young,R.A. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
5. Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J., Mermod,N. and Bucher,P. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831–835.
6. Bulyk,M.L., Huang,X., Choo,Y., Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
7. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Is there a code for protein–DNA recognition? Probab(ilistical)ly. *Bioessays*, **24**, 466–475.
8. Matthews,B.W. (1988) Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
9. Choo,Y. and Klug,A. (1997) Physical basis of a protein–DNA recognition code. *Curr. Opin. Struct. Biol.*, **7**, 117–125.
10. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.
11. Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein–DNA binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.
12. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
13. Suzuki,M., Brenner,S.E., Gerstein,M. and Yagi,N. (1995) DNA recognition code of transcription factors. *Protein Eng.*, **8**, 319–328.
14. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
15. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
16. Man,T.-K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new Quantitative Multiple Fluorescence Relative Affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
17. Bulyk,M., Johnson,P.L. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
18. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
19. Gregoret,L.M. and Sauer,R.T. (1993) Additivity of mutant effects assessed by binomial mutagenesis. *Proc. Natl Acad. Sci. USA*, **90**, 4246–4250.
20. Raumann,B.E., Brown,B.M. and Sauer,R.T. (1994) Major groove DNA recognition by β-sheets: the ribbon–helix–helix family of gene regulatory proteins. *Curr. Opin. Struct. Biol.*, **4**, 36–43.

21. Youderian,P., Vershon,A., Bouvier,S., Sauer,R.T., and Susskind,M.M. (1983) Changing the DNA-binding specificity of a repressor. *Cell*, **35**, 777–783.
22. Raumann,B.E., Knight,K.L. and Sauer,R.T. (1995) Dramatic changes in DNA-binding specificity caused by single residue substitutions in an Arc/Mnt hybrid repressor. *Nature Struct. Biol.*, **2**, 1115–1122.
23. Silbaq,F.S., Ruttenberg,S.E. and Stormo,G.D. (2002) Specificity of Mnt 'master residue' obtained from *in vivo* and *in vitro* selections. *Nucleic Acids Res.*, **30**, 5539–5548.
24. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
25. Fields,D.S., He,Y-Y., Al-Uzri,A.Y. and Stormo,G.D. (1997) Quantitative specificity of Mnt repressor. *J. Mol. Biol.*, **271**, 178–194.
26. Stormo,G.D. and Fields,D.S. (1998) Specificity, energy and information in DNA–protein interactions. *Trends Biochem. Sci.*, **23**, 109–113.
27. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–57.
28. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.*, **13**, 583–586.
29. Raumann,B.E., Rould,M.A., Pabo,C.O. and Sauer,R.T. (1994) DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. *Nature*, **367**, 754–757.