A Statistical Model for Investigating Binding Probabilities of DNA Nucleotide Sequences Using Microarrays

Mei-Ling Ting Lee[1,2,3*], Martha L. Bulyk[4,5], G.A. Whitmore[6], and George M. Church[4,5]

1: Channing Laboratory, Brigham & Women's Hospital, 181 Longwood Ave, Boston, MA 02115
2: Harvard Medical School, Boston, MA, USA
3: Biostatistics Department, Harvard School of Public Health, Boston, MA, USA
4: Graduate Biophysics Program, Harvard University, Cambridge, MA, USA
5: Department of Genetics, Harvard Medical School, Boston, MA, USA
6: McGill University, Montreal, Quebec, Canada
∗: email: meiling@channing.harvard.edu

SUMMARY: There is considerable scientific interest in knowing the probability that a site-specific transcription factor will bind to a given DNA sequence. Microarray methods provide an effective means for assessing the binding affinities of a large number of DNA sequences as demonstrated by Bulyk *et al* (2001) in their study of the DNA-binding specificities of Zif268 zinc fingers using microarray technology. In a follow-up investigation, Bulyk, Johnson and Church (2002) studied the interdependence of nucleotides on the binding affinities of transcription proteins. Our paper is motivated by this pair of studies. We present a general statistical methodology for analyzing microarray intensity measurements reflecting DNA-protein interactions. The log-probability of a protein binding to a DNA sequence on an array is modeled using a linear ANOVA model. This model is convenient because it employs familiar statistical concepts and procedures and also because it is effective for investigating the probability structure of the binding mechanism.

KEYWORDS: Binding probability, dependence, multi-stage ANOVA, log-linear model, log-probability model, microarrays, protein, transcription factor

1. Background

Genes are regulated in part through the action of site-specific transcription factors. The DNA sequences at which these proteins bind are not unique. Binding can also occur at variants of the optimal site with some variants being more preferential for binding than others. There is considerable scientific interest in knowing the binding affinity of a sequence, i.e. the probability that a protein will bind to a given sequence. Micoarray methods permit the assessment of the binding affinities of a large number of DNA sequences in a single study. They allow the estimation of binding affinities across the full range from very low to very high affinities. This estimation is valuable because it is possible that even low-affinity DNA sites are functionally important in transcriptional regulation of gene expression.

Bulyk, Huang, Choo and Church (2001) studied the DNA-binding specificities of Zif268 zinc fingers using microarray technology in an experiment that tested the feasibility of using microarrays for analyzing large numbers of DNA-protein interactions. They employed the full set of 64 possible 3-bp (base pair) sequences in their study but noted that "A full set of sequences spanning all possible 8-bp binding sites would consist of roughly 65,000 spots which could fit onto a single microscope slide" (p. 7163). Thus, the cost of DNA synthesis would appear to be the only practical limitation to generalizing the experimental method from a scientific point of view. We return to consider the implications of large scale studies for statistical analysis later.

In a follow-up investigation, Bulyk, Johnson and Church (2002) studied the interdependence of nucleotides in the binding affinities of transcription proteins using several techniques, including t-tests and a hidden Markov model. Two statistical problems are of interest in this kind of context. First, it is of interest to test whether the probability of a protein binding to any sequence of nucleotides, representing a transcription site, equals the mathematical product of the probabilities for the individual nucleotides in the sequence. This equality would indicate that binding is probabilistically independent from one nucleotide to another. Second, where the hypothesis of independence is rejected, it is of interest to study the type of dependence in the sequence that is predictive of binding affinity. The analysis of interdependence for transcription binding sites has received attention from a number of scientific investigators. Man and Stormo (2001), for example, found dependence for nucleotides at positions 16 and 17 in the 21-bp binding site of *Salmonella* bacteriophage repressor Mnt. Related background work on methodologies for studying DNA binding site preferences may be found in Staden (1988), Stormo, Schneider and Gold (1986), Zhang and Marr (1993), Ponomarenko *et al* (1999), Wingender, Karas and Knuppel (1997) and Quandt, Frech, Karas, Wingender and Werner (1995).

We present a general statistical methodology for analyzing microarray intensity measurements reflecting DNA-protein interactions. The log-probability of a protein binding to a DNA sequence on an array is modeled using a linear analysis of variance (ANOVA) model. This model is convenient because it employs familiar statistical concepts and procedures and also because it is especially effective for investigating the probability structure of the binding mechanism. For example, the hypothesis of whether DNA binding sites are mainly composed of independent nucleotide selections is easily tested by our model. We show how the dependence structure of the binding probability distribution can be explored by statistical methods. We propose several measures for describing the binding affinity and dependence structure. The strategic adaptation of the methodology to cope with large scale investigations (long nucleotide sequences) is discussed.

2. Log-probability ANOVA Model

Each nucleotide in a sequence defining a potential binding site is represented by one letter from the nucleotide alphabet $\{A, C, G, T\}$. We denote the DNA nucleotide sequence at a given spot on the array by index $\boldsymbol{a} = \{a_1, a_2, \cdots, a_l\}$, where a_j denotes the nucleotide at position j in the sequence. Each sequence is assumed to be of length l, where $l > 1$. The sequence index $\boldsymbol{a}$ ranges over a set $\mathcal{A}$, which may contain up to 4^l possible sequences. Some or all of the possible sequences may appear on an array, including replicates. In developing the model initially, we assume that the array contains all possible sequences, possibly replicated in equal numbers, so the experimental design is balanced. Alternative designs that relax this requirement are described in the last section.

Our modeling begins with the observation that microarray intensity measurement at a spot (for example, a fluorescence measure) varies in proportion to the number of fundamental binding events that occur between the target protein and the DNA sequence on the spot. The number of binding events occurring on a spot containing sequence $\boldsymbol{a}$ is denoted by random variable $N_{\boldsymbol{a}}$. We assume that the mean of $N_{\boldsymbol{a}}$ equals $\mathcal{N}\ P_{\boldsymbol{a}}$, where $\mathcal{N}$ is a fixed large count that can be interpreted as the maximum or *saturation* number of potential binding events that might occur at any spot on the array, and $P_{\boldsymbol{a}} = P(\boldsymbol{a})$ is the probability that the protein molecule binds on sequence $\boldsymbol{a}$. We assume that the experiments are performed in a way that avoids the saturation threshold for any spot, i.e., that $P_{\boldsymbol{a}}$ is not too large. We do not specify an exact distribution form for $N_{\boldsymbol{a}}$ but we anticipate that it is related to the binomial distribution family.

We denote the microarray intensity at the spot with sequence $\boldsymbol{a}$ by $W_{\boldsymbol{a}}$ and assume that the intensity is proportional to $N_{\boldsymbol{a}}$, i.e., $W_{\boldsymbol{a}} = c\ N_{\boldsymbol{a}}$ for some constant c. Thus, $W_{\boldsymbol{a}}$ has mean value $E(W_{\boldsymbol{a}}) = c\ \mathcal{N}\ P_{\boldsymbol{a}}$. Letting $Y_{\boldsymbol{a}}$ denote the natural log-intensity, it follows that $Y_{\boldsymbol{a}}$ and binding

probability $P_{\boldsymbol{a}}$ for sequence $\boldsymbol{a}$ are related as follows under our proposed model.

$$Y_{\boldsymbol{a}} = \ln W_{\boldsymbol{a}} = \ln\ (c\,\mathcal{N}) + \ln P_{\boldsymbol{a}} + \epsilon_{\boldsymbol{a}} \tag{1}$$

Here $\ln(c\,\mathcal{N})$ is a parameter that is assumed to be invariant with respect to $\boldsymbol{a}$ and $\epsilon_{\boldsymbol{a}}$ is a statistical error term which we take to be independent for different sequences $\boldsymbol{a}$. The error term $\epsilon_{\boldsymbol{a}}$ will not be exactly normally distributed with zero mean and constant variance for all sequences $\boldsymbol{a}$ but we anticipate that these conditions will hold to an adequate approximation for reliable inferences to be drawn from the ANOVA model. In any case, the conditions can be verified in each application, as we will demonstrate later with our case application.

We use the following log-linear ANOVA model to describe the relation between the binding log-probability $\ln P_{\boldsymbol{a}}$ and the nucleotide structure of sequence $\boldsymbol{a} = \{a_1, \ldots, a_l\}$ (Agresti, 1996; McCullagh and Nelder, 1989). We refer to this model as the *full dependence* ANOVA model.

$$\ln P_{\boldsymbol{a}} = \beta_0 + \sum_{i=1}^{l} \beta_{i,a_i} + \sum_{i=1}^{l-1} \sum_{j>i}^{l} \beta_{ij,a_i a_j} + \sum_{i=1}^{l-2} \sum_{j>i}^{l-1} \sum_{k>j}^{l} \beta_{ijk,a_i a_j a_k} + \cdots \tag{2}$$

The successive terms on the right-hand side of the equality sign here represent a constant term (β_0), main effects (β_{i,a_i}) and interaction effects ($\beta_{ij,a_i a_j}, \beta_{ijk,a_i a_j a_k}$, etc.) of all possible orders for sequences of length l. The index notation, (i, a_i), $(ij, a_i a_j)$, $(ijk, a_i a_j a_k)$, etc. refers to the particular nucleotides in sequence $\boldsymbol{a} = \{a_1, \ldots, a_l\}$ corresponding to positions i, j, k, etc. For instance, if $l = 3$ nucleotides, sequence $\boldsymbol{a} = \{G, A, T\}$ and $(i, j) = (2, 3)$ then $\beta_{ij,a_i a_j} = \beta_{23,AT}$ represents the two-way interaction effect of nucleotides A and T at positions 2 and 3 in sequence $\boldsymbol{a} = \{G, A, T\}$.

The ANOVA model may be fitted by ordinary least squares. The sets of least squares estimates of the main effects and interaction effects in (2) are subject to estimability constraints, such as the sum-to-zero constraint or set-last-to-zero constraint. The constant term of the fitted model will be an estimate of the sum of the leading constant in (1) and the leading constant β_0 in (2), i.e., an estimate of $\ln(c\mathcal{N}) + \beta_0$. The presence of the constant $\ln(c\mathcal{N})$ implies that a binding probability can be determined from the microarray intensity data only up to an unknown multiplicative constant. Thus, ratios of probabilities can be estimated but not the actual probabilities themselves.

With a balanced experimental design, where all sequences of length l are present in equal numbers, the design components corresponding to the sets of regression coefficients of successively higher order in model (2) are orthogonal. Therefore, they can be estimated successively, proceeding from lower to higher orders. For example, the sets of main effects β_{i,a_i} and second- and third-order interaction effects, $\beta_{ij,a_i a_j}$ and $\beta_{ijk,a_i a_j a_k}$, will have independent estimators with a balanced design. We will demonstrate the value of this property later.

The full dependence ANOVA model (2) has 4^l parameters. Referring to the constant term, main effects and successively higher interaction terms as effects of order $0, 1, 2, \ldots, k, \ldots, l$, respectively, we can count parameters of each order using the following binomial expansion.

$$4^l = (3+1)^l = \sum_{k=0}^{l} \binom{l}{k} 3^k$$

A brief tabulation of these counts is found in Table 1. We shall refer to this table later in discussing strategies for dealing with large scale studies.

Model (2) describes the *complete* joint binding probability distribution for a protein with respect to all nucleotide sequences of length l. The model is therefore more comprehensive than Markov models of such sequences. Once the parameters of the ANOVA model are estimated, the corresponding binding probability of any sequence of length l can be estimated, up to an unknown constant. The full dependence ANOVA model corresponds to an l-way contingency table having 4 levels (the four possible nucleotides) for each of l factors (the l nucleotide sites). The literature on log-linear models contains an extensive discussion of how the model captures various degrees and forms of independence and conditional independence in the joint probability distribution. For some discussion, refer to McCullagh and Nelder (1989:215-217) and Agresti (1996:146-152). It is precisely the estimation and examination of the dependence structure of this joint distribution that interests us in this paper. We turn to this topic in the next section.

Before leaving the presentation of the log-probability ANOVA model, we note that in some studies there may be a need to incorporate other factors in model (2) to account for identifiable sources of variability such as pin tip or array effects. These kinds of factors can be added to the model as additional main effects as required. We shall not consider this type of extension further.

3. Examining Dependence in Protein Binding Affinity

If the binding probability for a whole sequence $\boldsymbol{a}$ equals the product of the probabilities for the individual nucleotides making up the sequence, i.e., $P_{\boldsymbol{a}} = \prod_{i=1}^{l} P_{a_i}$ where $\boldsymbol{a} = \{a_1, \cdots, a_l\}$ and $P_{a_i} = P(a_i)$, then ANOVA model (2) for binding probability reduces to the following form, which contains only the main effects in model (2). We refer to this model as the *independence* ANOVA model.

$$\ln P_{\boldsymbol{a}} = \beta_0 + \sum_{i=1}^{l} \beta_{i,a_i} \tag{3}$$

Thus, the test of whether the nucleotides making up a binding sequence are probabilistically independent is equivalent to testing if all interaction parameters in (2) are zero.

3.1 Testing for Independence

To implement our methodology, we first fit regression model (1) with component $\ln P_{\boldsymbol{a}}$ having the independence ANOVA form (3). This first-stage ANOVA analysis might be viewed as a normalization of the microarray intensities that takes account of the relative abundance of bindings for each nucleotide in a given sequence position. For any sequence $\boldsymbol{a}$, the fitted response of the log intensity $Y_{\boldsymbol{a}}$ from this ANOVA model is denoted by $\widehat{Y}^I_{\boldsymbol{a}}$, where superscript I denotes *independence*. Next, in a second-stage analysis, we fit regression model (1) with component $\ln P_{\boldsymbol{a}}$ having the full dependence ANOVA form in (2) with all main effects and higher-order interaction effects. The fitted response for sequence $\boldsymbol{a}$ for this model is denoted by $\widehat{Y}^D_{\boldsymbol{a}}$, where superscript D denotes *dependence*.

A comparison of the ANOVA results for the two models allows us to test whether one or more interaction effects are significantly different from zero and, hence, if there is some probabilistic binding dependence among the nucleotides for any sequence. Simultaneous inference methods, such as Bonferroni procedures, can then be used to identify which particular interaction terms (or, possibly, combinations of interaction terms) is producing the significant test result. In this way, the dependence can be isolated to particular sequences.

3.2 Dependence Probability Ratio of a Sequence

A sequence $\boldsymbol{a}$ that has a binding probability larger than expected under the independence assumption, i.e., $P_{\boldsymbol{a}} > \prod_{i=1}^{l} P_{a_i}$ where $\boldsymbol{a} = \{a_1, \cdots, a_l\}$, might be viewed as a binding *attraction* of the protein to the sequence as a whole. Similarly, a sequence $\boldsymbol{a}$ with a binding probability that is smaller than expected under independence, i.e., $P_{\boldsymbol{a}} < \prod_{i=1}^{l} P_{a_i}$, might be viewed as a binding *aversion* by the protein for the sequence. Both attraction and aversion are indications of dependence and both may be scientifically important.

If there is dependence then it can be studied for selected sequences of interest by calculating the following ratio, which we refer to as a *dependence probability ratio*.

$$\text{Dependence Probability Ratio} = \frac{\widehat{P}^D_{\boldsymbol{a}}}{\widehat{P}^I_{\boldsymbol{a}}} = \exp(\widehat{Y}^D_{\boldsymbol{a}} - \widehat{Y}^I_{\boldsymbol{a}}) \tag{4}$$

The symbols $\widehat{P}^D_{\boldsymbol{a}}$ and $\widehat{P}^I_{\boldsymbol{a}}$ denote estimates of the binding probability for sequence $\boldsymbol{a}$ under the full dependence and independence models, respectively. Although the leading unknown constant in model (1) makes it impossible to estimate the binding probability of an individual sequence, the ratios of such probabilities can be estimated and that is precisely what is given here in (4). The difference $\widehat{Y}^D_{\boldsymbol{a}} - \widehat{Y}^I_{\boldsymbol{a}}$ equals the sum of the estimated interaction terms of various orders, provided the independent and full dependence models employ the same estimability constraints. Equivalently, $\widehat{Y}^D_{\boldsymbol{a}} - \widehat{Y}^I_{\boldsymbol{a}}$ is the fitted value for sequence $\boldsymbol{a}$ when the residuals from the independence model are fitted with an ANOVA model containing the interaction terms of the full dependence model. This

fact makes it easy to compute an estimated standard error for the difference and, hence, for the dependence probability ratio itself.

3.3 Comparing Binding Affinities of Two Sequences

Fitting the full dependence model also allows us to compare the binding affinities of different DNA sequences. Specifically, for any pair of sequences that may interest an investigator, say $\boldsymbol{a}_1 = \{a_{11}, \cdots, a_{1l}\}$ and $\boldsymbol{a}_2 = \{a_{21}, \cdots, a_{2l}\}$, we can estimate the ratio of their binding probabilities as follows.

$$\text{Relative Affinity} = \frac{\widehat{P}^D_{\boldsymbol{a}_2}}{\widehat{P}^D_{\boldsymbol{a}_1}} = \exp(\widehat{Y}^D_{\boldsymbol{a}_2} - \widehat{Y}^D_{\boldsymbol{a}_1}) \tag{5}$$

We refer to this ratio as the *relative affinity* of sequence $\boldsymbol{a}_2$ to sequence $\boldsymbol{a}_1$. The difference $\widehat{Y}^D_{\boldsymbol{a}_2} - \widehat{Y}^D_{\boldsymbol{a}_1}$ is a comparison of two 'treatments', $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, in the usual experimental sense. Hence, an estimated standard error can be constructed for the difference in the conventional manner for comparisons and this standard error, in turn, can be converted into one for the relative affinity itself.

The reference sequence for calculating relative affinities, namely, sequence $\boldsymbol{a}_1$ in (5), may be chosen in many ways. One choice that we use in our case application is to select that sequence for which $Y^D_{\boldsymbol{a}_1}$ is largest among all sequences. Then the relative affinity for any sequence will describe the probability of binding for that sequence relative to the 'most successful' sequence.

3.4 Conditional Probability of Binding

A closely related measure of interest is the *estimated conditional probability of binding* among a reference set of sequences. If $\boldsymbol{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\}$ is a set of m distinct sequences of interest, where $\boldsymbol{b}_j = (b_{j1}, \cdots, b_{jl})$, then the following formula gives the estimated conditional probability that sequence $\boldsymbol{b}_j \in \boldsymbol{B}$ will be the actual bound sequence in a given binding event, conditional on the binding being one of the sequences in $\boldsymbol{B}$.

$$\widehat{P}(\boldsymbol{b}_j | \boldsymbol{B}) = \frac{\exp(\widehat{Y}^D_{\boldsymbol{b}_j})}{\sum_{i=1}^m \exp(\widehat{Y}^D_{\boldsymbol{b}_i})} \tag{6}$$

4. Case Application

4.1 The Experiment

As a case application, we consider the study described in Bulyk *et al* (2001) and Bulyk, Johnson and Church (2002). In this study, a wild-type and four mutant Zif268 zinc fingers were bound to microarrays. The mutant variants are abbreviated here as KASN, RGPD, LRHN and REDV. The DNA spotted onto the the microarrays were 37 bp long with only the three central nucleotides varying. It was the $4^3 = 64$ possible arrangements of these three central nucleotides that were

examined in this study. Nine replicates were spotted on the slide for each triplicate sequence with the exception of the KASN variant where only five replicates were used.

DNA fluorescence intensities were obtained for all possible triplets of nucleotides as follows. R-phycoerythrin was used for detection of bound proteins. It is an excitable protein which emits fluorescently. Except for measurement variability, the fluorescence level will be directly proportional to the number of bound proteins. The microarrays were scanned at multiple laser power settings to ensure that fluorescence intensities remained below the saturation level of the microarray scanner. Background levels were first subtracted from all raw signal intensities. Then, the relative signal intensity of each of the spots within a replicate was calculated as a fraction of the highest signal intensity for a spot containing one of the 64 different 37 bp sequences. To normalize for possible variability in the DNA concentrations of the different DNA samples that were spotted onto the microarrays, each of the average relative signal intensities from zinc finger phage binding was divided by the respective average relative signal intensities from SybrGreen I staining. The resulting measure, referred to as *frac SI* (fractional signal intensity) in Bulyk *et al* (2001), is proportional to binding affinity. Hence, for our purposes here, we can take $Y = \log(\textit{frac SI})$ as our response variable in the ANOVA model.

4.2 Two-Stage ANOVA

To illustrate the methodology, we consider the REDV variant. According to Choo and Klug (1994), nucleotide sequences GCG and GTG were used to select this variant from the phage display library. We thus expect these sequences to exhibit high binding probabilities. We wish to establish that this is the case and also to check for probabilistic binding dependence.

Proceeding with the two stages of analysis, we first estimate the independence ANOVA model using (3) and then the full dependence ANOVA model using (2). Table 2 shows the combined ANOVA table. Lines 1-4 and 11 give the ANOVA results under independence. Lines 1-3, 5-7 and 9-11 give the ANOVA results under full dependence. For the moment, ignore the intermediate Error (level 2) given in line 8. The results for the two ANOVA models can be combined in a single table here because adding the second- and third-order interaction effects $a_1 * a_2$, $a_1 * a_3$, $a_2 * a_3$ and $a_1 * a_2 * a_3$ to the model leaves the sums of squares for the main effects a_1, a_2 and a_3 undisturbed. This reflects the orthogonality of the sets of estimates for main effects and interaction coefficients which results from the balanced design used in the study. Indeed, the addition of each set of interaction terms of successively higher order leaves the preceding ANOVA estimates and sums of squares unchanged.

Checking the statistical results in Table 2, it is found that all F-statistics for the interaction sums of squares are large (these are not shown in the table). In this case the F-statistics are computed

using MSE for Error (level 3) as a divisor. The P-values for these F-statistics all equal 0.0000 to four decimal places, from which we may infer that one or more interaction terms of each order are non-zero. The result confirms binding dependence among both pairs and triplets of nucleotides for this variant. We can say, therefore, that no intermediate order of independence is present and that the binding probability depends on the complete central triplet, imbedded within the 37 bp string of nucleotides.

4.3 Relative Affinities: REDV Variant

We now look at some relative affinities of sequences for the REDV variant, using the most successful sequence GCG as the reference. The six largest of these relative affinities are shown in Table 3(a). The fitted values for the full dependence model reveal that sequence GCG has a binding probability that is almost double that of GTG and more than five times that of GAG, the two sequences with the next strongest affinities. Among the top six sequences, the estimated conditional probability that GCG will bind in lieu of any of the other five top contenders can be calculated from (6) to be 0.46. Estimated standard errors can also be calculated for these relative affinities as well as approximate confidence intervals. For example, the relative affinity for GTG (with GCG as the reference) has the approximate 95 percent confidence interval $[0.542, 0.754]$.

For sequences whose fitted values under the full dependence model are at the signal detection threshold, we find that the relative affinity is 0.0109. The true binding affinities for these sequences are likely to be at or below the limit of detection of the experiment but we cannot establish their values with precision when the signal intensity is below the threshold.

4.4 Dependence Probability Ratios: REDV Variant

We next look at dependence probability ratios to see the impact of dependence for different sequences. Recall that this ratio compares the estimated binding probabilities under the full dependence and independence ANOVA models – see (4). The three largest and three smallest of these ratios are shown in Table 4 for the REDV variant. The table also shows the relative affinities for these sequences to assist with interpretation. To illustrate the results, the ratio 7.341 for sequence GTG indicates that the estimated binding probability is a multiple 7.341 higher than expected if the three nucleotides were to occur independently with their average frequencies at the three positions in the sequence. Thus, this sequence has a binding probability that is larger than expected under the independence assumption and, hence, exhibits binding attraction. Similarly, the ratio 0.213 for sequence TCG indicates that the estimated binding probability is a multiple 0.213 lower than expected under independence and, hence, reflects binding aversion. Estimated standard errors can also be calculated for these dependence probability ratios as well as approximate confidence intervals. For example, the approximate 95 percent confidence interval for the ratio for sequence

GTG is $[6.531, 8.251]$.

We note for the smallest ratios in Table 4 that the estimated relative affinity is at the detection level, 0.0109. Thus, true binding affinities for such sequences are likely at or below the detection threshold and binding aversion would appear to be at an extreme in these instances.

4.5 Results for Other Mutant Variants and Wild-type

We have also conducted similar analyses for the other three mutant variants and the wild-type. In all cases, dependencies of all orders were found to be significant (the ANOVA tables are not presented here). Table 3 (b)-(e) show the sequences with the largest relative affinities for each case (with the most successful sequence as the reference). The mutants LRHN and KASN had been isolated repeatedly (Choo and Klug, 1994) after independent sets of in-vitro selections by using many different 3-bp binding sites for the second of three zinc fingers (ACT, AAA, TTT, CCT, CTT, TTC, AGT, CGA, CAT, AGA, AGC, and AAT). Therefore, since the in-vitro selections resulted in only poorly characterized sequence specificities for these mutants, the protein-binding microarray approach was used to determine their binding-site preferences (Bulyk *et al*, 2001). KASN appears to be quite unspecific with relative affinities that are fairly large across many sequences, as illustrated by the six sequences listed in Table 3(b). The results in Table 3(c) show that LRHN is quite specific for TAT with other sequences showing up with lower, but material, relative affinities. Like the REDV variant, the RGPD variant was selected from the phage display library using two DNA sequences that were almost identical. For RGPD, these sequences were GCG and GCT. The aim was to see if the microarray technology could distinguish proteins with similar binding-site preferences. The results suggest it can. The results for RGPD in Table 3(d) confirm that the experiment has correctly identified the binding-site preferences, although GCT has a markedly lower affinity than GCG and CCG has about the same affinity as GCT. Both GCT and CCG differ by one nucleotide from GCG. Finally, the results for the wild-type protein in Table 3(e) show specific preferences for TAG and TGG.

We have also conducted a study of the dependence probability ratios for the other mutants and wild-type but do not report the results here as they are qualitatively similar to what we have shown for the REDV variant.

4.7 Checking the Error Assumptions

The error terms in (1) must be independent and identical centered normal random variables for ordinary least squares fitting to be efficient and for conventional normal-based tests, such as the F test, to apply. We have examined normal probability plots and other diagnostic plots for residuals from the full dependence ANOVA model (2) in the case application. These plots give confidence that the error assumptions are reasonable in this case application. Some moderate distortion of

the normal probability plots (except for the KASN variant) is produced by the presence of multiple readings at the detection threshold for some sequences. The result is a saw-toothed linear pattern for the plot, such as that for REDV variant shown in Figure 1.

The presence of non-zero residuals in this case application is due entirely to the presence of replicated spots for the 64 sequences. Without replication, the full dependence model would be a saturated model and fit the 64 intensity readings exactly. We note that a set of main effects for replicates does not account for significant variation when added to the model (P-value = 0.852).

The detection threshold for array technology introduces the problem that small affinities are censored at this threshold. ANOVA estimation with a left censored response might be employed to overcome this artifact. We have not taken into account the censored nature of intensity readings below the detection threshold. Estimation methods for censored responses have been employed and give only slightly different results in this study.

5. Strategies for Studying Dependence in Large-scale Studies

Large-scale microarray investigations of protein binding require special strategies for the statistical component of the work because of the potentially large number of parameters in the full model. Table 1 shows the escalating numbers as the nucleotide sequence lengthens. We now discuss some of these strategies and show that large-scale studies are indeed manageable.

In a balanced design, as we have noted already, the addition of each set of interaction terms of successively higher order leaves the preceding estimates and sums of squares unchanged. This observation offers the simple strategy of exploring successively higher orders of dependence without necessarily estimating the full model. The fitting of the interaction effects of each order can be done in separate ANOVA computer runs by using residuals from the preceding level to fit the next level. As computerized ANOVA routines can handle several hundred parameter estimates at once, it can be seen from Table 1 that interaction effects of at least order 2 can be estimated for nucleotide sequences as large as $l = 8$. With the use of various automated model selection routines, such as stepwise regression, and sequence decomposition strategies, even larger sequences can be handled.

Tests of dependence can be conducted at intermediate stages of estimation by using either an intermediate MSE or the MSE for pure error, denoted later by MSPE. Table 2 illustrates the us of an intermediate MSE. Observe that if only the main effects and second-order interactions were included in the ANOVA model, an ANOVA table containing lines 1-3, 5-8 and 11 could be constructed. The MSE for Error (level 2) will be an upper bound on the MSPE, which is Error (level 3) in line 10 here. F-tests based on an inflated MSE value, if significant, give corect test conclusions. An alternative approach for the F-tests is to calculate the MSPE directly from the full (saturated) model without estimating the individual model parameters. With replication, the fitted value of the full ANOVA

model for a sequence is the mean reading taken over the set of replicates for that sequence. These mean values can be computed without estimating the individual ANOVA parameters. MSPE is the pooled variance of the replicated readings about their respective means and is quite easy to compute.

For studies involving very long sequences, the use of fractional factorial designs might be advisable for exploring high-order interactions without constructing numerous arrays. The basic idea would be that with a selective choice of sequences for the arrays, all sets of interaction terms up to a given order (less than the full order) could be estimated from the ANOVA model. Thus, some partial degrees of dependence could be studied without using all 4^l sequences.

The theory of log-linear models provides strategic guidance for exploring dependencies and for decomposing the problem of discovering dependencies (more,precisely, independencies) in large-scale studies. We now give two theoretical results that illustrate how one can obtain insight into the dependence structure. For the presentation of these results, we consider, without loss of generality, a simple partition of sequence $\boldsymbol{a}$ as follows $\boldsymbol{a} = \boldsymbol{a}_1\boldsymbol{a}_2$ where $\boldsymbol{a}_1 = \{a_1 \cdots a_r\}$ and $\boldsymbol{a}_2 = \{a_{r+1} \cdots a_l\}$. In other words, we break the nucleotide sequence into two segments of length r and $l-r$, respectively, with $r < l$.

(a) If segments $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ act independently in determining binding for the target protein then we have

$$\ln P_{\boldsymbol{a}} = \ln P(\boldsymbol{a}) = \ln P(\boldsymbol{a}_1\boldsymbol{a}_2) = \ln [P(\boldsymbol{a}_1)P(\boldsymbol{a}_2)] = \ln P(\boldsymbol{a}_1) + \ln P(\boldsymbol{a}_2) \tag{7}$$

Each of $\ln P(\boldsymbol{a}_1)$ and $\ln P(\boldsymbol{a}_2)$ can be expanded in main effects and interaction effects as in (2). We then can see that the ANOVA model represented by the sum $\ln P(\boldsymbol{a}_1) + \ln P(\boldsymbol{a}_2)$ will not have any interaction effects for nucleotides that are drawn from both $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. Thus, the non-zero interaction effects in (2) give an immediate picture of the dependence structure for the binding of the target protein.

(b) Conditional independence is also potentially important. For example, if the nucleotides of segment $\boldsymbol{a}_1 = \{a_1, \ldots, a_r\}$ bind independently, provided segment $\boldsymbol{a}_2 = \{a_{r+1}, \ldots, a_l\}$ is in place, then we have

$$\ln P(\boldsymbol{a}_1|\boldsymbol{a}_2) = \sum_{i=1}^{r} \ln P(a_i|\boldsymbol{a}_2). \tag{8}$$

In this case, among the subset of sequences having segment $\boldsymbol{a}_2$, the independence ANOVA model (3), with index l replaced by r, will fit the log-intensity data.

ACKNOWLEDGEMENTS

M.-L.T. Lee was supported in part by NIH grant NCI-R03-CA89756-01. M. Bulyk and G. Church were supported in part by grants from the Office of Naval Research (N00014-99-1-0783) and the Department of Energy (DE-FG02-87ER60565). G. A. Whitmore was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*, Wiley.

Bulyk, M. L., Huang X., Choo Y. and Church G.M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proceedings of the National Academy of Sciences*, **98**, 7158-7163.

Bulyk, M. L.,Johnson, P.L.F., and Church, G.M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors, *Nucleic Acids Res.*(in print).

Choo, Y. and Klug, A. (1994). Toward a code for the interactions of zinc fingers with DNA: Selection of randomized fingers displayed on phage. *Proceedings of the National Academy of Sciences, USA*, **91**, 11163-11167.

Lee, M.-L.T., Kuo, F.C., Whitmore G.A., and Sklar J.L. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations, *Proceedings of the National Academy of Sciences*, **97**, 9834-9839.

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.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Second Edition, Chapman & Hall.

Ponomarenko, M.P., Ponomarenko, J.V., Frolov, A.S., Podkolodnaya, O.A., Vorobyev, D.G., Kolchanov, N.A. and Overton, G.C. (1999). Oligonucleotide frequency matrices addressed to recognizing functional DNA sites, *Bioinformatics*, **15**, 631-643.

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucleic Acids Res.*, **23**, 4878-4884.

Staden, R. 1988). Methods to define and locate patterns of motifs in sequences, *Comput. Appl. Biosci.*, **4**, 53-60.

Stormo, G.D., Schneider, T.D. and Gold, L. (1986). Quantitative analysis of the relationship between nucleotide sequence and functional activity, *Nucleic Acids Res.*, **14**, 6661-6679.

Wingender, E., Karas, H. and Knuppel, R. (1997). TRANSFAC database as a bridge between sequence data libraries and biological function, *Pac. Symp. Biocomput.*, 477-485.

Zhang, M.Q. and Marr, T.G. (1993). A weight array method for splicing signal analysis, *Comput. Appl. Biosci.*, **9**, 499-509.

Table 1. Numbers of main-effect and interaction parameters of different orders for the full-dependence ANOVA model.

Sequence Length	Total Parameters	Main Effects	Interaction Effects						
			2nd	3rd	4th	5th	6th	7th	8th
2	16	6	9						
3	64	9	27	27					
4	256	12	54	108	81				
5	1,024	15	90	270	405	243			
6	4,096	18	135	540	1,215	1,458	729		
7	16,384	21	189	945	2,835	5,103	5,103	2,187	
8	65,536	24	252	1,512	5,670	13,608	20,412	17,496	6,561

Table 2. ANOVA table for the REDV mutant variant, showing intermediate error *SS*, *df* and *MS* values for main effects and second-order interaction terms.

Line	Source	*SS*	*df*	*MS*
1	a_1	102.51	3	34.170
2	a_2	87.61	3	29.205
3	a_3	125.42	3	41.806
4	Error (level 1)	345.90	566	0.61113
5	$a_1 * a_2$	57.69	9	6.410
6	$a_1 * a_3$	164.49	9	18.277
7	$a_2 * a_3$	16.77	9	1.863
8	Error (level 2)	106.94	539	0.19840
9	$a_1 * a_2 * a_3$	90.64	27	3.357
10	Error (level 3)	16.30	512	0.03183
11	Total	661.44	575	1.150

Table 3. Estimated relative affinities for selected sequences for each of four mutant variants and the wild-type. Relative affinity here is the ratio of the fitted binding probability for the sequence to the largest fitted binding probability for any sequence.

(a) REDV		(b) KASN		(c) LRHN		(d) RGPD		(e) wild-type	
Seq.	Est. Rel. Aff.	Seq.	Est. Rel. Aff.	Seq.	Est. Rel. Aff.	Seq.	Est. Rel. Aff.	Seq.	Est. Rel. Aff.
GCA	0.0993	CGT	0.8688	TGT	0.3751	GTG	0.1664	GAG	0.1435
GGG	0.1134	ACT	0.9265	CAT	0.3990	GCC	0.1994	AGG	0.2272
CCG	0.1617	TAT	0.9401	AGT	0.4318	GCA	0.2053	CGG	0.3119
GAG	0.1808	ATT	0.9605	GAT	0.4923	CCG	0.3146	GGG	0.3547
GTG	0.6393	AAT	0.9944	AAT	0.5651	GCT	0.3907	TAG	0.7170
GCG	1.0000	CCT	1.0000	TAT	1.0000	GCG	1.0000	TGG	1.0000

Table 4. Dependence probability ratios for selected sequences for the REDV variant. Estimated relative affinities are also shown for the same sequences. The dependence probability ratio compares the estimated binding probabilities of a sequence under the full dependence and independence ANOVA models.

Nucleotide Sequence	Type of Binding	Dependence Probability Ratio	Estimated Relative Affinity
TCG	Aversion	0.213	0.0109
CTG		0.270	0.0109
GTC		0.347	0.0109
ATA	Attraction	4.876	0.0566
GCG		6.395	1.0000
GTG		7.341	0.6393