doi:10.1016/S0022-2836(02)00917-8 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) **323**, 701–727

JMB

Probabilistic Code for DNA Recognition by Proteins of the EGR Family

Panayiotis V. Benos[1], Alan S. Lapedes[2] and Gary D. Stormo[1*]

[1]*Department of Genetics*
School of Medicine
Washington University
Campus Box 8232, St. Louis
MO 63110, USA

[2]*Theoretical Division, Los*
Alamos National Laboratories
Los Alamos, NM 87545, USA

**Corresponding author*

A recognition code for protein–DNA interactions would allow for the prediction of binding sites based on protein sequence, and the identification of binding proteins for specific DNA targets. Crystallographic studies of protein–DNA complexes showed that a simple, deterministic recognition code does not exist. Here, we present a probabilistic recognition code (P-code) that assigns energies to all possible base-pair–amino acid interactions for the early growth response factor (EGR) family of zinc-finger transcription factors. The specific energy values are determined by a maximum likelihood method using examples from *in vitro* randomisation experiments (namely, SELEX and phage display) reported in the literature. The accuracy of the model is tested in several ways, including the ability to predict *in vivo* binding sites of EGR proteins and other non-EGR zinc-finger proteins, and the correlation between predicted and measured binding affinities of various EGR proteins to several different DNA sites. We also show that this model improves significantly upon the prediction capabilities of previous qualitative and quantitative models. The probabilistic code we develop uses information about the interacting positions between the protein and DNA, but we show that such information is not necessary, although it reduces the number of parameters to be determined. We also employ the assumption that the total binding energy is the sum of the energies of the individual contacts, but we describe how that assumption can be relaxed at the cost of additional parameters.

Keywords: DNA–protein interactions; recognition code; DNA-binding specificity; zinc-finger proteins

Introduction

Unravelling the rules that govern the recognition of the target DNA sequences by transcription factors is one of the great challenges in computational biology today. The regulation of the expression of any gene in a cell is initiated by the specific binding of one or more transcription factors in its promoter region. Revealing the mechanisms of this process would constitute a key step towards understanding a cell's regulation and its response to various factors. This can lead to tools for engineering gene regulation, with many potential therapeutic applications.

Research in this field was initiated by Seeman *et al.* over 25 years ago,[1] at a time when the term computational biology sounded quite exotic (if not self-contradictory) to most people. The analysis of the structure of the amino and nucleic acid residues led them to postulate that specificity can be achieved through a network of hydrogen bonds that can be formed between amino acid residues and bases. They concluded that two or more bonds per base–amino acid pair are required for efficient discrimination.

That study raised the hopes that a simple, deterministic model (or "recognition code") might exist in nature that will adequately explain the protein–DNA interactions.[2] It took 12 more years of research and a handful of protein–DNA co-crystal structures to realise that, in the course of evolution, nature employs a variety of strategies to achieve protein–DNA recognition. On the basis of the

Present address: P.V. Benos, Department of Human Genetics, Center for Computational Biology and Bioinformatics and Cancer Institute, University of Pittsburgh, PA 15261, USA.

Abbreviations used: EGR, early growth response.

E-mail addresses of the corresponding authors: stormo@genetics.wustl.edu; benos@pitt.edu differences between the 3D structures of various transcription factors, Matthews claimed in 1988 that there is "no code for (protein–DNA) recognition",[3] although he made it clear that he was referring to a simple, deterministic code. In the years that followed this publication, the existence of a recognition code became a highly debatable issue. We know now that there are clear preferences for base–amino acid contacts.[4–9] Thus, although the concept of a universal, deterministic recognition code has largely been abandoned, a two-way "probabilistic code" (P-code) constitutes a promising approach to this problem.

There have been a number of attempts to mathematically model protein–DNA interactions, which have met with variable success.[7,10–14] All of these methods are based on empirical observations or semi-arbitrarily determined "scores" to evaluate the DNA specificity of proteins. The currently available models can be classified into two main categories. We call a model qualitative if it uses binary values (e.g. 1 or 0) to describe the interactions between bases and amino acid residues. The qualitative models usually consist of a look-up table that associates amino acid residues in certain positions in the protein ("contacting" amino acid positions) with particular bases in certain positions in the DNA target.[15,16] By contrast, a quantitative model provides a measure for the binding affinity for any given protein–DNA pair. The quantitative models usually consist of a weight matrix that assigns a score to the base–amino acid contacts. The score can be position-dependent (as used by Suzuki *et al.*[10]) or position-independent (as used by Mandel-Gutfreund & Margalit[7,13] and Kono & Sarai[12]). A qualitative model can be viewed as a degenerate quantitative model, where each of the weights has a value set to 1 ("permissible") or 0 ("non-permissible"). A quantitative model can be transformed into qualitative by setting a score threshold that will separate the permissible from the non-permissible contacts.

One feature common to all existing models is that they consider the individual contacts to be independent from each other and hence have an additive contribution to the total binding energy/affinity score. It is known that the additivity assumption is not altogether correct. In fact, the violation of this assumption in certain situations constitutes one of the main arguments against a recognition code. But, for prediction purposes, the additivity assumption needs to hold only for high-affinity protein–DNA pairs.[17] In a few cases where it has been examined closely, additivity between positions in the binding sites does not hold exactly,[18,19] but the data in these studies show that it provides a good approximation to the true binding energies.[20]

Here, we present a P-code for modelling the protein–DNA interactions, which can be divided into two parts. First, we present a theoretical framework of how the protein–DNA specificity can be described probabilistically. On the basis of this framework, which derives from the statistical mechanics theory, we develop an algorithm that can estimate the base–amino acid energetic potentials from *in vitro* randomisation experimental data (SELEX and/or phage display, see below). We implemented this algorithm into a program called Statistical Algorithm for Modelling Interaction Energies (SAMIE). The description of the algorithm can be found in Materials and Methods. We use the early growth response (EGR) protein family as a test case for our algorithm. SAMIE is trained on published data from *in vitro* selection experiments and specifies the position-specific energetic potentials of the base–amino acid contacts. The program determines the values of the parameters that maximise the probability of obtaining the observed data. For the training, we use knowledge of structural details of the protein–DNA interface. However, we describe how a model could be developed without using such knowledge at the expense of additional parameters that needed to be determined. Also, for the training, we invoke the assumption of additivity, but we describe how more complex models without that assumption can be used, again at the expense of additional parameters, which would require more data to determine.

Secondly, we evaluate the accuracy of the derived model in several ways. In order to choose the model that best describes the protein–DNA interactions of the EGR family, we do several trainings using different partitions of the dataset (SELEX data only, phage display data only or both) and employing different contacting schemes. The resulting models are compared in terms of prediction efficiency on their own training sets in order to select the one that will be used for subsequent analysis/evaluation of the algorithm. The evaluation is performed in various ways, including a comparison of predicted relative binding constants to those measured for several proteins. We compare it to previous models, both qualitative and quantitative, and demonstrate its improved accuracy.

The EGR protein family

There are a number of protein families that have provided the basis for previous studies of protein–DNA interactions. Among them, the EGR factor family is probably the most extensively studied and therefore we use it here as a test case for our algorithm, SAMIE.

Members of this family were initially identified in mammals;[21–23] recently they were discovered in a variety of other species, including *Xenopus laevis*[24] and zebrafish.[25] The EGR proteins contain three zinc-finger regions, a domain common to many eukaryotic transcription factors. They belong to the Cys_2His_2 subfamily of zinc-fingers, which derives its name from the residues that are used for the coordination of the zinc ion. Cys_2His_2 is a

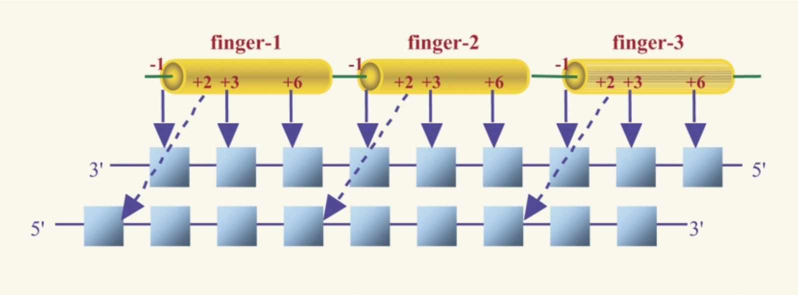


Figure 1. Representation of the binding of the EGR protein to its DNA target. According to crystallographic studies, each of the three zinc-finger domains of the EGR protein contacts four bases in an antiparallel fashion. There is one base overlap in the target sequence between any two adjacent fingers. The numbering of the amino acid residues is with respect to the beginning of the α-helix. Amino acid residues −1, +3 and +6 contact bases at positions 3, 2 and 1, respectively, whereas amino acid residue +2 contacts the complementary base at position 4 (overlapping base).

eukaryotic motif very common to many transcription factors, 2719 members in Pfam v7.2.[26] The structure of the three zinc-fingers of the mouse protein Zif268 (or EGR1) bound to its consensus DNA sequence was initially solved crystallographically at 2.1 Å resolution[27] and subsequently refined to 1.6 Å resolution.[28] The crystal structure showed that each of the three fingers contacts its DNA target in a modular, antiparallel fashion. Initially, it was believed that each finger recognises three bases, but the refinement of the structure showed that a fourth base is contacted as well. In each finger there are four amino acid positions involved in contacts with four DNA positions. For some DNA positions, only one base is contacted, whereas in others both the forward and the complementary base are contacted by amino acid residues in adjacent fingers (overlapping base positions). In particular, the amino acid residues at positions −1, +3 and +6 in each finger (with respect to the beginning of the α-helix) contact the bases at positions 3′, middle and 5′, respectively;[27] the amino acid residue at position +2 of the helix contacts the complementary strand on the fourth position (overlapping base).[28] This contacting scheme is illustrated in Figure 1. We note that some of the EGR variants have been shown to deviate from this pattern of contacts.[29]

Data from selection experiments

Protein–DNA interaction data can generally be divided into three categories. One refers to pairs of sequences (protein and DNA) that are known to bind to each other with some affinity. Typically, these examples are the result of experiments that aimed to measure directly the binding affinity of the protein (or its mutants) to particular DNA targets. This category includes the wild-type protein sequences bound to their *in vivo* targets.

The remaining two categories refer to data derived from *in vitro* selection experiments, namely SELEX and phage display. In a SELEX experiment, a protein of known sequence is used to select DNA targets from a pool of randomised oligonucleotides.[30] This procedure usually yields more than one DNA target. Although one would assume that all selected targets exhibit high affinity towards the protein, the highest affinity target might not be selected at all (for purely stochastic reasons).

In a phage display experiment, the reverse randomisation/selection procedure applies.[31] A recombinant DNA library is constructed that consists of variants of the cDNA sequence of a known DNA-binding protein. The nucleotides that code for certain amino acid positions are randomised (usually, these are the positions that are known to contact the DNA or they are assumed to do so). Upon expression, the polypeptides are displayed on the outer coat of the phages. Thus, proteins can be selected that bind with high affinity to a certain (fixed) DNA target. As in the case of SELEX, multiple proteins are usually selected for any given DNA target and the protein that exhibits the highest affinity towards the particular DNA sequence might not be among those selected.

Theory

The P-code model for protein–DNA recognition

In this section, we present the theoretical framework of the protein–DNA recognition upon which our method is based.

Thermodynamic aspects of the protein–DNA recognition

The recognition of specific DNA sequences by the corresponding transcription factors can be a complicated, multi-step process.[32] Nonetheless, if we assume that a protein comes into contact with various DNA sequences *via* diffusion, then in equilibrium we would expect that the time that it interacts with each of them will be inversely proportional to their dissociation constants K_D.

Consider a DNA-binding protein, A, that has N_{tot} possible targets. For example, if the DNA target for this protein is L bases long, then there will be $N_{tot} = 4^L$ possible targets. Each of these targets will have a relative frequency $P_n({}_kN)$ among all possible binding sites, for example in the genome. Assuming equilibrium, the conditional probability that this protein will be bound to a particular DNA target, ${}_kN$, is given by the following equation:

$$P({}_kN|A) = P_n({}_kN)\mathrm{e}^{-H({}_kN,A)} \bigg/ \left(\sum_{k'} P_n({}_{k'}N)\mathrm{e}^{-H({}_{k'}N,A)} \right) \tag{1}$$

where $H({}_kN,A)$ is the binding energy of the interaction and the sum in the denominator is the partition function over all possible DNA target sequences, ${}_{k'}N$ ($k' = 1,\ldots,N_{\mathrm{tot}}$). This formula derives from the Boltzmann distribution of the statistical mechanics theory[33] and it relates the energy values with the binding probabilities.

As evident from equation (1), the DNA specificity of a given protein is determined by the relative energy, and the absolute energy values are not important. (This assumes that the protein is always bound to DNA, which is essentially true *in vivo* unless it is prohibited from binding by some mechanism, such as binding to an inhibitory protein.) One can subtract any constant from all of the energy values and the probability of binding will be unchanged. This formula follows the convention that lower energy values correspond to stronger binding (or longer dissociation times). Sometimes, the value of zero is assigned to the lowest-energy state and all other states have positive values. In this study, we allow the energy values to take both positive and negative values; more negative values correspond to stronger binding (higher affinity). We arbitrarily assign an energy of zero to the average affinity, so that those sequences that bind better than average have negative energy and those that bind worse are positive.

Methods for modelling the protein–DNA interactions

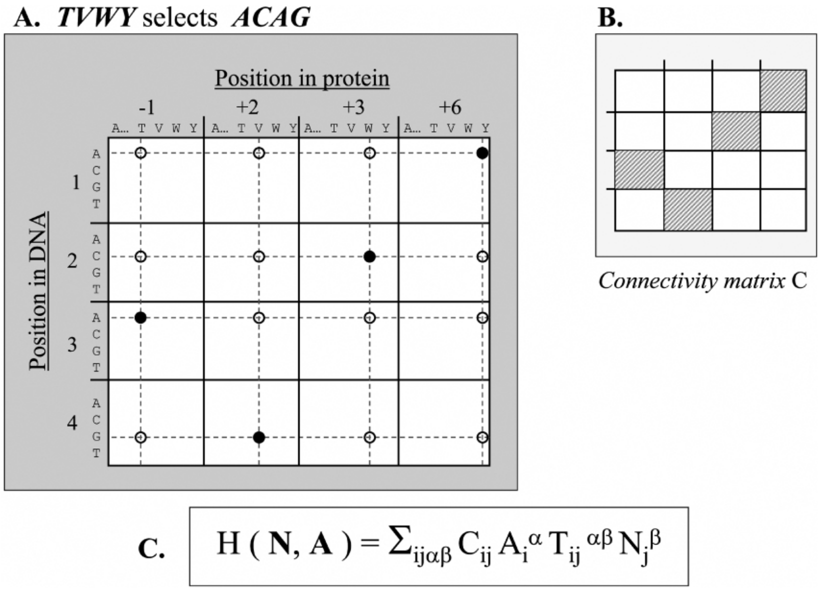
In theory, one could determine the binding energy for a protein of interest to all possible target sequences. In this case, equation (1) would be an accurate description of the distribution of the protein on the various sequences at equilibrium. Note, however, that this would not constitute a model for the protein–DNA interaction, but simply a look-up table that stores the measured values. And of course, for a complete representation of a given protein family, one should determine such a table for all possible protein sequences (by which we mean all possible variants at the residues involved in the sequence-specific DNA binding). For example, a typical DNA-binding protein has binding sites of about ten base-pairs. There are over 10^6 possible sequences for a 10 bp site. If the protein uses ten amino acid residues to recognise its DNA target, then the number of all possible proteins (with respect to these positions) is 20^{10}, which is greater than 10^{13}. A complete look-up table for this protein family would contain the affinity values of each of these proteins to each of these DNA targets ($>10^{19}$ values). Having a mathematical model for the interaction, some kind of code, allows one to predict the interaction energy for all possible combinations based on measurements for only a fraction of them.

In the simplest model, only certain combinations of amino acid residues and base-pairs have favourable affinity and those combinations are used exclusively for the protein–DNA contacts. This was the type of code sought by Seeman *et al.*,[1] but only a few crystal structures of protein–DNA complexes were sufficient to determine that such a simple code does not exist.[3] Recent qualitative models for the EGR family are similar, but they allow for multiple amino acid residues to interact favourably with each base-pair and multiple base-pairs to interact favourably with each amino acid.[14–16,31,34] Thus they are simple, degenerate codes in which the amino acid and base-pair combinations are categorised as either permissible or not. Such codes have been useful in modelling certain protein–DNA interactions, and even in designing proteins with high affinity to specific sequences. However, by design they cannot predict the affinities of different sequence combinations and, as we show below, about 40% of the interactions obtained in SELEX and phage display experiments are not accounted for by these models.

The next more complicated models are quantitative, with a score associated with each base-pair–amino acid combination, and additive. That is, the base-pair–amino acid contacts are considered to be energetically independent and therefore, the total energy of the interaction is merely the sum of energies over all contacts. The number of parameters in the additive models is proportional to the number of contacts, not the number of sequence combinations. Each contact has (at most) 80 parameters, for all possible combinations of base-pairs with amino acid residues. Using the example above, the number of parameters one needs to estimate is 800 for the 10 bp DNA target (assuming that each of the ten amino acid residues contacts one base only). This is a significantly smaller number than the list of $>10^{19}$ possible combinations. However, the effort required to determine 800 energy values experimentally is still large.

Representation of the model

Figure 2 illustrates the model for protein–DNA interaction using a single zinc-finger as an example. In Figure 2(a), the four positions of the binding site are listed along the left side of the Table, with each of the possible bases for each position shown. The "positions" in the protein are listed across the top of the Table. These are only the protein residues that are directly involved in



$$H(\mathbf{N}, \mathbf{A}) = \Sigma_{ij\alpha\beta} C_{ij} A_i^{\alpha} T_{ij}^{\alpha\beta} N_j^{\beta}$$

Figure 2. A representation of the energy calculation. SAMIE exploits *in vitro* randomisation experiments and calculates the weight matrix *T* that maximises the likelihood of the training set according to equation (1) or (3). (a) A graphical representation of such a weight matrix for one zinc-finger of the EGR protein family. Each finger of the EGR proteins uses four amino acid residue positions (numbered −1, +2, +3 and +6 from the beginning of the α-helix) to recognise a 4 bp DNA target. Each small sub-matrix of matrix *T* consists of 80 values, which correspond to single base–amino acid energetic potentials (20 amino acid residues × 4 bases). For calculating the predicted energy of the interaction of a given protein sequence of this family (e.g. TVWY) to a given DNA sequence (e.g. ACAG), we encode the two sequences in two unary vectors, *A* and *N*, as we describe in Materials and Methods. When multiplied with matrix *T*, the two unary vectors, *A* and *N*, select the appropriate columns and rows, respectively, that define the base–amino acid energetic potentials of the interaction. Under the additivity rule, the total energy of these individual predicted energetic potentials is summed (circles). (b) Additional structural details of the interaction can be imposed by terms of the connectivity matrix, *C*. In this matrix, the positions that are known to interact have the value of 1 (shaded boxes) and all the others are set to 0. Multiplication of this matrix with the rest depicts only the relevant energetic potentials for the final sum (filled circles). The mathematical representation of this procedure (see equation (2)) is shown in (c).

contacting the bases in the binding sites, as determined from crystallography, and are positions −1, +2, +3 and +6 relative the α-helix of the zinc-finger. Each position can be occupied by any of the 20 amino acid residues, but only a few of them are listed to save space. The elements of the Table are energies of interactions between specific amino acid residues and base-pairs, and it allows the calculation of interaction energy for any protein sequence with any DNA sequence, using the additivity approximation. For a specific protein binding to a specific DNA, the interaction energy is simply the sum of the contacts for those particular sequences. For example, if the protein sequence is TVWY (at positions −1, +2, +3 and +6 in the zinc-finger, respectively), then only the columns corresponding to those amino acid residues, indicated in the Figure, are relevant to the energy calculation. Likewise, if the binding site contains the sequence ACAG, then only the rows corresponding to that sequence are relevant to the binding energy. So the total binding energy might be the sum of the 16 values where those rows and columns intersect. However, for this protein we know that the amino acid residue at position −1 interacts only with the 3′ DNA position, the residue at position +3 interacts with the middle position, etc., so that only the intersections shown with filled circles contribute to the binding energy. If we did not know which amino acid residues interacted with which DNA positions, we would have to include all of the possible interactions in calculating the energy (all of the circles), but presumably with enough training data the interacting positions would become clear as the only significant contributions to specificity.

The interaction energies are denoted by $T_{ij}^{\alpha\beta}$, where *i* and *j* are the positions in the amino acid and nucleotide sequences, respectively; α and β are the specific amino acid residues and bases that occur at those positions; and *C* is the “connectivity matrix” that indicates which positions in the protein contact which positions in the DNA (Figure 2(b)). Matrix element C_{ij} is “1” for those contacting amino acid and base positions and “0” elsewhere. As such, only the positions in contact contribute to the final energy. Together the *T* and *C* matrices allow the calculation of binding energy for any zinc-finger protein with any four-long DNA sequence. To get the specific interaction between a particular protein and a particular DNA, one just sums the elements of *T* that correspond to those interactions, as shown in the Figure. In the equation (Figure 2(c)), *N* is a unary vector that contains 1 for the base that occurs at each position, and 0 elsewhere. Likewise, *A* is a unary vector for the amino acid sequence. It contains a 1 for the amino acid at each position and 0 elsewhere. Multiplying them with the *T* matrix serves to select out only those elements that correspond to the interacting positions, as shown in the Figure, the sum of which constitute the predicted binding energy of that interacting pair. Additional details of the encoding are provided in Materials and Methods. Note that for a specific protein it is possible to determine the "weight matrix"[35] that predicts the binding energy to all possible sites for that protein by just combining the rows that correspond to that protein sequence, as shown in Figure 9 in Materials and Methods. That allows one to easily calculate the probability of the protein binding to any particular sequence. The same thing was done to generate a weight matrix that described the energy of any particular DNA sequence to all possible protein sequences (not shown), which is used in the calculation of probabilities for the phage display data.

Using data from SELEX randomisation experiments

Instead of measuring the relative energy of the interaction of a systematic set of protein and DNA combinations designed to obtain all of the desired parameters, one could perform SELEX randomisation experiments with many (but not all) of the proteins. Each of the proteins is allowed to select its preferred target(s) from a pool of randomised oligonucleotides. Of course, due to the stochastic nature of the selection process, the target with the lowest binding energy (i.e. highest affinity) might not be among those that are finally selected. Nevertheless, we would expect that the targets that will be selected would bind to the protein with high affinity. Moreover, we can assume that the stronger that a base–amino acid contact is, the higher the probability that it will be selected.

The probability that a particular nucleotide target ${}_kN$ would be selected by the protein A will be given by the same formula (equation (1)). This time, $P_n({}_kN)$ is the relative frequency of the selected target ${}_kN$ in the oligonucleotide pool; the denominator is, again, the partition function, now calculated over all DNA target sequences present in the oligonucleotide pool. $H({}_kN, A)$ is the total binding energy potential of the DNA target to the protein A, which assuming additivity over all contacts, can be calculated from the formula:

$$H({}_kN, A) = \sum_{ij\alpha\beta} C_{ij} A_i^{\alpha} T_{ij}^{\alpha\beta} N_j^{\beta} \qquad (2)$$

where the variables are as described above (see also Figure 2).

Using data from phage display randomisation experiments

In the previous section, we focused on the problem of "DNA recognition" by a particular protein. Similar arguments and formulae can be stated and written for the reverse problem: i.e. "protein recognition" by a given DNA target. Phage display randomisation experiments can be performed, in which a fixed DNA target will be used to select some protein sequences that bind to it with high affinity, from a pool of randomised proteins. In this case, the probability that a given ("fixed") DNA sequence will select a particular protein is given by:

$$P({}_kA|N) = P_a({}_kA)\mathrm{e}^{-H(N,{}_kA)} \Big/ \left(\sum_{k'} P_a({}_{k'}A)\mathrm{e}^{-H(N,{}_{k'}A)} \right) \qquad (3)$$

where now the sum in the partition function is calculated over all possible protein variants. The randomised amino acid positions are usually limited to those assumed to make direct contacts with the bases.

SAMIE: maximising the probability of the observed interaction data

Using the statistical mechanics theory just described, the problem of modelling the protein–DNA interactions for a given protein family consists of estimating a number of parameters that correspond to position-specific energetic potentials of single contacts. By assuming that the interactions are additive, the problem is simplified significantly, because each contact will require $20 \times 4 = 80$ parameters to be modelled. For example, the modelling of a single zinc-finger of the EGR family bound to a tetranucleotide target would consist of specifying the values of a weight matrix like that presented in Figure 3 (framed submatrix). If we did not know the exact pattern of contacts and we had to allow for any combination between the four bases and the four amino acid residues, then we would have to use the "all-to-all" model and the number of parameters would be $16 \times 4 \times 20 = 1280$ (i.e. all the 16 submatrices within the framed area in Figure 3). Restricting the model solely to the contacts known from the crystal structure[28] reduces this number to 320 (i.e. "one-to-one" model, shown as shaded areas in the framed submatrix in Figure 3). The "many-to-one" model includes two additional contacts from amino acid residues in the adjacent fingers to the overlapping bases (i.e. the external shaded areas in Figure 3), but the number of parameters remains 320. This is because the two additional contacts are identical with those included in the "one-to-one" model (dotted arrows, Figure 3), so they can be linked during training.

The algorithm we developed, SAMIE, estimates these parameters from SELEX and/or phage display data. Essentially, SAMIE calculates the matrix T that maximises the probability (or log-probability) of observing the data. To this extent, it can be viewed as a maximum likelihood estimator of the interaction energy parameters. A detailed description of the algorithm is provided in Materials and Methods.

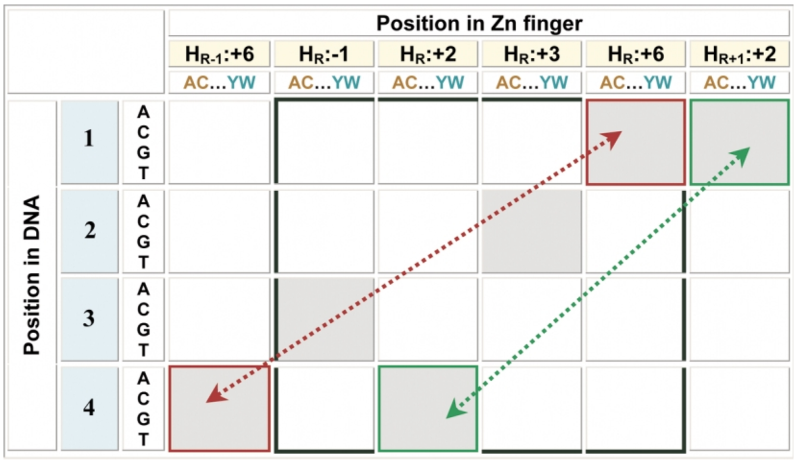


Figure 3. Models of interaction. This weight matrix represents the parameters that SAMIE needs to estimate for the modelling of a single finger target site (EGR protein family). Each small sub-matrix consists of 4 × 20 energy values that relate the amino acid residues of a particular contacting position in the protein with the bases in a DNA target position. The framed area constitutes the full model of the four amino acid residues of a single finger contacting four bases in the DNA target. The shaded sub-matrices correspond to the contacts observed in the co-crystal structure of the EGR protein.[27,28] The four shaded submatrices in the framed area (320 parameters in total) constitute the one-to-one model of interactions that we refer to in the text. The many-to-one model of interactions includes the two external shaded submatrices. These submatrices (i.e. contacts from the adjacent fingers) contain 80 parameters each (20 amino acid residues × 4 bases). However, due to the fact that the additional contacts are (stereochemically) identical with others included in the main matrix (connected by broken-line arrows in the Figure), the number of parameters to be estimated remains $4 \times 4 \times 20 = 320$ in this model too.

Results

Training datasets and models calculated by SAMIE

In a preliminary report, we used solely SELEX data from the EGR protein family to train SAMIE.[36] The training vectors in that set were constructed according to the "one-to-one" model of interactions, consisting of the four "contacting" amino acid residues and their corresponding (four) target bases (see Figure 3).[27,28]

In the present study, we use an expanded dataset of the same protein family. All data are collected from the literature and stored in a database.[36] They are organised into six datasets that differ in the type of data they contain (SELEX, phage display or both) as well as the model of interactions they are built upon one-to-one or many-to-one (see Figure 3). Both models consider only the contacts that have been observed in the co-crystal structure (Figure 3, shaded submatrices). The many-to-one model consists of a superset of the one-to-one. For each finger modelled (H_R), it contains two additional contacts: the amino acid residue at position +2 of the following finger (H_{R+1}) that contacts base 1 of the DNA target and amino acid residue +6 of the preceding finger (H_{R-1}) that contacts base 4 (overlapping base). The two additional contacts are linked to their "identical" ones (see Figure 3, dotted lines) during training.

The purpose of the training on multiple sets is to determine how different are the models derived from SELEX data only or phage display data only when compared to the models derived from the combined datasets. Also, we would like to see whether there is any difference in the models resulting from training on the one-to-one or many-to-many modes of interactions.

The three datasets that are made according to the one-to-one model are named SELEX_4, PHAGE_4, and COMBINED_4, whereas those that are made according to the many-to-one model are named SELEX_6, PHAGE_6 and COMBINED_6. The detailed description of the construction of the datasets is presented in Materials and Methods. SAMIE was trained on each of these six datasets, resulting in six different weight matrices (or models) with the base–amino acid energetic potentials for the EGR family. We call these matrices SAMIE_S4, SAMIE_P4, SAMIE_C4, SAMIE_S6, SAMIE_P6, and SAMIE_C6. The letters S, P and C are indicative of the type of data in the training set (i.e. SELEX, phage display and composite datasets, respectively) and the numbers of the model of interaction that was used; 4 for the one-to-one; 6 for the many-to-one.

In the following, we use the evaluation measures success rate and specificity index that we define in Materials and Methods and some others where appropriate, in order to evaluate the ability of SAMIE to describe protein–DNA interactions in five ways. (a) First, by predicting the data of its own training set, we show that there is no internal inconsistency in the model. (b) Second, by predicting known *in vivo* EGR binding sites, SAMIE is tested on its potentials as a "genomic scanner" for the EGR proteins. (c) Third, predicting known *in vivo* binding sites of other transcription factors indicates the extent to which SAMIE constitutes a good representation for other Cys_2His_2 zinc-finger proteins. (d) Fourth, the correlation between predicted energy values with measured affinities tests how well the frequency-derived weights of SAMIE correspond to real energy potentials. It shows to what extent potential "docking rearrangements” might affect SAMIE’s predictions. (e) Finally, we compare SAMIE with the other existing qualitative and quantitative models to indicate the strengths and weaknesses of each one.

Evaluation on the training sets

There are three reasons that we are interested in evaluating SAMIE on its training sets (self-test). First, we would like to check whether the algorithm is self-consistent. An algorithm that is self-consistent should be able to predict quite accurately, at least its own training set. The second reason is that we would like to measure how much the various training sets and modes of interactions are affecting the predictive power of SAMIE. Finally, we would like to select the model that performs best for further analysis (i.e. prediction of *in vivo* binding sites, prediction of binding energies, etc.).

We use the parameters of each of these six weight matrices to evaluate SAMIE on the corresponding SELEX and phage display training vectors. For each of the fixed sequences of the vectors in the evaluation set, all possible randomised sequences were ranked according to their probabilities of selection, as they are calculated by SAMIE (equations (1) and (3); see also Figure 2), using the corresponding weight matrix. For the SELEX-derived vectors, an equal reference probability of 0.25 was assigned for all bases; whereas for the phage display vectors, the particular experiment-specific randomisation scheme was adopted. This is important, because some phage display randomisation schemes, such as VNN and VNS†, exclude certain amino acid residues from the selection procedure. Others, like NNK and NNS, simply alter the amino acid frequencies.

The results of the self-tests are presented in Table 1 and they show that all SAMIE’s models are self-consistent ($SR_{0.1}$ values of 0.854–0.929). In general, training according to the one-to-one model of interactions seems to give slightly better predictions on phage display data, whereas the many-to-one model of interactions predicts the SELEX data better. The training on the combined data sets resulted in predictions of the SELEX and the phage display datasets almost as accurate as the training on these sets individually. On average, the model of interactions we used for the training on combined datasets did not affect much the prediction capabilities of SAMIE ($SR_{0.1}$ values of 0.907 and 0.902 for the SAMIE_C4 and SAMIE_C6, respectively).

The results show that there is no internal consistency in the method. Since the two SAMIE models that are trained on composite sets (SAMIE_C4 and SAMIE_C6) perform about the same overall, we chose to use SAMIE_C6 for the subsequent analysis, because we feel it is the more complete. The weight matrices of the four principal contacts of this model are presented in Table 2.

Table 1. Evaluation of SAMIE trained on various datasets and two models of interactions: evaluation of SAMIE trained according to the (A) one-to-one model of interactions, and (B) many-to-one model of interactions

		SAMIE model		
Evaluation set	No. of examples	SAMIE_S4	SAMIE_P4	SAMIE_C4
A.				
SELEX 4	96	83 (0.865)	n/a	82 (0.854)
PHAGE 4	311	n/a	288 (0.926)	287 (0.923)
Total	407	n/a	n/a	369 (0.907)
B.				
		SAMIE_S6	SAMIE_P6	SAMIE_C6
SELEX 6	99	92 (0.929)	n/a	89 (0.899)
PHAGE 6	279	n/a	256 (0.918)	252 (0.903)
Total	378	n/a	n/a	341 (0.902)

The names of the evaluation sets are indicative of the type of data they contain (i.e. SELEX data only, phage display data only or both sets combined). The number of examples contained in each evaluation set is shown in the second column. The number of the examples that the various SAMIE models ranked in the top 10% of all the predictions is reported, together with the $SR_{0.1}$ (number in parentheses).

Evaluation on known *in vivo* binding sites

The database used to train SAMIE contains results from *in vitro* selection experiments involving the proteins of the EGR family, and it contains EGR protein–DNA binding pairs known to occur naturally *in vivo*. The latter were not used in training SAMIE. It is therefore of interest to see how well SAMIE evaluates the known, *in vivo* binding sites for the EGR family. Furthermore, we report results on how well SAMIE ranks known, *in vivo* binding sites for three yeast proteins that are not homologues of EGR (i.e. MIG1, MIG2 and ADR1), but that contain two Cys_2His_2 domains each. Similar analysis is performed with the Drosophila Cys_2-His_2 protein Tramtrack.

For these tests, we use SAMIE_C6, which is trained on the combined dataset according to the many-to-one model of interactions. All rankings are based on the probabilities as they are calculated by SAMIE (equation (1)). For the prediction of the natural binding sites of the yeast proteins, we used the GC content of the organism to define the prior probabilities, P_n, in equation (1). This reflects a “protein’s view” of the genome, where the different binding sites “compete” for that protein. Similar results were obtained when an equal probability of 0.25 was used for all bases.

Evaluation on known in vivo EGR-binding sites

Apart from the SELEX and phage display examples, our database contains 24 *in vivo* DNA

† Nucleotide representation according to IUPAC: V: G or C or A; K: G or T; S, G or C; N: A or C, or G or T; the triplet refers to the codon.

Table 2. The energy matrix as it was calculated by SAMIE when trained on the COMBINED_6 dataset

	Finger position = −1; base position = 3				Finger position = +2; base position = 4				Finger position = +3; base position = 2				Finger position = +6; base position = 1			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	1.29	0.92	1.72	6.42	0.25	−0.87	1.32	0.58	0.54	2.12	2.27	−1.14	−0.33	1.38	−0.13	0.72
C	5.21	4.71	6.29	4.51	3.86	−0.93	6.92	4.94	4.16	1.89	6.27	−0.08	4.68	5.07	5.11	5.39
D	1.93	−2.35	0.33	−0.68	2.27	0.42	−0.44	−0.29	5.82	−1.32	7.20	6.32	1.75	0.64	1.17	−0.21
E	0.34	−1.06	−0.05	−0.98	6.34	−1.72	1.88	0.29	1.30	0.43	1.98	1.16	0.09	1.68	1.26	−0.08
F	5.21	4.71	1.33	4.51	3.86	5.55	1.36	4.94	4.16	6.92	6.27	5.13	4.68	0.42	5.11	5.39
G	2.45	0.44	1.07	0.47	0.23	−1.58	0.39	−0.55	0.27	1.43	2.27	−0.51	2.62	2.62	1.61	1.54
H	1.41	−0.10	1.42	−0.85	−0.38	−0.99	0.20	−1.78	0.28	2.64	−1.31	6.72	1.73	6.10	−0.08	0.33
I	1.55	1.58	1.16	1.77	1.05	5.55	8.45	5.90	5.68	2.58	7.23	1.23	1.81	1.77	5.78	6.46
K	0.33	−0.37	−1.04	−1.14	6.59	−0.94	2.70	6.47	5.54	7.47	0.77	1.04	0.87	6.99	−1.01	−1.38
L	1.18	0.19	7.89	−0.46	7.05	6.63	3.50	1.14	6.93	1.54	8.16	0.89	1.09	2.44	0.79	7.94
M	6.52	4.99	6.55	−1.55	6.33	5.55	2.41	0.28	0.26	7.32	2.06	−0.29	6.91	5.71	5.48	6.27
N	0.04	0.73	0.06	−1.33	1.14	−1.72	0.75	1.21	−3.73	0.86	−0.005	−0.80	−0.81	0.48	−1.00	1.75
P	7.33	5.89	7.36	0.75	7.01	0.17	2.27	6.58	6.20	8.16	2.85	1.73	7.70	2.18	0.21	0.74
Q	−2.50	6.32	0.72	−0.52	0.20	−1.32	1.78	0.30	−0.26	1.34	0.77	0.35	−0.92	6.10	−0.08	0.33
R	1.01	3.26	−1.36	0.46	1.66	−1.05	2.11	7.37	6.59	3.54	2.56	2.14	0.38	0.55	−3.33	1.29
S	1.06	0.73	0.61	−0.57	2.29	−1.45	0.28	−0.91	6.06	1.15	1.69	−1.65	−0.21	0.03	−0.40	−0.39
T	0.27	1.72	0.59	−1.33	−0.18	−1.52	1.29	1.23	7.50	−0.04	3.26	0.26	−1.19	0.20	−0.63	−0.46
V	7.59	6.70	1.57	0.92	7.01	0.17	1.76	0.57	8.01	0.79	8.52	0.80	1.13	1.98	0.46	0.78
W	2.21	4.71	6.29	4.51	3.86	−0.52	0.67	−0.71	4.16	6.92	6.27	−0.08	4.68	5.07	5.11	−0.27
Y	5.21	4.71	−0.05	4.51	4.95	−1.25	1.43	5.31	4.16	6.92	1.27	5.13	−1.37	5.85	5.23	5.61

The COMBINED_6 dataset consists of both SELEX and phage display data, according to the many-to-one model of interactions. The energy values presented here have been normalised so that the average binding constant for each contact is 1.0. This normalisation does not affect the calculation of probabilities, as described in the text. The normalised matrix shows that the energetic potential of a base–amino acid contact varies, depending on its position in the DNA target and the protein.

targets (10-mers) of the EGR proteins. These data are not included in the training sets, partly because the choice of prior probabilities for both the bases and the amino acid residues would be arbitrary. Furthermore, SAMIE was trained on the 4 bp (sub)targets and not on the complete sequences. Nevertheless, parts of these sequences had been recovered in SELEX or phage display experiments. The 24 naturally occurring binding sites consist of 31 different tetranucleotides, 14 of which are included in the COMBINED_6 training set.

SAMIE is able to rank 21 out of the twenty-four 10 bp natural sites at positions 1–2000 (i.e. in the top 0.2% of all possible 10 bp targets). The remaining three are sites that have been found in the promoter regions of the human basic fibroblast growth factor[37] and the human interleukin 2 gene,[38] and they rank at positions 11,766, 51,436 and 62,580 (or in the top 6% of all possible targets). It is of some interest that at least one of these three sites is not unambiguously a target for the EGR genes. In the case of the site found in the promoter of the human basic fibroblast growth factor,[37] there were two additional putative target sites in the proximity of the genomic region, and these additional sites were ranked by SAMIE_C6 at positions 258 and 661, respectively.

Evaluation on known in vivo MIG-binding sites

In yeast, there is no EGR homologue. However, yeast has a number of proteins that contain the Cys_2His_2 zinc-finger domain. Transcription factors MIG1, MIG2 and ADR1 are three known examples. All three of them contain two fingers that are quite similar to those of the EGR. However, there is no global similarity between the yeast proteins and the EGR, or between the MIG proteins and ADR1. On the contrary, there are some major differences: (1) the three yeast proteins contain one zinc-finger domain less than the EGR; (2) the domains are located in the amino terminus of the yeast proteins, whereas EGR has them at the carboxy terminus; (3) the ADR1 is twice the size of the others and it is known to require an additional 20 amino acid residues to recognise its DNA target efficiently.[39]

The MIG transcription factors are known to regulate a number of yeast genes. One of their primary targets is SUC2-A and it has been shown that the MIG proteins bind to its promoter region with high affinity. The target site for the SUC2-A promoter is believed to be 5′-GCGGGGA-3′.[40]

MIG1 and MIG2 are identical with respect to the amino acid residues at the "contacting" positions of the two fingers. Assuming that the two MIG proteins, similar to EGR, bind the DNA in an antiparallel fashion, we used SAMIE_C6 to predict their composite 7 bp long target sequence. The weighted LOGO of their possible target sites was calculated by program ENOLOGOS (P.V.B. *et al.*, unpublished results) and is presented in Figure 4. ENOLOGOS uses the algorithm presented by Schneider *et al.*,[41] but each sequence is weighted according to the predicted probability of binding.

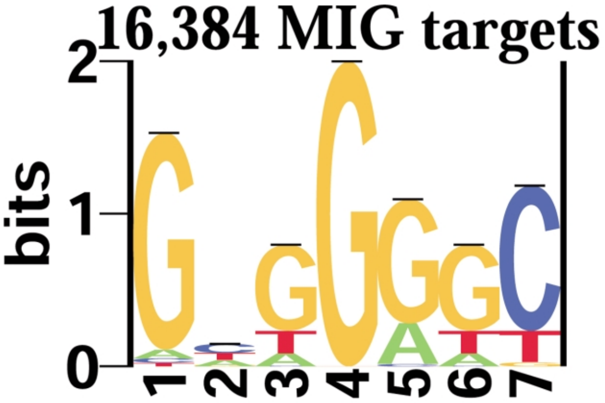


Figure 4. The weighted LOGO of the yeast zinc-finger proteins MIG1 and MIG2. This LOGO plots all possible nucleotide targets, weighted by the probability of interaction, as it is calculated by SAMIE. The naturally found binding site with the highest affinity to the MIG proteins is 5′-GCGGGGA-3′.[40]

Compared to the binding site found in the promoter of the SUC2-A gene, SAMIE predicts correctly all but the last nucleotide position (where it predicts C or T instead of A). The 6 bp (sub)sequence 5′-GCGGGG-3′ of the natural site was SAMIE's topmost prediction (out of a total of 4096 possible sites). Lutfiyya *et al.* reported the results of SELEX experiments that they performed using these proteins.[40] These results agree with SAMIE that A is anti-selected at the last position and they show that MIG proteins prefer G in this position, with T or C being their second preference. This observation raises the point that other factors (besides the affinity) might play an important role in the *in vivo* selection of the binding sites. The transcription factors might recognise genomic sequences that do not exhibit the highest affinity, although one would not expect them to differ too much from the highest-affinity ones. The SELEX results show that the second position is the least conserved; an observation that agrees with predictions made by SAMIE. In the same paper, Lutfiyya *et al.* reported the binding sites for three other genes that are regulated by MIG1 and/or MIG2. The measured binding affinity of MIG proteins to these sites is not as high as to that of SUC2-A. All the three sites are predicted to be in the top 5% of SAMIE's list of targets.

Evaluation on the known in vivo ADR1-binding site

We repeated the above analysis for the binding site of the ADR1 protein, but the results were not as good as with the MIG proteins. In particular, although the assumed binding site in the regulatory region is 5′-TTGGAGA-3′, our consensus sequence matches only four of these seven nucleotides (i.e. positions 2–4 and 6). Yet, the natural binding site ranked at position 1151 or in the top 7% of all possible 16,384 heptanucleotides. NMR studies on ADR1 (free and bound to specific DNA) indicate that this transcription factor utilises only amino acid positions −1, +3 and +6 of finger 1 and amino acid position −1 of finger 2 to target the 5′-GGAG-3′ subsite. According to SAMIE_C6, this is the second preferred subsite of this protein (among 256 possible; the first has G instead of A at the third position).

The fact that SAMIE is not able to predict the complete ADR1-binding site effectively may reflect the strong deviation from the EGR pattern of contacts. In fact, it is known that a region amino-terminal to the first finger is essential for the binding of the protein.[42] Nonetheless, SAMIE predicts correctly the 4 bp subsite for which the zinc-finger contacts are conserved. This observation indicates that its underlying model (or recognition code) should be consistent and its predictions could be useful even for more distantly related zinc-finger proteins.

Evaluation on the known in vivo Tramtrack-binding site

Tramtrack is another member of the $\mathrm{Cys_2His_2}$ protein family and it is probably one of the best studied genes in Drosophila. It regulates the developmental gene *fushi-tarazu*. Like the MIG and the ADR1 proteins, it contains two zinc-fingers. Tramtrack provides an interesting test case, mainly because its co-crystal structure has been solved and we know that its contacting scheme is slightly different from that of the EGR.[43] In particular, the amino acid position +6 of finger 2 as well as the amino acid position −1 of finger 1 do not participate in the binding. The base at position 6, which is normally contacted by the amino acid residue at position −1 of finger 1, in the case of Tramtrack is contacted by the amino acid residue at position +2. The protein in the crystallised complex is bound to the sequence 5′-AGGAT-3′, which is contained in the promoter region of the *fushi-tarazu* gene. The heptanucleotide sequence around this region is 5′-AAGGATA-3′.

We used SAMIE_C6 to predict the 7 bp binding site of Tramtrack under an assumed pattern of contacts identical with that of the EGR. The resulting weighted consensus sequence is presented in Figure 5. The prediction agrees with the naturally occurring binding site (i.e. 5′-AGGAT-3′)[43] and the agreement even extends to the base that is upstream of that in the genome. The prediction disagrees with the base downstream of the binding site in the genome (SAMIE_C6 predicts C or T instead of the naturally found A). But the crystal structure shows that this last base does not interact with any amino acid residue. Nonetheless, the naturally found 7 bp site ranked at position 67 of SAMIE's list (or in the top 0.4%).

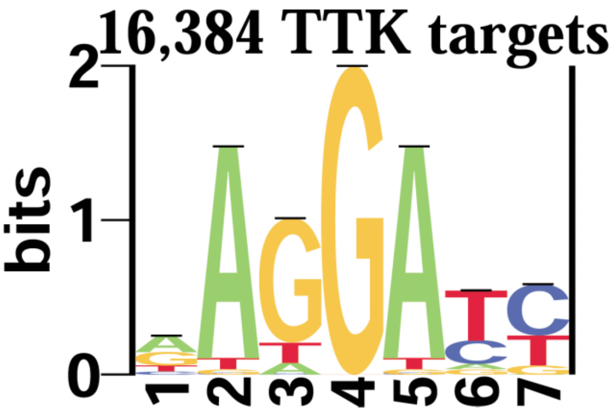


Figure 5. The weighted LOGO of the Drosophila zinc-finger protein Tramtrack. This LOGO plots all possible nucleotide targets weighted by the probability of interactions, as it is calculated by SAMIE_C6. Interestingly, TTKB recognises only the 5′-AGGAT-3′ subsequence *in vivo*.[43]

Interestingly, Figure 5 depicts the subsequence 5′-AGGA-3′ (in the middle) as the part with the strongest signal. These are the exact base positions where the two fingers have an identical pattern of contacts to the EGR protein. The first base position does not exhibit any strong preference, agreeing with the observation that amino acid position +6 of finger 2 does not contact the base at this position (which, in fact, happens to be A in the natural site). Similarly, the crystal structure shows there is no amino acid residue contacting the last base (predicted to be C or T by SAMIE; naturally found to be A).

All the above observations show that SAMIE can predict the natural-binding site of a protein very accurately, even if it deviates slightly from its learned pattern of contacts. It indicates that minor docking rearrangements, although they change the overall pattern of contacts, can still allow for a reasonable prediction of a binding site.

Correlation with measured energy data

One of SAMIE's characteristics is that it associates the frequencies of the observed contacts to the binding energies of the corresponding interactions. If we assume that the weight matrix that is calculated by SAMIE corresponds to the real energetic potentials of the base–amino acid contacts†, then we would expect that the measured binding constants, K_A, will be related to the energy values, H,

† There are two points worth noting, here. First, equations (1) and (3) lack the temperature parameter that is present in the Boltzmann distribution: i.e. we assume a constant temperature for all experiments. Second, the energy values of each base–amino acid contact have been adjusted to an arbitrarily defined zero point; but this does not alter the probabilities calculated from these equations.

according to the formula:

$$K_{\mathrm{A}} \propto \mathrm{e}^{-H} \tag{4}$$

Thus, if SAMIE's values correspond to the real binding energetic potentials, then the correlation coefficient between the measured association constants and SAMIE's predicted probabilities should be high. There are a number of examples in the literature where binding constants have been measured for EGR-derived proteins and various DNA targets, and we use them for further evaluation of SAMIE.

We generally use the correlation coefficient statistic to measure the agreement of the observed *versus* the predicted values. When the measured data are limited and they consist mainly of high-affinity pairs, then the correlation of the measured *versus* predicted energy is evaluated. When both high-affinity and low-affinity pairs are included, though, the evaluation will be done on the corresponding probability (or, equivalently, K_{A}) values. The reason we prefer to use the probabilities in this case is that it is not very important for a prediction method to be accurate in the low-affinity states, which have large energies and diminish the correlation coefficient unduly.

Comparison with the data from Hamilton et al.[44]

As a first test, we use the data reported by Hamilton *et al.* on the wild-type EGR1 protein bound to various DNA targets.[44] Should our model be consistent, one would expect that it will be able to predict well at least the specificity of the wild-type protein to these targets. Hamilton *et al.* reported the ratios of the dissociation constants to 15 DNA targets relative to the wild-type DNA target. We use these ratios to calculate the corresponding energy differences (see equation (4)). Then we calculate the predicted energy differences for the same targets, using the SAMIE_C6 weight matrix (Table 2). We find that the correlation coefficient between the two sets of values is $R = 0.75$, $P < 0.01$. This represents quite a good fit to the data, especially given the fact that accurate measurement of K_{D} was not possible for some of the targets.[44] The R value for the K_{A} terms on EGR1 is 0.69.

In the same paper, Hamilton *et al.* reported the ratios of the dissociation constants for the WT1 wild-type protein to 18 DNA targets. WT1 protein belongs to the $\mathrm{Cys_2His_2}$ zinc-finger family but, compared to EGR, it has one additional finger. The sequence of the three last fingers of WT1 is very well conserved compared to the three fingers of the EGR1 proteins. We repeated the above analysis using the WT1 data and we found that our predictions correlate strongly with them too ($R = 0.76$, $P < 0.001$). The R value for the K_{A} terms on WT1 is 0.65.

The strong correlation between SAMIE's predictions and measured energy values for the two wild-type proteins and various targets is indicative of the potential of our model. Indeed, it seems that, given sufficient training data, which are entirely qualitative, SAMIE can determine a quantitative model, or P-code, that can predict the corresponding energy values fairly well.

Comparison with the data from Segal et al.[45]

We extend the analysis in predicting energy values for variants of the EGR family. Segal *et al.* reported the K_{D} values for a number of such variants.[45] For nine of these proteins, K_{D} was measured for two alternative DNA sites. In two cases, the selection data in the training set contradict the measured values directly or they are not sufficient to model the corresponding interactions accurately. In the case of the protein *s***r***s***dd***lv***r** bound to gcg and gag targets, the measured K_{D} values are 9 and 6, respectively. In other words, this protein binds the same or stronger to the latter target. According to the crystal structure, this means that the Asp at position +3 of the finger contacts the A in the middle position the same or more strongly than a C at the same position. Our combined dataset that was used for the training of SAMIE (i.e. COMBINED_6) contains 82 examples with Asp at position +3 and a C is found in the second base position in all cases (Table 5). We do not know the source of this discrepancy between the data from the randomisation experiments and the measured values. It may be that the particular protein variant deviates strongly from the pattern of contacts observed in the crystal structure. In any case, we excluded this example from further analysis.

In the second case (i.e. protein *s***r***s***dd***lv***r** bound to ggg and gtg), the relative K_{D} value between the two targets is greater than 233 (the reported K_{D} values are 6 and >1400, respectively). That order of preference agrees with our dataset, where Lys at position +3 clearly favours G over T in the middle position (14 G *versus* one T), but the magnitude cannot be estimated accurately from our current training set. Thus, this example was excluded from further analysis.

For the remaining 14 sequences (seven pairs) with reported K_{D} values, the difference between the experimentally measured relative energies and the SAMIE's predictions correlate very well ($R = 0.82$, $P < 0.025$). The R value for the K_{A} terms on all sequences (not pairs) is 0.79.

Comparison with the data from Miller & Pabo[16]

More recently, Miller & Pabo used the wild-type EGR protein and the mutant D20A to measure the relative binding affinity for the trinucleotide targets GNG and GCN.[16] D20A has an Asp to Ala replacement at position +2 of finger 1 with respect to the wild-type protein. The amino acid replacement is expected to affect the overall binding affinity of the protein due to the contact of the replaced amino acid to the overlapping base (G). However, if the two proteins contact the DNA in the same way (i.e. the pattern of contacts for the D20A is that observed in the crystal structure), then the replacement should not affect the specificity of this protein to the DNA targets used in the study. This is because, according to the crystal structure, the nucleotide positions that vary are not contacted by the amino acid at position +2 of finger 1. Consistent with that, SAMIE_C6 predicts that the Asp to Ala replacement at position +2 of this finger should result in a sixfold increase of the dissociation constant for all seven studied targets. However, the K_D values reported in that paper for the two proteins are practically the same. This can be explained if one assumes that the amino acid residue at position +2 of finger 1 does not contact the base at position 10 of the DNA target, thus not contributing to the binding energy at all. In fact, the refinement of the crystal structure has shown this amino acid position to be at marginal distance from base position 10 with respect to hydrogen bond contacting potentials.[28]

Furthermore, their data show that protein D20A has a considerably low dissociation constant towards DNA target GCT. This fact cannot be predicted by a simple recognition code if one assumes a conserved pattern of contacts. The authors also solved the co-crystal structures of the D20A protein bound to two different DNA targets (GCG and GCT). They found that, overall, the D20A maintains the same contacting scheme, although they found a less ordered complex around bases 10 and 11 (note that base 10 is the base that is expected to be affected most by the Asp to Ala replacement). The authors conclude that their findings cannot be explained by a "simple recognition code".

We use equation (4) to predict the association constants for the two proteins and the seven DNA targets, using SAMIE's energy values (Table 2; Figure 2). Then we compare these predictions with the observed values (i.e. association constants) and we find them to be highly correlated (R values of 0.93 and 0.81 for the wild-type and D20A protein, respectively; $P < 0.01$ in all cases). This might seem surprising; however, we note that the published K_D values for the two proteins are also highly correlated ($R = 0.87$, $P < 0.01$). In fact, apart from the GCT target, the two proteins agree very well on their K_D values ($R = 0.94$, $P < 0.005$). The R value for the energy terms on D20A is 0.62; on Zif268 is 0.56, and on both is 0.58.

Comparison with the data from Bulyk et al.[19,46]

Determination of the dissociation constants of various EGR-derived proteins against all possible trinucleotide targets has been performed with the use of microarray technology.[46] In this study, the wild-type EGR protein and four variants with amino acid substitutions on the middle finger of the protein were used to bind to a microarray that contained all possible trinucleotide targets for the middle finger. Assuming that the intensity of the signals corresponds to the affinity of the interactions, the authors calculated the dissociation constants using these intensity values. We analysed these data and compared their results with SAMIE's predictions.

In their first study,[46] there are ten targets reported for the wild-type protein, but only seven of them exhibited high affinity (according to the authors, the other three that were reported exhibit lower affinity and/or they were used as negative controls). SAMIE_C6 ranks the first six of the seven high affinity targets at position 8 or higher (out of the total 64 possible trinucleotides). The correlation coefficient between the energies predicted by SAMIE_C6 and the logarithms of the ten measured K_D values is 0.61, $P < 0.05$; the conversion of the K_D values to energies was done according to equation (4). This result also shows very good correlation between experimental energy measurements and SAMIE's predictions for the high-affinity sites. Furthermore, if one compares the measured K_A values for the entire set of 64 triplets[19] with the predictions from SAMIE, the correlation is 0.61. While not as high as some of the examples described previously, this is still a very good correlation between measured and predicted binding constants over the entire range of highest to lowest-affinity sites.

One of the four mutant proteins analysed in that paper is RGPD (the amino acid residues correspond to positions −1 to +3 of the helix of the second finger). The authors reported six high-affinity DNA targets for this protein in their first study. The first five of them rank at positions 1–5 according to SAMIE, although SAMIE's ranking order is different from that reported. The correlation coefficient, calculated like before over the measured and predicted energy data is 0.73, $P < 0.10$. The LOGO of all binding sites weighted by SAMIE's probability of interaction (Figure 6) is very similar to the one presented in that paper. The correlation of the measured K_A values to those predicted for all 64 triplets is 0.99. Notably, finger 2 of the protein RGPD (i.e. the finger under study) differs from the corresponding finger of the wild-type protein in all "contacting" amino acid residues, except position −1 of the helix (Arg). The fact that SAMIE can predict the binding affinities of this mutant protein to the entire collection of low and high-affinity sites extremely well demonstrates the potential of the probabilistic algorithms for modelling the protein–DNA interactions.

The results of the analysis are similar for the mutant protein REDV. The authors report the K_D values for six targets,[46] but only two of them really exhibit high affinity (GCG and GTG). SAMIE_C6 ranks these two targets at the two topmost positions of its list. Furthermore, we find that SAMIE's predicted energy values for the six reported targets correlate very well with the experimental values; $R = 0.80$, $P < 0.05$. The six

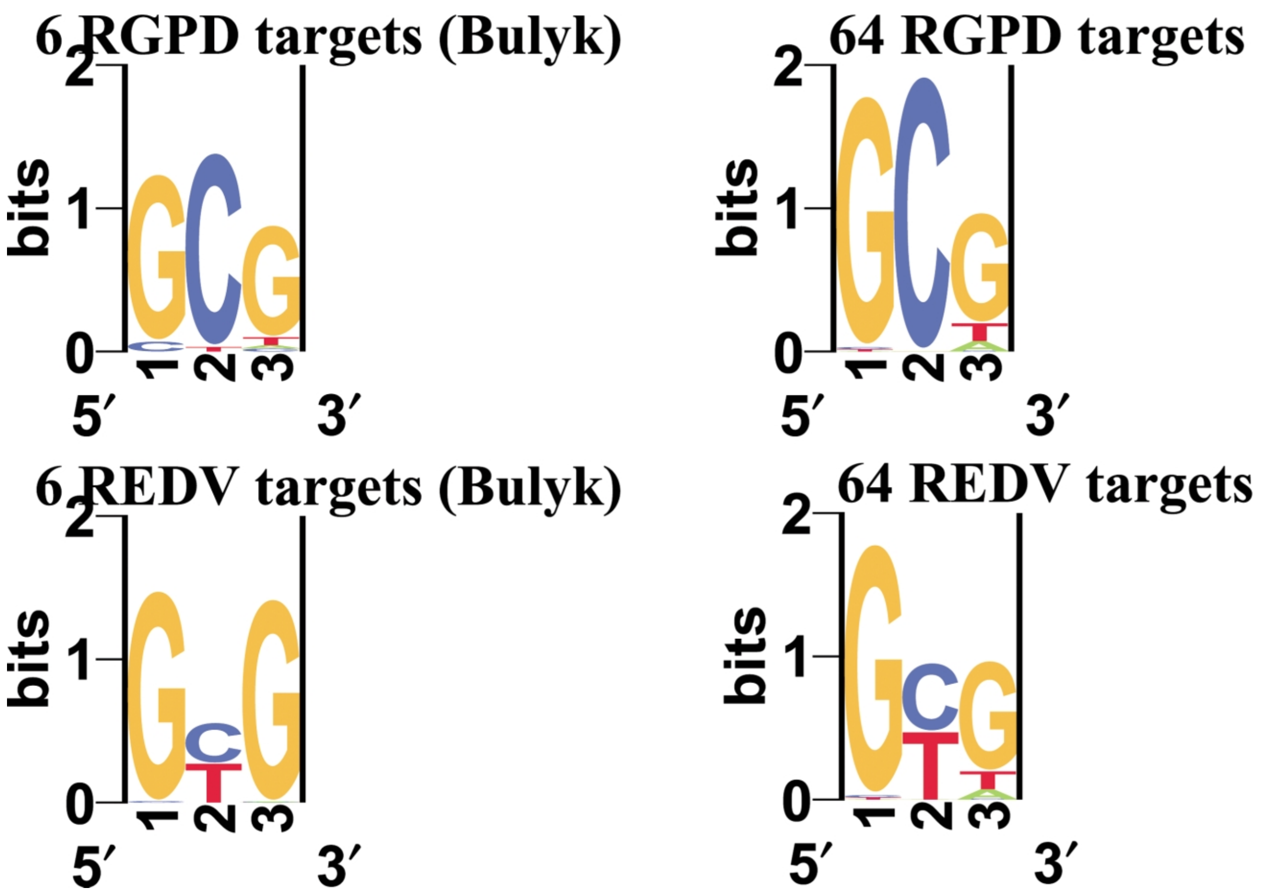


Figure 6. Weighted LOGOs of the DNA-binding sites of two mutated EGR fingers. The fingers correspond to protein variants RGPD and REDV presented by Bulyk *et al.*[46] For the LOGOs, we plot all possible nucleotide targets, weighted by the probability of interaction as it is calculated using the SAMIE_C6 weight matrix. These two LOGOs are very similar to those presented by Bulyk *et al.*[46] (and reconstructed here) that are weighted by relative binding affinity.

reported targets ranked in the top nine positions of SAMIE's list. We calculated the LOGO of all predictions (weighted by SAMIE's probabilities; Figure 6) and we find it to be very similar to that presented (weighted by relative affinity of the high-affinity targets; see Bulyk *et al.*[46]). Furthermore, the correlation between measured and predicted K_A values for all 64 triplets is 0.73.

The LRHN shows less specificity than the others, with 13 high-affinity sites. Even so, the predicted consensus sequence matches that derived from the experiments, TAT, and the overall correlation between predicted and measured K_A values is 0.56.

The protein KASN shows practically no specificity, and no consensus binding sequence emerged from the experiments.[46] All of the measured K_D values were at least 83 times higher than that of the wild-type protein to its preferred site. Not surprisingly, SAMIE does not predict the binding affinities well in this case, with a correlation of only 0.16 for all 64 triplets. This result is consistent with the notion that the predictions from SAMIE are best for those proteins with high specificities, which includes all natural transcription factors.

Comparison of SAMIE with other methods

In this section, we compare SAMIE's prediction accuracy with that of other quantitative and qualitative models in order to assess its strengths and its possible weaknesses. For the evaluation, we use the SELEX and phage display dataset(s).

Comparison with other quantitative models

We compare SAMIE_C6 with the two currently available quantitative models: that presented by Suzuki and colleagues[10,47] and that from Margalit's group.[7,13] A third approach, followed by Kono & Sarai,[12] is very similar to that of Margalit's,[7,13] in the sense that the scores for the base–amino acid contacts are calculated to be proportional to the logarithm of the frequencies in the training set. The training sets for the two groups consist of 53 (Margalit's) and 52 (Sarai's) non-redundant co-crystal structures. The main difference is that Margalit's group is considering contacts between bases and amino acid side-chains only, whereas Sarai's model considers also the contacts to the DNA and protein backbone. However, when examining the additional data provided by Kono

& Sarai,[12] we did not find any significant base preference of the amino acid residues towards DNA in them. Another difference between Sarai's and Margalit's models is in the normalisation of the base–amino acid frequencies. Kono & Sarai use the amino acid frequencies in the training set as prior probabilities, whereas Margalit's group derives theirs from the SWISS-PROT database. Unlike Suzuki's and Margalit's groups, Kono & Sarai do not provide the final scoring matrix for their method. Thus, it is very difficult for us to evaluate their model and compare its predictions with those made by SAMIE.

For the comparison of SAMIE with the other two methods, we use the SAMIE_C6 weight matrix and the latest models published by these groups. We (re)form their data into weight matrices equivalent to SAMIE's, according to the many-to-one model of interactions. Then, for each vector in the SELEX_6 and PHAGE_6 datasets, we rank all possible variable sequences with respect to their score towards the fixed sequence (see Figure 9). The score is calculated as the sum of scores of individual base–amino acid contacts. For SAMIE_C6, this score represents an estimate of the binding energy of the interaction, which constitutes one of the advantages of our method. The specificity index and the success rate are calculated as before (see Materials and Methods). The only difference is that in this case the corresponding rankings are based on the sum of weights instead of the probability values. This results in an underprediction of the phage display data (e.g. SAMIE's $SR_{0.1}$ value drops to 0.875 from 0.903; see Table 1). This is unavoidable, though, since the other two models do not provide a method for calculating the binding probabilities. The reason for the underprediction is that ranking phage display data according to score treats all amino acid residues as equiprobable. This is not true, even if the randomisation scheme for each amino acid position in the phage display experiments was NNN, where N stands for A or C or G or T and each N refers to a codon position. The results are presented in Table 3.

Table 3. Prediction scores of the available quantitative models

	SAMIE_C6		SUZUKI		MARG-ALIT_01	
Evaluation set	$SR_{0.1}$	SI	$SR_{0.1}$	SI	$SR_{0.1}$	SI
SELEX 6	0.896	95.4	0.917	91.2	0.365	80.9
PHAGE 6	0.875	96.3	0.620	83.7	0.570	85.3
COMBINED 6	0.880	96.1	0.696	85.6	0.517	84.2

Currently available quantitative models are compared in terms of their ability to predict SELEX and phage display data. The models are: (a) SAMIE_C6, (b) the one presented by Suzuki and colleagues[10,47] and (c) the one from Margalit's group. For the evaluation of the accuracy of the predictions the success rate and the specificity index (mean value over all predictions) are used, as they are defined in Materials and Methods.

In terms of success rate, Suzuki's model is predicting the SELEX data better than the other two methods. It ranks correctly 88 vectors in the top 10% of their list ($SR_{0.1} = 0.917$). In comparison, SAMIE_C6 ranks correctly 86 vectors ($SR_{0.1} = 0.896$) and Margalit's model[35] ($SR_{0.1} = 0.365$). However, on average, SAMIE_C6 has a better specificity index (95.4 compared to 91.2 of Suzuki's model and 80.9 of Margalit's model). On predicting phage display data, SAMIE_C6 out-performs the other two methods with $SR_{0.1}$ value of 0.875 and SI_{avg} of 96.3. Interestingly, although Suzuki's model has a higher success rate than Margalit's on phage display data, the latter has better average specificity index value. Finally, on the overall performance SAMIE_C6 is clearly better than the other two. Suzuki's model has a better $SR_{0.1}$ value than Margalit's (0.696 compared to 0.517), but they both have about the same average specificity (SI of 85.6 for Suzuki's model compared to 84.2 for Margalit's).

SAMIE_C6 and Suzuki's model predict SELEX data with similar high levels of accuracy. This is interesting, since SAMIE is a data-driven statistical mechanical approach, whereas Suzuki's model is based on the chemical and stereochemical properties of the base–amino acid contacts. The fact that the two can predict well the SELEX data supports the idea that common principles might exist that govern these interactions; and, if so, we might be able to formulate them in a mathematical way. This is supported further by the fact that the submatrices of SAMIE_C6 and Suzuki's model for the contact between amino acid position -1 and base position 3 are correlated ($R = 0.56$, $P < 0.10$)†. The same is true (although the R value is smaller) for the weight submatrices that correspond to the contact between amino acid position $+3$ and base position 2 ($R = 0.41$, $P \sim 0.05$). These are the only two single contact positions (see Figure 1).

Suzuki's model for the EGR protein family predicts that the submatrices of certain base–amino acid contacts should be related. In particular, the submatrix for amino acid position -1 of the helix should be identical with that of position $+6$ (both contacts involve large amino acid residues only). According to SAMIE_C6, the weight matrices for these contacts have the highest correlation coefficient value (i.e. $R = 0.4$, $P < 0.005$). Although the probability value is low, the correlation coefficient is not high enough to consider these two contacts identical. All coefficients between the submatrices of SAMIE_C6 that correspond to the modelled contacts show positive correlations (R values between 0.18 and 0.4, and $P < 0.05$ and < 0.001). This reflects the fact that each amino acid has preferences for only certain bases, and *vice versa*, but the

† The correlation coefficient was calculated only on the base–amino acid pairs that Suzuki's model permits contact.

correlations are low because each position also has differences in their contacts. Those similarities and differences are embodied in Suzuki's chemical and stereochemical merit points and allow it to do a fairly good job of predicting SELEX results. But SAMIE has the advantage of optimising the parameters of the model on the basis of experimental data.

We note that Margalit's model, by design, does not take into consideration the position-variation of the base–amino acid contacts. It consists of a single weight matrix, which is used to model any contact, regardless of its position in the protein/DNA. We believe that this "averaging over all contacts" constitutes one of the limitations of this model and is responsible for the somewhat lower performance compared to the other two.

Comparison with the qualitative model(s)

We implemented the qualitative model for the EGR protein family, which can be viewed as a list of position-specific "permissible" base–amino acid contacts. This model was initially presented by Choo & Klug[15,31] and Pabo and colleagues later.[14,16,34] The models proposed by these two groups, however, are different with respect to the set of observed/permissible contacts they include. In fact, the contacts included in both models are less than half of the total number. Giving the "benefit of the doubt" to the qualitative modelling, we used the composite set of permissible base–amino acid contacts (Table 4). This set is based on the Tables presented by Choo & Klug[15] and by Miller & Pabo,[16] respectively. We tested this model on our COMBINED_6 training set and we found that it can predict only 61.1% of all observed base–amino acid combinations (i.e. 1228 out of the total of 2009). In other words, about 40% of the experimental data included in our set are not catalogued in either of the two qualitative models. We must emphasise, though, that the combinations contained in the COMBINED_6 dataset are only "inferred" contacts; that is, if one assumes a pattern of contacts identical with that found in the EGR co-crystal structure.[28] Of course, this is not the case for all of them. For example, in some cases the amino acid residue might be too far away to form a bond with the base. However, the fact remains that all of these combinations are observed in selection experiments, yet the qualitative models were not able to predict them. Finally, we found that the composite qualitative model is able to predict correctly only 195 of the 820 single-finger examples (or 23.8%) in the COMBINED_6 set. In this case, we consider a prediction "correct" if all observed base–amino acid combinations are listed in the composite qualitative model; each of the individual models does worse than this. This shows that the majority of interacting fingers and sites obtained in *in vitro* selection experiments are not accounted for by the simple, qualitative models. In contrast, the quantitative P-code model allows for all possible combinations of fingers and sites, and the predicted rankings match the observed data quite well overall (Table 3).

Discussion

The present study addresses the problem of modelling protein–DNA interactions. Our thesis is that a probabilistic "recognition code" (or P-code) can model such interactions accurately enough for prediction purposes. We develop a probabilistic algorithm, SAMIE, that can "learn" the contact-specific, base–amino acid energetic potentials from data derived from randomisation experiments (SELEX and/or phage display).

The underlying theoretical model of SAMIE is based on statistical mechanics theory. According

Table 4. The composite qualitative model

	Position in DNA											
	1			2			3			4		
	C	b	P	C	b	P	C	b	P	C	b	P
A			Q		N	S H		Q		A		
C				L	D T V			D				S
G	S T	R	K		H	K		R			D	S F
T	S T		K		A S V	T	Q	N	L T	S	D	

The qualitative model can be viewed as a list of position-specific, permissible base–amino acid contacts. Two qualitative models have been proposed so far: one by Choo & Klug[15,31] and another by Pabo and colleagues[14,16,34] This Figure presents the permissible contacts that each group proposes. The contacts proposed by one of the groups only are in the columns marked C and P for Choo's and Pabo's models, respectively. The contacts that are present in both models are in the columns marked with b. Interestingly, less than half of the total number of valid contacts are proposed by both groups.

to our model, the probability that a fixed component (protein or DNA) will select a given variable component (DNA or protein, respectively) from a pool of randomised molecules is defined by the Boltzmann distribution (equations (1) and (3)). The energies constitute a weight matrix, which associates every base at a particular DNA target position with every amino acid at the corresponding "contacting" position(s) of the protein. The values of the weights in this matrix (i.e. the parameters of the model) are estimated from the base–amino acid frequencies observed in randomisation experiments (SELEX and/or phage display), following the steepest ascents method. SAMIE constitutes essentially a maximum likelihood approach for the estimation of the parameters (see equation (6)). The exact "contacting" amino acid positions need not be known *a priori*, although knowledge of them helps to reduce the number of parameters of the model. Here, we restrict the training to those contacts that have been observed in the co-crystal structures of the EGR protein.[27,28]

We further make the assumption that the base–amino acid interactions are additive over all contacts. In other words, we assume that the contacts are independent and therefore contribute additively to the total binding energy. We know that this assumption is not exactly valid,[18,19] but even in those studies the data show that additive models can be very good approximations.[20] For practical purposes, it is usually sufficient for it to hold only for the high-affinity states, where it tends to hold fairly well.[17,20] Nevertheless, our model is fairly general and it can represent both additive and non-additive interactions adequately at the cost of an increased number of parameters that needed to be estimated.

The performance of the model derived from SAMIE was evaluated on several different datasets.

Self-test

The training on each of the six datasets resulted in a different SAMIE model. Each of these six models is tested initially on predicting their training datasets. All SAMIE models were able to predict "correctly" 85% or more of the corresponding training vectors. In this context, correctly refers to the success rate that we define in Materials and Methods. Practically, it means that in 85% of the training vectors, the randomised counterpart of the vector was ranked by SAMIE in the top 10% of all possible randomised targets. Of course, testing performance on the training data runs the risk of over-fitting the parameters and obtaining artificially good results. We do not think that is a problem, because the number of examples greatly exceeds the number of parameters; in the COMBINED_4 and the COMBINED_6 datasets there are 2009 and 3612 base–amino acid combinations, respectively, and the SAMIE_C4 and SAMIE_C6 models have only 320 parameters. Nevertheless, the most important performance evaluations are on data not included in the training set.

The results of the self-test prove that our model is internally consistent, and confirm our notion that training on composite dataset(s) (i.e. SELEX and phage display vectors) results in a generally better model.

Evaluation on naturally found binding sites

We tested the ability to predict the naturally found binding sites of EGR as well as other zinc-finger proteins. Out of the twenty-four 10 bp long *in vivo* DNA sites present in our database for the EGR protein, SAMIE was able to rank 21 in the top 0.2% of all potential binding sites. The remaining three putative target sites rank only in the top 6%, but at least one of those is unconfirmed and there are higher-ranked alternative sites in the proximity in the genome.

We tested the prediction ability for proteins that are not related to the mammalian EGR family, apart from the fact that they all contain the Cys_2-His_2 motif. We used three well-characterised proteins, the yeast transcription factors MIG1/MIG2 and ADR1 and the Drosophila TTKB (Tramtrack). All of these proteins contain two zinc-fingers (compared to the three present in the EGR proteins) and in the case of ADR1 and TTKB a different pattern of contacts has been observed for one of the two fingers (there are no structural data available for the MIG proteins). In the case of the yeast transcription factors MIG1 and MIG2 (they are identical with respect to the "contacting amino acid residues"), SAMIE_C6 is able to predict correctly all but the last nucleotide of the 7 bp long DNA consensus target site. All of the known naturally occurred MIG target sites are ranked in the top 5% of SAMIE's list, and the rankings of bases in each position correspond closely to those obtained in SELEX experiments.[40] In the case of ADR1, SAMIE_C6 was able to predict correctly the 4 bp subsite that is believed to be contacted by one of the two fingers, and the *in vivo* target site ranks in the top 1% of all possible sites. In the case of TTKB, SAMIE_C6 predicts correctly all but the last nucleotide of the 7 bp long target. Moreover, it depicts correctly the 4 bp long subsite that is known to be significant for the binding, and the *in vivo* consensus sequences ranks in the top 0.4%.

These examples show that SAMIE's predictions can be very useful for estimating the binding specificities of EGR proteins, and even for proteins that are not members of the EGR family but use the Cys_2His_2 motif for DNA binding. There are several reasons why the predictions are not even better than these results. First, even though there are many examples of SELEX and phage-display combinations reported in the literature, more data should provide more accurate estimates of the parameters. Second, we know the additivity assumption is not completely accurate and so limits our prediction ability. But the fact that we do fairly well indicates that the additive model is a reasonable approximation. Third, we know that for some proteins the contacting interactions vary, which will limit our accuracy. And in some cases, such as ADR1, additional amino acid residues outside the zinc-finger contribute to binding specificity. And fourth, the *in vivo* sites that we attempt to predict may be influenced by other constraints, such as overlapping sites for other proteins or cooperative interactions that modify the interaction of the EGR protein. But, despite these potential problems, the probabilistic model, and the SAMIE method of determining the parameters, appear quite promising for providing reasonably reliable, quantitative predictions of binding site specificities, at least for EGR proteins and perhaps, with appropriate data, for other families as well.

Correlation with measured binding energy data

One of the advantages of the P-code model is that it provides quantitative predictions for relative binding energies. This allows us to predict the effects of variations, in either the DNA or protein sequence, on the interaction energy. We compared SAMIE's predictions of the binding energy changes with experimentally determined values for DNA targets bound by the EGR protein and its variants. In most cases, the SAMIE predictions are in close agreement with the experimental values. Correlation coefficients (R) between the measured K_A terms and those predicted by the SAMIE_C6 model are generally greater than 0.7, and sometimes much higher. The cases with lower correlations are usually due to an overall low specificity of the protein to the DNA targets (non-specific binding).

Comparison with other models

Other currently available protein–DNA interaction models include two† that are quantitative and two that are qualitative. These terms refer to the way that the models represent the base–amino acid interactions. The quantitative models assign a numeric value (“weight”) to these interactions, whereas the qualitative models catalogue them as permissible or non-permissible. The qualitative models can be considered as a degenerate form of quantitative, where the weights of the interactions have been adjusted to either 1 (permissible) or 0 (non-permissible)‡. Any quantitative model can be transformed into qualitative, by setting a threshold to separate the permissible from the non-permissible contacts.

All currently available models assume energetic additivity of the individual contacts. The most important reasons are that: (a) in the majority of the cases the interactions are approximately additive (especially in the high-affinity states); (b) data limitations make modelling of non-additive interactions impractical; (c) simplicity of the additive “codes”. However, we note that one of the advantages of SAMIE is that its underlying principles are quite general and thus it can accommodate non-additive interactions at a cost, which are more parameters to be estimated.

The first quantitative model, developed by Suzuki's group,[10,47] is based on a set of chemical rules for each base–amino acid pair (that are independent of the family modelled) and a set of stereochemical rules derived from co-crystal structures (protein family-specific). The combination of these sets of rules generates a score for any given protein–DNA pair. This score can then be used to rank DNA targets according to their binding probabilities and predict possible binding sites. The limitations of this model are the requirement for the crystal structure and the fact that the rules (or weights) have been assigned in a semi-arbitrary way. Comparison of SAMIE_C6 with this model showed that, in terms of success rate, Suzuki's model is slightly better in predicting SELEX data, and SAMIE_C6 was significantly better on phage display data and the combined data. In terms of specificity index (an evaluation measure introduced by Suzuki *et al.*[10]) SAMIE_C6 was substantially better on all sets.

The second quantitative model we compared SAMIE with was that developed by Margalit's group.[7] For this model, the score of each base–amino acid pair is derived from the observed frequency of this pair in a non-redundant set of crystal structures (training set). The score of the individual contacts is calculated according to the formula:

$$S_{ij} = \ln[f_{ij}/(f_i f_j)] \qquad (5)$$

where f_{ij} is the pair frequency of amino acid i and base j, f_i is the frequency of amino acid i ($i = 1,\ldots,20$) in the SWISS-PROT database and f_j is set to 0.25 for each base j ($j = 1,\ldots,4$). The original model was recently refined by including more contacts in the training set.[13] We compare SAMIE_C6 with the latter model because, in agreement with the authors, we found that it performs better than their initial model. In all cases, SAMIE_C6 performs better than Margalit's model. However, in terms of specificity index, Margalit's model is better than Suzuki's in predicting the phage display data.

One disadvantage of Margalit's method is that it assigns very high “penalties” (i.e. very low weights) to certain base–amino acid pairs on the

† The available quantitative models are, in fact, three but two of them, that proposed by Margalit[7,13] and that due to Kono & Sarai,[12] are very similar to each other.

‡ The representation of permissible and non-permissible contacts with 1 and 0, respectively, corresponds to assigning probability values. In order to be consistent with our previous notation (i.e. weights correspond to energy values), we can assign the value of 0 (or any number) to the permissible contacts and the value of $+\infty$ to the non-permissible contacts.

Table 5. Table of EGR frequency data

	nt = 3; aa = −1				nt = 4; aa = +2				nt = 2; aa = +3				nt = 1; aa = +6			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	3	1	2	0	25	24	35	14	3	18	2	24	17	11	9	11
C	0	0	0	0	0	3	0	0	0	1	0	1	0	0	0	0
D	1	28	5	7	33	42	431	121	0	82	0	0	1	4	1	7
E	4	4	6	6	0	7	7	2	4	73	3	5	10	10	1	7
F	0	0	1	0	0	0	2	0	0	0	0	0	0	1	0	0
G	1	2	4	3	11	16	51	15	4	7	2	13	2	7	4	9
H	3	9	3	21	11	8	28	21	3	100	84	0	1	0	2	3
I	2	1	3	1	2	0	0	0	0	5	0	1	1	1	0	0
K	4	2	16	7	0	5	2	0	0	0	4	1	4	0	12	29
L	5	6	0	14	0	0	2	2	0	8	0	4	7	2	4	0
M	0	0	0	8	0	0	2	2	1	0	1	3	0	0	0	0
N	6	1	6	11	2	10	15	2	78	24	9	7	19	9	8	1
P	0	0	0	2	0	2	7	0	0	0	1	1	0	1	3	4
Q	99	0	4	8	15	16	8	8	2	3	4	2	14	0	2	3
R	14	1	337	18	7	15	11	0	0	3	2	1	12	13	801	4
S	4	3	9	12	3	47	59	52	0	50	3	27	10	30	11	30
T	15	2	11	49	19	28	22	10	0	79	1	11	47	37	55	125
V	0	0	3	3	0	2	11	3	0	24	0	10	10	15	6	12
W	0	0	0	0	0	2	4	2	0	0	0	1	0	0	0	2
Y	0	0	4	0	0	7	2	0	0	4	1	0	8	0	0	0

The data for this Table are derived from the COMBINED_6 data set. Only the combinations for the contacts found in the crystal structure are presented. Although the two contacts from the neighbouring fingers (see also Table 2) have been added to their connected ones, the data are limited in parts of the matrix.

grounds that they do not have contacting potentials. We believe that this might affect the predictive power of this model in some cases. For example, Gly with no hydrogen bond donor/acceptor potential has been assigned the most negative score. This prevents it from appearing in many of the model's predictions. Yet, by being a small, non-polar amino acid, Gly can be tolerated in most cases. In our training set, there are 151 examples with Gly present in a "contacting" position (Table 5). Another important limitation of Margalit's model is that it assigns the same score to each base–amino acid contact, irrespectively of their position in the DNA and protein. In other words, it does not consider any stereochemical rules (as in Suzuki's model). The advantage of this model over Suzuki's, though, is that it uses a probabilistic view of the data to assign the scores (i.e. it "learns" from the data).

In terms of quantitative modelling, SAMIE combines characteristics of both these models. Its weights are position-specific, so they are adjusted during training in a way that reflects the chemical and the stereochemical rules. In this case, a sufficiently large dataset is required for adequate training. However, unlike Suzuki's model, there is no requirement for *a priori* knowledge of the contacting positions or the structural details. We expect that given big datasets, the "neutral" positions (in terms of binding specificity) will be evident during the training process. Also, like Margalit's model, SAMIE uses a probabilistic method to assign the scores in a non-arbitrary way.

Taking into consideration the experiment-specific reference probabilities of the randomised molecules is an essential part of our method and, in our opinion, one of its major advantages over the other quantitative models (see Comparison of SAMIE with other methods). The other two quantitative models either do not consider reference probabilities at all (Suzuki's model) or they use a fixed reference probability for all data in their training set (Margalit's and Sarai's models).

The two qualitative models have the objective of identifying the preferred combinations of bases and amino acid residues at the interacting positions. Such information can be used to predict the binding sites for particular proteins and to design proteins that will bind to particular sites.[31] But they do not account for the fact that a given protein will generally bind to a family of similar sequences with K_A values that are not too different; or that a given DNA sequence may be bound by many different proteins. In fact, 40% of the combinations of binding site and protein sequences that have been reported from either SELEX or phage display experiments are not included in the set of interactions that constitute those models. And they do not attempt to predict changes in affinity associated with sequence variations. While the relative affinities predicted from the current SAMIE_C6 model are not perfect, they do show a reasonable correlation that can be useful in ranking different binding sites and predicting the effects of mutations.

Conclusions

The exploitation of the statistical mechanics theory for the calculation of the individual base–amino acid "energy" values gives SAMIE a strong theoretical basis and constitutes one of its major advantages. Moreover, the consideration of the background probabilities for the selected sequences makes the whole approach more robust. Additionally, the scores that SAMIE calculates for each protein–DNA pair reflect directly the probability (or specificity) of the binding, which is an additional advantage of our method. In all cases, we find a positive correlation between the observed data and SAMIE's predictions. Moreover, in most of the cases, the predicted order in the preference of the different nucleotides agrees with that indicated by the measured K_D values. The quantitative parameters of the model are obtained solely from qualitative data, the reported SELEX and phage display combinations of binding sites and protein sequences, using a maximum likelihood approach. A modified algorithm could take into account quantitative measurements of relative affinity for different sequences, and thereby improve the estimates of the parameters. Currently there is not enough quantitative data to make a significant difference, but high-throughput methods for relative affinity measurements will make such data available in the future.[18,46]

Additivity

We would like to emphasise that energetic additivity over contacts as well as knowledge of the structural details of the interaction are not prerequisites for our method. The underlying statistical framework that SAMIE is based upon is quite general. Non-additive energetic contributions can be modelled easily within the same theoretical framework, by an increase in the number of parameters. For example, if the mononucleotide positions deviate strongly from the additivity, whereas dinucleotide interactions do not, then we would need a 16 × 20 weight matrix to model each dinucleotide contact (or equivalently this could be modelled with a first-order Markov chain).

Additivity is clearly not exactly true,[18,19] but it can be a very good approximation to the true energy contributions,[20] with higher correlations than those obtained with SAMIE in this study. We think, therefore, that the current model is limited by lack of sufficient data, and especially quantitative data, more so than by the additivity approximation. It remains to be determined how much information is lost in this additive model, and how much better the predictions will be with more complex models. If the predictions can be made significantly better, then it would be worthwhile collecting enough data to determine the additional parameters. The current data are sufficient only for the additive model, and we are encouraged by the results obtained, although there is clearly room for improvement.

Docking rearrangements and other protein families

SAMIE uses a general algorithm that can model interactions for any protein family, given sufficient data. It does assume that the binding between DNA sites and variants of the protein under study use essentially a conserved pattern of contacts. That is, each protein family uses a particular way to contact the DNA that does not depend on the exact identity of those amino acid residues. Obviously, this is not exactly true: some amino acid replacements might lead to slightly different "dockings" and, in extreme cases, some amino acid changes might abolish the specific binding completely (e.g. KASN protein in the study by Bulyk *et al.*[19]). The question of how well slight changes can be approximated by a P-code needs further investigation. Some of the EGR variants that have been crystallised do show minor changes in contacting positions,[16,29] but the SAMIE predictions were not affected drastically. The crystal structure of another Cys_2His_2 protein that is otherwise unrelated to EGR (i.e. TTKB) showed a striking conservation with respect to the pattern of contacts in one of its two fingers, but changes were observed in the second one. Still, SAMIE was able to predict very accurately all but the last base of the genomic target, and it was able to depict the most significant base positions. We do not know at this time how similar the P-codes will be for other families of DNA-binding proteins. Most such families use α-helices for their DNA recognition domains, and we might expect that very similar energy relationships will be obtained for a variety of different protein families. The model described by Suzuki[10] implies only a few different classes of interactions, and the results from the EGR family indicate that we can do somewhat better than his model by allowing for more position-specificity in the parameters. But it may be that only a few significantly different base–amino acid energy matrices exist, and that the specificities of new protein families can be determined much more efficiently by extrapolation from existing ones.

Beyond DNA binding to gene regulation

This study developed a model, a probabilistic recognition code (P-code), for DNA recognition and binding by the EGR family of proteins. Such a P-code allows for the prediction of binding sites within a genome for all such proteins encoded in that genome. If we had similar P-codes for every DNA-binding protein family, it would be possible to simply examine the genomic sequence, predict the transcription factor proteins (which can be done fairly easily and reliably), and then to predict which transcription factors regulate which genes. But it is clear that transcription factor binding to specific DNA targets is required, but not sufficient, for gene regulation. At least in higher eukaryotes, most genes are regulated by multiple factors acting in concert, so the effect on gene expression of any one factor may depend on what other factors are in the same cells. And, of course, there can be competition between different factors for binding to the same sites. In addition, there are controls of gene expression at the level of chromatin structure. The transcription factors must have access to the DNA in order to bind and effect expression, and that access may be controlled by other factors, perhaps acting over long regions. And, finally, the transcription factors themselves must be expressed and active in the cells where their effects are manifest. Many regulatory events are post-transcriptional, and even post-translational, and must be accounted for to get accurate models of gene regulation. The ability to predict DNA-binding sites from protein sequences solves only one small part of the total regulatory system. But it is an important and essential part, and accurate, or even partially accurate, predictions can serve to focus further experimentation and analysis so as to accelerate the deciphering of the entire system.

Materials and Methods

The data

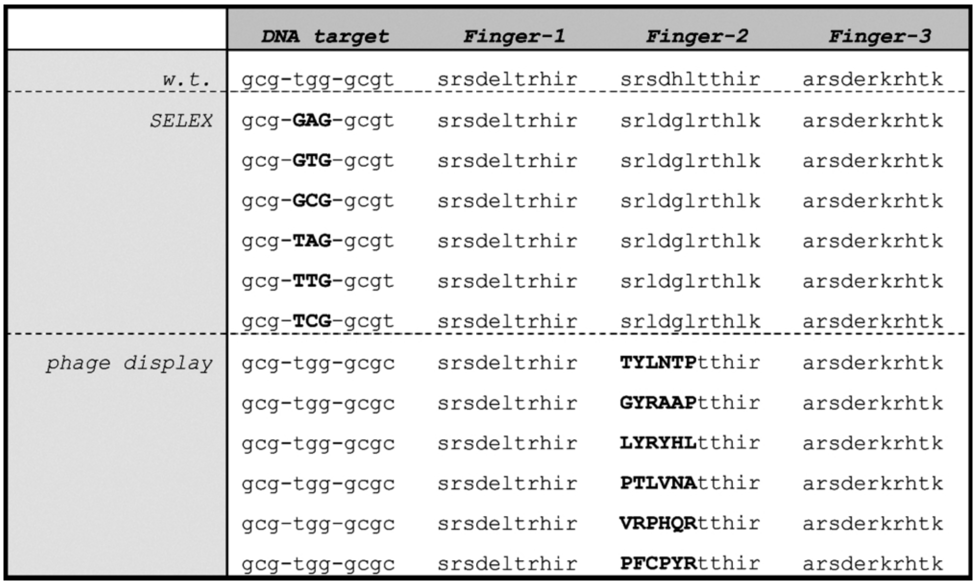
Data of DNA bound by variants of the EGR proteins were collected from the literature.[36] The original set was expanded to include more recently published data. It now contains a total of 1033 examples of protein–DNA interactions, 919 of which are non-redundant. From these data, 322 are derived from SELEX experiments (304 non-redundant) and 431 from phage display (399 non-redundant). In the remaining 280 experiments, neither the protein nor the DNA was randomised.

An example of our dataset is presented in Figure 7. Essentially, each line corresponds to a protein–DNA interaction "experiment" and it contains the 10 bp long target and the sequence of the three α-helices (amino acid positions −2 through +9 with respect to the beginning of the helix). Capital and small letters designate randomisation and fixation of the corresponding position, respectively.

EGR proteins contain three zinc-finger motifs of the Cys_2His_2 type. Each finger of this type is believed to bind the DNA in a modular fashion, independently of the others (except for the overlapping base).[30,48] Thus, we decided to focus our analysis on the interactions of a single finger. Considering the interaction of a single finger reduces the number of parameters that one needs to estimate. Under the additivity assumption, the contacts can be modelled independently and we need a maximum of 80 parameters to model each contact. Thus, for modelling all 12 contacts of the EGR protein we would need 960 parameters, whereas restricting the model in a single finger reduces this number to 320. The six datasets we used for training SAMIE (see the section on Training datasets and models calculated by SAMIE) are derived from the master database.

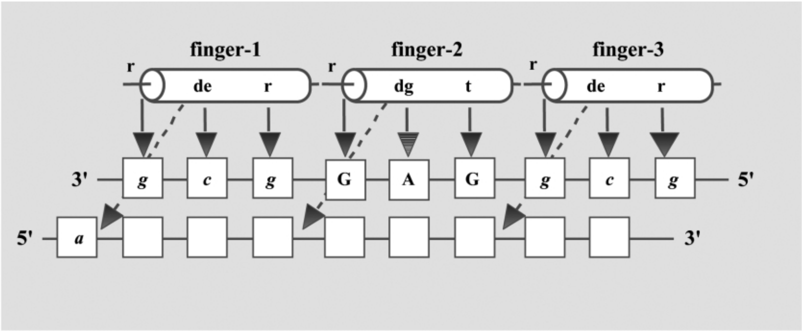
There are two ways one can model the interactions of a tetranucleotide target of a single Cys_2His_2 finger. One is to consider the contacts from the primary contacting finger (e.g. finger 2 for the base positions 4–7 of the 10 bp long target) and the other is to consider the two additional contacting amino acid residues in the neighbouring fingers. We call the first model one-to-one and the second many-to-one. For a more detailed discussion of these models, we refer the reader to the section Training datasets and models calculated by SAMIE.

For the construction of the datasets, we pooled the corresponding single finger vectors from the master



	DNA target	Finger-1	Finger-2	Finger-3
w.t.	gcg-tgg-gcgt	srsdeltrhir	srsdhltthir	arsderkrhtk
SELEX	gcg-**GAG**-gcgt	srsdeltrhir	srldglrthlk	arsderkrhtk
	gcg-**GTG**-gcgt	srsdeltrhir	srldglrthlk	arsderkrhtk
	gcg-**GCG**-gcgt	srsdeltrhir	srldglrthlk	arsderkrhtk
	gcg-**TAG**-gcgt	srsdeltrhir	srldglrthlk	arsderkrhtk
	gcg-**TTG**-gcgt	srsdeltrhir	srldglrthlk	arsderkrhtk
	gcg-**TCG**-gcgt	srsdeltrhir	srldglrthlk	arsderkrhtk
phage display	gcg-tgg-gcgc	srsdeltrhir	**TYLNTP**tthir	arsderkrhtk
	gcg-tgg-gcgc	srsdeltrhir	**GYRAAP**tthir	arsderkrhtk
	gcg-tgg-gcgc	srsdeltrhir	**LYRYHL**tthir	arsderkrhtk
	gcg-tgg-gcgc	srsdeltrhir	**PTLVNA**tthir	arsderkrhtk
	gcg-tgg-gcgc	srsdeltrhir	**VRPHQR**tthir	arsderkrhtk
	gcg-tgg-gcgc	srsdeltrhir	**PFCPYR**tthir	arsderkrhtk

Figure 7. Example of the EGR data file. Each line contains the single result of an experiment. The target DNA is followed by the amino acid sequence of the three fingers (positions −2 to +9 with respect to the beginning of the α-helix). Capital letters denote randomisation and small letters denote fixation at the corresponding position.



SELEX_4	
gcgt	*F1*:rder
GAGg	*F2*:rdgt
gcg**G**	*F3*:rder

SELEX_6	
gcgt	*F1*:rder, *F2*:d
GAGg	*F1*:r, *F2*:rdgt, *F3*:d
gcgg	*F2*:t, *F3*:rder

Figure 8. An example of the selection of the training vectors. In this particular SELEX example, only the three middle bases had been randomised. For the SELEX_4 training set, the three four-base targets and their contacting fingers consist three independent vectors. Note that the first of these vectors was subsequently excluded from the SELEX_4 training set, since both the DNA and the amino acid residues have fixed values. For the SELEX_6 training set, the same DNA (sub)targets were considered, but the two amino acid residues from the neighbouring fingers were added. In the case of the third vector, the information for the contact of its randomised fourth base (G) is included in vector-2, thus we consider this position as non-randomised on vector 3 (represented by the last g in the gcgg sequence). Hence, both vectors number 1 and 3 were excluded from the SELEX_6 training set.

database, excluding those where both protein and DNA sequences were fixed (see Figure 8; and see Evaluation on known *in vivo* binding sites). After eliminating redundancy, we had 293 SELEX and 319 phage display vectors for the one-to-one model of interaction, and 366 SELEX and 444 phage display vectors for the many-to-one model. These datasets are named SELEX_4, PHAGE_4, SELEX_6 and PHAGE_6, respectively. The corresponding composite sets are named COMBINED_4 and COMBINED_6.

An example of the construction of a training set is presented in Figure 8. In the 293 vectors of the SELEX_4 dataset, 116 different "proteins"† were used to select 79 different tetranucleotides. For the 366 vectors of the SELEX_6 dataset, the corresponding numbers are 180 and 70. For the 319 vectors of the PHAGE_4 dataset, 56 different tetranucleotides were used to select 252 different proteins. For the 444 vectors of the PHAGE_6 dataset, the corresponding numbers are 59 and 361.

Note that in the case of the one-to-one model of interaction, the number of all possible combinations is 256 for the DNA and 160,000 for the protein sequences. In the case of the many-to-one model, the number of all possible amino acid combinations is even higher (20^6 or 64×10^6), whereas the increase of the dataset size is very modest. Thus, it is obvious that even with the reduction of the analysis to single finger interactions, our datasets are still far from complete. This is illustrated in Table 5, which presents the frequencies of particular base–amino acid contacts for the many-to-one model of interaction (COMBINED_6 data set). This Table might look like the one presented by Mandel-Gutfreund *et al.*,[49] with the frequencies derived from selection experiments (instead of crystal structures) and partitioned with respect to the contact (i.e. the position(s) of the base and amino acid residue, respectively). We must therefore emphasise the fact that these numbers are derived from randomisation experiments that generally differ on their randomisation scheme. One of the novelties of our method is that it incorporates these experiment-specific randomisation schemes into the calculation of the probabilities.

Data representation

Each training vector can be encoded into two sparse unary vectors (${}_xN$ and ${}_yA$), which consist of the binary representation of the target DNA and the contacting amino acid residues, respectively. For the binary representation of the four bases, we use the following notation:

$$\mathrm{A} = (1000),\ \ \mathrm{C} = (0100),\ \ \mathrm{G} = (0010),\ \ \mathrm{T} = (0001)$$

Thus, the DNA sequence AGGA can be written as:

$${}_{\mathrm{AGGA}}N = (1000\ 0010\ 0010\ 1000)$$

or:

$$N_j^{\beta} = \begin{cases} 1 & (j = 1,4 \wedge \beta = \mathrm{A}) \vee (j = 2,3 \wedge \beta = \mathrm{G}) \\ 0 & \text{otherwise} \end{cases}$$

An amino acid sequence is represented in an analogous way, as a vector A_i^{α}, using a similar set of 20 binary digits for the representation of the amino acid residue in each position.

The mode of interaction between protein and DNA residues is encoded in a "connectivity matrix", C. Matrix C consists of binary values: if the amino acid residue at position i is assumed to contribute to the affinity of interaction by contacting the base at position j, then $C_{ij} = 1$, otherwise it is 0.

Finally, using similar symbolism, we can represent the base–amino acid energetic potential as a weight matrix $T_{ij}^{\alpha\beta}$, where i and j denote the amino acid residue and base positions, respectively, whereas α and β are the residues of the protein and the DNA target in these positions. Figure 3 illustrates the matrix T used for modelling the interactions of a single finger of the EGR protein family. Using this formalism, the calculation of the additive total energy of an interaction between a protein ${}_yA$ and a DNA target ${}_xN$ would be given by equation (2).

† Proteins here simply refers to the maximum of six contacting amino acid residues.

Some algorithmic aspects

Maximising the log-odds

SAMIE is an algorithm that uses data from randomisation experiments (SELEX and/or phage display) to specify the energy matrix, T, of base–amino acid energetic potentials that maximise the log-odds of the data. Consider a training vector ($_kV$, F) that consists of a variable ($_kV$) and a fixed (F) component†. We assume that the variable component was selected by the fixed one from a pool of randomised molecules. The reference probability of every molecule in the pool ($P_{\mathrm{ref}}(_{k'}V)$, $k' = 1,\ldots,N_{\mathrm{tot}}$) is known. The log-odds of this selection can be written in a way similar to equations (1) and (3):

$$\log(P(_kV|F)) = \log(P_{\mathrm{ref}}(_kV)) - H(_kV, F) - \log(Z) \quad (6)$$

where Z is the experiment-specific partition function, which is calculated over all possible molecules in the randomised pool:

$$Z = \sum_{k'} P_{\mathrm{ref}}(_{k'}V)\mathrm{e}^{-H(_{k'}V,F)} \quad (7)$$

and $H(_{k'}V, F)$ is the energy function, which in the simplest case is additive over all contacts (see equation (2)).

For specifying the energy matrix, T, which maximises the objective function $\log(P(_kV|F))$ over all data SAMIE follows the steepest ascent procedure. Every free parameter is incremented iteratively by a value proportional to the gradient of the function with respect to the parameter.

Considering equation (6), for all data, we get:

$$\log(P(\mathrm{data})) = \sum_{\mathrm{data}} \log(P(_kV|F))$$

$$= \sum_{\mathrm{data}} \log(P_{\mathrm{ref}}(_kV)) - \sum_{\mathrm{data}} H(_kV, F) - \sum_{\mathrm{data}} \log(Z) \quad (8)$$

Considering the sums on the right side of this equation, we note that the first term is independent of the parameters, thus its derivative would be zero. The second term, $H(_kV, F)$ is linear with respect to the parameters (equation (2)), therefore its derivative would be equal to the total number of examples (N) times the average of the energy in the dataset. Hence differentiation of this term with respect to a parameter required in the gradient merely selects the corresponding frequency as calculated in the given data set, since all parameters occur linearly. The last term is the partition function for a Boltzmann distribution, and it contains a sum over all possible variable counterpart sequences of an exponential of energy. Differentiating a partition function with respect to parameters appearing linearly in the energy results in the negative of the expectation as computed within the distribution. Hence the steepest ascent process will reach a fixed point, i.e. zero gradient), when the expectations as calculated within the distribution match the frequencies as calculated within the given data set. This intuitive result is basically the Boltzmann machine algorithm for neural network training,[50] an observation that results in some additional insights into the algorithm we propose, but will not give in detail here.

Calculating the partition function

Finally, we need to calculate the expectation within the distribution at any trial setting of the parameters. This needs to be done at each step of the iterative steepest ascents process. Naively, this would involve a sum over all possible variable counterpart sequences, which for an L long sequence grows exponentially as 4^L for nucleotides in SELEX data, or as 20^L for amino acid residues in phage display data. Depending on L, this simple approach can be impractical. In Boltzmann machine training, this issue is addressed by computing the expectations *via* Monte Carlo averaging, which approximates the sum over an exponential state space by importance sampling. Closely related importance sampling methods, such as the Gibbs sampler arise in standard biological sequence analysis.[51] However, we can calculate the required expectations exactly, without recourse to importance sampling, because our energy function has a very simple structure.

The trick to computing the expectations exactly is to notice that in equation (2) either the amino acid residues (for SELEX data) or the nucleotides (for phage display data) are fixed, leaving the nucleotides (respectively amino acid residues) appearing linearly in H. Hence, the sums, in e.g. the SELEX partition function, factorise across positions. Thus the sum over 4^L states reduces to the product of sums over only four states at each position (similarly, over 20 states for phage data). Hence, the required expectations can be computed exactly. If more complicated energy functions are found to be required later, we can fall back on importance sampling to compute the required distribution averages.

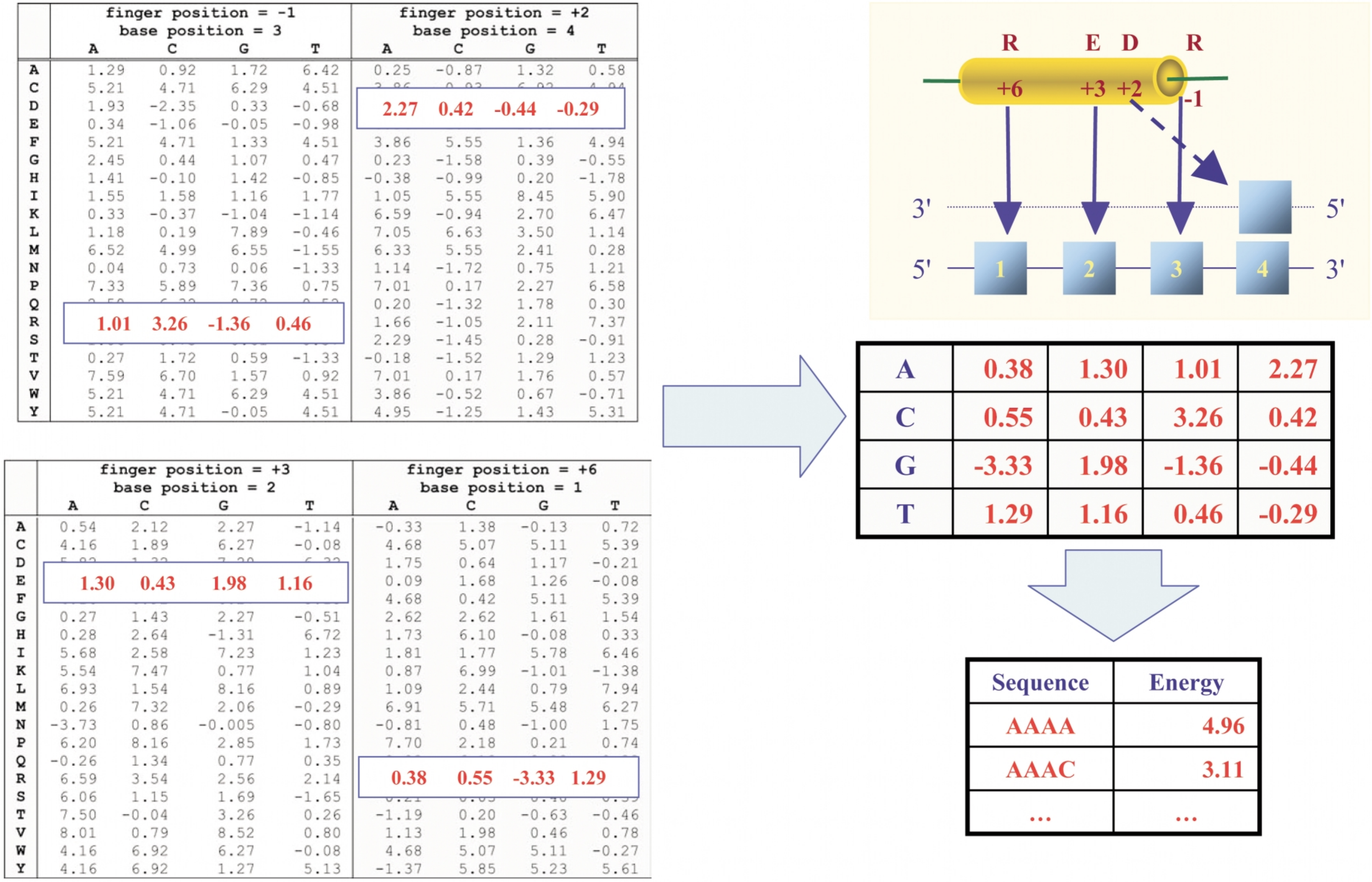
Prediction of probabilities

The calculation of the success rate and specificity index is based on the ranking of the DNA targets of a particular single finger sequence or *vice versa* according to the predicted probabilities of the SAMIE's model that we examine. In order to calculate the predicted probabilities of all possible tetranucleotide targets of a given zinc-finger in a SELEX training vector, we do the following (see also Figure 9 for an example). First, for each amino acid residue in each contacting position of the zinc-finger sequence, we extract the corresponding energetic potentials from the weight matrix T, which comprises the SAMIE's model. The 16 resulting values constitute a position-specific weight matrix for this finger. Using this weight matrix, we can calculate the energetic potentials of the interaction of each tetranucleotide target by summing the values of each base. The total predicted energy is used in equation (1) to give an estimate of the probability of the interaction to every DNA target. Similar procedure is followed for the calculation of the probabilities in a phage display experiment. For SELEX experiments, we normally use equiprobable nucleotide background (i.e. $P_n(_kN) = 0.25^4$ for every nucleotide sequence $_kN$), except in the case of MIG proteins, where the yeast GC content was used instead. For evaluation of phage display vectors, the background probabilities, P_a, in equation (3) depend on the randomisation scheme that was used in the particular experiment.

Methods of evaluation

The evaluation of a prediction method is the assessment of how accurately it can predict a given dataset (evaluation set). The most interesting evaluation sets are

† In the case of a SELEX, the variable component, $_kV$, is the DNA and the fixed component, F, is the amino acid sequence. In the case of a phage display, it is *vice versa*.


	finger position = -1 base position = 3				finger position = +2 base position = 4			
	A	C	G	T	A	C	G	T
A	1.29	0.92	1.72	6.42	0.25	-0.87	1.32	0.58
C	5.21	4.71	6.29	4.51	[illegible]	[illegible]	[illegible]	[illegible]
D	1.93	-2.35	0.33	-0.68	2.27	0.42	-0.44	-0.29
E	0.34	-1.06	-0.05	-0.98	[illegible]	[illegible]	[illegible]	[illegible]
F	5.21	4.71	1.33	4.51	3.86	5.55	1.36	4.94
G	2.45	0.44	1.07	0.47	0.23	-1.58	0.39	-0.55
H	1.41	-0.10	1.42	-0.85	-0.38	-0.99	0.20	-1.78
I	1.55	1.58	1.16	1.77	1.05	5.55	8.45	5.90
K	0.33	-0.37	-1.04	-1.14	6.59	-0.94	2.70	6.47
L	1.18	0.19	7.89	-0.46	7.05	6.63	3.50	1.14
M	6.52	4.99	6.55	-1.55	6.33	5.55	2.41	0.28
N	0.04	0.73	0.06	-1.33	1.14	-1.72	0.75	1.21
P	7.33	5.89	7.36	0.75	7.01	0.17	2.27	6.58
Q	[illegible]	[illegible]	[illegible]	[illegible]	0.20	-1.32	1.78	0.30
R	1.01	3.26	-1.36	0.46	1.66	-1.05	2.11	7.37
S	[illegible]	[illegible]	[illegible]	[illegible]	2.29	-1.45	0.28	-0.91
T	0.27	1.72	0.59	-1.33	-0.18	-1.52	1.29	1.23
V	7.59	6.70	1.57	0.92	7.01	0.17	1.76	0.57
W	5.21	4.71	6.29	4.51	3.86	-0.52	0.67	-0.71
Y	5.21	4.71	-0.05	4.51	4.95	-1.25	1.43	5.31

	finger position = +3 base position = 2				finger position = +6 base position = 1			
	A	C	G	T	A	C	G	T
A	0.54	2.12	2.27	-1.14	-0.33	1.38	-0.13	0.72
C	4.16	1.89	6.27	-0.08	4.68	5.07	5.11	5.39
D	[illegible]	[illegible]	[illegible]	[illegible]	1.75	0.64	1.17	-0.21
E	1.30	0.43	1.98	1.16	0.09	1.68	1.26	-0.08
F	[illegible]	[illegible]	[illegible]	[illegible]	4.68	0.42	5.11	5.39
G	0.27	1.43	2.27	-0.51	2.62	2.62	1.61	1.54
H	0.28	2.64	-1.31	6.72	1.73	6.10	-0.08	0.33
I	5.68	2.58	7.23	1.23	1.81	1.77	5.78	6.46
K	5.54	7.47	0.77	1.04	0.87	6.99	-1.01	-1.38
L	6.93	1.54	8.16	0.89	1.09	2.44	0.79	7.94
M	0.26	7.32	2.06	-0.29	6.91	5.71	5.48	6.27
N	-3.73	0.86	-0.005	-0.80	-0.81	0.48	-1.00	1.75
P	6.20	8.16	2.85	1.73	7.70	2.18	0.21	0.74
Q	-0.26	1.34	0.77	0.35	[illegible]	[illegible]	[illegible]	[illegible]
R	6.59	3.54	2.56	2.14	0.38	0.55	-3.33	1.29
S	6.06	1.15	1.69	-1.65	[illegible]	[illegible]	[illegible]	[illegible]
T	7.50	-0.04	3.26	0.26	-1.19	0.20	-0.63	-0.46
V	8.01	0.79	8.52	0.80	1.13	1.98	0.46	0.78
W	4.16	6.92	6.27	-0.08	4.68	5.07	5.11	-0.27
Y	4.16	6.92	1.27	5.13	-1.37	5.85	5.23	5.61

A	0.38	1.30	1.01	2.27
C	0.55	0.43	3.26	0.42
G	-3.33	1.98	-1.36	-0.44
T	1.29	1.16	0.46	-0.29

Sequence	Energy
AAAA	4.96
AAAC	3.11
…	…

Figure 9 (*legend opposite*)

those where their values have been derived/confirmed experimentally. For SAMIE's self-evaluation and for the comparison of SAMIE with the other modelling methods, we use two evaluation measures. One is the success rate ($SR_{0.1}$), which we define as the percentage of vectors in the evaluation set that are ranked in the top 10% of all the predictions:

$$0 \leq SR_{0.1} \equiv S_{0.1}/N_{\text{tot}} \leq 1 \qquad (9)$$

where $S_{0.1}$ is the number of vectors that are ranked in the top 10% of all the predictions and N_{tot} is the total number of vectors. This is similar to the evaluation method that was used previously to assess the accuracy of the binding interaction predictions.[7]

In order to calculate the success rate for the SELEX vectors we do the following: for each vector in the evaluation set, all possible randomised DNA sequences are ranked with respect to their predicted probability of being selected by the (fixed) amino acid sequence. The probability is calculated using equation (1) (see below). If the DNA (selected) sequence of this vector ranks in the top 10% of the list, then we consider this a success. The percentage of successes in the evaluation set is the success rate. For example, in the self-test on the SELEX_4 dataset (Table 1), 83 of the 96 DNA sequences predicted in the top 10% of the lists of the corresponding amino acid sequences, resulting in an $SR_{0.1}$ value of 0.865. For the phage display data set, the procedure is similar, but equation (3) is used for the calculation of the probabilities.

The success rate can be viewed as the probability of ranking correctly a particular selected sequence (from the randomised pool) given the fixed sequence. An $SR_{0.1}$ value of 85% means that for any fixed sequence, the trained model has 85% probability of ranking a selected sequence correctly, i.e. in the top 10% of the list of all possible randomised sequences.

The second measure of evaluation is the specificity index, which has been used in the past. It is defined by Suzuki & Yagi as:[47]

$$0 \leq SI \equiv 100 - n - m/2 \leq 100 \qquad (10)$$

In this formula, n and m are the percentages of the target sequences that score higher than, and the same as, the real ones, respectively. For example, if the selected sequence ranks second in the list and its score is equal to two more sequences, then SI would be $100 - (1 + 3/2) \times 100/K$, where K is the number of sequences in the list (e.g. for a trinucleotide target K is 64 and the SI of the example would have been 96.1). This measure was used in addition to the success rate as an independent method of evaluation for two reasons: (a) it had been used in the past,[10,47] so the comparison with previous studies would be straight-forward; and (b) the cutoff of 10% for the success rate is somewhat arbitrary. We note, though, that in all cases the specificity index yields higher scores than the success rate (see Table 3).

Acknowledgements

This work was supported by NIH grant HG00249 to G.D.S. The research of A.S.L. was supported by the Department of Energy under contract W-7405-ENG-36. The authors thank the Santa Fe Institute, where part of this work was performed.

References

1. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
2. Pabo, C. O. & Sauer, R. T. (1984). Protein–DNA recognition. *Annu. Rev. Biochem.* **53**, 293–321.
3. Matthews, B. W. (1988). Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
4. Nardelli, J., Gibson, T. J., Vesque, C. & Charnay, P. (1991). Base sequence discrimination by zinc-finger DNA-binding domains. *Nature*, **349**, 175–178.
5. Desjarlais, J. R. & Berg, J. M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl Acad. Sci. USA*, **89**, 7345–7349.
6. Jamieson, A. C., Wang, H. & Kim, S. H. (1996). A zinc finger directory for high-affinity DNA recognition. *Proc. Natl Acad. Sci. USA*, **93**, 12834–12839.
7. Mandel-Gutfreund, Y. & Margalit, H. (1998). Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. *Nucl. Acids Res.* **26**, 2306–2312.
8. Wolfe, S. A., Nekludova, L. & Pabo, C. O. (2000). DNA recognition by Cys_2His_2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 183–212.
9. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860–2874.
10. Suzuki, M., Brenner, S. E., Gerstein, M. & Yagi, N. (1995). DNA recognition code of transcription factors. *Protein Eng.* **8**, 319–328.
11. Choo, Y. & Klug, A. (1997). Physical basis of a protein–DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117–125.
12. Kono, H. & Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. *Proteins: Struct. Funct. Genet.* **35**, 114–131.
13. Mandel-Gutfreund, Y., Baron, A. & Margalit, H. (2001). A structure-based approach for prediction of protein binding sites in gene-upstream regions. *Pac. Symp. Biocomput.* **6**, 139–150.
14. Wolfe, S. A., Greisman, H. A., Ramm, E. I. & 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.

Figure 9. Example of calculation of the predicted energies for all tetranucleotide targets of finger 1 of the wild-type EGR protein. Given a specific protein sequence (e.g. in our case the protein sequence is RDER, which corresponds to positions −1, +2, +3 and +6 of the zinc-finger, respectively), the appropriate energy values are extracted from weight matrix T (i.e. SAMIE's calculated model). A position-specific weight matrix[35] is created with the energetic potentials for each base in each target position for this protein sequence. The total predicted energy of each target is the sum of the energies of the individual bases. The total predicted energy is used in equation (1) for the calculation of the corresponding predicted probabilities.

15. Choo, Y. & Klug, A. (1997). Physical basis of a protein–DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117–125.
16. Miller, J. C. & Pabo, C. O. (2001). Rearrangement of side-chains in Zif268 mutant highlights the complexities of zinc-finger DNA recognition. *J. Mol. Biol.* **313**, 309–315.
17. Benos, P. V., Lapedes, A. S. & Stormo, G. D. (2002). Is there a code for protein–DNA recognition? *Probab(istical)ly Bioessays*, **24**, 66–75.
18. Man, T.-K. & Stormo, G. D. (2001). Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucl. Acids Res.* **15**, 2471–2478.
19. Bulyk, M. L., Johnson, P. L. F. & Church, G. M. (2002). Nucleotides of transcription factor binding sites exert inter-dependent effects on the binding affinities of transcription factors. *Nucl. Acids Res.* **30**, 1255–1261.
20. Benos, P. V., Bulyk, M. L. & Stormo, G. D. (2002). Additivity on protein–DNA interactions: how good an approximation is it? *Nucl. Acids Res.* In press.
21. Milbrandt, J. (1987). A nerve growth factor-induced gene encodes a possible transcriptional regulatory factor. *Science*, **238**, 797–799.
22. Sukhatme, V. P., Cao, X., Chang, L. C., Tsai-Morris, C-H., Stamenkovich, D., Ferreira, P. C. P. *et al.* (1988). A zinc finger-encoding gene coregulated with c-fos during growth and differentiation, and after cellular depolarization. *Cell*, **53**, 37–43.
23. Shimizu, N., Ohta, M., Fujiwara, C., Sagara, J., Mochizuki, N., Oda, T. & Utiyama, H. (1992). A gene coding for a zinc finger protein is induced during 12-*O*-tetradecanoylphorbol-13-acetate-stimulated HL-60 cell differentiation. *J. Biochem.* **111**, 272–277.
24. Bradley, L. C., Snape, A., Bhatt, S. & Wilkinson, D. G. (1993). The structure and expression of the Xenopus Krox-20 gene: conserved and divergent patterns of expression in rhombomeres and neural crest. *Mech. Dev.* **40**, 73–84.
25. Drummond, I. A., Rohwer-Nutter, P. & Sukhatme, V. P. (1994). The zebrafish EGR1 gene encodes a highly conserved, zinc-finger transcriptional regulator. *DNA Cell Biol.* **13**, 1047–1055.
26. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263–266.
27. Pavletich, N. P. & Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 Å. *Science*, **252**, 809–817.
28. Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996). Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. *Structure*, **4**, 1171–1180.
29. Elrod-Erickson, M., Benson, T. E. & Pabo, C. O. (1998). High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition. *Structure*, **6**, 451–464.
30. Choo, Y. & Klug, A. (1994). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl Acad. Sci. USA*, **91**, 11163–11167.
31. Choo, Y. & Klug, A. (1994). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl Acad. Sci. USA*, **91**, 11168–11172.
32. von Hippel, P. H. & Berg, O. G. (1989). Facilitated target location in biological systems. *J. Biol. Chem.* **264**, 675–678.
33. Chandler, D. (1987). *Introduction to Modern Statistical Mechanics*, Oxford University Press, New York.
34. Greisman, H. A. & Pabo, C. O. (1997). A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science*, **275**, 657–661.
35. Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
36. Benos, P. V., Lapedes, A. S., Fields, D. S. & Stormo, G. D. (2001). SAMIE: statistical algorithm for modeling interaction energies. *Pac. Symp. Biocomput.* **6**, 115–126.
37. Biesiada, E., Razandi, M. & Levin, E. R. (1996). Egr-1 activates basic fibroblast growth factor transcription. *J. Biol. Chem.* **271**, 18576–18581.
38. Skerka, C., Decker, E. L. & Zipfel, P. F. (1995). A regulatory element in the human interleukin 2 gene promoter is a binding site for the zinc finger proteins sp1 and EGR-1. *J. Biol. Chem.* **270**, 22500–22506.
39. Thukral, S. K., Eisen, A. & Young, E. T. (1991). Two monomers of yeast transcription factor ADR1 bind a palindromic sequence symmetrically to activate ADH2 expression. *Mol. Cell. Biol.* **11**, 1566–1577.
40. Lutfiyya, L. L., Iyer, V. R., DeRisi, J., DeVit, M. J., Brown, P. O. & Johnston, M. (1998). Characterization of three related glucose repressors and genes they regulate in *S. cerevisiae*. *Genetics*, **150**, 1377–1391.
41. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.
42. Young, E. T., Kacherovsky, N. & Cheng, C. (2000). An accessory DNA binding motif in the zinc finger protein Adr1 assists stable binding to DNA and can be replaced by a third finger. *Biochemistry*, **39**, 567–574.
43. Fairall, L., Schwabe, J. W. R., Chapman, L., Finch, J. T. & Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature*, **366**, 483–487.
44. Hamilton, T. B., Borel, F. & Romaniuk, P. J. (1998). Comparison of the DNA binding characteristics of the related zinc finger proteins WT1 and EGR1. *Biochemistry*, **37**, 2051–2058.
45. Segal, D. J., Dreier, B., Beerli, R. R. & Barbas, C. F., III (1999). Toward controlling gene expression at will: selecting and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl Acad. Sci. USA*, **96**, 2758–2763.
46. 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.
47. Suzuki, M. & Yagi, N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.
48. Pomerantz, J. L., Sharp, P. A. & Pabo, C. O. (1995). Structure-based design of transcription factors. *Science*, **267**, 93–96.
49. Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds

in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.

50. Hertz, J., Krogh, A. & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA.

51. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Edited by J. Thornton

(Received 2 May 2002; received in revised form 26 August 2002; accepted 27 August 2002)