Assessing effects of symmetry on motif finding

lala Motlhabi
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/etd

Recommended Citation
https://openscholarship.wustl.edu/etd/476
WASHINGTON UNIVERSITY

The Division of Biology and Biomedical Sciences

Molecular Genetics & Genomics

ASSESSING EFFECTS OF SYMMETRY ON

MOTIF FINDING

By

Lala Maipelo Motlhabi

A thesis presented to the
Graduate School of Arts and Sciences
Of Washington University in
Partial fulfillment of the
Requirements for the
Degree of Master of Arts

December 2010
Saint Louis, Missouri

©2010 Lala Maipelo Motlhabi
Abstract

Identifying protein DNA binding sites like transcription factor binding sites is a key component of gene regulatory network. The challenging task of predicting and identifying DNA binding sites both experimentally, and computationally suffers from high false positive due to various contributing factors, including misinterpretation of DNA binding site sequence symmetry. Our study seeks to model and compare the ability of three methods of a motif-finding program, consensus, taking into account orientation to accurately predict variable symmetric/asymmetric true binding site models. The three consensus methods included the c0 method which ignores the complement of a sequence. In this particular experiment it is given the correct binding orientation and so it’s output reflects the true model limited only by sample size. The second method, -c2, takes into account both strands and includes them as a single strand, as a result, either orientation can be chosen as correct. The third method, c3, makes the assumption that the pattern is symmetrical and includes both orientation of each site in the model. Our results show that for a given asymmetric site, the c2 method is quite accurate in predicting the true model, while the c3
model results in a very poor true model prediction. On the other hand if
the site is symmetric, the c3 method gives a very accurate model, but
now the c2 model is inaccurate, predicting more information content
(IC) than is actually there. The results demonstrate that either method
can lead to inappropriate models if the underlying assumption of
symmetry is incorrect, resulting in high levels of false positives and false
negative motif predictions.
# TABLE OF CONTENTS

## INTRODUCTION

1. REGULATORY ELEMENTS AND DISEASES .......................... 2
2. EXPERIMENTAL APPROACHES ...................................... 3
3. PROTEIN DNA BINDING SITES: TFBS .............................. 7
4. MOTIF DISCOVERY .................................................. 8
5. MOTIF REPRESENTATIVE MODEL ................................. 9
6. MOTIF DISCOVERY METHODS ................................... 10

## METHODS AND RESULTS

15. TRUE BINDING SITE MODELS ................................. 15
16. TRUE MODEL DNA SAMPLING .................................. 16
17. CONSENSUS ...................................................... 17
18. SEQUENCE LOGOS ............................................... 21
19. BINDING PROBABILTY COMPARISONS ..................... 22

## DISCUSSION

28
REFERENCES

35
INTRODUCTION

Identifying protein DNA binding sites that control gene expression in a cell remains an important challenge in Molecular Biology. Gene expression regulation requires precise spatial and temporal coordination of a multitude of general and specific transcription factors at cis-regulatory elements, including enhancers, silencers, insulators and promoters (Orphanides and Reinberg, 2002; Fraser et al., 2005; Kadonaga, 2003). Recognition and binding of these sequences by transcription factors, for example, occurs within the context of chromatin, which plays a significant role in regulating gene expression (Mellor 2005). In order to understand how these DNA binding proteins regulate gene expression, one of the major tasks is accurately identifying and characterizing their specific DNA binding sites in a given genome. An understanding of such a gene regulatory network can be attempted using computational approaches (Covert et al. 2004). DNA-binding site information based off experimentally derived example binding sites can be modeled, searched, and identified in the non-coding sequence through the use of position weight matrices (PWMs) and DNA pattern recognition programs (Stormo 2000).
REGULATORY ELEMENTS AND DISEASES

Genetic abnormalities due to genetic mutations/disorders are more commonly reported in protein coding genes like in breast cancer (Narod et al., 1993), cystic fibrosis (Bobadilla et al., 2002), Parkinson’s disease (Valente et al., 2006)) and across many other diseases. In contrast, mutations in regulatory elements, transcription factor binding sites for example, are assumed less likely to have a pronounced phenotypic impact because they tend to affect gene expression pattern, and rarely gene product structure (Narlikar et al., 2009). However, contrary to these findings, non-coding mutations linked to human diseases have been characterized. For example, HTRA1 promoter mutation has been linked to macular degeneration (Dewan et al., 2006), KLR promoter mutation to pyruvate kinase deficiency function (Manco et al., 2010) and many other diseases (Wrary et al., 2007). Also Genome-wide association studies (GWAS) provide a high-throughput approach to rapidly identify disease-causing polymorphisms by scanning markers across genomes of many people. Results of many GWAS are available in the database of genotypes and phenotypes (dbGaP)(Mailman et al., 2007).
Overall various studies have demonstrated that mutations in many of the known regulatory elements are associated with diseases, thus indicating the importance of regulatory elements in disease diagnostics and drug discoveries (Maston et al., 2006).

The discovery of the lac operon and how its expression is regulated by a protein factor has since incited curiosities in deciphering gene regulatory networks. In spite of being laborious, and very costly, classical experimental approaches including sequencing made major breakthroughs in identifying and characterizing sequences of the lac operator, λ operators and other promoter sites (Gilbert and Maxam, 1973; Maniatis et al.; 1974, Pribnow, 1975). Advances in experimental techniques including whole genome sequencing have resulted in scaling up of the experimental approaches to high throughput studies resulting in an increase in example binding sites.

**EXPERIMENTAL APPROACHES**

The following are examples of experimental approaches taken to identify protein DNA binding sites. Advances in their experimental techniques have resulted in scaling up to high throughput studies,
resulting in databases of binding site information, along with refined computational tools aimed towards identifying the candidate "motifs".

**Gene expression assay** is one approach employed to examine cis regulatory signals. For instance, a promoter sequence placed upstream of a green fluorescent protein can be introduced in a sample of cultured cells, and assayed in 24-48hr-time period. Enhancers and promoters can be tested in transient or stable transfection manner. Where as the plasmid in transient transfection remains episomal, it is integrated into the genome in stable transfection. Pennacchio et al., carried out a large high throughput study involving human genomic regions tested in mouse embryos and the study indentified 75 enhancers active at a particular time point (Pennacchio L., et al., 2006).

**DNASE Hyper Sensitivity**: DNaseI hypersensitivity method maps changes in chromatin structure where by the degree of the response of the DNA sequence to DNase is classified as generalized sensitivity to hypersensitivity (Gazit and Cedar 1980). For instances, hypersensitivity refers to regions showing extreme sensitivity marking functional regions that fall in non coding sequences as in promoters, enhancers and (Gross and Gerrard 1988).
To date, high throughput approaches are employed to address the appearance and disappearance of functional sites on a genome wide scale to assess DNase hypersensitivity. For example, Quantitative chromatin profiling (Dorshcner et al. 2004) and massively parallel signature sequencing (Crawford et al. 2005) are some of the high throughput genome wide approaches from the ENCODE project taking such an approach (ENCODE project Consortium 2004).

**Classical approaches to defining protein-DNA interactions** include in vitro binding assays EMSA, the electrophoretic mobility shift assay which utilizes non-denaturing polyacrylamide gels to separate a protein bound DNA molecule from an unbound. DNase I footprinting is another approach that precisely seeks to identify localization of protein binding site without prior knowledge of the protein. The method combines binding reaction of EMSA with the cleavage reaction of DNaseI such that the radionuclide labeled probe is visualized on a denaturing polyacrylamide gel, sites protected from cleavage create a blank image in the semicontinuous ladder of nucleotide (Narlikar and Ivan Ovcharenko et al., 2009). Currently, technical advances relating to in vitro binding assays include fluorescence labeling, and are applied in
high throughput approaches. SELEX, (systematic evolution of ligands by exponential enrichment (Tuerk and Gold 1990) and Casting (cyclic amplification and selection of targets) (Wright et al. 1991) are two examples. These in vitro assay approaches screen large pools of short random oligonucleotide probes for a recognition by a specific protein.

CHIP-ASSAYS. Currently, binding assays that simultaneously capture protein DNA binding sites and the identity of the involved protein in vivo are quickly replacing gel shift assays and in vitro DNase footprinting. Chromatin immnuprecipitation (ChIP) assay is such an approach whereby the protein of interest is cross-linked to the chromatin in the cells, which are then lysed and the DNA is sheared into pieces of desired size. Using an antibody specific to the protein of interest, protein–DNA complexes are precipitated from the mixture. High throughput variations of the chip technique use ligation mediated PCR to amplify the pool of DNA, generating many copies of all genomic binding sites for a given protein (The identity of DNA regions that are part of the complex can be determined either by using microarrays (ChIP-chip) or by high-throughput sequencing (ChIP-seq) (Taslim et al., 2009; Johnson et al., 2007).
PROTEIN DNA BINDING SITES: TFBS

Large scale In vivo experimental approaches like Chip–Chip or chip-Seq, and the other above mentioned high throughput experimental methods reproduce large qualitative protein binding data sets, from which specificity of transcription factors can be inferred by computational methods. Also, development of computational methods for regulatory binding sites like transcription factor binding sites (TFBS) recognition has traditionally been based on the use of training sets of sequences that are known from such laboratory experimentation to interact with the TF in question. But to date, only a few transcription factors have been experimentally characterized well enough to know which binding sites have sufficient binding sites to be used as regulatory sites in vivo (Homsi et al., 2009).

Several specialized Databases like TRANSFAC (Matys et al., 2006), JASPAR (Sandelin et al., 2004), ARTSITE (Khlebodarova et al., 2006) contain several experimentally verified Transcription factor binding information in ready made transcription factor Binding (TFBs) weight matrices. These matrices can be divided into two types: those
constructed based on natural genomic sequences and those from the artificially selected techniques in vitro.

Given a set of genomic sequences, for example, protein DNA binding site information from several types of experimental reports, the next task is identifying a common overrepresented signature or ‘motif’ from these regions, and de novo motif discovery programs have proven to be invaluable (Elnitski et al., 2006).

**MOTIF DISCOVERY**

A motif is a regulatory site defined as a nucleic acid sequence pattern that has some biological significance such as being a DNA binding site for a regulatory protein (Das and Dai, 2007). It is a fairly short (5-20bp in length) pattern often associated with structural motifs found in proteins. Motifs can occur on both strands of DNA, and for example, a transcription factor binds directly on the double stranded DNA. Various forms of DNA Motifs have been characterized and include asymmetric, palindromic or symmetric motifs. Also included are spaced dyad motifs consisting of two smaller conserved sites separated by a spacer which are usually targeted and bound by Transcription factors as a dimer. A dyad motif implicates a target protein consisting of two subunits with
two separate contact points separated by a non conserved spacer, and usually has a fixed length but might be variable (Rombouts S, et al., 1999).

**MOTIF REPRESENTATIVE MODEL**

There are many ways of representing the sequence specificity of a DNA-binding protein. Firstly, consensus methods have proven to be the simplest way of representing a motif. Consensus in such a context is generally referred to as a sequence that matches all of the example sites closely, but not exactly (Stormo 2000). As Stormo (2000) states, although consensus sequence is the simplest way to represent a collection of sites, it is not optimal for predicting the occurrence of new sites (Stormo 2000). A number of consensus methods for generating consensus sequences from data are possible, and Day and Mc McMorris (1993) have compared several of the methods. Secondly, a position weight matrix (PWM, is another employed way of representing binding sites (Stormo, 2000; methods). For a given matrix, there is a matrix element for all possible bases at every position in the site. The score of any particular site is the sum of matrix values for that site (Stormo 2000; methods).
MOTIF DISCOVERY METHODS

There are several ways by which the task of motif discovery can be formulated. One way is by selecting a motif finding algorithm that best fits the type of DNA sequences which could either be promoter sequences from coregulated genes from a single genome, or orthologous promoter sequences of a single gene from multiple species (i.e., phylogenetic footprinting) and promoter sequences of coregulated genes as well as phylogenetic footprinting (Das Modan K and Ho-Kwok Dai, 2007).

But traditionally, motif discovery methods employed can be broadly divided into two categories: (1) word based (Van Helden et al., 1998) algorithm: enumerative methods which typically involve examining the frequency of all DNA strings and compute overrepresented strings to form a PWM. They are best suited to consensus specific sequence motif models (van Helden J, et al., 1998, 2000; Jensen L et al., 2000; Sinha S et al., 2000).

Alternatively, there are probabilistic approaches that involve representation of the motif model by a position weight matrix (Stormo 2000). These models include methods like expectation–maximization
(Cardon L et al., 1992), Gibbs sampling or greedy approaches (Hertz G et al. 1990).

Notably, The greedy probabilistic sequence based model is one of the first implementations (Consensus motif finding program, see methods) for finding a matrix representation of protein DNA Binding site, for example a transcription factor binding sites (Stormo and Hartzell 1989; Hertz et al., 1990). The criterion by which the algorithm identifies the best alignment of potential sites is by choosing the one with the highest information content (Stormo 2000).

Experimental approaches like chip-chip assay that allow identification of the genomic region to which a protein is bound have assay limitations. In addition to the challenging task of identifying the exact protein bound site (specificity), the assays do not specifically reveal the orientation/complementarity (forward or reverse complement) of the bound sequence. Other experimental methods like gene expression assays, DNaseI protection and Hypersensitivity assays to name a few, may identify DNA regulatory sites without prior knowledge of the binding protein. The challenge then becomes identifying the exact binding site orientation especially if given a symmetric site.
Regardless of which model one decides to employ, the task of motif discovery is that one starts with a collection of genomic sequences known to contain binding sites information for either a known common factor or unknown factor but neither the specificity of the protein, nor the positions or orientation of the binding sites is known.

In our study, we seek to demonstrate that taking orientation into account in predicting high affinity binding sites can affect the accuracy of the predictions and inattention to this detail can lead to false positives and false negatives. We posit that given a set of binding site information with no prior knowledge of the transcription factor to infer binding site sequence orientation, failure to take into account orientation in employing a motif discovery program is more than likely to give false positives.

As such, we sought to model and compare the ability of three methods of a motif finding program, consensus, taking into account the orientation of sampled symmetric/asymmetric binding sites input to accurately predict variable generated true binding models. The first method, denoted –c0 ignores the complement (no flipping) of a given sequence. In this particular experiment it is given the correct binding
orientation and so its output reflects the true matrix limited only by the sample size. The second method, -c2 takes into account both strands, and includes them as a single strand, hence allows flipping. This means that it could select an asymmetric site in the wrong orientation, or for symmetric sites it is forced to pick one of the two possible orientations for each site. The third method, -c3 makes the assumption that the pattern is symmetrical and includes both orientations of each site in the model. This doubles the sample size. Notably, we will test for artifacts on symmetric sites when symmetry is not accounted for properly with the goal of showing how to best recognize symmetry by taking into account the e-values of the statistical scoring criteria. The initial true binding models were derived from the 7-long half site for the Mnt protein. Mnt protein of salmonella phage 22 is a repressor for which the binding specificity has been experimentally characterized (Stormo et al., 1997). Mnt is a site-specific DNA binding repressor and it exists as a tetrramer, with each dimer binding to a nearly symmetric half site (Fields et al., 1997).

Although the pattern is that of a restriction enzyme, HincII site, Figures 1 vividly demonstrates the argument we posit that failure to take into
consideration orientation and symmetry when running a motif
discovery program, is more than likely to lead to a false motif model.
HincII restriction site, GTYRAC is a degenerate site which allows
ambiguity (NEB) and figure 1, depicts the sequence logos of the true
HincII model compared to computational models generated by motif
discovery program either set with orientation specified (flipping),
HincII_c2, and with symmetry specified, HincII_c3. Clearly, even though
the HincII_c2 model has deviated from the true model, it has the highest
Information content compared to the HincII_c3 model and one might
assume that it was the correct model (figure 1).

Again, we posit that given a set of binding site information with no prior
knowledge of the transcription factor to infer binding site sequence
orientation, failure to take into account orientation in employing a motif
discovery program is more than likely to give false positives if the true
model is symmetric.
METHODS and RESULTS

TRUE_BINDING_SITE_MODELS

The true binding site models employed to compare the three consensus methods were derived from the experimentally measured and characterized binding site and affinity of the operator of the Mnt protein of salmonella phage P22 (Stormo et al., 1997). Mnt is a site-specific DNA binding repressor. It exists as a tetramer, with each dimer binding to a nearly symmetric half site (Stormo et al., 1997).

In order to model and compare the ability of three methods of a motif finding program, consensus, to accurately predict variable generated true binding models, the two sets of true binding site models derived were 7 long asymmetric sites namely mnu/mnx, and the symmetric binding sites denoted mnus/mnxs (Table 1). The second sets of the true binding site models were the 6 long Mnt protein derived asymmetric binding half sites denoted mnv/mny and the symmetric sites mnvs/mnys (Table 1). The rationale is to generate variable true model binding sites to compare across the three consensus methods taking into account symmetry versus asymmetry, odd symmetry versus even symmetry.
Each true model binding site and specificity is represented by a position weight matrix that has an element for all possible bases at every position in the site (Stormo 2000). For a given matrix, each row represents one of the four possible bases (A, C, G, T), while each column represents one of the positions of the binding site, hence equal to the motif length (Stormo 2000). A score for a particular site is the sum of matrix values for that site sequence (Stormo et al., 2000).

**True Model DNA sampling**

For our simulations, we make the assumption that each position contributes independently and therefore additively to the total binding. Given each energy matrix, E (labeled as the RE matrices in figures2.1E-H, figure2.2E-H), the binding energy of a sequence $S_i$ is $E \cdot S_i$, whereby $S_i$ is a matrix of the same form as E with a 1 for each base that occurs at each position and 0 for the other bases (Stormo, 2000). For each of the true model sites sampling (sample size = 500), we assumed the Boltzmann distribution where the probability of site being sampled depends on its binding energy. Table 2 lists the sampled sites and their energies for each model; note that lower energy sites occur more frequently than higher energy sites, and most of possible 6- or 7-long
sites do not occur at all in these samples. The probability of site being sampled is:

\[
P(s=1|S) = 1/1+e^{(E_{i}-\mu)} \quad (1)
\]

where \(E_{i}\) is the relative energy score at each site(figure 2.1E-H,2.2E-H), and \(\mu\) is a chemical potential, which corresponds to a free energy for the collection of all the states with e.g. a transcription factor (TF) not bound to \(S_{i}\) (Zhao et al., 2009). For our simulations we set the \(\mu\) value to -0.5.

**6mer and 7mer True Model Sampled Sites**

The true binding site models (figure2.1-2.2) were sampled (n=500) from all possible 7mers and 6mers, and all sampled sites are depicted in graphs corresponding to their binding energies distribution (figures 3A-H). The first graph in each set is the distribution of binding energies drawn from all possible sites from each of the true relative energy matrices. The second graph is the binding probability curve for \(\mu=-0.5\), and it depicts the probability of drawing a sequence with a specific energy score (Yue Zhao et al 2009). The third graph is the posterior distribution of binding energies from the sites in Table 2, (and it tends to be the normalized product of the first two plots (Zhao, Yue et al., 2009).
CONSENSUS:

Each set of sequences listed in table 2 generated from the various true model binding sites were run through a motif finding program, Consensus (Hertz et al., 1990). We used the three methods of consensus of which all take into account orientation of each sequence in different ways (see below). Although our initial input sample size was 500 sequences for each of the three methods, the c0 and c2 methods most statistically significant top matrices (with the lowest probability of occurring by chance) were generated from a sequence range 450-500. For the c3 method which assumes each sequence is symmetric, despite the initial input sample size of 500 sequences, the method takes into account both the sequence and it’s complement resulting in a total analysis of a total of 1000 sequence fragments. As a result the top most statistically significant matrices (with the lowest probabilities were generated from a range of 778 sequences to 1000 sequences (table 3). Table 3 lists all the generated computational models across the three methods (c0-blue, c2-red, c3-green) along with the length and orientation of each model-binding site, which is either a 7mer or 6 mer symmetric, or asymmetric binding site model.
Pertaining to our project, we sought to compare three methods of Consensus taking into account the orientation of sampled symmetric/asymmetric binding sites input. The first method, denoted -c0 ignores the complement (no flipping) of a given sequence. In this particular experiment it is given the correct binding orientation and so its output reflects the true matrix limited only by the sample size. The second method, -c2 takes into account both strands, and includes them as a single strand, hence allows flipping (either orientation can be chosen as correct). This means that it could select an asymmetric site in the wrong orientation, or for symmetric sites it is forced to pick one of the two possible orientations for each site.

The third method, -c3 makes the assumption that the pattern is symmetrical and includes both orientations of each site in the model. This doubles the sample size.

Each generated model is named according to the true model it was derived from along with the consensus method by which it was generated, for instance “mnu_c0” was derived from the mnu computational models generated and summarizes the consensus statistical analysis output: the number of sampled sites used to build the
model, the Information Content of the model and the p-value and e-value for each model. Figures 4-7 show graphically the information contained in Table 3., the consensus output for each experimental model using the three methods (-c0, -c2, -c3) includes two different lists of matrices. We carried forth the first count matrix from the first list having the highest Information content from each cycle, ordered by decreasing statistical significance (Stormo et al., 1989).

Figures 8-11 shows the entire consensus generated count matrices (from the alignments of the sites by each consensus method), which were converted to the relative affinity matrices and relative energy matrices. For each generated count matrix, we first added a pseudocount of 1 to each position to avoid zero counts in generating a frequency matrix. At each individual position, dividing the frequency of each base relative to the ‘consensus’ base generated the relative affinity matrix, such that the consensus base at each position has a score of 1, and other sites are between 0 and 1(Stormo et al., 2000). We represent the specific binding energy contribution of a base at each position in a relative energy matrix by taking the negative natural log of each score represented in a relative affinity matrix. The rationale is that
the best binding base will have a zero energy score while maintaining the differences with the other bases (Berg and von Hippel, 1987).

**SEQUENCE LOGOS: For a graphical representation view**

For each generated relative energy matrix (figures 8-11), we employed the web-based tool, Enologos (Workman et al., 2005), to generate sequence logos of the true model and the computational model binding sites. Figures 12-15 depict the sequence logos of all the true models and the generated computational models. The height (y-axis) of each column corresponds to the information content (IC) of each position (horizontal axis) represented in the relative energy matrix, while the size of each base in each column represents the nucleotide frequency, and positions are in a 5’ to 3’ orientation (Workman et al., 2005). In each figure (12-15), figures A-D shows the logos of the asymmetric true model and the computational models generated from the denoted consensus method (c0, c2, c3), and in contrast to each asymmetric model, figures E-H show the logos of the symmetric true model along with the corresponding computational models across the three consensus methods (Figures E-H). (Workman et al., 2005)
A comparison of the logos allows one to draw the main conclusion of this work. Because of how the sites were sampled the c0 is always very similar to the true model, the differences due entirely to the limited sample size, so it serves as a control for how well Consensus can work when given prior information about the orientation. The other two methods, c2 and c3, test how well consensus can work when the orientation is not known and the sites may, or may not, be symmetric. As can be seen, if the site is asymmetric, the c2 model is quite accurate while assuming symmetry (c3 model) results in a very poor model (quite different from the true model). On the other hand, if the site is symmetric c3 gives a very accurate model but now the c2 model is inaccurate, predicting more information content than is actually there (as in the HincII example in the introduction). These results show that either method can lead to inappropriate models if the underlying assumption of symmetry or asymmetry is incorrect. If the true model is unknown both methods should be tried and the results compared.

**BINDING PROBABILITY COMPARISONS**

Given the computational models for each dataset, and using each motif finding method, we can compare their predicted binding probabilities.
For each of the computational models we sample sequences and generate probability distributions, as we did for the true models in Figure 3. Figures 16-19 show the probability distributions from each of the computational models (Figures 8-11) now using a sample size of 1000 and μ=0 (sampled sites are available in supplementary file 1). Consistent with graphical views of the Logos (Figures 12-15), the probability distributions of the c0 models are quite similar to the true distributions (as expected since this is a control) and the c2 distributions are accurate when the true model is asymmetric and the c3 distributions are accurate when the true model is symmetric.

**Effects On Binding Searches**

Lastly, we want to determine what the effects of the incorrect models are when searching a genome for new binding sites. Given a PWM a person will search for new potential binding sites by scanning a genome and finding high scoring sites. To see whether the computational models are likely to make false positive and false negative predictions we can compare the scores of all the predicted sites with their true binding energies. As shown in Figures 16-19, for each computational model we can predict the binding energy distributions, and furthermore we can rank each possible 7mer and 8mer for their predicted binding
probability and compare them to the true binding probabilities. We do this using the following two approaches: 1) we will take the total sum of the binding probabilities of each sequence and it’s reverse complement; 2) instead of summing up and obtaining the total binding energy of each sampled sequence and it’s reverse complement, we will compare the two sequences and take the maximum binding probability of the two. Most search programs for predicting new sites use the second approach, but the first one is really more accurate and so we make comparisons using both approaches.

Figures 20A-H to 27A-H are scatter plots of relative binding energies of the sampled sites generated from the three consensus methods(-c0,c2,c3) binding site models versus the relative binding energy scores of the true binding model sampled sites(figures 19A-26H). Each figure includes plots generated from the relative binding energies as stated above. The plots denoted with the inclusion of “ss” are from the second approach whereby we took the maximum binding probability score between the forward sequence and it’s reverse complement. For each plot, on the y-axis is the calculated relative binding probability of each experimental model against the relative binding probability of the
corresponding true model (horizontal axis). Furthermore for the linear regression analysis, we used the linear model (lm) function in R, and the abline function to extract information from the linear model (lm) fitted to each of the data points as depicted by the regression lines throughout our plotted data (R statistical program, 2010). Also derived from the lm function and included in all the plots is the squared Pearson correlation coefficient, $R^2$. The plots are only shown up to energies of 5, which constitute all of the high affinity binding sites but ignores the low affinity sites.

**7-long Asymmetric/ symmetric sampled sites:**

The relative binding energy plots of both Mnu/Mnx c0 and c2 methods generated models depict a higher positive association relative to the true model sites with much higher $R^2$ values than the c3 model which assumes the sampled sites are symmetric, clearly deviates from the Mnu/Mnx sites true model sites (figures 20-21 D, H).

Figures 22A-23H are plots of the of the relative binding energies of the 7-long symmetric Mnus/Mnxs c0,c2, c3 methods generated sample sites relative to the Mnus/Mnxs True model sampled sites. All the three methods across all the plots show high positive correlation with True
model sampled sites relative energies. The Mns/Mnx c3 method shows the best linear fit compared to the other two methods, with the c2 method showing a greater deviation from the true model although it is surprisingly good considering the inaccurate Logo.

6-Long Asymmetric/Symmetric Sampled sites

Figures 24A-25H show relative binding energy scatter plots of sampled sites from the 6-long asymmetric Mnv/Mny c0, c2, and c3 methods. Like the 7 long asymmetric sampled sites from the c0, c2 methods (figures 20-21), the 6 long asymmetric sampled sites from Mnv and Mny c0, c2 methods show a higher positive association relative to the sampled sites of Mnv and Mny true models. Despite the overfitting of the relative binding energy scores relative to the True model relative binding energy cutoffs of less than or equal to 5 (figures A,B,E,F). The relative binding energy scores of the sampled sites from the Mnv and Mny c3 method clearly deviate from those generated from the Mnv and Mny True models.

Figures 26-27 are plots of the relative energy binding of the sampled sites from the 6-long symmetric Mnvs/Mnys c0, c2, and c3 methods. Notably, and unlike the observed 7-long symmetric c2 method
generated models, the 6-long symmetric sampled sites from the c2 method generated models show a significantly lower positive correlation compared to the other two methods (figures 25B, 25F and 25O). The c0 and c3 methods show an almost perfect linear correlation relative to the true model (figures 26-27 A, C, E, F). Figures 26-27 D, H is a plot of all the sampled sites derived from the three experimental models relative to the true binding models, and it shows that the sampled sites from the c2 generated method are the most deviated from the true model and the c0 and c3 methods.
DISCUSSION

Availability of Genomic sequences along with advancements in experimental techniques and computational methods are significantly contributing to our ability to identify and characterize regulatory elements (Elnitski L et al., 2006).

Transcription is the hallmark of gene expression and it’s regulation. For instance, transcription initiation reaction is facilitated by transcription factors’ recognition of cis-regulatory regions like promoters. Within these cis-regulatory regions lies DNA sequence motifs of which are target sites for general and/or specific transcription factors (Patshne et al., 1997). In order for the right gene to be expressed at the right place, at the right moment, at the right level, a high degree of specificity during protein DNA recognition events is required to directly or indirectly recruit the transcriptional machinery (Ptashne et al., 1997). The challenging task of predicting and identifying cis-regulatory elements both from an experimental and computational stand points suffers from high false positive due to various contributing factors including misinterpretation of DNA binding site sequence orientation.
Given a newly sequenced genome, the primary goal is to decipher cis regulatory elements that regulate gene expression. The convergence of in vivo experimental approaches and computational methods could help in identifying motifs with high confidence for a particular transcription factor (Elnitski L et al., 2006).

Large genomic scale experimental approaches like chip-chip assay strategies that detect protein DNA interactions in vivo give a broad spectrum of binding site information, but do not readily reveal orientation of binding sequence sites. In such studies, a specific DNA binding protein under study like a transcription factor’s DNA binding domain structure may not be fully understood to accurately predict the target DNA binding site orientation, including if it’s symmetric or asymmetric. Furthermore, even if the DNA binding protein (e.g., transcription factor) structure is known to dimerize, a property known to be required for high-affinity and sequence-specific DNA binding (Lewin, B. 2000), such studies do not reveal the types of homodimers, or heterodimers it forms, hence making it a challenge to confidently characterize the orientation of the DNA binding site and whether it’s symmetric or asymmetric.
For example, approximately 80% of the few fully characterized TFs in higher eukaryotes contain a helix-turn-helix domain, a basic helix-loop-helix, a zinc finger domain, or a leucine zipper (Lewin, B. 2000). Helix-loop-helix and leucine zipper domains allow dimerization of factors that contain an N-terminal helix that ensures interactions with DNA (Lewin, B. 2000). By fully knowing the Tfs binding domain and its structure gives one a better spectrum of characterizing the DNA binding site as symmetric, asymmetric and overall orientation. Studies have demonstrated the various forms of homodimers with Examples of their various configurations shown to occur in vivo. For example the two main ways of how homodimers formed include the most frequently form where each monomer has a domain that interacts with the same domain of the other monomer (e.g., a leucine zipper) as is the case with the estrogen receptor DNA (ER) binding domain and it’s bipartite TFBS (Schwabe, J. et al., 1993). In these cases, the resulting dimers could be predicted as symmetric, with the ideal arrangement of the two binding sites being a palindrome, a geometry that also displays a central symmetry (Georges et al., 2010). The other way occurs when one domain of one monomer recognizes a different domain of the other monomer. A TF complex formed according to such a pattern would
preferably recognize half-binding sites arranged in a head-to-tail configuration (Georges et al., 2010).

Where experimental data does not suffice, computational approaches, including motif discovery methods have been instrumental in modeling protein DNA binding interactions, their affinity and specificity. Studies by von Hippel and Berg (1986) provided the rationale for converting the biophysical problem of TF–DNA affinity into a pattern matching and pattern discovery problem (Stormo 2000; D’haeseleer, 2006; Djordjevic et al., 2003). As such, the preferential binding of some TFs to certain DNA sequences can be expressed in terms of a sequence motif or position weight matrix (Wasserman and Sandelin, Zhao et al., 2009).

Scoring criteria for motifs have been developed, and most motif discovery methods have their own preferred metric for scoring. Most scores involve a measure of information content or statistical overrepresentation. Motif discovery computational methods like all computational approaches suffer from high false positive rates due to various contributing factors. Regardless of which motif discovery program is employed, In our study we sought to demonstrate how if orientation is not taken into account when modeling DNA binding sites,
a motif discovery method may generate high scoring and statistically significant DNA binding site model/models which may either be false positive/negative, hence inconclusive. As depicted in Tables 1 and 3 we demonstrated this possibility by taking true binding sites models which were either asymmetric or odd symmetric with one spacer region or even symmetric with no spacer and simply testing how well a motif discovery program performs in generating the true models when taking into consideration orientation.

As depicted in table 3, running a motif finding program with the option of taking into consideration only the c2, complement (flipping) option and not the c3, symmetry option, one could easily predict either a false positive model as the most statistically significant model (Table 3, Mnnus/Mnxn models) if based only on high information content and lower P-value. We tested this possibility by testing how well the three orientation options perform in predicting the true symmetric binding site model (Methods). All three methods generated significantly higher IC and lower p-value, with the c2 method scoring the highest. Interestingly though, the e-value scores seemed to reveal the c3 method generated model as the most significant. This observation leads us to
conclude that the e-value scoring criterion gives a more accurate prediction (table 3).

If one has a set of genomic sequences consisting of either DNA binding sites information from a high throughput experimental methods, or from a newly sequenced genome yet to be gene annotated, and without knowing beforehand the symmetric or asymmetric orientation of the putative cis regulatory regions, as depicted in table 3 there is a possibility of predicting a false positive model hence a false positive sequence logos (figures 12E_G, 13E_G, 14E_G, 15E_G).

We further validated our observation by taking each of our generated experimental DNA binding model and sampled all possible 7mer or 6mer DNA binding half sites. Furthermore we calculated binding probability of each site with respect to the Boltzmann distribution. As described in the result we took two major approaches in calculating the binding energy of each sampled site. In the first approach we took the total sum of the binding probabilities of each sequence and it’s reverse complement. In our second approach, we compared the two sequences and took the maximum binding probability of the two sequences.
Although there wasn’t a significant difference between the output of the above mentioned approaches we took to calculate the binding energy of each sampled site, we validated our observations by linear regression analysis(Figures20-27) methods.

This work has several main conclusions. The first is that if asymmetric sites are assumed to be symmetric the resulting models are generally very poor. But it is also true, although less obvious, that if symmetric sites are not assumed to be symmetric the resulting model can also be rather poor. In both cases the incorrect models can lead to high levels of false positive and false negative predictions of true binding sites. Judging the best models by their information content, or even by their p-values, does not resolve which is the best model. However, but taking into account the symmetry constraints when computing the e-values, the best model can be obtained routinely.
REFERENCES


10. Dorschner MO, Michael Hawrylycz, Richard Humbert, James C Wallace, Anthony Shafer, Janelle Kawamoto, Joshua Mack, Robert Hall, Jeff Goldy,


31. Manco L, Vagace JM, Relvas L, Rebelo U, Bento C, Villegas A, Letícia Ribeiro M. “Chronic haemolytic anaemia because of pyruvate kinase (PK) deficiency in a child heterozygous for haemoglobin S and no clinical features of sickle cell


1183–1187


HincII Restriction site Logos: GTRYAC

A. TRUE-SITE

B. c2-FLIP

C. c3-Symmetry

Figure 1. HincII restriction site: GTRYAC. A. Sequence logo of the true site model. B. Sequence logo of the consensus c2(flipping) method generated Model. C. Sequence logo of the consensus c3(symmetry) method generated model.
### 7-LONG SYMMETRIC/ASYMMETRIC BINDING MATRICES

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th></th>
<th>B</th>
<th></th>
<th>C</th>
<th></th>
<th>D</th>
<th></th>
<th>E</th>
<th></th>
<th>F</th>
<th></th>
<th>G</th>
<th></th>
<th>H</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mnu_T_RA</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Mnu_T_RA</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Mnu_T_RA</td>
</tr>
<tr>
<td>A</td>
<td>0.187</td>
<td>0.086</td>
<td>0.197</td>
<td>0.25</td>
<td>1</td>
<td>0.105</td>
<td>0.053</td>
<td>A</td>
<td>1.677</td>
<td>2.454</td>
<td>1.625</td>
<td>1.387</td>
<td>0</td>
<td>2.254</td>
<td>2.938</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>0.093</td>
<td>0.143</td>
<td>0.013</td>
<td>1</td>
<td>0.2</td>
<td>1</td>
<td>1</td>
<td>C</td>
<td>2.376</td>
<td>1.945</td>
<td>4.343</td>
<td>0</td>
<td>1.61</td>
<td>0</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0.2</td>
<td>1</td>
<td>1</td>
<td>0.143</td>
<td>0.013</td>
<td>0.093</td>
<td>G</td>
<td>0</td>
<td>1.61</td>
<td>0</td>
<td>0</td>
<td>1.945</td>
<td>4.343</td>
<td>2.376</td>
<td>G</td>
</tr>
<tr>
<td>T</td>
<td>0.053</td>
<td>1</td>
<td>0.105</td>
<td>0.25</td>
<td>0.086</td>
<td>0.197</td>
<td>0.187</td>
<td>T</td>
<td>2.938</td>
<td>0</td>
<td>2.254</td>
<td>1.387</td>
<td>2.454</td>
<td>1.625</td>
<td>1.677</td>
<td>T</td>
</tr>
</tbody>
</table>

**Figure 2.1A-H.** **Relative affinity and energy binding matrices.** A-H. Relative frequency of each type of DNA base at each position of asymmetric (A, C) and symmetric true model 7-long half site. The highest affinity binding sites are listed in table 1. E-H. The relative energy scores at each position which are the negative natural log of the values in A-H. Each score position represents the binding energy contributed by a particular base at that position in the site to the total binding energy. The relative energy matrix represents the “true” binding model.
**6-long symmetric/asymmetric binding matrices**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mnv_T_RA</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0.187</td>
<td>0.086</td>
<td>0.197</td>
<td>1</td>
<td>0.105</td>
<td>0.053</td>
<td>1.677</td>
<td>2.454</td>
</tr>
<tr>
<td>C</td>
<td>0.093</td>
<td>0.143</td>
<td>0.013</td>
<td>0.2</td>
<td>1</td>
<td>1</td>
<td>2.376</td>
<td>1.945</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0.2</td>
<td>1</td>
<td>0.143</td>
<td>0.013</td>
<td>0.093</td>
<td>0</td>
<td>1.61</td>
</tr>
<tr>
<td>T</td>
<td>0.053</td>
<td>1</td>
<td>0.105</td>
<td>0.086</td>
<td>0.197</td>
<td>0.187</td>
<td>2.938</td>
<td>0.2254</td>
</tr>
</tbody>
</table>

**Figure 2.2 A-H.** 6-long Relative affinity and Energy binding matrices. A-H. Relative frequency of each type of DNA base at each position of asymmetric (A,C) and symmetric true model 6-long half site. The highest affinity binding sites are listed in table 1. E-H. The relative energy scores at each position which are the negative natural log of the values in A-H. Each score position represents the binding energy contributed by a particular base at that position in the site to the total binding energy. The relative energy matrix represents the “true” binding model.
Figure 3A-H. Distribution of binding energies of all the sampled 7mers and 6mers from the true model half-site energy matrices. (A.1-H.1). The binding probability curves for $\mu = -0.5$ (3A.2-H.2) show the posterior distribution of binding energies whereby each plot is the normalized product of the plots in, figures 3A-H and figures 3A.1-3H.1 (Yue Zhao et al 2009)
Figure 4A-C. 7-long asymmetric experimental Binding Models Consensus statistical output. The three consensus methods are denoted as follows: c0-blue, c2-red, c3-green. A. Information content scores across the three consensus methods. B. The ln(p-value) scores. C. ln(e-value) scores.
Figure 5A-C. 7-long symmetric experimental Binding Models Consensus statistical output. The three consensus methods are denoted as follows: c0-blue, c2-red, c3-green. A. Information content scores across the three consensus methods. B. The ln(p-value) scores. C. ln(e-value) scores.
Figure 6A-C: Mnv/MnyMn 6-long asymmetric experimental Binding Models Consensus

**statistical Outputs.** The three consensus methods are denoted as follows: c0-blue, c2-red, c3-green. A. Information content output across the three consensus methods. B. The ln(p-value) scores. C. ln(e-value) scores.
Figure 7A-C: Mnvs/MnysMn6-long symmetric experimental Binding Models Consensus statistical Output. The three consensus methods are denoted as follows: c0-blue, c2-red, c3-green. A) Information content output across the three consensus methods B. The ln(p-value) scores C. ln(e-value) scores.
LONG ASYMMETRIC AND SYMMETRIC CONSENSUS GENERATED COUNT MATRICES ACROSS THE THREE METHODS: CO-BLUE, C-RED, C-GREEN.

<table>
<thead>
<tr>
<th>Method</th>
<th>Count Matrix</th>
<th>Feature Names</th>
</tr>
</thead>
</table>
| A | MnuT_RA | RELATIVE AFFINITY
| B | MnuT_RE | RELATIVE ENERGY
| C | Mnu_T_CO | SAME NAME?
| D | Mnu_T_FR | SAME NAME?
| E | Mnu_C CO | SAME NAME?
| F | Mnu_C FR | SAME NAME?
| G | Mnu_C CO | RELATIVE AFFINITY
| H | Mnu_C FR | RELATIVE ENERGY

**Figure 8A-H.** Mnu/Mnus 7-long asymmetric and symmetric consensus generated count matrices across the three methods: CO-BLUE, C-RED, C-GREEN. For each count matrix is a corresponding frequency (name_FR), relative affinity (name_RA), relative energy (name_RE). (See Results for details)
Figure 9A-H. Mnx/Mnx$_T$-long asymmetric and symmetric consensus generated count matrices across the three methods (c-blue, c-red, c-green). For each count matrix is a corresponding frequency(name_FR), relative affinity(name_RA), relative energy(name_RE). (See Results for details)
Figure 10A-H. **Mnv/MvS** 6-long asymmetric and symmetric consensus generated count matrices across the three methods: **c0-blue, c2-red, c3-green**. For each count matrix is a corresponding frequency(name_FR), relative affinity(name_RA), relative energy(name_RE). (See Results for details)
LONG ASYMMETRIC AND SYMMETRIC CONSENSUS 

GENERATED COUNT MATRICES ACROSS THE THREE METHODS

![Image showing the table and diagram](image)

**Figure 11A-H.** Mny/Mnns 6-long asymmetric and symmetric consensus generated count matrices across the three methods: c0-blue, c2-red, c3-green. For each count matrix is a corresponding frequency (name_RA), relative affinity (name_RA), relative energy (name_RE). (See Results for details)
Figure 12A-H. Sequence logos of the 7 long symmetric and asymmetric experimental binding half sites in comparison to the true model binding sites. The experimental models correspond to the three consensus methods: c0 ignores the complement, c2 includes both strands as a single sequence, c3 assumes symmetric. The height of the stacked bases corresponds to the Information content (IC) measured in bits at each highlighted position. (A) Sequence logo derived from the asymmetric Mnus_true model binding site. (B), (C), (D) Sequence logos of the asymmetrical Experimental binding sites corresponding to the three consensus methods: c0, c2, c3 respectively. Figures E-H. (E) Sequence logo generated from the Symmetrical Mnus_true model. (F), (G), (H) Sequence logos of the symmetrical Experimental binding site Mnus corresponding to the Consensus methods c0, c2, and c3 respectively.
Figure 13A-H. Sequence logos of the Mnx/Mnxs 7 long symmetric and asymmetric experimental binding half sites in comparison to the Mnx/Mnxs true model binding sites. The Experimental models correspond to the three consensus methods: c0, ignores the complement; c2-includes both strands as a single sequence, c3- assumes symmetric. The height of the stacked bases corresponds to the Information content (IC) measured in bits at each position. (A) sequence logo derived from the asymmetric Mnx True model binding site. (B),(C),(D) Sequence logos of the asymmetrical Experimental binding sites corresponding to the three consensus methods; c0, c2, c3 respectively. Figures E-H. (E) Sequence logo generated from the Symmetrical Mnx true model. (F),(G),(H) Sequence logos of the symmetrical experimental binding site Mnxs corresponding to the Consensus methods c0, c2, and c3 respectively.
Figure 14A-H. Sequence logos of the 6 long symmetric and asymmetric experimental binding half sites in comparison to the true model binding sites. The Experimental models correspond to the three consensus methods: c0, ignores the complement, c2-includes both strands as a single sequence, c3-assumes symmetric. The height of the stacked bases corresponds to the Information content (IC) measured in bits at each position. (A) sequence logo derived from the asymmetric Mnv True model binding site. (B),(C),(D) Sequence logos of the Mnv asymmetrical experimental binding sites corresponding to the three consensus methods; c0,c2,c3 respectively. Figures E-H. (E) Sequence logo generated from the Symmetrical Mnv true model. (F),(G),(H) Sequence logos of the symmetrical experimental binding site Mnv corresponding to the Consensus methods c0,c2, and c3 respectively.
Figure 15A-H. Sequence logos of the 6 long symmetric and asymmetric experimental binding half sites in comparison to the true model binding sites. The experimental models correspond to the three consensus methods: c0, ignores the complement, c2-includes both strands as a single sequence, c3-assumes symmetry. The height of the stacked bases corresponds to the Information content (IC) measured in bits at each position. (A) sequence logo derived from the symmetric Mny True model binding site. (B),(C),(D) Sequence logos of the Mny asymmetrical Experimental binding sites corresponding to the three consensus methods; c0, c2, c3 respectively. Figures E-H. (E) Sequence logo generated from the Symmetrical Mny True model. (F),(G),(H) Sequence logos of the Mny symmetrical experimental binding site Mnyus corresponding to the Consensus methods c0, c2, and c3 respectively.
For each model, the first graph in red depicts the distribution of binding energies across all the sampled 6mers. The second graph in blue is the binding probability curve for $\mu = 0$. The third graph in green is the posterior distribution of binding energies and it is the normalized product of the first two graphs (Yue Zhao et al., 2009).
Figure 17A-H. Mny/Mnys 6-long asymmetric/symmetric true and experimental binding models. "Binding energy distribution plots". For each model, the first graph in red, depicts the distribution of binding energies across all the sampled 6mers. The second graph in blue, is the binding probability curve for μ = 0. The third graph in green is the posterior distribution of binding energies and it is the normalized product of the first two graphs (Yue Zhao et al., 2009)
Figures 18A-H. Mnu/Mnus-7-long asymmetric/symmetric true and experimental binding models. Binding energy distribution plots. For each model, the first graph in red, depicts the distribution of binding energies across all the sampled 6mers. The second graph in blue, is the binding probability curve for \( \mu = 0 \). The third graph in green is the posterior distribution of binding energies and it is the normalized product of the first two graphs (Yue Zhao et al., 2009).
Figures 19A-H: Mnx/MnxS 7-long asymmetric/symmetric true and experimental binding models. Binding energy distribution plots. For each model, the first graph in red, depicts the distribution of binding energies across all the sampled 6mers. The second graph in blue, is the binding probability curve for $\mu = 0$. The third graph in green is the posterior distribution of binding energies and it is the normalized product of the first two graphs (Yue Zhao et al., 2009).
Figure 20A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 7-long Mnu asymmetrical models versus the sampled sites from Mnu true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the Mnu c0 method generated models versus the Mnu true model. C and D are all the plots of sampled sites from Mnu c2 method generated models versus the true model. E and F are sampled sites from the Mnu c3-method generated models versus the sampled sites from the Mnu true models. D and H plots of all the sampled sites from the three consensus methods versus the Mnu true binding model sampled sites. Included in each plot is squared Pearson correlation, R² value.
Figure 21A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 6-long Mnx asymmetrical models versus the sampled sites from Mnx true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the Mnx c0 method generated models versus the Mnx true model. C and D are all the plots of sampled sites from the Mnx c2 method generated models versus the Mnx true model. E and G are sampled sites from the Mnx c3-method generated models versus the sampled sites from the Mnx true models. D and H plots of all the sampled sites from the three consensus methods versus the Mnvts true binding model sampled sites. Included in each plot is the squared Pearson correlation, $R^2$ value.
Figure 22A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 7-long Mnas symmetrical models versus the sampled sites from the symmetrical Mnas true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the c0 method generated models versus the Mnas true model. C and D are all the plots of sampled sites from c2 method generated models versus the true model. E and F are sampled sites from the c3-method generated models versus the sampled sites from the Mnas true models. D and H plots of all the sampled sites from the three consensus methods versus the Mnas true binding model sampled sites. Include in each plot is the R² value, the squared Pearson correlation.
Figure 23A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 7-long MnxS symmetrical models versus the sampled sites from MnxS true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H, the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the MnxS 0 method generated models versus the MnxS true model. C and D are all the plots of the sampled sites from the MnxS c2 method generated models versus the MnxS true model. E and F are sampled sites from the MnxS c3-method generated models versus the sampled sites from the MnxS true models. G and H plots of all the sampled sites from the three consensus methods versus the MnxS true binding model sampled sites. Included in each plot is the squared Pearson correlation, R² value.
Figure 24A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 6-long Mnv asymmetrical models versus the sampled sites from Mnv true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the c0 method generated models versus the Mnv true model. C and D are all the plots of sampled sites from c2 method generated models versus the true model. E and F are sampled sites from the c3-method generated models versus the sampled sites from the Mnu true models. D and H plots of all the sampled sites from the three consensus methods versus the Mnv true binding model sampled sites. Include in each plot is the $R^2$ value, the squared Pearson correlation.
Figure 25A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 6-long Mny asymmetrical models versus the sampled sites from Mny true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the Mny c0 method generated models versus the Mny true model. C and D are all the plots of sampled sites from the Mny c2 method generated models versus the Mny true model. E and F are sampled sites from the Mny c3-method generated models versus the sampled sites from the Mny true models. D and H plots of all the sampled sites from the three consensus methods versus the Mny true binding model sampled sites. Included in each plot is the squared Pearson correlation, R² value.
Figure 26A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 6-long Mnvs symmetrical models versus the sampled sites from Mnvs true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the c0 method generated models versus the Mnvs true model. C and D are all the plots of sampled sites from Mnvs c2 method generated models versus the mnvs true model. E and F are sampled sites from the Mnvs c3-method generated models versus the sampled sites from the Mnvs true models. G and H plots of all the sampled sites from the three consensus methods versus the Mnvs true binding model sampled sites. Included in each plot is the squared Pearson correlation, R² value.
Figure 27A-H: Scatter plots of the relative binding energy of the sampled sites from the experimental 6-long Mnyx symmetrical models versus the sampled sites from Mnyx true binding model A-D. The relative binding energy of each sampled site was derived from the total sum of the binding probability of both the forward strand and its reverse complement (see methods). E-H the relative binding energy of each sampled site is the maximum binding probability between the forward sequence and its reverse complement (see methods). A and B are all the plots of the Mnyx0 method generated models versus the Mny true model. C and D are all the plots of sampled sites from the Mnyx c2 method generated models versus the Mny true model. E and F are sampled sites from the Mnyx c3-method generated models versus the sampled sites from the Mnyx true models. G and H plots of all the sampled sites from the three consensus methods versus the Mnyx true binding model sampled sites. Included in each plot is the squared Pearson correlation, R² value.
<table>
<thead>
<tr>
<th>Model_name</th>
<th>Highest_Affinity site</th>
<th>Length</th>
<th>Symm/Asymm</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-long Mnt-half site</td>
<td>5'-GTGGGACC-3'</td>
<td>7</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>Mnu</td>
<td>5'-GTGSCACC-3'</td>
<td>7</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>Mnus</td>
<td>5'-GTGSCAC-3'</td>
<td>7</td>
<td>Symmetric</td>
</tr>
<tr>
<td>Mnx</td>
<td>5'-GTGSCACC-3'</td>
<td>7</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>Mnxs</td>
<td>5'-GTGSCAC-3'</td>
<td>7</td>
<td>Symmetric</td>
</tr>
<tr>
<td>Mnv</td>
<td>5'-GTGACC-3'</td>
<td>6</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>Mnvs</td>
<td>5'-GTGCAC-3'</td>
<td>6</td>
<td>Symmetric</td>
</tr>
<tr>
<td>Mny</td>
<td>5'-GTGACC-3'</td>
<td>6</td>
<td>Asymmetric</td>
</tr>
<tr>
<td>Mnys</td>
<td>5'-GTGCAC-3'</td>
<td>6</td>
<td>Symmetric</td>
</tr>
</tbody>
</table>

Table 1. A list of all the true model binding sites derived from the 7-long Mnt protein half site. The highest affinity sites of each model, the length, and symmetric/asymmetric orientation are highlighted in blue (7-long Mnt half site) and red (6 long-half site).
<table>
<thead>
<tr>
<th>Model</th>
<th>Length</th>
<th>Symmc/Asymm</th>
<th>Sample Size</th>
<th>Info_Content</th>
<th>(\ln(p\text{-value}))</th>
<th>(\ln(e\text{-value}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mnu_C0</td>
<td>7</td>
<td>Asymmetric</td>
<td>500</td>
<td>3.20677</td>
<td>-1547.98</td>
<td>-1547.98</td>
</tr>
<tr>
<td>Mnu_C2</td>
<td>7</td>
<td>Asymmetric</td>
<td>500</td>
<td>3.25667</td>
<td>-1572.88</td>
<td>-1227</td>
</tr>
<tr>
<td>Mnu_C3</td>
<td>7</td>
<td>Asymmetric</td>
<td>1000</td>
<td>2.08044</td>
<td>-1015.09</td>
<td>-1015.09</td>
</tr>
<tr>
<td>Mnus_C0</td>
<td>7</td>
<td>Symmetric</td>
<td>463</td>
<td>3.57558</td>
<td>-1600.54</td>
<td>-1600.54</td>
</tr>
<tr>
<td>Mnus_C2</td>
<td>7</td>
<td>Symmetric</td>
<td>455</td>
<td>4.32738</td>
<td>-1914.32</td>
<td>-1599.63</td>
</tr>
<tr>
<td>Mnus_C3</td>
<td>7</td>
<td>Symmetric</td>
<td>926</td>
<td>3.56121</td>
<td>-1623.15</td>
<td>-1623.15</td>
</tr>
<tr>
<td>Mnx_C0</td>
<td>7</td>
<td>Asymmetric</td>
<td>446</td>
<td>3.56493</td>
<td>-1535.39</td>
<td>-1535.39</td>
</tr>
<tr>
<td>Mnx_C2</td>
<td>7</td>
<td>Asymmetric</td>
<td>425</td>
<td>3.72657</td>
<td>-1529.62</td>
<td>-1235.73</td>
</tr>
<tr>
<td>Mnx_C3</td>
<td>7</td>
<td>Asymmetric</td>
<td>828</td>
<td>2.17905</td>
<td>-877.612</td>
<td>-877.612</td>
</tr>
<tr>
<td>Mnxs_C0</td>
<td>7</td>
<td>Symmetric</td>
<td>448</td>
<td>3.54323</td>
<td>-1532.73</td>
<td>-1532.73</td>
</tr>
<tr>
<td>Mnxs_C2</td>
<td>7</td>
<td>Symmetric</td>
<td>450</td>
<td>4.21324</td>
<td>-1841.35</td>
<td>-1530.13</td>
</tr>
<tr>
<td>Mnxs_c3</td>
<td>7</td>
<td>Symmetric</td>
<td>896</td>
<td>3.5283</td>
<td>-1555.11</td>
<td>-1555.11</td>
</tr>
<tr>
<td>Mnv_C0</td>
<td>6</td>
<td>Asymmetric</td>
<td>500</td>
<td>3.11601</td>
<td>-1511.15</td>
<td>-1511.15</td>
</tr>
<tr>
<td>Mnv_C2</td>
<td>6</td>
<td>Asymmetric</td>
<td>500</td>
<td>3.16433</td>
<td>-1535.3</td>
<td>-1189.42</td>
</tr>
<tr>
<td>Mnv_c3</td>
<td>6</td>
<td>Asymmetric</td>
<td>1000</td>
<td>2.07989</td>
<td>-1017.66</td>
<td>-1017.66</td>
</tr>
<tr>
<td>Mnvs_C0</td>
<td>6</td>
<td>Symmetric</td>
<td>500</td>
<td>3.09582</td>
<td>-1501.06</td>
<td>-1501.06</td>
</tr>
<tr>
<td>Mnvs_C2</td>
<td>6</td>
<td>Symmetric</td>
<td>500</td>
<td>3.43655</td>
<td>-1671.48</td>
<td>-1325.6</td>
</tr>
<tr>
<td>Mnvs_C3</td>
<td>6</td>
<td>Symmetric</td>
<td>1000</td>
<td>3.08518</td>
<td>-1519.87</td>
<td>-1519.87</td>
</tr>
<tr>
<td>Mny_C0</td>
<td>6</td>
<td>Asymmetric</td>
<td>458</td>
<td>3.46987</td>
<td>-1543.1</td>
<td>-1543.1</td>
</tr>
<tr>
<td>Mny_C2</td>
<td>6</td>
<td>Asymmetric</td>
<td>422</td>
<td>3.75072</td>
<td>-1537.51</td>
<td>-1245.69</td>
</tr>
<tr>
<td>Mny_C3</td>
<td>6</td>
<td>Asymmetric</td>
<td>778</td>
<td>2.3514</td>
<td>-893.074</td>
<td>-893.074</td>
</tr>
<tr>
<td>Mnys_C0</td>
<td>6</td>
<td>Symmetric</td>
<td>449</td>
<td>3.54278</td>
<td>-1544.8</td>
<td>-1544.8</td>
</tr>
<tr>
<td>Mnys_C2</td>
<td>6</td>
<td>Symmetric</td>
<td>430</td>
<td>4.07547</td>
<td>-1707.36</td>
<td>-1410</td>
</tr>
<tr>
<td>Mnys_C3</td>
<td>6</td>
<td>Symmetric</td>
<td>904</td>
<td>3.50071</td>
<td>-1560.05</td>
<td>-1560.05</td>
</tr>
</tbody>
</table>

Table 3. A summary table of Consensus statistical output highlighting the different methods as follows: -c0 (no flipping) in blue, -c2 (flipping) method in red, and -c3 (assuming symmetric) method in green.