Identification of Structural Signatures Within Transcription Factor Binding Sites and Their Flanking Regions in Saccharomyces cerevisiae Enhanced Through ΔTRX Based Multiple Sequence Alignment

Katharina Schulze

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation

This Thesis is brought to you for free and open access by the Thesis/Dissertation Collections at RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact rtscholarworks@rit.edu.
Identification of Structural Signatures Within Transcription Factor Binding Sites and Their Flanking Regions in *Saccharomyces cerevisiae* Enhanced Through ΔTRX Based Multiple Sequence Alignment

by

Katharina Schulze

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Bioinformatics

Thomas H. Gosnell School of Life Sciences
College of Science

Rochester Institute of Technology
Rochester, NY
May 19, 2014
Thesis Advisor:

Gregory A. Babbitt, Ph.D.

Thesis Committee Members:

Gregory A. Babbitt, Ph.D.

Feng Cui, M.D., Ph.D.

Gary R. Skuse, Ph.D.
# Table of Contents

Abstract ........................................................................................................................................... 4

Introduction ...................................................................................................................................... 5

Methods and Materials ..................................................................................................................... 11
  Transcription Factor Binding Site Consensus Pre-Processing ..................................................... 12
  Consensus Sequence Matching ...................................................................................................... 13
  Sequence Alignment ...................................................................................................................... 14
    Center Alignment ....................................................................................................................... 15
  Consensus Match Alignment Based on ΔTRX ............................................................................. 15
  Information Content Comparison ................................................................................................. 16
  Enhanced Consensus Logos .......................................................................................................... 19

Results .............................................................................................................................................. 20

Discussion ......................................................................................................................................... 26

Conclusion ......................................................................................................................................... 31

References .......................................................................................................................................... 33
Abstract

Traditional transcription factor binding site analyses focus solely on the nucleotide composition of site despite the fact that more recent studies have shown transcription factors to rely on the DNA structural features within and surrounding their binding sites. In this study a metric of intrinsic DNA flexibility referred to as the TRX scale is used to assess the structural features within functionally annotated binding sites and their up- and downstream flanking regions based on their Shannon information content (IC). Two methods of sequence alignment, center and a novel ΔTRX based multiple sequence alignment, are compared. The results show that at least 95% of all up- and downstream flanking regions contained more IC in their structural signature as defined by the TRX scale. Between 23% and 35% (excluding and including bridging phosphate bonds, respectively) of flanking regions also showed significant differences between the sets of confirmed and non-confirmed matches. However, few to no significant differences in IC were observed in consensus match regions where sequence dependent major groove contacts are most likely to occur. These findings support the notion that structural context is highly important in the distinction between true and false binding sites. Enhanced consensus logos are demonstrated for the visualization of these structural signatures. While ΔTRX based multiple sequence alignment appeared to be superior in flanking regions when compared to center alignment, further analyses are needed in order to increase the confidence in these findings.
Introduction

One of the most crucial processes in any organism, including *Saccharomyces cerevisiae*, is that of utilizing the genetic information stored in its DNA through gene expression. Transcription factors are vital to this effort as they assist in the regulation of gene expression by recruiting or deterring RNA polymerase. These transcription factors typically attach to specific transcription factor binding sites (TFBSs). These sites can vary from one class of transcription factors to the next and many transcription factors are known to bind to more than one specific type of site.

Despite the fact that transcription factors tend to target particular sequence regions, many allow a certain degree of variability with regard to the nucleotide sequence within their binding region. Naturally, these variations reflect in the experimentally obtained binding sites and have traditionally been summarized in consensus sequences. Aside from the four characters A, C, G, and T, representing the nucleotides adenine, cytosine, guanine, and thymine, respectively, these consensus sequences can also be written to include less restrictive single letter notations for all possible variations at a given site. These include, for example the letter 'R' which signals that only purines are tolerated while 'N' on the other hand specifies that any nucleotide can be found at this location.

To represent site-specific frequencies of nucleobases in TFBS, consensus sequences can be represented as consensus logos (see Figure 1). These logos are typically made up of stacked letters, representing the possible nucleotides for each position within the binding site. The size of each letter is proportionate to the relative frequency with which it is found at a given site and derived from a position
weight matrix. All frequencies are adjusted to represent the Shannon information content, measured in bits, for each location [1].

**Figure 1:** Consensus logo for the consensus sequence TGTTKACHNW. Information content is measured in bits. The maximum information content for a 4 letter alphabet, as seen above, is $\log_2(4)=2$ bits. [2]

The information content is a measure of how much knowledge can be derived from the collection of symbols at the given location. If all letters are distributed randomly then no conclusion can be drawn about which letter is preferred, i.e. the information content is zero. Conversely, if only one type of letter appears at the current position then one can conclude all binding sites must include this particular nucleotide at the specified location in order for the transcription factor to recognize the site. Thus the information is at its maximum. Any intermediate bit values suggest that there is a bias towards a selection of multiple nucleotides. With the inclusion of information content in the consensus logos these biases in sites of variation become more visually apparent than they would in consensus sequences. However, one of the main drawbacks of both consensus
sequences and logos is that they only capture single nucleotide sequence specific information and completely disregard any nucleobase context as well as any underlying structural information of the DNA macromolecule that may be affected by more than one A, T, C, or G.

This stands in contrast with findings, which demonstrate that the yeast [3-5] as well as the human genome [6, 7] are subject to selective pressures acting on the physical structure of their DNA, particularly minor groove width (MGW). Moreover, Parker et al. [6] showed that these structural signatures in MGW were more helpful in the prediction of functional elements within the DNA than its nucleotide sequence. This supports the notion that the molecular interactions between proteins and DNA are not solely limited to the chemical properties of the four different nucleotides which characterize DNA. Research has shown that nucleosome positioning is highly dependent on structural DNA patterns [8, 9]. Whereas the positioning of nucleosomes in turn can determine how accessible a TFBS is in S. cerevisiae [10-13]. Tirosh et al. [3], for example, hypothesize that the relatively inflexible nature of yeast promoter regions results in a reduced nucleosome occupancy, which may facilitate transcription initiation.

Yet, this is not the only way in which transcription factor - DNA interactions are influenced by their chromatin context. Many transcription factors rely on the shape readout of their TFBS in combination with their base readout, as review by Rohs et al. [14]. While the base readout is commonly dependent on how the transcription factor's amino acids interact with the base-specific hydrogen bonds, which are most often accessible through the major groove, the shape readout is
defined by the binding site’s physical structure. Such shape readout can either be local through the electrostatic potential found in the minor groove and kinks in the DNA helix or global through the DNA curvature and overall helix shape (e.g. A-DNA, B-DNA, Z-DNA).

Yang et al. [15] took advantage of these structural features and developed the database of structural TFBS motifs TFBSshape. These features were defined based on a combination of measures of their minor groove width, roll, and propeller and helix twist. The researchers also included the 2bp long up- and downstream flanking regions surrounding the TFBS, whenever these were available. Gordân et al. [16] showed that even when the transcription factor binding region was removed from its chromatin context the nucleotide regions directly flanking the TFBSs differentially influenced the binding of two different transcription factors (Cbf1 and Tye7) in S. cerevisiae. Both Yang et al. [15] and Gordân et al. [16] were able to improve the precision of computational TFBS prediction based on the structural feature definitions and up- and downstream flanking regions. As discussed by Bansal et al. [17] the most statistically informative structural features identified thus far for TFBS prediction in eukaryotic as well as prokaryotic promoter regions have been superhelical structures induced by DNA destabilization [18], nucleosome positioning [19], stacking energy [20], and DNA rigidity and curvature [21-23].

While some of these features are difficult to quantify given a DNA sequence, Heddi et al. [24] have developed a metric of intrinsic B-DNA flexibility using double-stranded dinucleotides which can be applied to assess DNA rigidity. This metric is referred to as the TRX (twist, roll, and x-displacement) scale and is based on the
ratio of percentages of two possible states, BI and BII, of the B-DNA phosphate backbone which differ between their torsion angles $\varepsilon$ and $\zeta$. Effectively, it quantifies how often a phosphate link connecting a dinucleotide pair switches between the two states, such that frequent transitions have high TRX scores and are considered flexible bonds, while infrequent switches result in low TRX scores, indicating a rigid backbone structure between the two nucleotides (see Table 1).

**Table 1: TRX score distribution**

<table>
<thead>
<tr>
<th>Complimentary Dinucleotides</th>
<th>TRX Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CpG • CpG</td>
<td>43</td>
</tr>
<tr>
<td>CpA • TpG</td>
<td>42</td>
</tr>
<tr>
<td>GpG • CpC</td>
<td>42</td>
</tr>
<tr>
<td>GpC • GpC</td>
<td>25</td>
</tr>
<tr>
<td>GpA • TpC</td>
<td>22</td>
</tr>
<tr>
<td>TpA • TpA</td>
<td>14</td>
</tr>
<tr>
<td>ApG • CpT</td>
<td>9</td>
</tr>
<tr>
<td>ApA • TpT</td>
<td>5</td>
</tr>
<tr>
<td>ApC • GpT</td>
<td>4</td>
</tr>
<tr>
<td>ApT • ApT</td>
<td>0</td>
</tr>
</tbody>
</table>

Adapted from Heddi et al. [11], where the TRX score is the half sum of the BII percentages of the two phosphate links opposite of each other in a complementary dinucleotide pairing. Consequently, maximum flexibility would be represented by a TRX score of 50, while minimum flexibility has a score of 0.

Given that the B-DNA backbone properties are closely linked with major and minor groove dimensions, DNA curvature and winding, all of which are important features for protein-DNA interactions [24], the TRX score lends itself to the extraction of B-DNA structural features affecting transcription factor binding both within the binding site as well as its surrounding up- and downstream flanking regions. Thus this project aims to identify structural signatures within TFBSs and
their flanking regions in the *S. cerevisiae* genome which help distinguish true binding sites from spurious ones and can be characterized with the help of the TRX scale.

This analysis is supported by two different alignment methods: center and ∆TRX based multiple sequence alignment. While most sequence alignment methods thus far have focused solely on the DNA, RNA or amino acid sequence level by only taking into account the composition of the individual entities, few approaches to date rely on other underlying structural features. Similar to the approach taken by Salama and Stekel [25], who used base stacking free energies, we will use change in TRX (∆TRX) measures to guide the alignment of TFBSs. Instead of comparing sequences based on their single nucleotide properties, ∆TRX based multiple sequence alignment aligns the given sequences based on the characteristics of their DNA backbone phosphate links as captured by the TRX score. By aligning the sequences based on their structural properties we hope to better highlight their shared characteristic structural signatures. This method of alignment will then be contrasted to the more commonly used center alignment of TFBSs.

In order to visualize and assess the structural signatures incorporated in TFBSs, the traditional consensus logos will be enhanced such that they include not only sequential but also structural information.
Methods and Materials

In order to identify structural features within and surrounding transcription factor binding sites in *S. cerevisiae*, TFBS consensuses and gene regions were preprocessed before they could be used to find TFBS consensus matches. These matches were subsequently classified as confirmed or non-confirmed transcription factor - gene interactions, before the confirmed sites were isolated, aligned, and comparatively analyzed based on the information captured in their DNA sequence and TRX level. The overall workflow is illustrated in Figure 2 while each intermediate step is described in greater detail in the following sections.

![Diagram](image.png)

**Figure 2:** Work flow of finding and evaluating transcription factor binding site consensus matches based on their nucleotide sequence and structural features as captured by the TRX metric.
Transcription Factor Binding Site Consensus Pre-Processing

The YeastRact database [28] provides consensus sequences for 307 transcription factors, all of which were manually curated from 1,337 manuscripts. All consensus sequences available from YeastRact were experimentally characterized using Chromatin Immunoprecipitation (ChIP), its high throughput applications ChIP-on-chip and ChIP-seq, or Electrophoretic Mobility Shift Assays (EMSAs). Their effect on gene expression in turn was evaluated using northern blotting, quantitative RT-PCR, microarray analysis or expression proteomics [28].

Due to the fact that many of the 307 transcription factors recognize different DNA sequences and their variations, there are a total of 732 consensus sequences. A selection of these sequences allow a variation in length of unspecified nucleotides generally noted in brackets such as N\{(8,35)\}, which means that the stretch of unspecified nucleotides can range from 8 to 35. After expanding this notation into all possible sequence lengths there were 777 TFBS consensus sequences. Each occurrence of the IUPAC ambiguity code 'X', which stands for unknown nucleotide, was replaced with the code for any nucleotide 'N', since they both allow all nucleotides to match. A certain number of consensus sequences also included lowercase nucleotide notations. These lowercase letters are representative of nucleotide biases, such that a lowercase 'a', for example, indicates that while all nucleotides may occur at this position, adenines are favored. Since each nucleotide is still allowed to occur these lower letters were also replaced with 'N' to facilitate string matching. The resulting consensus sequences which were shorter than 5 letters long or those which included less than 5 non-N characters were removed.
from the list of consensuses to facilitate processing as they were expected to result in a magnitude of false positive matches.

**Consensus Sequence Matching**

Although they have been found in trans-regulatory [26] as well as within exonic regions [27], to date TFBSs have most frequently been identified in promoter regions, upstream of those genes, which they help regulate. Thus the promoter regions and open reading frames (ORFs), which were retrieved from Yeastract [28] in separate files, were merged into one file per gene to include any potential exonic binding sites. Since Yeastract's gene and ORF sequences are only given in one orientation the reverse complements for each set of sequences was created. Next the sequences were queried for consensus matches. And identified matches were stored alongside 25 bases of their immediate upstream and 25 bases of their immediate downstream regions. All matches were subsequently flagged as confirmed or non-confirmed TFBSs using the two column regulatory table provided by Yeastract. This table was created based on experimentally defined transcription factor gene interactions and lists which transcription factors are known to act on which genes. Thus confirmed matches are found on those genes, which the transcription factor in question has been shown to target. Non-confirmed matches on the other hand are either spurious sequence matches or true transcription factor - gene interactions, which thus far have not been found experimentally. However, the regulatory table does not specify at which specific location or region these interactions occur.
All confirmed matches with more than three and less than two thousand hits were then further analyzed using center and ΔTRX based multiple sequence alignment. The lower boundary was placed to justify the use of multiple sequence alignment (as opposed to pairwise alignment) and eliminate extremely small sample sizes, as they are not expected to have a lot of statistical power. The upper limit was selected based on the rate at which the multiple sequence alignment algorithm was able to process the given samples.

**Sequence Alignment**

Two methods of sequence alignment were applied to the collection of confirmed consensus matches and their up- and downstream flanking regions. The method of center alignment was also applied to the non-confirmed matches in order to compare their structural signatures to those of the confirmed matches to identify statistically significant differences between the two classes. This was done despite the fact that the non-confirmed samples are likely to contain true matches and are thus expected to differ very little from the set of confirmed matches.

Both center and ΔTRX based sequence alignment methods were used separately to compare the information contained on the sequence level to that on the TRX level. Additionally, both methods were tested against each other to determine which produced more informative results.
Center Alignment

During center alignment all matches for a specific consensus sequence were lined up with each other along the length of their sequences. All corresponding up- and downstream flanking regions fell into place according to how they were connected with their consensus sequence. Thus the alignment was produced from the center outward.

Consensus Match Alignment Based on ΔTRX

For each TFBS consensus the confirmed sequence matches were aligned using a progressive multiple sequence alignment algorithm provided as part of MATLAB’s Bioinformatics Toolbox software package [29]. This algorithm performs all possible pairwise comparisons between the given sequences, which are then used to establish the distances need to create a guide tree for the multiple sequence alignment. Using this guide tree those sequences, which are most similar to each other, will be aligned first according to the neighbor-joining principle thereby forming intermediate alignments. The algorithm then continues to join these alignments until one final multiple sequence alignment is generated.

An advantage of this algorithm provided by MATLAB is that it allows its users to provide their own scoring matrix to be used in order to determine the match and mismatch scores in the initial pairwise comparison. To align sequences based on the property of their intrinsic, structural flexibility a custom scoring matrix was created based on the absolute value of the difference in TRX (|ΔTRX|) found between opposing dinucleotides (d₁ and d₂) from each of the two sequences involved in the
pairwise comparison. The change in TRX is used as a score \( s \) normalized to range between -1 and 1 according to the following equation such that the mismatch penalty score is intrinsic to the scoring matrix:

\[
s_{d1,d2} = \left| \Delta TRX_{d1,d2} \right| \times \frac{-1}{21.5} + 1
\]  

(1.1)

where 21.5 lies halfway between the minimum and maximum recorded TRX score. As a result base changes, which do not lead to the same or similar measure of flexibility are rewarded, while those causing great differences in flexibility are penalized. Intrinsic to the algorithm, gaps which are newly introduced to the alignment are penalized with a score of five times that of the average match score, while the extension of existing gaps is penalized with a fourth of the average match score. Terminal gaps are not penalized.

Since this algorithm was created to perform multiple sequence alignment between sequences of individual amino acids or nucleotides, all dinucleotides were translated into a single letter alternative code, to ensure that the alignment was performed on the phosphate linkage between nucleotides rather than the nucleotides themselves. Once the multiple sequence alignment was completed the alternative code was translated back into a nucleotide sequence to allow further analysis and comparison.

**Information Content Comparison**

In order to test whether or not confirmed consensus match sites and their flanking regions contain more information when captured at the TRX level
compared to their DNA sequence equivalent, the information content (IC) measured in bits for both was calculated at each position $i$ within the matched consensuses and their up- and downstream flanking regions according to the following equation:

$$IC_i = \log_2 k - H_i$$

(2.1)

where $k$ is the number of total possible symbols for each representation. Consequently, when calculating IC for the nucleotide sequence $k=4$ in order to account for the four possible nucleotides, which can potentially be found at each location. Since there are nine distinct scores among the ten TRX scores seen in Table 1, $k=9$ when analyzing each matched region by its flexibility. $H_i$ in turn is defined as:

$$H_i = -\sum_{s=1}^{k} p_{s,i} \times \log_2 (p_{s,i})$$

(2.2)

where $p_{s,i}$ is the relative frequency of the symbol $s$ appearing among all samples at the given location $i$. As an example, if the nucleotide symbol A were to occur ten times at position $i$ among fifty samples then $p_{A,i} = 10/50 = 0.2$.

Due to the fact that the single nucleotide representation has a smaller bit range than the dinucleotide notation (maximum information for each is $\log_2(4)=2.0$ bits and $\log_2(9)=3.2$ bits, respectively), all IC scores were normalized to percent information content (%IC). Whenever a gap was found at a specific location in an alignment the %IC was calculated as %IC $\times$ (1 - frequency of gaps at location $i$). As a result if there were only gaps present at location $i$ then the IC would be zero. Conversely, if no gaps were present, then this adjustment has no effect on the IC.
In the case of center alignment such gaps could occur due to the consensus match being located close to either end of the gene sequence such that only less than 25 bases of the flanking region could be stored. In ∆TRX based multiple sequence based alignment these gaps were introduced to improve the alignment of neighboring patterns. The correction of %IC was chosen so that locations with many gaps didn’t have too much weight when gaps were ignored. For example, a location which may have four gaps and only one nucleotide would have 100% IC at that location, despite the fact that there is only one nucleotide. By correcting the %IC based on the frequency of gaps these locations are then reduced in meaning since they don’t have as many samples to support the signature.

The two notations, i.e. sequence and TRX based, were subsequently compared according to their difference in average %IC for both consensus match regions and their up- and downstream flanking sites according to Equation 2.3:

$$d = \frac{%IC_{TRX} - %IC_{DNAsequence}}{}$$

(2.3)

Negative $d$ values indicate that there is more information held in the DNA sequence than in the TRX notation. Conversely, positive $d$ values suggest more information stored in the TRX representation than in the nucleotide sequence for the given TFBS consensus. A Student’s t test was performed for each comparison in order to test the statistical significance of the calculated difference using a Bonferroni multiple test correction of $0.05/416=0.00012$.

A similar analysis was conducted in order to compare the TRX-based center and multiple sequence alignments. In each comparison consensus and flanking
regions are compared separately to each other. The maximum span of the consensus match found in the multiple sequence alignment was used to compare consensus matches, due to the fact that this form of alignment may result in the introduction of gaps within the consensus match region. For example, if a consensus match region were to be extracted from an alignment then the consensus would start at the earliest occurrence of the first nucleotide matched to the original consensus sequence among all matched sequences, reading from up- to downstream. This matched region would expand to the last occurrence of the last nucleotide matched to the original consensus sequence among all matched sequence, again reading from up- to downstream. The flanking regions are then defined as those regions up- and downstream of this maximum consensus span.

**Enhanced Consensus Logos**

The enhanced consensus logos were created with the help of code developed by Connor Fortin, which calculates the Shannon information content for the different nucleotides and phosphate linkages, where the latter is normalized to match the nucleotide scale of information content (0 to $\log_2 4$). This resulted in one merged logo, where bars signifying the information carried on the phosphate links are interspersed between the traditional consensus logo. Each of these bars is shaded in gray, the hue of which becomes lighter with increasing flexibility.
Results

After data preprocessing 6436 gene sequences and their reverse complements consisting of merged promoter and ORF regions were queried using 696 consensus sequences from 180 transcription factors. This resulted in matches for 663 consensus sequences, 416 of which had between 3 and 2000 hits per consensus. The matched sequences and their up- and downstream flanking regions of these 416 consensus sequences from 163 transcription factors were subsequently used for both center and ∆TRX based multiple sequence alignments.

The final normalized scoring matrix used for the ∆TRX alignment can be seen in Table 2.

**Table 2:** Normalized scoring matrix for ∆TRX based multiple sequence alignment.

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AC</th>
<th>AG</th>
<th>AT</th>
<th>CA</th>
<th>CC</th>
<th>CG</th>
<th>CT</th>
<th>GA</th>
<th>GC</th>
<th>GG</th>
<th>GT</th>
<th>TA</th>
<th>TC</th>
<th>TG</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.7</td>
<td>-0.7</td>
<td>-0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>-0.2</td>
<td>-0.7</td>
<td>1.0</td>
<td>0.6</td>
<td>0.2</td>
<td>-0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>AC</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.8</td>
<td>1.0</td>
<td>0.5</td>
<td>0.2</td>
<td>-0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>AG</td>
<td>0.8</td>
<td>0.8</td>
<td>1.0</td>
<td>0.6</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.6</td>
<td>1.0</td>
<td>0.4</td>
<td>0.3</td>
<td>-0.5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.4</td>
<td>-0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>AT</td>
<td>0.8</td>
<td>0.8</td>
<td>0.6</td>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>0.6</td>
<td>0.0</td>
<td>-0.2</td>
<td>-1.0</td>
<td>0.8</td>
<td>0.3</td>
<td>0.0</td>
<td>-1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>CA</td>
<td>-0.7</td>
<td>-0.8</td>
<td>-0.5</td>
<td>-1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>-0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>1.0</td>
<td>-0.8</td>
<td>-0.3</td>
<td>0.1</td>
<td>1.0</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>-0.7</td>
<td>-0.8</td>
<td>-0.5</td>
<td>-1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>-0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>1.0</td>
<td>-0.8</td>
<td>-0.3</td>
<td>0.1</td>
<td>1.0</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>CG</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-0.6</td>
<td>-1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>-0.6</td>
<td>0.0</td>
<td>0.2</td>
<td>1.0</td>
<td>-0.8</td>
<td>-0.3</td>
<td>0.0</td>
<td>1.0</td>
<td>-0.8</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>0.8</td>
<td>0.8</td>
<td>1.0</td>
<td>0.6</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.6</td>
<td>1.0</td>
<td>0.4</td>
<td>0.3</td>
<td>-0.5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.4</td>
<td>-0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>GA</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.4</td>
<td>1.0</td>
<td>0.9</td>
<td>0.1</td>
<td>0.2</td>
<td>0.6</td>
<td>1.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>GC</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.3</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.9</td>
<td>1.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.9</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>GG</td>
<td>-0.7</td>
<td>-0.8</td>
<td>-0.5</td>
<td>-1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>-0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>1.0</td>
<td>-0.8</td>
<td>-0.3</td>
<td>0.1</td>
<td>1.0</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>GT</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.8</td>
<td>1.0</td>
<td>0.5</td>
<td>0.2</td>
<td>-0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>TA</td>
<td>0.6</td>
<td>0.5</td>
<td>0.8</td>
<td>0.3</td>
<td>-0.3</td>
<td>-0.3</td>
<td>-0.3</td>
<td>0.8</td>
<td>0.6</td>
<td>0.5</td>
<td>-0.3</td>
<td>0.5</td>
<td>1.0</td>
<td>0.6</td>
<td>-0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>TC</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.4</td>
<td>1.0</td>
<td>0.9</td>
<td>0.1</td>
<td>0.2</td>
<td>0.6</td>
<td>1.0</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>TG</td>
<td>-0.7</td>
<td>-0.8</td>
<td>-0.5</td>
<td>-1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>-0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>1.0</td>
<td>-0.8</td>
<td>-0.3</td>
<td>0.1</td>
<td>1.0</td>
<td>-0.7</td>
<td></td>
</tr>
<tr>
<td>TT</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
<td>-0.7</td>
<td>-0.7</td>
<td>-0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>-0.7</td>
<td>1.0</td>
<td>0.6</td>
<td>0.2</td>
<td>-0.7</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The average %IC for both alignment methods was compared for each set of consensus sequence matches in order to assess which method was superior for detecting structural features captured in the TRX metric. The summary statistics for
this comparison can be found in Table 3. Since the TRX measure is based on the phosphate link between two nucleotides the bonds connecting the consensus match to its flanking regions were either counted towards the consensus or the flanking sequence range. Approximately one-fourth of all aligned sequences showed significant differences between the alignment methods. All significantly different alignments of consensus match regions proved to be represented better using center alignment. However, all consensuses with significant differences in their flanking regions held more information in the $\Delta$TRX based than the center alignment.

**Table 3: Summary of alignment method comparison between $\Delta$TRX based multiple sequence alignment and center alignment.**

<table>
<thead>
<tr>
<th>Sequence Range</th>
<th>Number consensuses with significant difference between methods ($p &lt; 0.00012$)</th>
<th>Number of consensuses with significantly more information in the $\Delta$TRX based multiple sequence alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus match sequence (including bridge bonds to flanking regions)</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Consensus match sequence (excluding bridge bonds to flanking regions)</td>
<td>102</td>
<td>0</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (excluding bridge bonds to flanking regions)</td>
<td>106</td>
<td>106</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (including bridge bonds to flanking regions)</td>
<td>103</td>
<td>103</td>
</tr>
</tbody>
</table>

A Student's t test was used to calculate whether there was a statistically significant difference in %IC in the consensus matched sequences and their flanking regions. Using Bonferroni multiple testing correction $p$-values below 0.00012 were found to be significant. All numbers are out of a maximum of 416.

Table 4 outlines the results of the comparison between the information content contained in the DNA sequence and its phosphate bonds, the latter of which
were represented using the corresponding TRX scores. All mean differences in %IC favored more IC in the TRX notation.

**Table 4:** Summary of comparison between information held in sequence to that held in TRX notation using ∆TRX based multiple sequence alignment.

<table>
<thead>
<tr>
<th>Sequence Range</th>
<th>Number consensuses with significant difference between methods (p &lt; 0.00012)</th>
<th>Number of consensuses with significantly more information in the TRX notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus match sequence (including bridge bonds to flanking regions)</td>
<td>146</td>
<td>146</td>
</tr>
<tr>
<td>Consensus match sequence (excluding bridge bonds to flanking regions)</td>
<td>142</td>
<td>142</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (excluding bridge bonds to flanking regions)</td>
<td>349</td>
<td>349</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (including bridge bonds to flanking regions)</td>
<td>358</td>
<td>358</td>
</tr>
</tbody>
</table>

A Student’s t test was used to calculate whether there was a statistically significant difference in %IC in the consensus matched sequences and their flanking regions. Using Bonferroni multiple testing correction p-values below 0.00012 were found to be significant. All numbers are out of a maximum of 416.

The results for a similar comparison of information content between nucleotide and TRX sequence using center alignment instead of ∆TRX based multiple sequence alignment can be found in Table 5. None of the consensus regions showed a significant difference in information content between the two sequence notations, while approximately 96% of up- and downstream flanking regions were shown to have more information in their TRX than in their nucleotide sequence.
Table 5: Summary of comparison between information held in sequence to that held in TRX notation using center alignment.

<table>
<thead>
<tr>
<th>Sequence Range</th>
<th>Number consensuses with significant difference between methods (p &lt; 0.00012)</th>
<th>Number of consensuses with significantly more information in the TRX notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus match sequence (including bridge bonds to flanking regions)</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Consensus match sequence (excluding bridge bonds to flanking regions)</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (excluding bridge bonds to flanking regions)</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (including bridge bonds to flanking regions)</td>
<td>399</td>
<td>399</td>
</tr>
</tbody>
</table>

A Student's t test was used to calculate whether there was a statistically significant difference in %IC in the consensus matched sequences and their flanking regions. Using Bonferroni multiple testing correction p-values below 0.00012 were found to be significant. All numbers are out of a maximum of 416.

The confirmed consensus matches were subsequently compared to the non-confirmed matches from the equivalent TFBS consensus based on the average %IC in their TRX scores. The summary of this comparison is shown in Table 6. Once again, most significant differences were observed in the flanking regions (151 excluding and 107 including bridging phosphate links). 140 and 95 of these regions, respectively, also contained more information in their TRX notation as compared to their nucleotide sequence.
Table 6: Summary of comparison of information held in TRX notation between confirmed and non-confirmed consensus matches using center alignment.

<table>
<thead>
<tr>
<th>Sequence Range</th>
<th>Number of consensuses with significant difference between methods (p &lt; 0.00012)</th>
<th>Number of consensuses with significantly more information in the confirmed consensus</th>
<th>Number of consensuses with significantly more information in the confirmed consensus AND more information in the TRX sequence compared to the nucleotide sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus match sequence (including bridge bonds to flanking regions)</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Consensus match sequence (excluding bridge bonds to flanking regions)</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (excluding bridge bonds to flanking regions)</td>
<td>162</td>
<td>151</td>
<td>140</td>
</tr>
<tr>
<td>Up- and downstream flanking sequences (including bridge bonds to flanking regions)</td>
<td>117</td>
<td>107</td>
<td>95</td>
</tr>
</tbody>
</table>

A Student’s t test was used to calculate whether there was a statistically significant difference in %IC in the consensus matched sequences and their flanking regions. Using Bonferroni multiple testing correction p-values below 0.00012 were found to be significant. All numbers are out of a maximum of 416.

Enhanced sequence logos of the consensus match (Figure 4) and up- and downstream flanking regions (Figures 3 and 5) were created for one of the Gat3’s TFBS consensus NNNNBRGATCTACNNNNNNN, which showed significant differences in the confirmed and non-confirmed matches as well as between the TRX and nucleotide notation.
Figure 3: Enhanced sequence logo of Gat3’s upstream flanking region for the consensus NNNNBGRGATCTACNNNNNN. Maximum information is 2 bits. Bars between letters show the information content of the connecting phosphate links while their level of gray shading indicates the bond’s flexibility ranging from high to low as light to dark.

Figure 4: Enhanced sequence logo of Gat3’s consensus match region for the consensus NNNNBGRGATCTACNNNNNN. Maximum information is 2 bits. Bars between letters show the information content of the connecting phosphate links while their level of gray shading indicates the bond’s flexibility ranging from high to low as light to dark.
Figure 5: Enhanced sequence logo of Gat3’s downstream flanking region for the consensus NNNNBRGATCTACNNNNNN. Maximum information is 2 bits. Bars between letters show the information content of the connecting phosphate links while their level of gray shading indicates the bond's flexibility ranging from high to low as light to dark.

Discussion

For both methods of sequence alignment it was determined that more structural information encoded in the TRX scale can be found in the flanking regions of TFBSs as compared to the information derived from the nucleotide sequence alone (see Tables 4 and 5). In the case of center alignment there were no significant differences between the two carriers of information. This suggests that in most cases they are either equally important or closely interdependent.

Babbitt [30] describes a relaxed selection on TFBSs with conserved flanking regions, which agrees nicely with the findings presented here. The increased levels
of information in the TRX notation of the flanking regions is likely used to distinguish between which transcription factor if any is attracted to the general TFBS region as also shown by Gordân et al. [16] and Rohs et al. [31]. This would allow the binding sites themselves to be made up of somewhat precise and balanced (with regard to sequence and local structure) yet relatively small DNA sequences, which are likely to be encountered fairly frequently.

Approximately 35% of consensus sequence matches contained more information in their structural notation as compared to their nucleotide sequence using ΔTRX based multiple sequence alignment (Table 4). While these signatures could be important for the attraction and binding of certain classes of transcription factors it is unclear to what extent these results may be an artifact of the maximum span definition of the consensus region applied to those sequences aligned based on ΔTRX. For example, in certain cases, in which a consensus match may have only a short or non-existent flanking region due to the fact that it was located on the perimeters of the available gene sequence, a particular sequence match might become misaligned with the remaining set of sequences such that it is very far displaced from all other consensus match start sites. This in turn could result in an exaggerated maximum consensus span, which ultimately captures more information that should be attributed to the flanking as opposed to the consensus match region, thereby potentially giving a false positive result for the comparison of information contained on the TRX and nucleotide sequence level in the consensus region. A future analysis of TFBS flanking and consensus match regions aligned separately from each other may give more insight into how much of the findings presented
here are a result of such an exaggerated maximum span. Additionally, an isolated alignment of flanking regions may further improve the identification of structural features in these regions when compared to a nucleotide based multiple sequence alignment of their DNA sequence.

No improvement was made with regard to the information content retrieved by the ΔTRX based multiple sequence alignment method in consensus match regions in comparison to the center alignment. As a matter of fact, for those consensuses, which did have a significant difference between the two groups, center alignment proved to be superior to ΔTRX based alignment (Table 2). However, for approximately one fourth of the consensuses, ΔTRX based multiple sequence alignment was able to capture significantly more information in the flanking regions than center alignment.

A closer look at the topography of a selection of TFBS consensus matches with significantly more information in the flanking regions shows that these DNA regions can have varying structural patterns. Figure 6 shows the signature of TRX scores found in the 620 sequences matched to the transcription factor Fkh1’s consensus sequence RYAAACAWW. Almost all phosphate links found in the up- and downstream flanking regions have a mean TRX score of approximately 9, indicating that these regions tend to be rather rigid and are likely AT-rich. These findings pair well with the Tirosh et al.’s [3] observations of inflexible promoter regions, which allow transcription factor binding due to low nucleosome occupancy. Additionally, these rigid regions also likely result in a narrowing of the minor groove width and thereby cause an increase in the DNA molecule’s negative electrostatic potential,
which as shown by Rohs et al. [31] may be used to for the initial attraction of the transcription factor to the general binding region only to then allow it to hone in and bind to its specific target site.

![Figure 6: TRX score analysis in consensus match and up- and downstream flanking region of the transcription factor Fkh1’s consensus site RYAAACAWW. All bars colored in light blue are flanking regions while dark blue marks the consensus match region.](image)

However, not all flanking regions show such orderly patterns. Gat3’s consensus region NNNNBRGATCTACNNNNNN, for example, appears to vary greatly from one location in the flanking region to the next, switching between regions of very low flexibility to those of very high flexibility within one or two nucleotide steps (see Figure 7). The exact purpose of these patterns is unclear at this time.

![Figure 7: TRX score analysis in consensus match and up- and downstream flanking region of the transcription factor Gat3’s consensus site NNNNBRGATCTACNNNNNN. All bars colored in light blue are flanking regions while dark blue marks the consensus match region.](image)

Other transcription factors appear to prefer more periodic differences in their consensus as well as their flanking regions, such as Put3 (Figure 8) and Ppr1
(Figure 9). Both transcription factors exhibit a preference for palindromic transcription factor binding sites. An illustration by Bansal et al. [17] seen in Figure 10 shows that such palindromes can form a stem loop based cruciform structure, which may in turn account for the need of a certain symmetric and periodic structural pattern.

**Figure 8:** TRX score analysis in consensus match and up- and downstream flanking region of the transcription factor Put3’s consensus site CGGNNNNNNNNNCCG. All bars colored in light blue are flanking regions while dark blue marks the consensus match region.

**Figure 9:** TRX score analysis in consensus match and up- and downstream flanking region of the transcription factor Prp1’s consensus site CGGNNNNNNNCCG. All bars colored in light blue are flanking regions while dark blue marks the consensus match region.
The enhanced sequence logos (Figures 3-5) provide a way in which the phosphate bond preferences of these structures can be visualized. In the future it would be very helpful to test 1) how far the structural signatures corresponding to the binding sites extend up- and downstream of the consensus match site and 2) which of these signatures are distinctly associated with confirmed but not non-confirmed sites. Once this was defined the enhanced logos could be translated into enhanced consensus sequences, which include information on the preferred TRX values at certain locations, and could be used to improve the precision of computational TFBSs queries and predictions. Such a notation could, for example, look similar to the following: A[4-9]N[22-25]GCT, where the values in the brackets capture the allowed TRX score range.

**Conclusion**

This study demonstrated that, while there are generally few to no significant differences in information content contained in the consensus match regions between their TRX and nucleotide notations, at least 95% of all up- and downstream
flanking regions contained more information in their structural signature as defined by the TRX scale. Between 23% and 35% (excluding and including bridging phosphate bonds, respectively) of these also showed significant differences between the sets of confirmed and non-confirmed matches. These findings support the notion that structural context is highly important in the distinction between true and false binding sites. While ΔTRX based multiple sequence alignment appeared to be superior in flanking regions when compared to center alignment, further analyses are needed in order to increase the confidence in these findings. Enhanced consensus logos represent an excellent visualization of these structural signatures in combination with their nucleotide sequence and provide a stepping-stone to create more defined structural motifs to improve TFBS queries and predictions.
References


[27] Stergachis, Andrew B., Eric Haugen, Anthony Shafer, Wenqing Fu, Benjamin Vernot, Alex Reynolds, Anthony Raubitschek et al. "Exonic Transcription Factor


