Sequence analysis

Position dependencies in transcription factor binding sites

Andrija Tomovic and Edward J. Oakeley*

Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Maulbeerstrasse 66, CH-4058 Basel, Switzerland

Received on October 29, 2006; revised on January 17, 2007; accepted on February 9, 2007

Advance Access publication February 18, 2007
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Most of the available tools for transcription factor binding site prediction are based on methods which assume no sequence dependence between the binding site base positions. Our primary objective was to investigate the statistical basis for either a claim of dependence or independence, to determine whether such a claim is generally true, and to use the resulting data to develop improved scoring functions for binding-site prediction.

Results: Using three statistical tests, we analyzed the number of binding sites showing dependent positions. We analyzed transcription factor–DNA crystal structures for evidence of position dependence. Our final conclusions were that some factors show evidence of position dependencies. It does not require complex possibilities within our model. Our method for modeling position dependencies. The related methods include Bayesian networks (Barash, 2003), permuted Markov models (Zhao et al., 2002b), matrix profiles (Stormo et al., 1982) and sequence logos (Schneider and Stephens, 1990); for a review see (Wasserman and Sandelin, 2004). Methods based on the assumption of independence between positions are simple with small numbers of parameters, making them easy to implement. These methods are widely used and often considered as acceptable models for binding-site predictions (Bens et al., 2002b). However, recent experimental evidence (Bens et al., 2002b; Bulyk et al., 2002; Man and Stormo, 2001; Udalova et al., 2002; Wolfe et al., 1999) has prompted the development of models which incorporate position dependencies. The related methods include Bayesian networks (Barash, 2003), permuted Markov models (Zhao et al., 2005), Markov chain optimization (Ellrott et al., 2002), hidden Markov models (Marinescu et al., 2005), non-parametric models (King and Roth, 2003), and generalized weight matrix models (Zhou and Liu, 2004). Methods based on position-dependency models usually have better binding site prediction accuracy with lower false positive rates. But these methods require more complicated mathematical tools, with more parameters to estimate, and require more experimental data than are typically available (Barash, 2003; Ellrott et al., 2002; King and Roth, 2003; Marinescu et al., 2005; Zhao et al., 2005; Zhou and Liu, 2004). The purpose of this work is to investigate whether or not TFs show position dependencies in their binding sites. We suggest a rigorous statistical approach for testing dependencies. Our findings indicate that there is no universal answer. Some factors seem to show dependencies whereas others do not. We, therefore, decided to allow both possibilities within our model. Our method for modeling dependencies is simply an extension of methods which assume position independencies. It does not require complex methods. Scanning methods infer binding sites from known, experimentally verified binding sequences. Example tools include ConSite (Sandelin et al., 2004), Match (Kel et al., 2003), Mapper (Marinescu et al., 2005), Patser (Hertz et al., 1990), and rVista (Loots and Ovcharenko, 2004; Loots et al., 2002). Ab initio approaches infer specificities without any prior knowledge of binding sites, based on sequence homology. Example tools include Gibbs sampler (Lawrence et al., 1993), MEME (Bailey and Elkan, 1994), Bioprospector (Liu et al., 2001), Yeast motif finder (Sinha and Tompa, 2003) and ANN-Spec (Workman and Stormo, 2000).

Availability: http://promoterplot.fmi.ch/cgi-bin/dep.html
Contact: edward.oakeley@fmi.ch
Supplementary information: Supplementary data (1, 2, 3, 4, 5, 6, 7 and 8) are available at Bioinformatics online.

1 INTRODUCTION

The transcription of genes is controlled by transcription factor proteins (TFs) which bind to short DNA sequences known as transcription factor binding sites (also known as DNA-binding motifs or cis-regulatory sequences). TF-binding sites are usually very short and highly degenerate, and such short sequences are expected to occur at random every few hundred base pairs. This makes their prediction extremely difficult. An important task in the computational prediction of TF-binding sites is reducing the false positive rate while still retaining a high sensitivity. Currently, predictions rely on either scanning or ab initio methods. Scanning methods infer binding sites from known, experimentally verified binding sequences. Example tools include ConSite (Sandelin et al., 2004), Match (Kel et al., 2003), Mapper (Marinescu et al., 2005), Patser (Hertz et al., 1990), and rVista (Loots and Ovcharenko, 2004; Loots et al., 2002). Ab initio approaches infer specificities without any prior knowledge of binding sites, based on sequence homology. Example tools include Gibbs sampler (Lawrence et al., 1993), MEME (Bailey and Elkan, 1994), Bioprospector (Liu et al., 2001), Yeast motif finder (Sinha and Tompa, 2003) and ANN-Spec (Workman and Stormo, 2000).

*To whom correspondence should be addressed.
2 METHODS

2.1 Testing dependencies

In this section, we describe methods to test dependencies in binding sites.

Let us suppose that we have n binding sites of length k for a given TF:

\[ b_1^i \ b_2^i \ \ldots \ b_k^i \]  

where, \( b_i \in \{a, c, g, t\} \), and 1 \( \leq i \leq k \), and 1 \( \leq j \leq n \). We introduce the notation: \( B_i \) and \( B_j \) to represent random variables which can take values from the set \( \{a, c, g, t\} \). indices \( i \) and \( j \) represent positions in the binding sites and 1 \( \leq i, j \leq k \) and \( i \neq j \),

\[ B_i \left( \frac{a}{P(a,i)}, \frac{c}{P(c,i)}, \frac{g}{P(g,i)}, \frac{t}{P(t,i)} \right) \]  

and likewise for \( B_j \).

Let \( N(i) \) be a vector of the frequencies \( N(i)=[N(a,i), N(c,i), N(g,i), N(t,i)] \) where, \( N(a,i) \) is the frequency of base \( a \) at position \( i \) and so on. Similarly, for column \( j \) we introduce a frequency vector \( N(j) \). Using a maximum likelihood approach and the method of Lagrange multipliers, we can estimate probabilities:

\[ P(b_i, i) = \frac{N(b_i, i)}{n} \quad P(b_j, j) = \frac{N(b_j, j)}{n} \]  

where \( b \) is one of the bases \( \{a, c, g, t\} \).

First, we can calculate mutual information (Chiu and Kolodziejczak, 1991), a quantitative measure of pairwise sequence covariation. The mutual information between positions \( i \) and \( j \) is given by:

\[ M_{ij} = \sum_{b_i b_j} P(b_i, b_j, i, j) \log \frac{P(b_i, b_j, i, j)}{P(b_i, i)P(b_j, j)} \]  

where, the probability \( P(b_i, b_j, i, j) \) can be estimated using the maximum-likelihood method and the method of Lagrange multipliers:

\[ P(b_i, b_j, i, j) = \frac{N(b_i, b_j, i, j)}{n} \]  

where, \( N(b_i, b_j, i, j) \) is the frequency of base pairs \( b_i b_j \) at positions \( i \) and \( j \). This is a descriptive measure of divergence from independence of \( i \) and \( j \). \( M_{ij} \) varies between 0 and 2 bits. It is maximal when \( i \) and \( j \) are perfectly correlated. If \( i \) and \( j \) are uncorrelated, the mutual information is zero. Very often we do not have extreme values of \( M_{ij} \) and we cannot deduce if \( i \) and \( j \) are independent using only the value of \( M_{ij} \). In order to identify positions that may not be highly correlated as measured by \( M_{ij} \), but are as correlated as they can be given the limited variability of the individual positions, we can calculate two other values (Gutell et al., 1992):

\[ R_1(i, j) = \frac{M_{ij}}{H_i} \quad R_2(i, j) = \frac{M_{ij}}{H_j} \]  

where \( H_i \) and \( H_j \) are entropies for positions \( i \) and \( j \), respectively.

\[ H_i = -\sum_{b_i} P(b_i, i) \log_2 P(b_i, i) \quad H_j = -\sum_{b_j} P(b_j, j) \log_2 P(b_j, j) \]  

Both R values vary between 0 and 1 and, in general, they are not equal. Therefore, if we use only \( M_{ij} \) we may miss some correlated positions, but some of these may be detected using R-values. However, it should be emphasized that we cannot easily estimate the significance of R-values. So, we will have false positives as well as true correlated positions. R-values are also descriptive measures of dependencies between two positions. A more formal way to test dependencies is hypothesis testing:

\[ H_0: \text{positions } i \text{ and } j \text{ are independent} \]  

\[ H_1: \text{otherwise.} \]  

To test this hypothesis, we can use a \( \chi^2 \)-test of independence (Ellrott et al., 2002) on each pair of positions \( i \) and \( j \):

\[ \chi^2 = 2 \sum_{b_i b_j} \frac{(P(b_i, b_j, i, j) - P(b_i, i)P(b_j, j))^2}{P(b_i, i)P(b_j, j)} \]  

The distribution of \( \chi^2 \) statistics is close to a \( \chi^2 \) distribution with \( (|b_i| - 1) \times (|b_j| - 1) \) degrees of freedom, where \( |b_i| \) is the number of bases for which \( P(b_i, i) \) is not zero, and likewise for \( |b_j| \). So, using \( \chi^2 \) statistics and \( \chi^2 \) distributions we can test the hypothesis at a given significance level e.g. 0.05. This hypothesis may also be tested using a G-test of independence (log-likelihood ratio test) (Sokal and Rohlf, 2003). For each pair of positions \( i \) and \( j \), we can calculate G statistics:

\[ G = 2 \sum_{b_i b_j} P(b_i, b_j, i, j) \log_2 \frac{P(b_i, b_j, i, j)}{P(b_i, i)P(b_j, j)} \]  

where, \( a=(|b_i| - 1) \times (|b_j| - 1) \), and \( b=a-1 \) as this provides a better approximation to the \( \chi^2 \) distribution. Conahan found that if expected frequencies are higher than 10, G statistics approximate well to the exact multinomial probability distribution (Conahan, 1970). She found that G statistics were adequate and better than \( \chi^2 \) statistics, where there are more than five degrees of freedom and expected frequencies greater than or equal to 3. In all other cases she recommends the exact test. Larntz, in his comparison of G and \( \chi^2 \) statistics, did not consider the corrections of G statistics when drawing his conclusion that \( \chi^2 \) statistics fits the theoretical chi-squared distribution better than G statistics do (Larntz, 1978). Sokal et al. (Sokal and Rohlf, 2003) showed that G statistics with William’s correction approximates to the \( \chi^2 \) distribution more closely than they do without the correction. It is very difficult to find a single rule to cover all cases when the observed distributions of G statistics and \( \chi^2 \) statistics are close to real \( \chi^2 \) distributions, if we have small expected
There is always a danger of type I errors (rejecting the null hypothesis when in fact it is true) when applying multiple tests to data. These may be minimized with Bonferroni’s correction or its extensions/variants (e.g. Dunn–Šidák, Holm’s, Simes–Hochberg or Hommel’s method). The Bonferroni adjustment of $P$-value ($0.05/k$, where $k$ is the number of tests) is very stringent and can introduce type II errors, which are also important. The use of Bonferroni is much debated (Perneger, 1998).

As a compromise, in the case of the Bayesian test, we propose that a more stringent BF factor $BF(H_{i}, H_{j}) < 0.1$ could be used to report stronger evidence against the null hypothesis.

### 2.2 New scoring function

Any existing scoring function which works with models that assume independence between positions within binding sites, can easily be modified to incorporate dependencies. These new functions do not have dramatically more parameters, and do not require additional data or complex mathematical approaches.

If we have $n$ binding sites of length $k$ for a given TF and sequence $l$ with length $k$, then to determine if a putative-binding site is for a given TF we will follow the notation of (Wasserman and Sandelin, 2004) where, $w_{b,i}$ is a position weight matrix (PWM) value of base $b$ in position $i$, calculated by:

$$W_{b,i} = \log_{2} \frac{P(b, i)}{P(b)}$$

where $P(b)$ is the background probability of base $b$, $P(b,i)$ is a corrected probability of base $b$ at position $i$, and is calculated by:

$$P(b,i) = \frac{N(b,i)}{n} + a(b)$$

where $a(b)$ is smoothing parameter ($a(b) = 0.01$). The fit of any given DNA sequence can be quantitatively scored by summing all the values of $W_{b,i}$ for every base in the sequence (hereafter, we will refer to this ‘old’ scoring function as $S_{old}$):

$$S_{old} = \sum_{i=1}^{k} W_{b,i}$$

For a large set of well-characterized binding sites, these scores are proportional to the factor-binding energies (Stormo, 2000).

To incorporate position dependencies, we will extend this function and this model for the representation of the TF-binding sites in the following way.

First, we will introduce a corrected probability for the bases $b_{1}, \ldots, b_{m}$ in $i_{1}, \ldots, i_{m}$ dependent positions.

$$P(b_{1}, \ldots, b_{m}, i_{1}, \ldots, i_{m}) = \frac{N(b_{1}, \ldots, b_{m}, i_{1}, \ldots, i_{m})}{n} + a(b_{1}, \ldots, b_{m})$$

$a(b_{1}, \ldots, b_{m})$ is a smoothing parameter and can be calculated by:

$$a(b_{1}, b_{2}, \ldots, b_{m}) = a(b_{1}) \cdots a(b_{m})$$

Then we can calculate values which correspond to PWM values:

$$W_{b_{1}, \ldots, b_{m}, i_{1}, \ldots, i_{m}} = \log_{2} \frac{P(b_{1}, \ldots, b_{m}, i_{1}, \ldots, i_{m})}{P(b_{1}) \cdots P(b_{m})}$$

Finally, the new scoring function ($S_{new}$), which incorporates dependencies, can be expressed thus:

$$S_{new} = \sum_{i=1}^{k} W_{b,i} + \sum_{i_{1}=1}^{k_{1}} \sum_{i_{2}=1}^{k_{2}} \cdots \sum_{i_{m}=1}^{k_{m}} W_{b_{1}, \ldots, b_{m}, i_{1}, \ldots, i_{m}} +$$

$$+ \sum_{i=1}^{k_{1}} W_{b_{1}, \ldots, b_{k_{1}}, i_{1}, \ldots, i_{m}}$$

where, $k_{1}$ is the number of independent positions, $k_{2}$ is the number of dependent positions order 2, $k_{3}$ the number of dependent positions order 3 ($k_{m}$ nucleotides at positions $j_{1}$ and $j_{1+k_{1}}$ and $k_{m}$ the number of dependent positions order $m$ (nucleotides at positions $j_{1}, j_{1+k_{1}}, \ldots, j_{1+k_{m}-1}$). Higher-order dependencies can be constructed from
the second-order dependencies in the following way: if we analyze three positions \(i_1, i_2\) and \(i_3\) and if every two combinations \((i_1 - i_2, i_1 - i_3\) and \(i_2 - i_3)\) are dependent, then we can claim that positions \(i_1, i_2\) and \(i_3\) show third-order dependencies. This approach may be extended to find \(m\)th-order dependencies between \(k_m\) positions of a binding site. For the new scoring function (22), higher order dependencies can be constructed in a less stringent way: if we find when analyzing three positions \(i_1, i_2\) and \(i_3\) that only two combinations \((i_1 - i_2, i_2 - i_3\) or \(i_1 - i_3)\) are dependent, we can say that there are third order dependencies among positions \(i_1, i_2\) and \(i_3\). This will not have any influence on the final results (because of equation (12)) and the logarithm property \((\log(P(B_i, B_j))\) will tend towards \((\log(P(B_i)) + \log(P(B_j)))\). Small differences may be observed because of the smoothing parameters, but this helps in the practical implementation of new scoring function.

Binding scores calculated by the scoring functions \(S_{\text{old}}\) and \(S_{\text{new}}\) can be normalized according to (Bucher, 1990; Tsunoda and Takagi, 1999):

\[
S'_{\text{old}} = \frac{S_{\text{old}} - S_{\text{old}}^{\text{min}}}{S_{\text{old}}^{\text{max}} - S_{\text{old}}^{\text{min}}} , \quad S'_{\text{new}} = \frac{S_{\text{new}} - S_{\text{new}}^{\text{min}}}{S_{\text{new}}^{\text{max}} - S_{\text{new}}^{\text{min}}}
\]  

(23)

where \(S_{\text{old}}^{\text{min}}, S_{\text{old}}^{\text{max}}\) are the hypothetical minimum and maximum for \(S_{\text{old}}\) and \(S_{\text{new}}^{\text{min}}, S_{\text{new}}^{\text{max}}\) are the hypothetical minimum and maximum for \(S_{\text{new}}\) (analytic formula for their calculation is given in Supplemental Material 1).

For the final implementation of the function (22), it is useful to construct sequence dependency corrected matrices of TFs. However, in practice, this can be very inefficient because the dimensions of these matrices can be very high with a lot of zeros. Because of this, we provide a database (available at http://www.fmi.ch/members/andrija.tomovic/database.txt) with sequences and dependent positions written below a database (available at http://www.fmi.ch/members/andrija.tomovic/database.txt).

### 3 RESULTS AND DISCUSSION

#### 3.1 Distributions of transcription factors with dependent positions

To determine the distributions of TFs with dependent positions, we used the public database JASPAR (Lenhard and Wasserman, 2002; Sandelin et al., 2004b) which contains experimentally determined, high-quality binding sites. The JASPAR database represents a curated and non-redundant data-set (Lenhard and Wasserman, 2002; Sandelin et al., 2004b). We selected all TFs for which there were binding sequences (not only matrix profiles) and the final data set contained 107 TFs with 3239 binding sites. We applied three different tests (Section 2.1) to each of these binding sites to establish how many factors showed position dependencies (Table 1). We also show the effect of either applying Bonferroni’s corrections or using the more stringent Bayesian factor \((BF)\).

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>TFs with dependent positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>74.77%</td>
</tr>
<tr>
<td>B</td>
<td>49.52%</td>
</tr>
<tr>
<td>C</td>
<td>62.62%</td>
</tr>
<tr>
<td>D</td>
<td>38.32%</td>
</tr>
<tr>
<td>E</td>
<td>23.26%</td>
</tr>
<tr>
<td>F</td>
<td>26.17%</td>
</tr>
</tbody>
</table>

A—Monte Carlo randomization test without Bonferroni’s correction; B—Exact multinomial ‘goodness-of-fit’ test without Bonferroni’s correction; C—Bayesian hypothesis testing \((BF(H_0, H_1))<0.01\); D—Monte Carlo randomization test with Bonferroni’s correction; E—Exact multinomial ‘goodness-of-fit’ test with Bonferroni’s correction; F—Bayesian hypothesis testing \((BF(H_0, H_1))<0.01\).

In addition, we looked to see if the length and number of known binding sites were different between the groups of TFs with and without dependent positions (Table 2). The variances of these two groups are not statistically different (tested by Bartlett’s test). Using Student’s \(t\)-test, we tested the null hypothesis that mean length and number of binding sites between the two groups are equal against a one-tailed alternative hypothesis that TFs without dependent positions have shorter lengths and smaller numbers of known binding sites. In each case, we obtained \(P\)-values less than 0.05 and thus we should reject the null hypothesis and accept the alternative.

---

Table 1. Distributions of TFs with dependent positions tested

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>TFs with dependent positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>74.77%</td>
</tr>
<tr>
<td>B</td>
<td>49.52%</td>
</tr>
<tr>
<td>C</td>
<td>62.62%</td>
</tr>
<tr>
<td>D</td>
<td>38.32%</td>
</tr>
<tr>
<td>E</td>
<td>23.26%</td>
</tr>
<tr>
<td>F</td>
<td>26.17%</td>
</tr>
</tbody>
</table>

A—Monte Carlo randomization test without Bonferroni’s correction; B—Exact multinomial ‘goodness-of-fit’ test without Bonferroni’s correction; C—Bayesian hypothesis testing \((BF(H_0, H_1))<0.01\); D—Monte Carlo randomization test with Bonferroni’s correction; E—Exact multinomial ‘goodness-of-fit’ test with Bonferroni’s correction; F—Bayesian hypothesis testing \((BF(H_0, H_1))<0.01\).

---

Table 2. Variances of TFs with and without dependent positions tested

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>TFs with dependent positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>74.77%</td>
</tr>
<tr>
<td>B</td>
<td>49.52%</td>
</tr>
<tr>
<td>C</td>
<td>62.62%</td>
</tr>
<tr>
<td>D</td>
<td>38.32%</td>
</tr>
<tr>
<td>E</td>
<td>23.26%</td>
</tr>
<tr>
<td>F</td>
<td>26.17%</td>
</tr>
</tbody>
</table>

A—Monte Carlo randomization test without Bonferroni’s correction; B—Exact multinomial ‘goodness-of-fit’ test without Bonferroni’s correction; C—Bayesian hypothesis testing \((BF(H_0, H_1))<0.01\); D—Monte Carlo randomization test with Bonferroni’s correction; E—Exact multinomial ‘goodness-of-fit’ test with Bonferroni’s correction; F—Bayesian hypothesis testing \((BF(H_0, H_1))<0.01\).
Position dependencies in transcription factor binding sites

Table 2. Average length and number of binding sites between a group of TFs with dependent positions and a group of TFs without dependent positions

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>Average length of TFs binding sites</th>
<th>Average number of known binding sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>A</td>
<td>11.67</td>
<td>8.25</td>
</tr>
<tr>
<td>B</td>
<td>12.15</td>
<td>9.43</td>
</tr>
<tr>
<td>C</td>
<td>11.66</td>
<td>9.3</td>
</tr>
<tr>
<td>D</td>
<td>12.19</td>
<td>9.89</td>
</tr>
<tr>
<td>E</td>
<td>11.92</td>
<td>10.265</td>
</tr>
<tr>
<td>F</td>
<td>12.00</td>
<td>10.34</td>
</tr>
</tbody>
</table>

I—group with dependent positions; II—group without dependent positions; A, B, C, D, E, F — notation the same as in Table 1.

These results imply that more factors may show dependencies once additional binding-site data becomes available.

Based on the second-order dependencies (dinucleotide dependencies), it is possible to construct higher order dependencies, as explained in section 2. It is clear that when we have such dependencies, there are lower order dependencies in all combinations. Because of this, it is useful to analyze distributions of dependencies of different orders \( k_m \) (\( 2 \leq k_m \leq 9 \)) constructed in a more stringent way (Supplemental Material 3). We analyzed the distributions of TFs with dependent positions in structural classes of TF–DNA-binding domains. We wanted to investigate whether there is any tendency for certain folds to have position dependencies (Supplemental Material 4). We noticed that some structural classes contain TFs with position dependencies in their binding sites detected by almost all statistical tests, such as: T-BOX, P53, AP2, TRP, CAAT-box and MADS. Other classes contain TFs without dependent positions like: ZH-FINGER-DOF, ZH-FINGER-GATA, HOMEO/CAAT and ‘Unknown’ class. However, the major structural classes contain TFs with and without dependent positions (bZIP, nuclear receptor, etc.).

3.2 Do position dependencies relate to 3D structures?

We wanted to investigate possible biological explanations of the dependent positions we predicted. We investigated this by examining 3D crystal structures when available. Possible explanations of dependency include:

- active amino acids might interact with dependent nucleotides either singly or in pairs via hydrogen bonds or salt bridges;
- conformational changes in the structure of DNA caused by one dependent base may alter the accessibility of the other dependent bases to the binding site;
- something else.

We selected 32 TF–DNA co-crystal pairs of structures from the PDB database at resolutions better than 3.0 Å (Berman et al., 2000) corresponding to TFs with published binding sites in JASPAR (September 2006) (Table Sup3-1 in Supplemental Material 3). Direct contacts between bases and amino acids were investigated (Table Sup3-2 in Supplemental Material 3).

Next, we wanted to check whether there were any relationships between dependent positions and conformational changes of the DNA. We could calculate structural parameters to describe the 3D nucleic acid structures using the software package 3DNA (Lu and Olson, 2003), but there are many parameters (shift, slide, rise, tilt, roll and twist) to describe the structure of DNA, and because we have relatively few sequences in our data set it is difficult to identify significant effects. Similarly, if we want to investigate spatial distribution patterns of neighboring amino acids around dependent positions, we will have a data-mining problem.

We decided to use the energy Z-scores (Ahmad et al., 2006; Gromiha et al., 2004; Kono and Sarai, 1999) for TF–DNA complexes for both ‘direct’ and ‘indirect’ readouts. The energy Z-score for direct readouts quantifies the spatial distributions of side chains around base pairs, and represents the base–amino acid interaction energy. The energy Z-score for indirect readouts quantifies DNA conformation, and represents the conformational energy of DNA. The more negative the Z-score, the better the target sequence fits into a given structure (Ahmad et al., 2006). The list of all Z-score values can be found in Supplemental Material 4. We tested the Z-scores using a one-tailed Student’s t-test (Table 3). The direct readout showed no difference between TFs with dependent or independent positions (\( P > 0.1 \)). However, the conformational energy (indirect readout) was always significantly lower for TFs with dependent positions (\( P < 0.02 \)). This means that TFs with dependent positions fit their target DNA motifs better than those without. These results suggest a possible relationship between position dependencies and the 3D structure of TFs.
We investigated if DNA sequence length influences the conformational energy. In the 32 cases we studied where we have both a 3D crystal structure and a JASPAR matrix ID, we performed one- and two-tailed \( t \)-tests on the lengths of sequences found to show dependencies and those without dependencies from each of the six dependency tests we investigated. These results showed that five out of six of the tests (two-tailed) and three out of six (one-tailed) do not show significant differences in sequence length between the two groups for which we have conformational energies (Supplemental Material 6). If the conformation of the DNA fragment is not sequence specific, then the conformational energy is expected to fluctuate independently of fragment size. But, if the conformation is sequence specific, then the total energy should decrease with the size although the average energy per base will not decrease if the energy distribution is uniform (A. Sarai, personal communication). For these reasons, we believe that sequence length is not the major factor contributing to the significantly lower conformational energies we found for the group of TFs with dependent positions.

We analyzed relationships between dependent position and DNA stiffness to show the influence of DNA stiffness on protein–DNA binding specificity (Gromiha, 2005). We calculated the average stiffness of DNA using the structure-based sequence-dependent stiffness scale (Gromiha, 2005) for binding sites with and without position dependencies (Supplemental Material 7). In two cases, we found that the average stiffness values are significantly larger (one-tailed Student’s \( t \)-test \( P < 0.028 \)) for sites with dependent positions (detected by Bayesian hypothesis testing in both variants) than without dependent positions. However, in the other four cases no significant differences were found.

### 3.3 Evaluation of a new scoring function for the prediction of TF-binding sites

The evaluation of ab initio methods for the prediction of TF-binding sites is described in (Tompa et al., 2005). Here, we will perform a slightly different validation. In order to evaluate the new scoring function given by (22) and (23), we performed a validation using both synthetic and experimentally verified data.

#### Table 3. Average Z-score for direct and indirect readout for: I—a group of TFs with dependent positions; and II—a group of TFs without dependent positions

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>Average Z-score (direct readout)</th>
<th>Average Z-score (indirect readout)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>A</td>
<td>–2.5</td>
<td>–2.62</td>
</tr>
<tr>
<td>B</td>
<td>–2.67</td>
<td>–2.42</td>
</tr>
<tr>
<td>C</td>
<td>–2.667</td>
<td>–2.25</td>
</tr>
<tr>
<td>D</td>
<td>–3.054</td>
<td>–2.26</td>
</tr>
<tr>
<td>E</td>
<td>–3.44</td>
<td>–2.3</td>
</tr>
<tr>
<td>F</td>
<td>–3.1025</td>
<td>–2.32</td>
</tr>
</tbody>
</table>

\( *P < 0.05, **P < 0.01, ***P < 0.001. \)

A, B, C, D, E, F—notation the same as in Table 1. The variances of groups I and II are not statistically different (Bartlett’s test).
old scoring functions are the same. We currently only apply our correction for positions that show statistically significant dependencies. If, instead, we factor the observed frequency \( P \)-scores for all bases, regardless of their significance, then the new function will tend towards the old function because of Equation (12), and the logarithm property (\( \log(P(B_i, B_j)) = \log(P(B_i)) + \log(P(B_j)) \)) will tend towards \( \log(P(B_i)) + \log(P(B_j)) \) but small differences may be observed because of the smoothing parameters. The price for doing this is computational time, and it does not appear to offer any great advantage over the solution we have implemented.

To further evaluate our new scoring function, we generated 1850 random sequences sampled from a third-order Markov model background distribution with lengths from 250 to 500. In 50, we planted binding sites for MA0041 Foxd3, and we then analyzed the true- and false-positive rates for different threshold values using the new and old scoring functions (Fig. 2 and Table Sup8-1 in Supplemental Material 8). Both functions have good scores for true positives, but the new scoring function gave better results. The biggest difference was in the false-positive rate which was much better with the new scoring function. Next, we generated five random sequences sampled from a third-order Markov model background distribution (with lengths from 400 to 600) in which we planted 0–3 binding sites from a set of 15 (all 15 contained dependent positions). The data set is given in Supplemental Material 8. We wanted to measure the accuracy of prediction with the new scoring function and compare it with other available tools and methods (PATSER, ConSite and the old scoring function). Given that almost all of the methods can detect true positives (i.e. they have a high sensitivity), the accuracy of each method should be estimated by its selectivity (false-positive rate). These results are shown in Figure 3 and Table Sup8-2 in Supplemental Material 8. Our new scoring function (22-23) performed best with the smallest number of false positives per nucleotide and per TF. Finally, we analyzed real experimental data. As ConSite had the next best prediction results with the synthetic data, we decided to use it for benchmark comparisons with the experimental data as well. We used a set of genes showing skeletal muscle-specific expression (Wasserman and Fickett, 1998). This set is an updated version from (Defrance and Touzet, 2006) which has been used to evaluate such tools in the past. This dataset includes upstream regions (2000 bp) of nine genes (see Table Sup8-3 in Supplemental Material 8) and six TFs from the JAPSAR database (MA0052, MA0055, MA0056, MA0057, MA0079 and MA0083) which are known to be involved in the regulation of skeletal muscle-specific expression. MA0055’s binding sites are not listed in JASP, so its detection will be unchanged from the old function (24). We scanned the upstream sequence of the nine genes using all of the TFs from JAPSAR. There are 16 TFs (including MA0055) for which there is no binding sequence information, only weight matrices. These will be treated as having independent binding (24), which will have a negative effect on the results from the new scoring function, but is more realistic. However, even with this limitation, the results from the new scoring function are slightly better than those from ConSite (TableSup 8-3 in Supplemental Material 8). The false-positive rate for all nine sequences is smaller with the new scoring function, and the true-positive rate is almost the same. ConSite detected one true positive hit more (for three sequences) than our scoring function with this data set.

### 4 CONCLUSIONS

In this work, we performed a detailed analysis of dependencies within TF-binding sites. Our conclusion is that we cannot assume that positions are either dependent or independent. This must be tested using one of three proposed statistical tests. Our structural analysis indicates that some of the predicted dependencies agree with 3D structure data from TF–DNA complexes. We propose that the dependencies we have identified should be used in binding-site predictions. Previous attempts at such modeling have required complex tools with many parameters which really require more training data than is currently available. Here, we present a simple way of modeling these dependencies. We demonstrated how to modify existing dependence-free scoring functions to consider dependencies. Such modifications improve prediction quality for TFs...
with dependent positions. Our technique does not require complex tools or more training data than scoring functions and models which assume independence. This approach can be used with any scoring function which assumes independence (one such is presented here). We demonstrated this approach using scanning methods for the prediction of TF-binding sites, but it can be applied to work with ab initio methods and different methods of prediction which incorporate comparative genomic analysis (phylogenetic footprinting conservation).

ACKNOWLEDGEMENTS
We would like to thank Prof Gill Bejerano, Prof Frank Hampel, Dr Hans-Rudolf Roth, Dr Michael Studler and Prof Akinori Sarai for useful discussions and advice. This work was supported by the Novartis Research Foundation. Funding to pay the Open Access publication charges was provided by the Novartis Research Foundation FMI.

Conflict of Interest: none declared.

REFERENCES
Conohan,M.A. (1970) The comparative accuracy of the likelihood ratio and Chi-squared as approximation to the exact multinomial test. Lehigh University, 64.