A non-independent energy based multiple sequence alignment improves prediction of transcription factor binding sites

Rafik A Salama¹,* and Dov J. Stekel²
¹School of Biosciences, University of Nottingham, Sutton Bonington Campus, Leicestershire, LE12 5RD, UK.
²Cancer Research UK, Cambridge Institute, University of Cambridge, Cambridge, CB2 0RE, UK.

ABSTRACT
Motivation: Multiple Sequence Alignments (MSAs) are usually scored under the assumption that the sequences being aligned have evolved by common descent. Consequently, the differences between sequences reflect the impact of insertions, deletions and mutations. However, non-coding DNA binding sequences, such as transcription factor binding sites (TFBS), are frequently not related by common descent, and so the existing alignment scoring methods are not well suited for aligning such sequences.

Results: We present a novel multiple MSA methodology that scores TFBS DNA sequences by including the interdependence of neighboring bases. We introduced two variants supported by different underlying hypotheses, one statistically and the other thermodynamically based. We assessed the alignments through their performance in TFBS prediction: both methods show considerable improvements when compared with standard MSA algorithms. Moreover, the thermodynamically based hypothesis (EDNA) outperforms the statistical one due to improved stability in the base stacking free energy of the alignment. EDNA can be downloaded from http://sourceforge.net/projects/msa-edna/.

1 INTRODUCTION

With the advent of third generation sequencing technology, the rate of genome sequencing is likely to continue to outstrip the capacity for experimental analysis of gene function and regulation. In addition, there is a similar growth in the availability of Transcription Factor Binding Site (TFBS) sequences through genome-wide ChIP-seq methods. Thus there is a continued need for more accurate computational methodologies. One particularly important set of methods are Multiple Sequence Alignment (MSA) programs, such as ClustalW (Thompson, et al., 2002; Thompson, et al., 1999) and Dialign (Morgenstern, 2007), that are widely used to identify sections of protein or DNA sequences that share similar consensus. These alignment programs generally work under the assumption that the sequences being aligned are evolutionarily related and that they have derived from a common ancestor, undergoing changes due to insertions, deletions and substitutions. This assumption is then built into scoring matrices which are used to find the best possible alignment, such as BLOSUM (Henikoff and Henikoff, 1992) and PAM (Dayhoff, 1978), as well as for gap penalties.

However, non-coding DNA regions, especially transcription factor binding sites, are rather more conserved and not necessarily evolutionarily related (Mukherjee R, 2013), and may have converged from non-common ancestors. Thus, the assumptions used to align protein sequences and DNA coding regions are inherently different from those that hold for TFBS sequences. While it is meaningful to align DNA coding regions for homologous sequences using mutation operators, alignment of binding site sequences for the same transcription factor cannot rely on mutation operators for an alignment. Similarly, the evolutionary operator of point mutations can be used to define an edit distance for coding sequences, but this has little meaning for TFBS sequences, since any sequence variation has to maintain a certain level of specificity for the binding site to function.

It has been long known that interdependencies between neighboring DNA bases have a significant impact on DNA topology. For example, the thermodynamic properties of base stacking interactions have been extensively measured, and are commonly used in computational methods for DNA secondary structure prediction (Mathews, et al., 1999). This was illustrated in work discussing the effect of DNA flexure on the binding site affinity (Calladine and Drew, 1986). Compensating mutations between neighboring DNA bases have been long known (Stormo, et al., 1986). Consideration of the co-dependencies between DNA bases in a TFBS have been proved to be efficient in modeling binding site specificity (Homsi, et al., 2009; Stormo and Zhao, 2010), also suggesting the possibility of co-evolution between the DNA bases (Mukherjee R, 2013). We have shown in previous work (Salama and Stekel, 2010) using mutual information analysis that there are dependencies between neighboring and distant positions of the TFBSs; many of the distant interactions reflect the palindromic nature of TFBSs.
The issues of effective MSAs for DNA binding sites become particularly important for supervised TFBS prediction. Supervised TFBS prediction algorithms take as input a set of known TFBS sequences, known as a “training set”, and use these to build an expected model of the TFBS that can be used to identify further sequences (Hertz and Stormo, 1999; Lee and Huang, 2012; Pauling et al., 2012; Salama and Stekel, 2010; Stormo, et al., 1986). These can be contrasted with de novo binding site prediction methods that analyze upstream regions of genes to find over-represented motifs (Bailey et al., 2006; Herz and Stormo 1999) and which are not the subject of this paper. We previously demonstrated, that the use of a non-optimal alignment for training can seriously disrupt TFBS prediction efficacy, even as compared with using a block alignment (Salama and Stekel, 2010). Moreover, effective supervised TFBS prediction relies on a suitably large training set of known binding sites and so predictions can only be made for “global” regulators. Most transcription factors only regulate a small number of genes and supervised prediction of those “non-global” regulators has been particularly hard. One solution would be to use binding sites from paralogous or orthologous transcription factors, as interdependence patterns are similar across related transcription factors (Mukherjee et al. 2013). In order to do so it becomes necessary to produce multiple sequence alignments for those binding sites that capture transcription factor binding site specificity.

In this paper, we aim to include the interdependencies in the MSA to provide a better representation for the DNA binding site specificity. We introduce a dinucleotide representation method of the TFBS sequence that captures the interdependencies. We then introduce two alignment cost metrics methodologies, the first is using a statistical approach similar to the one used to calculate BLOSUM (Henikoff and Henikoff, 1992); the second makes use of base stacking thermodynamics. We specifically used a Boltzmann distribution centered on the change in base stacking free energy as the null hypothesis for dinucleotide substitutions.

To validate our work, we compare both methods against each other, and against other commonly used alignment methods (ClustalW, Dialign), and against our previously published method using block alignments. There are a good range of benchmark alignments to test protein multiple sequence alignments methods including BAliBASE (Thompson, et al., 2005), OXBench (Raghava, et al., 2003), SABmark (Walle, et al., 2004) and SMART (Ponting, et al., 1999). For DNA coding regions, Carroll et al. (Carroll, et al., 2007) have developed a DNA reference benchmark based on the tertiary structure of encoded protein. However, no such benchmark alignments are available for non-coding DNA sequences. A score is often used to detect the accuracy of the MSA using the homology in the resulting alignment (Thompson, et al., 1999). The approach we take is to test the alignments for their capacity to act as a training set for a supervised first order Hidden Markov Model (Salama and Stekel 2010) to predict known transcription factor binding sites.

2 METHODS

2.1 Dinucleotide Substitution Matrix

The Substitution Matrix considered in this case is a 16×16 matrix representing the dinucleotide substitution rates for the TFBS. The dinucleotides are symbolized using the alphabet A to P to represent alphabetically each possible pair of neighboring nucleotides AA through to TT in alphabetical order. A DNA sequence is translated into our new alphabet in an overlapping manner. For instance, the sequence ACA would be represented in our new alphabet by the sequence BE: B for AC and E for CA. Every sequence in the training set is converted into this new alphabet for alignment with the other sequences.

The hypothesis behind this conversion is that the sequences are now forced into an interdependent representation of the binding site rather than an independent one. This representation captures the heart of the single nucleotide mutation effect on neighboring base interactions. For instance a single point mutation (transition) of the Cytosine to Thymine in this sequence ACA => ATA would result in a change of two neighboring base interactions AC => AT and CA=>TA, and so would be represented as two position mutations in dinucleotide representation BE => DM. When considering stacking interactions, this can be used to represent the change in free energy of both interactions.

A substitution matrix is highly specific for each transcription factor binding site set. The substitution matrix is then computed for each transcription factor binding site set. A problem is that to construct such a substitution matrix a valid alignment is required. This is inherently recursive, as obtaining a correct substitution matrix requires a valid multiple sequence alignment, and the multiple sequence alignment requires an optimized substitution matrix. This recursive nature of the problem have led us to devise a simple recursive algorithm similar to the Expectation Maximization algorithm devised by (Cao et al. 2009).

Initially, we define a random substitution matrix with costs drawn from independent uniform random variables between 0 and 100 (see supplementary methods). This is then used to generate the first alignment. In the maximization step, the resulting alignment is used to calculate the substitution matrix using the log odds cost matrix (Equation 1).

\[
\text{Lod}_{ij} = \ln \frac{O_{ij}}{E_{ij}}
\]

Where \(O_{ij}\) is the observed probability of aligning dinucleotide \(i\) with dinucleotide \(j\) and \(E_{ij}\) is the expected probability (Null Hypothesis) of aligning dinucleotide \(i\) with dinucleotide \(j\). In the expectation step the resulting substitution matrix is then used again to compute a new sequence alignment with optimum alignment cost. In all our implementations of this algorithm, it has converged; this is likely to be because of the analogy with EM algorithms (Wu, 1983). The final substitution matrix is selected through another iterative process optimizing gap penalties, as explained later. The final dinucleotide substitution matrix is then used to align the transcription factor known binding sites.

The observed probability \(O_{ij}\) is constructed using the following equations

\[
O_{ij} = \frac{Q_{ij}}{\sum_{j=1}^{16} \sum_{i=1}^{16} Q_{ij}}
\]
Where $Q_{ij}$ is given by equation (3),

$$Q_{ij} = \begin{cases} 
\sum_{x=1}^{N_i} N_i^x, & \text{if } i \neq j \\
\sum_{x=1}^{N_i} N_i^x(N_i^x - 1), & \text{if } i = j 
\end{cases}$$ (3)

Where, $N_i^x$ is the number of dinucleotides of type i at position x. Thus $Q_{ij}$ represents the number of times each dinucleotide as been aligned to each other dinucleotide across the whole alignment.

The log odds ratio optimization process is controlled by the null hypothesis used in the denominator of the equation. The null hypothesis in the log odd scoring controls the substitution cost as in equation (1). Hence in the algorithm devised; a dinucleotide alignment cost is rewarded if the observed cost is better than the expected cost and penalized if it is worse.

Accordingly, we have assessed two different null hypothesis distributions, statistical and thermodynamical, with associated cost matrices. These have been denoted as SDNMSA for the statistically generated null hypothesis and EDNA for the thermodynamically generated null hypothesis.

### 2.2 Statistically generated Null Hypothesis (SDNMSA)

The first null hypothesis is a purely statistical hypothesis representing the expected independent joint distribution of the dinucleotides which is generated following the equations as given in (Henikoff and Henikoff, 1992).

$$E_i = \left(\frac{Q_{ij} + \sum_{i=j}^{16} Q_{ij}}{2}\right) \sum_{i=1}^{16} \sum_{j=1}^{16} Q_{ij}$$ (4)

Where $E_i$ is the expected probability dinucleotide $i$

$$E_{ij} = \begin{cases} 
E_iE_j, & \text{if } i = j \\
2E_iE_j, & \text{if } i \neq j 
\end{cases}$$ (5)

From equation 2 and 5, a statistical cost matrix is then generated using

$$S_{ij} = K \left(1 - \ln \frac{Q_{ij}}{E_{ij}}\right)$$ (6)

Where $S_{ij}$ is the statistical cost of substituting dinucleotide $i$ with dinucleotide $j$ and $K$ is a scaling constant.

### 2.3 Thermodynamically generated Null Hypothesis (EDNA)

The second null hypothesis is thermodynamical, and assumes that the dinucleotide substitutions associated with a transcription factor binding site set are distributed using Boltzmann distributions, and that the joint distribution between two dinucleotides (e.g. representing a single base substitution) is energetically independent. The parameters for the Boltzmann distributions are derived from the dinucleotide stacking free energies (Allawi and SantaLucia, 1997; Allawi and SantaLucia, 1998; Allawi and SantaLucia, 1998; Allawi and SantaLucia, 1998; Allawi and SantaLucia, 1998; SantaLucia and Turner, 1997). First constructing a Boltzmann distribution of the existing dinucleotides in the training set using Equation 7, and then constructing a joint distribution assuming independence as in Equation 8 compute the null hypothesis distribution.

$$B_i = d_i e^{\Delta G_i / K_b T}$$ (7)

Where $B_i$ is the Boltzmann distribution of dinucleotide $i$, $\Delta G_i$ is the stacking free energy of the dinucleotide $i$, $T$ is temperature (298 Kelvin) and $K_b$ is the Boltzmann Constant equal to 0.00198721 kcal/mol/K

$$E_{ij} = \begin{cases} 
B_iB_j, & \text{if } i = j \\
2B_iB_j, & \text{if } i \neq j 
\end{cases}$$ (8)

The thermo dynamical cost can be constructed from equation 2 and 8.

$$T_{ij} = K \left(1 - \text{Lod}_{ij}\right)$$ (9)

Where $T_{ij}$ is the thermodynamical cost of substituting dinucleotide $i$ with dinucleotide $j$ and $K$ is a scaling constant. Reversing the sign of the odd score function is a pure technicality of the alignment program, which requires low cost values for expected substitutions and high cost for unexpected ones. Hence reversing the sign would penalize the negative odd score and reward the positive ones.

### 2.4 Gap Penalties

An affine gap penalty function is used with different sets of penalties for gaps within the sequence (internal) and for prefix and suffix gaps (Altschul and Erickson, 1986). Thus the total gap penalty is given by:

$$G = \sum_{i=1}^{n} (\gamma + \beta D_i) + \sum_{j=1}^{n} (\gamma' + \beta' D_j)$$ (10)

Where $\gamma$ is the internal gap open penalty, $\beta$ is the internal gap extension penalty, $D_i$ is the length of internal gap $i$, $\gamma'$ is the terminal
The gap penalty has to be tightly optimized with the cost matrix optimization process. Therefore the gap penalty constants are optimized iteratively as fractions of the average of the substitution matrix calculated as shown in equation 11.

$$G = \sum_{j=1}^{n}(\gamma \mu_c + \beta \mu_c D_j) + \sum_{j=1}^{n}(\gamma' \mu_c + \beta' \mu_c D_j)$$  \hspace{1cm} (11)

Where, $\mu_c$ is the average of the cost matrix previously optimized.

The acceptance function $\varepsilon$ for a chosen penalty is calculated for every computed G as follows:

$$\varepsilon = \frac{\Omega}{(L_A - L_B)}$$  \hspace{1cm} (12)

Where $\Omega$ is the number of columns in the alignment for which the highest frequency of symbols is greater than $\alpha$ as given by equation (13), $L_A$ is the length of the alignment and $L_B$ is the initial length of the binding site block alignment (ungapped).

$$\Omega = \text{count} \left( \frac{P_i}{N} \geq \alpha \right)$$  \hspace{1cm} (13)

Where $P_i$ is the number of the most frequent symbol in column $i$, and $\alpha$ is the maximal symbol frequency observed in the block alignment given by:

$$\alpha = \max \left( \frac{P_i}{N} \right)$$  \hspace{1cm} (14)

2.5 Validation

As a validation, we tested the ability of a supervised first order Hidden Markov Model (Salama and Stekel, 2010), trained using alignments from each of these methods, to predict known transcription factor binding sites. Accordingly, we considered 18 global transcription factors in E. coli K12 (Supplementary Table S1). For each factor, we used the RegulonDB 7.0 (Gama-Castro, et al., 2011) known binding sites, as these can be considered as a highly reliable ‘gold standard’. We aligned the binding sites for each transcription factor using the four different alignment methods compared in this paper. We used the alignments to build first order Hidden Markov Models (HMM) for each factor (see Supplementary Methods). We used a leave one out cross validation to determine the likelihood of each binding site within the transcription factor binding site set. Finally, we constructed a negative binding set using a moving window of the same length of the binding site in the intergenic regions in E.Coli genome, as described previously (Salama and Stekel, 2010). The collective predictive power of each alignment is then assessed using area under Receiver Operating Characteristic (ROC) curve (Zweig and Campbell, 1993).

3 IMPLEMENTATION

EDNA is implemented as an extension of the open source MSA program, Opal (Wheeler and Kececioglu, 2007). Our alignment methodology makes profound changes in the alphabet, substitution matrix computation and gap penalties optimization. We have optimized the specific alignment parameters in Opal. EDNA is currently available at source forge, which is currently open for use by the public, (http://sourceforge.net/projects/msa-edna/). Example data, namely the sequences from the binding sites used in this manuscript, are included with the code.

4 RESULTS

4.1 EDNA conserves the thermodynamic stability of multiple sequence alignments

This method is intended to primarily optimize the thermodynamic stability of the alignment. In other words, find the correct alignment where a dinucleotide substitution is possible thermodynamically without disrupting the free energy of neighboring interactions. Hence, we have measured the thermodynamic stability of each of the alignments for each of the TFBSs considered in the study (Figure 2). It can be seen that the thermodynamic stability is greatly improved even for the statistical dinucleotide alignment SDNMSA, with further improvement for the thermodynamically based alignment EDNA.
Fig. 2. 3D plots of the standard deviations in the ΔG for each column in the alignments for every binding site using the four gapped alignment methods (columns with more than 90% gaps have been excluded). The standard deviation vector associated with all of the binding sites in each alignment has been smoothed using non-parametric loess regression and then recalculated over 50 points to deal with variable alignment lengths. Note how EDNA produces the least variability of the four alignment methods, indicating that these alignments are the most thermodynamically stable.

As an illustrative example, greater detail is shown for the AraC TFBS alignments (Figure 3). In general, we noticed that ClustalW shows a better free energy conservation than Dialign. SDNMSA and EDNA on the other hand, showed an even better conservation than ClustalW, however, EDNA showed the most conserved column count. Dialign produced a highly unstable alignment with only a very small, stable core. Similarly, ClustalW produced an alignment with high stability in the core columns, but also a terminal stability is noticed. SDNMSA showed on the other hand a much better conservation of the free energy across the alignment, with conservation on both sides and the core. Finally, EDNA showed a similar but better conservation to SDNMSA with more columns conserved.

4.2 Dinucleotide-based alignments outperform independent column alignments for transcription factor binding site prediction

Dinucleotide representation of the binding site outperformed independent position alignments methods, notably Dialign and ClustalW as shown in Table 1 and summarized in Figure 4.

Table 1. Comparison between various MSA tools in TFBS prediction sensitivity and specificity.

<table>
<thead>
<tr>
<th></th>
<th>Dialign</th>
<th>ClustalW</th>
<th>ULPB</th>
<th>SDNMSA</th>
<th>EDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraC</td>
<td>0.29</td>
<td>0.88</td>
<td>0.83</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>ArcA</td>
<td>0.72</td>
<td>0.77</td>
<td>0.88</td>
<td>0.86</td>
<td>0.95</td>
</tr>
<tr>
<td>ArgR</td>
<td>0.97</td>
<td>0.97</td>
<td>0.95</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>CpxR</td>
<td>0.89</td>
<td>0.86</td>
<td>0.84</td>
<td>0.86</td>
<td>0.95</td>
</tr>
<tr>
<td>CRP</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Fis</td>
<td>0.74</td>
<td>0.79</td>
<td>0.82</td>
<td>0.85</td>
<td>0.86</td>
</tr>
<tr>
<td>FhlDC</td>
<td>0.48</td>
<td>0.71</td>
<td>0.85</td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td>FNR</td>
<td>0.98</td>
<td>0.98</td>
<td>0.95</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Fur</td>
<td>0.58</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>GlpR</td>
<td>0.98</td>
<td>0.97</td>
<td>0.93</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>IHF</td>
<td>0.40</td>
<td>0.72</td>
<td>0.84</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>LEXA</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Lrp</td>
<td>0.50</td>
<td>0.46</td>
<td>0.75</td>
<td>0.70</td>
<td>0.81</td>
</tr>
<tr>
<td>MetJ</td>
<td>0.28</td>
<td>0.36</td>
<td>0.67</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>NtcR</td>
<td>0.88</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>PhoP</td>
<td>0.91</td>
<td>0.95</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>SoxS</td>
<td>0.32</td>
<td>0.44</td>
<td>0.95</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>TyrR</td>
<td>0.78</td>
<td>0.98</td>
<td>0.84</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table listing the Area under ROC curves for TFBS prediction sensitivity and specificity corresponding to 4 alignment methods (Dialign, ClustalW, SDNMSA, EDNA) applied to 18 of the global regulators with at least 20 known binding sites in RegulonDB E.coli MG1655, along with the simple block alignment based prediction ULPB (Salama and Stekel, 2010).

The AUCs for EDNA range between 81% to 99% (Table 1); many of the AUCs for Dialign are considerably worse, with the lowest, MetJ, at 28%, and the AUCs for ClustalW are also considerably worse, with MetJ at 36%. Overall, EDNA out-performs Dialign (p=0.001 paired Wilcoxon test), with larger AUCs for 14 of the 18 TFs, and equal AUCs for the other 4 TFs. Notably, these TFs, CRP, GlpR, FNR and LexA all have highly conserved binding sequences (Supplementary Table S1), and both methods perform equally well (98% or 99% AUC). EDNA outperforms ClustalW (p=0.0008 paired Wilcoxon test), with greater AUCs for 15 of the 18 TFs. The AUCs for FNR and LexA are equal, ClustalW outperforms EDNA for TyrR, but the AUCs for the two methods are 97% and 98% respectively, so there is minimal material difference between the two.
Fig. 4. Scatter plot of alignment methods Area under ROC curve for each of Dialign, ClustalW, ULPB and SDNMSA against EDNA. Points above the line represent TFBS sequences for which EDNA outperforms the alternative alignment method. Note the differences in scale in each graph, indicative of the overall performance of the alternate method. SDNMSA shows the most fit with the expected fitness line. Dialign shows the worst fitness to the expected fitness line.

EDNA also outperforms our previously published method using a block alignment, ULPB (Salama and Stekel 2011) (p=0.0005). The range of AUCs for ULPB are considerably better than those using Dialign and ClustalW, with the lowest being 67% for MetJ. EDNA outperforms ULPB for 16 of the 18 TFs. The exceptions are NtrC for which both report an AUC of 99%, and SoxS for which EDNA reports 93% and ULPB 95%.

4.3 Thermodynamic multiple alignments outperforms statistically based multiple alignment method

The thermodynamically-based method EDNA also outperforms the statistically-based method SDNMSA (Figure 5). The AUCs for SDNMSA are considerably better than those for the column alignments, with the lowest, for LRP, being 70% (Table 1). Overall, EDNA outperforms SDNMSA (p=0.006 paired Wilcoxon test). EDNA outperforms SDNMSA for 10 of the 18 TFs studied, and they have equal AUCs for the other 8 TFs, all of which are in the range 91-99%.

5 DISCUSSION

We have developed a new methodology for multiple sequence alignments of non-coding DNA sequences that uses an alignment of dinucleotides. Two variants have been described, using two different null hypotheses, one statistically driven and the other thermodynamically driven. These have been compared with other alignment programs, exemplified by ClustalW and Dialign, which are designed with a null hypothesis derived from the evolution of peptide sequences. They have been evaluated by assessing transcription factor binding site prediction 18 global regulators from E. coli K12. The ROC curves for the first order supervised HMM prediction using the dinucleotide alignment demonstrated better alignment than the current alignment tools, irrespective of the null hypothesis used. Between the two variants, the use of thermodynamic driven null hypothesis proved to be statistically better.

The Boltzmann distribution derived provides a base stacking interaction dinucleotides distribution for each transcription factor binding site studied. An independent thermodynamic joint distribution of dinucleotides represents the random distribution of dinucleotide thermodynamic alignment. The log odd scoring system in this case would provide a negative or zero score if the observed dinucleotides alignment joint probability is less than or equal to the random joint probability and positive otherwise. Accordingly, scoring the observed joint distribution versus this null hypothesis distribution would indicate the odds of randomly aligning any dinucleotide pair under thermodynamic hypothesis. The higher these odds are, the less favored they are in the alignment. This would indicate that the resulting alignment from EDNA would favor to align dinucleotides that are more thermodynamically dependent than independent.

5.1 Thermodynamic null hypothesis is not governed by rarity of the dinucleotide

Another major advantage of using a thermodynamically driven null hypothesis is that it is not governed by the rarity of the substituions in the binding site training set (Eddy, 2004). Instead, it provides a consistent behavior for all the binding sites, expecting that the rarity of an alignment event be governed by unfavorable stacking free energy. Thus, EDNA avoids problems resulting from under-sampling in a set of transcription factor binding sites, which may be particularly valuable when building alignments for non-global regulators.

5.2 Base stacking interaction driven convergence

In general, the method provides a good performance for most binding sites relative to alternative methods. Nevertheless, false posi-
tives are still observed in some binding sites, including Lrp, SoxS, IHF, and MetJ. For very well conserved binding sites, such as LexA, ArgR, FNR, Fur, GfpR, PhoP, all methods show good performance; this can be attributed to the small solution space providing a set of constraint alignment solutions. On the other hand, the predictions for the less conserved binding sites, such as AraC, ArcA, CpxR, FihDC, TyrR and MetJ, have been particularly enhanced by our new method (Supplementary Figure S2). This may reflect the importance of thermodynamic interactions between neighboring DNA bases, which have been described well by this method.

5.3 Higher order matrices

The fact that we have used a first order substitution matrix to capture the binding site might be on its own providing an advantage for the alignment, since we account for the substitutions and aligning blocks of two bases rather one. It could be argued that higher-order matrices might provide further improvements. However, a major problem with such matrices is the complexity involved; a trinucleotide matrix would require a 64x64 substitution matrix with more than 4000 parameters, which would be challenging to compute and optimize, and most multiple sequence alignments are likely to be under-sampled in this context. Moreover, the base stacking free energies used have only been experimentally determined for first order interactions, and not for higher order interactions (Allawi and SantaLucia, 1997; Allawi and SantaLucia, 1998; Allawi and SantaLucia, 1998).

ACKNOWLEDGEMENTS

Funding: This work has been funded by the Darwin Trust of Edinburgh. The Opal libraries are included in our software with permission from their authors (Travis Wheeler and John Kececioglu) under a Creative Commons Non Commercial license.

REFERENCES


