BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 21 2013, pages 2699–2704
doi:10.1093/bioinformatics/btt463

Sequence analysis

Advance Access publication August 28, 2013

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

Rafik A. Salama[1] and Dov J. Stekel[2,*]

[1]Cancer Research UK, Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK and [2]School of Biosciences, University of Nottingham, Sutton Bonington Campus, Leicestershire LE12 5RD, UK

Associate Editor: Martin Bishop

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 (TFBSs), 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 null hypotheses, one statistically and the other thermodynamically generated. We assessed the alignments through their performance in TFBS prediction; both methods show considerable improvements when compared with standard MSA algorithms. Moreover, the thermodynamically generated null hypothesis outperforms the statistical one due to improved stability in the base stacking free energy of the alignment. The thermodynamically generated null hypothesis method can be downloaded from http://sourceforge.net/projects/msa-edna/

Contact: dov.stekel@nottingham.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 19, 2013; revised on July 26, 2013; accepted on August 5, 2013

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. Particularly important are multiple sequence alignment (MSA) programs, such as ClustalW (Thompson *et al*., 1999, 2002;) 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 been derived from a common ancestor, undergoing changes because of 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 TFBSs, are rather more conserved and not necessarily evolutionarily related (Mukherjee *et al*., 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. Although 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. 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 because 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 codependencies 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 coevolution between the DNA bases (Mukherjee *et al*., 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

*To whom correspondence should be addressed.

of genes to find overrepresented motifs (Bailey *et al.*, 2006; Herz and Stormo 1999) and which are not the subject of this article. 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). To do so, it becomes necessary to produce MSAs for those binding sites that capture TFBS specificity.

In this article, 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, against other commonly used alignment methods (ClustalW, Dialign) and against our previously published method using block alignments. There is a good range of benchmark alignments to test protein MSA 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.* (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 (HMM) (Salama and Stekel, 2010) to predict known TFBSs.

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 TFBS set. The substitution matrix is then computed for each TFBS 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 MSA, and the MSA requires an optimized substitution matrix. This recursive nature of the problem has 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)]

$$\mathrm{Lod}_{i,j} = \ln \frac{O_{i,j}}{E_{i,j}} \tag{1}$$

Where $O_{i,j}$ is the observed probability of aligning dinucleotide i with dinucleotide j and $E_{i,j}$ 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 $O_{i,j}$ is constructed using the following equations

$$O_{i,j} = Q_{i,j} \Big/ \sum_{i=1}^{16} \sum_{j=1}^{i} Q_{i,j} \tag{2}$$

Where $Q_{i,j}$ is given by Equation (3)

$$Q_{i,j} = \begin{cases} \sum_{x=1}^{n} N_i^x N_j^x, & if \quad i \neq j \\ \sum_{x=1}^{n} \frac{1}{2} N_i^x (N_i^x - 1), & if \quad i = j \end{cases} \tag{3}$$

Where N_i^x is the number of dinucleotides of type i at position x. Thus, $Q_{i,j}$ 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

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)

$$E_i = \left(Q_{i,i} + \sum_{i \neq j}^{16} Q_{i,j}/2\right) \Bigg/ \sum_{i=1}^{16} \sum_{j=1}^{i} Q_{i,j} \quad (4)$$

Where $\boldsymbol{E}_{\mathrm{i}}$ is the expected probability dinucleotide i

$$E_{i,j} = \begin{cases} E_i E_j, & if \quad i = j \\ 2E_i E_j, & if \quad i \neq j \end{cases} \quad (5)$$

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

$$S_{i,j} = K\left(1 - \ln \frac{O_{i,j}}{E_{i,j}}\right) \quad (6)$$

Where $S_{\mathrm{i,j}}$ 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 it assumes that the dinucleotide substitutions associated with a TFBS 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, 1998a, b, c, d; SantaLucia and Turner, 1997). First, a Boltzmann distribution is constructed for the existing dinucleotides in the training set:

$$B_i = d_i e^{-\Delta G_i / K_B T} \Bigg/ \sum_{i=1}^{16} \left(d_i e^{-\Delta G_i / K_B T}\right) \quad (7)$$

Where $\Delta \mathrm{G}_i$ is the stacking free energy of the dinulceotide i, T is temperature (298 K) and $\mathrm{K_B}$ is the Boltzmann Constant equal to 0.00198721 kcal/mol/K. Next, the expected values for the null hypothesis distribution are constructed as a joint distribution assuming independence:

$$E_{i,j} = \begin{cases} B_i B_j, & if \quad i = j \\ 2B_i B_j, & if \quad i \neq j \end{cases} \quad (8)$$

The thermodynamical cost of substituting dinucleotide *i* with dinucleotide *j* can be constructed from Equations 2 and 8:

$$T_{i,j} = K\left(1 - \mathrm{Lod}_{\mathrm{i,j}}\right) \quad (9)$$

Where K is a scaling constant (set to 100). 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

$$G = \sum_{i=1}^{n} (\gamma + \beta D_i) + \sum_{j=1}^{2} (\gamma' + \beta' D_j) \quad (10)$$

Where γ is the internal gap open penalty, β is the internal gap extension penalty, $\mathrm{D_i}$ is the length of internal gap i, γ' is the terminal (external) gap open penalty, β' is the terminal gap extension penalty and $\mathrm{D_j}'$ the length of the terminal gap extension j.

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

$$G = \sum_{i=1}^{n} (\gamma \mu_C + \beta \mu_C D_i) + \sum_{j=1}^{2} (\gamma' \mu_C + \beta' \mu_C D_j) \quad (11)$$

Where μ_c is the average of the cost matrix previously optimized.

The acceptance function ε for a chosen penalty is calculated for every computed G as follows

$$\varepsilon = \frac{\Omega}{(L_A - L_B)} \quad (12)$$

Where Ω is the number of columns in the alignment for which the highest frequency of symbols is greater than αas 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)

$$\Omega = count\left(\frac{P_i}{N} \geq \alpha\right) \quad (13)$$

Where $\mathrm{P_i}$ is the number of the most frequent symbols in column i, and the homology percentage α is the maximal symbol frequency observed in the block alignment given by

$$\alpha = \max_i \left(\frac{P_b}{N}\right) \quad (14)$$

The proposed alignment mainly operates by shifting the binding sites relative to each other, preferring terminal gaps over internal gaps, due to the steric constraints of protein–DNA interactions. Accordingly, the internal gap open/extension penalties (γ, β) are fixed at the high value of 5. The terminal open/extension gaps, on the other hand, are optimized by fixing γ' at a 0.1 and increasing the extension terminal gap penalty (β') iteratively until no gaps can be added to the sequence. In other words, the opening of a terminal gap is always allowed, although the extension is optimized. The best terminal extension gap penalty is defined as the one that corresponds to higher Ω (Fig. 1).

2.5 Validation

As a validation, we tested the ability of a supervised first-order HMM (Salama and Stekel, 2010), trained using alignments from each of these methods, to predict known TFBS. Accordingly, we considered 18 global transcription factors in *Escherichia 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 article. We used the alignments to build first-order 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 TFBS 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

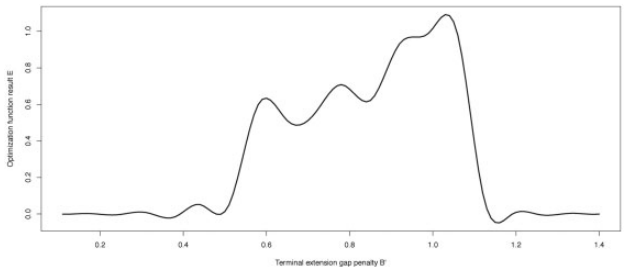

Fig. 1. Terminal extension gap penalty β' optimized against ε for AraC binding site, choosing the optimum penalty for the alignment that maximizes the optimization function—in this case the chosen terminal extension

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 MSA

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 (Fig. 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.

As an illustrative example, greater detail is shown for the AraC TFBS alignments (Fig. 3). In general, we noticed that ClustalW shows 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 small stable core. Similarly, ClustalW not only 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.

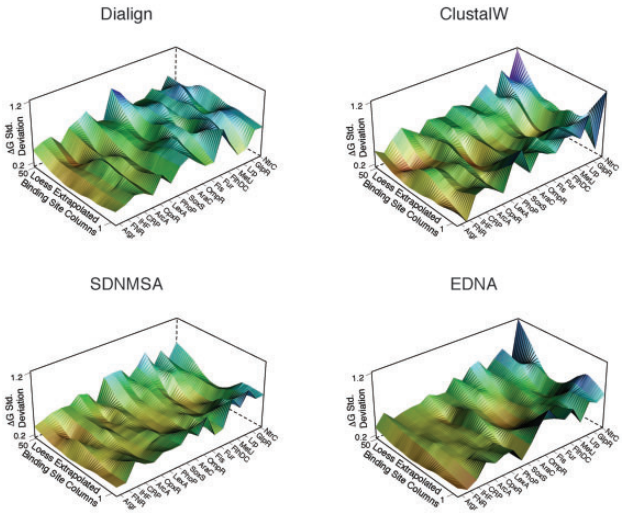

Fig. 2. The 3D plots of the standard deviations in the ΔG for each column in the alignments for every binding site using the four gapped alignment methods (columns with >90% gaps have been excluded). The standard deviation vector associated with all of the binding sites in each alignment has been smoothed using non-parametric Loess regression and then recalculated over 50 points to deal with variable alignment lengths. EDNA produces the least variability of the four alignment methods, indicating that these alignments are the most thermodynamically stable

4.2 Dinulecotide-based alignments outperform independent column alignments for TFBS 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.

The area under ROC curves (AUCs) for EDNA range between 81 and 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 outperforms 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.

EDNA also outperforms our previously published method using a block alignment, ungapped likelihood under a positional background (ULPB) (Salama and Stekel, 2010) ($P = 0.0005$). The range of AUCs for ULPB is 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

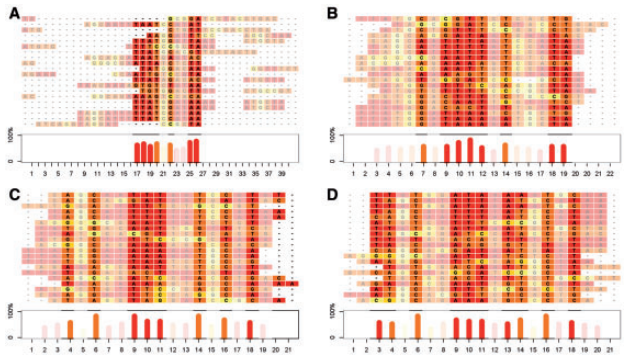

Fig. 3. Alignment of AraC binding site using the four gapped alignment methods studied. (**A**) Dialign, (**B**) ClustalW, (**C**) SDNMSA and (**D**) EDNA. Each square is assigned a color that follows a heat color palette representing red as the highest free energy cluster and yellow as the lowest free energy cluster. The histogram plot represents the most frequent color in each column; the height of column represents the frequency of that color in the binding site position. The strongly shaded columns are those with frequencies >50%. Dialign shows a poor alignment with little consensus; ClustalW shows a better alignment than Dialign but with an unwanted gap. SDNMSA and EDNA provide the longest alignments, with EDNA also having the most columns with highest frequency of conservation

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 outperforms the statistically based method SDNMSA (Fig. 5). The AUCs for SDNMSA fare 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 TFs studied, and they have equal AUCs for the other 8 TFs, all of which are in the range 91–99%.

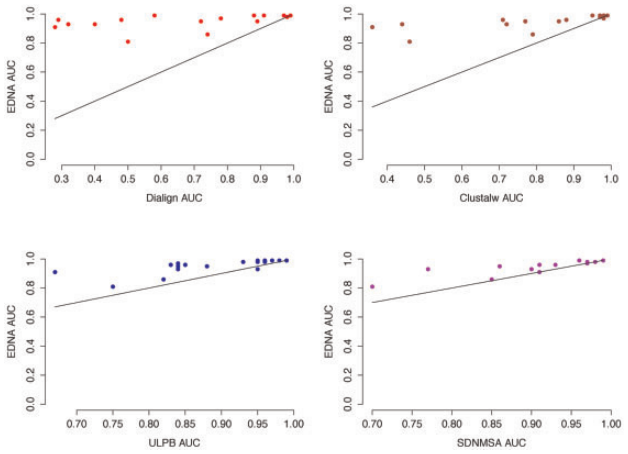

Fig. 4. Scatter plot of alignment methods. Area under ROC curve for each of Dialign, ClustalW, ULPB and SDNMSA against EDNA. Points above the line represent TFBS sequences for which EDNA outperforms the alternative alignment method. Differences in scale in each graph indicate the overall performance of the alternate method. SDNMSA shows the most fit to the expected fitness line. Dialign shows the worst fitness to the expected fitness line

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

Transcription Factor	Dialign	ClustalW	ULPB	SDNMSA	EDNA
AraC	0.29	0.88	0.83	0.93	0.96
ArcA	0.72	0.77	0.88	0.86	0.95
ArgR	0.97	0.97	0.95	0.99	0.99
CpxR	0.89	0.86	0.84	0.86	0.95
CRP	0.98	0.97	0.96	0.98	0.98
Fis	0.74	0.79	0.82	0.85	0.86
FlhDC	0.48	0.71	0.85	0.91	0.96
FNR	0.98	0.98	0.95	0.98	0.98
Fur	0.58	0.97	0.96	0.96	0.99
GlpR	0.98	0.97	0.93	0.97	0.98
IHF	0.40	0.72	0.84	0.90	0.93
LEXA	0.99	0.99	0.98	0.99	0.99
Lrp	0.50	0.46	0.75	0.70	0.81
MetJ	0.28	0.36	0.67	0.91	0.91
NtrC	0.88	0.98	0.99	0.99	0.99
PhoP	0.91	0.95	0.97	0.99	0.99
SoxS	0.32	0.44	0.95	0.77	0.93
TyrR	0.78	0.98	0.84	0.97	0.97

Table listing the area under ROC curves for TFBS prediction sensitivity and specificity corresponding to four 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).

5 DISCUSSION

We have developed a new methodology for MSAs 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 TFBS prediction for 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 dinucleotide distribution for each TFBS set 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 0 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.

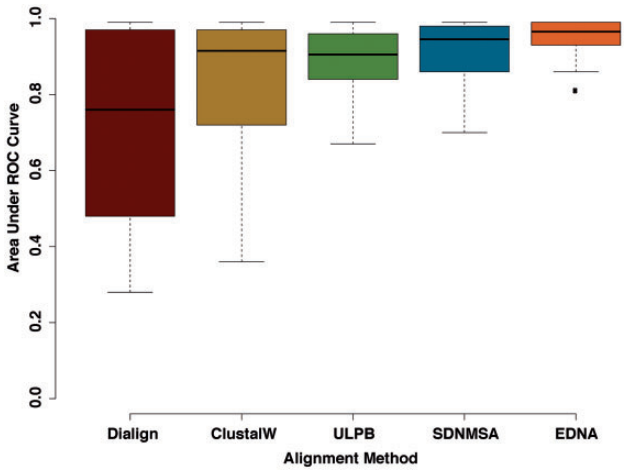

Fig. 5. Box plot for the area under curves for all five methods is compared indicating the significance of the TFBS prediction power for each of the methods

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 substitutions 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 TFBS, 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 positives are still observed in some binding sites, including Lrp, SoxS, IHF and MetJ. For well-conserved binding sites, such as LexA, ArgR, FNR, Fur, GlpR and 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, FlhDC, TyrR and MetJ, have been particularly enhanced by our new method (Supplementary Fig.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, as 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 64×64 substitution matrix with >4000 parameters, which would be challenging to compute and optimize, and most MSAs 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, 1998a, b, c, d).

ACKNOWLEDGEMENTS

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.

Funding: Rafik Salama was supported by the Darwin Trust of Edinburgh.

Conflict of Interest: none declared.

REFERENCES

Allawi,H.T. and SantaLucia,J. Jr (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry*, **36**, 10581–10594.

Allawi,H.T. and SantaLucia,J. Jr (1998a) Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry*, **37**, 2170–2179.

Allawi,H.T. and SantaLucia,J. Jr (1998b) Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. *Biochemistry*, **37**, 9435–9444.

Allawi,H.T. and SantaLucia,J. Jr (1998c) NMR solution structure of a DNA dodecamer containing single G.T mismatches. *Nucleic Acids Res.*, **26**, 4925–4934.

Allawi,H.T. and SantaLucia,J. Jr (1998d) Thermodynamics of internal C.T mismatches in DNA. *Nucleic Acids Res.*, **26**, 2694–2701.

Altschul,S.F. and Erickson,B.W. (1986) Optimal sequence alignment using affine gap costs. *Bull. Math. Biol.*, **48**, 603–616.

Bailey,T.L. *et al.* (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Calladine,C.R. and Drew,H.R. (1986) Principles of sequence-dependent flexure of DNA. *J. Mol. Biol.*, **192**, 907–918.

Cao,M.H. *et al.* (2009) Computing substitution matrices for genomic comparative analysis. *Adv. Knowl. Discov. Data Min.*, **5476/2009**, 647–655.

Carroll,H. *et al.* (2007) DNA reference alignment benchmarks based on tertiary structure of encoded proteins. *Bioinformatics*, **23**, 2648–2649.

Dayhoff,M.O. (1978) A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.*, **5**, 345–358.

Eddy,S.R. (2004) Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.*, **22**, 1.

Gama-Castro,S. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Homsi,D.S. *et al.* (2009) Modeling the quantitative specificity of DNA-binding proteins from example binding sites. *PLoS One*, **4**, e6736.

Lee,C. and Huang,C.H. (2012) Searching for transcription factor binding sites in vector spaces. *BMC Bioinformatics*, **13**, 215.

Mathews,D.H. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Morgenstern,B. (2007) Alignment of genomic sequences using DIALIGN. *Methods Mol. Biol.*, **395**, 195–204.

Mukherjee,R.E.P. *et al.* (2013) Correlated evolution of positions within mammalian cis elements. *PLoS One*, **8**, e55521.

Pauling,J. *et al.* (2012) On the trail of EHEC/EAEC - unraveling the gene regulatory networks of human pathogenic *Escherichia coli* bacteria. *Integr. Biol.*, **4**, 728–733.

Ponting,C.P. *et al.* (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.

Raghava,G.P. *et al.* (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.

Salama,R.A. and Stekel,D.J. (2010) Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. *Nucleic Acids Res.*, **38**, e135.

SantaLucia,J. Jr and Turner,D.H. (1997) Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*, **44**, 309–319.

Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.

Stormo,G.D. *et al.* (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.*, **14**, 6661–6679.

Thompson,J.D. *et al.* (2002) Multiple sequence alignment using ClustalW and ClustalX, *Curr. Protoc. Bioinformatics*, **Chapter 2**, Unit 2.3.

Thompson,J.D. *et al.* (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.

Thompson,J.D. *et al.* (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Walle,I. *et al.* (2004) Align-m—a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, **20**, 1428–1435.

Wheeler,T.J. and Kececioglu,J.D. (2007) Multiple alignment by aligning alignments. *Bioinformatics*, **23**, i559–i568.

Wu,C.F.J. (1983) On the convergence properties of the EM algorithm. *Ann. Stat.*, **11**, 95–103.

Zweig,M.H. and Campbell,G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, **39**, 561–577.