Published online 3 May 2010

Nucleic Acids Research, 2010, Vol. 38, No. 12 ***e135***
doi:10.1093/nar/gkq274

Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction

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

[1]Centre of Systems Biology, School of Biosciences, University of Birmingham, B15 2TT and [2]Multidisciplinary Centre for Integrative Biology, School of Biosciences, University of Nottingham, Sutton Bonington, Leicestershire LE12 5RD, UK

Received November 18, 2009; Revised March 22, 2010; Accepted April 3, 2010

ABSTRACT

Prediction of transcription factor binding sites is an important challenge in genome analysis. The advent of next generation genome sequencing technologies makes the development of effective computational approaches particularly imperative. We have developed a novel training-based methodology intended for prokaryotic transcription factor binding site prediction. Our methodology extends existing models by taking into account base interdependencies between neighbouring positions using conditional probabilities and includes genomic background weighting. This has been tested against other existing and novel methodologies including position-specific weight matrices, first-order Hidden Markov Models and joint probability models. We have also tested the use of gapped and ungapped alignments and the inclusion or exclusion of background weighting. We show that our best method enhances binding site prediction for all of the 22 *Escherichia coli* transcription factors with at least 20 known binding sites, with many showing substantial improvements. We highlight the advantage of using block alignments of binding sites over gapped alignments to capture neighbouring position interdependencies. We also show that combining these methods with ChIP-on-chip data has the potential to further improve binding site prediction. Finally we have developed the ungapped likelihood under positional background platform: a user friendly website that gives access to the prediction method devised in this work.

INTRODUCTION

Gene transcription is often controlled by transcription factors that bind to specific DNA-binding sites; these either promote (activate) or repress (inhibit) the binding of RNA polymerase. To fully understand a gene's functions, it is helpful to understand the regulatory network context in which the gene participates, and that includes identifying the transcription factors that regulate it. Transcription factor binding sites (TFBSs) can be determined experimentally, e.g. using DNA footprinting (1), or using high throughput techniques such as ChIP-on-chip (2) or ChIP-seq (3). However, with increased potential for high throughput genome sequencing (4), the availability of accurate computational methods for TFBS prediction has never been so important.

Computational methods for prediction of TFBSs fall into two broad classes: de novo methodologies, in which upstream regions of genes are analyzed for over-represented motifs; and training-based methodologies, in which a set of known binding sites is used to capture statistical information about a binding site in order to make predictions. De novo binding site prediction typically identifies binding site motifs without using prior knowledge of known binding sites (5). These methods can be classified as: (i) positional bias, using the concentration of a motif near the transcriptional start site (6), (ii) group specificity, comparing the localization of motifs in coding regions rather than non-coding regions (6) and (iii) least likelihood under background model (7).

On the other hand, training-based methods can be classified as: (i) consensus-based methods using the position weight matrix (8), (ii) Bayesian modeling of the binding site positions (9–11), (iii) Hidden Markov Models (HMMs) of binding site positions (12) or (iv) biophysical methods, as QPMEME (13). These methods mostly use

*To whom correspondence should be addressed. Tel: 0115 951 6294; Fax: +44 115 95 16292; Email: dov.stekel@nottingham.ac.uk

the position-specific weight matrix (PSWM) that describes the frequency of base occurrence (A, C, G and T) in each position of an alignment; QPMEME uses the binding energies between the amino acids and the DNA bases. The PSWM is computed as $P_i(x)$ for {A, C, G, T} at each position $\boldsymbol{i}$ from $f_i(x)$, the frequency of each base $\boldsymbol{x}$ among the sequences [that may include a pseudo-count to compensate for under sampling (12)]. Then if there are N sequences in the alignment (with appropriate pseudo-count correction), the proportion of symbol $\boldsymbol{x}$ in position $\boldsymbol{i}$ is given by $P_i(x) = f_i(x)/N$. Hence, given a new sequence of symbols $(x_i, \ldots, x_m)$, the simplest measure of position-specific probability associated with this sequence is:

$$\prod_{i=1}^{m} p_i(x_i) \quad (1)$$

This matrix can be also called the ungapped score matrix as it does not allow for evolutionary insertions or deletions represented by gaps in a multiple sequence alignment (MSA) into the computation of the score. The score will typically be calculated for all appropriate subsequences of an upstream region in order to identify the most likely binding sites.

Incorporating gaps into MSAs to allow representation of insertions or deletions has been found to increase the specificity of alignment models (12). Therefore, an evolutionary derived gapped model of the training sequences might provide a better prediction of the binding site likelihood. One way to achieve a gapped model of the binding site is with a HMM (12). HMMs have been used previously in research of binding site prediction to assess the likelihood of the binding site based on its statistical evolutionary profile. A zero order HMM models the sequence of bases as a Markov chain of three states (Match, Delete and Insert) as described by Durbin *et al.* (12). Transition and emission probabilities are calculated using an MSA of the training set of sequences.

Although current state-of-the-art TFBS prediction algorithms use position-specific methods, it has long been known that interactions 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 (14). This was illustrated in work discussing the effect of DNA flexure on the binding site affinity (15). Compensating mutations between neighboring DNA bases have been long known (16) and Tomovic and Oakeley have also shown that there are statistical dependences between bases and that they correlate with DNA structure (17). We have also shown using mutual information analysis that there are dependencies between neighboring and distant positions of the TFBSs that we study (see Results).

Similar ideas have been applied to analyze the splicing signals in eukaryotes (18–20). Other work includes the development of methodologies that can capture inter-position correlations using a set of training sequences (19,21,22) and apply these correlations to de novo TFBS searches (20). Bulyk *et al.* also showed that these correlations have an effect on the affinity of binding sites (23). Tomovic and Oakeley (17) have also introduced a statistical evaluator for the interdependence in binding site nucleotides based on ungapped joint probability (UJP) distributions.

The aim of this work is to improve the computational prediction of TFBSs by developing new training-based methods that incorporate base interdependencies in effective ways. We assess their performance by comparing them with position-specific approaches that are the most commonly used methods, hidden Markov models and the joint probability model of Tomavic and Oakeley. We provide further biological verification of the methods by showing that combining them with ChIP-on-chip data can improve binding site prediction. Finally, we have made this method accessible through an easy-to-use website.

MATERIALS AND METHODS

A novel method for TFBS prediction using base-pair dependencies

The core of this work is the development and evaluation of three novel TFBS prediction methods that extends position-specific methods by including information about correlated changes in neighboring DNA positions. The first method is an ungapped position-specific method that makes use of a block alignment without gaps (henceforth referred to as 'ungapped' methods); the second method is first-order HMM that is a modification of the zero order HMMs that use gapped MSAs to account for dependencies between neighboring positions in the binding sites; the third method is an enhancement to the ungapped model above by taking into consideration the positional background probability.

Ungapped likelihood

The ungapped scoring in this case is different from the normal position-specific scoring in the sense that the probability of a base in a certain position is conditional on the occurrence of the base in the previous position. That means, the probability of finding base x_{i+1} in position $i+1$ given x_i in position i is $\beta(x_{i+1}|x_i)$ which is computed as the frequency $f(x_i,x_{i+1})$ of finding the couple $x_i x_{i+1}$ at positions i and $i+1$, divided by the frequency $f(x_i)$ of finding x_i in position i, so that the conditional probability of finding a base x_{i+1} given the base x_i is given by:

$$\beta(x_{i+1}|x_i) = \frac{f(x_i,x_{i+1})+U}{f(x_i)+4U} \quad (2)$$

where U is a smoothing parameter that can also be thought of as a pseudo-count to compensate for under sampling (12). We have set $U = 0.25/n$, where n is the length of the alignment. The resulting matrix will contain $\beta(x_{i+1}|x_i)$ for all the 16 combinations of the four bases at every position. Using this model, we are able to calculate the conditional probabilities based on a training set of known binding sites and then use these probabilities to predict the binding sites in a new

sequence. Given a new sequence, the binding site likelihood is then a simple calculation of the probabilities computed over the binding site positions:

$$L_{\text{Ungapped}}(S) = p_1(x_1)\prod_{i=1}^{n-1}\beta(x_{i+1}|x_i) \quad (3)$$

where $L(S)$ is the likelihood of the sequence S of n bases, $p_1(x_1)$ is the PSWM probability of base x in position 1 (x is one of the DNA bases {A,C,G,T}).

Ungapped likelihood under positional background

The second model described in our work is an enhancement over the ungapped model that takes the background sequences into consideration. In this model the background ungapped conditional probabilities for the genome of interest (e.g. *Escherichia coli* K12 MG1655) is calculated using the entire genomic sequence so that:

$$\eta(y|x) = \frac{g(x,y)+U}{g(x)+4U} \quad (4)$$

where y and x are nucleotides {A, C, G, T}, $g(x)$ is the frequency of nucleotide x in the search sequence, and $g(x,y)$ is the frequency of nucleotides x and y at neighboring positions in the search sequence.

The binding sites likelihood ratio is now given as the ratio of the likelihood under the training sequence probabilities relative to the likelihood under the background model so that it becomes

$$\phi(x_{i+1}, x_i) = \frac{\beta(x_{i+1}|x_i)}{\eta(x_{i+1}|x_i)} \quad (5)$$

and the likelihood is given by:

$$L_{\text{ULPB}}(S) = p_1(x_1)\prod_{i=1}^{n-1}\varphi(x_{i+1}|x_i) \quad (6)$$

Throughout the work, we have used the log likelihood ratios for ease of calculation.

First-order gapped HMM

The HMM used in this work is a first-order HMM in which every match state emits only one base (i.e. probability of 1) and the transition probabilities capture all interdependencies between the binding-site bases. The HMM has the usual insert and delete states associated with Profile HMMs (Figure 1). The HMM transition and emission probabilities are calculated using training sequences with pseudo-counts and the Viterbi algorithm.

The hidden Markov state sequence for a given observation can be best found by finding the most probable path of states for a given observation, as defined by the Viterbi algorithm. Formally, the most probable path π can be found recursively. If we suppose that the probability $v_k(i)$ of the most probable path ending in state k with observation i is known for all states k, then such probabilities can be calculated for observation x_{i+1} as

$$v_l(i+1) = e_l(x_{i+1})\max_k(v_k(i)a_{kl}) \quad (7)$$

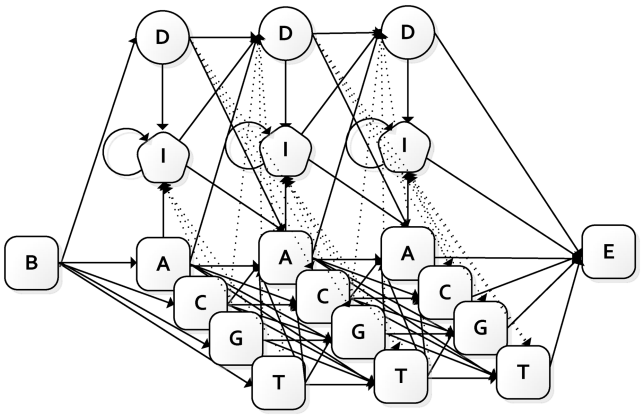

Figure 1. First-order HMM states for the DNA sequence, with four match states {A, C, G, T} emitting A, C, G or T, respectively, with probability 1. D is the delete state/silent state emitting no bases and I is the insert state which emits either A, C, G or T with equal probability. B and E denotes the beginning and end states of the HMM.

where: $e_l(x_{i+1})$ is the emission probability at state l for observation, $i+1$ a_{kl} the transition probability between state k and state l.

The Viterbi algorithm uses a dynamic programing approach to solve this problem, using the optimal substructure solution as the partial state sequence as a part of the observation. Applying the above general Viterbi equation to our first-order Markov model, then the resulting set of equations for the states in our model is as follows:

$$V_j^A(i) = 1 + \max\begin{bmatrix} V_{j-1}^C(i-1) + \log a_{C_{j-1}A_j} \\ V_{j-1}^G(i-1) + \log a_{G_{j-1}A_j} \\ V_{j-1}^T(i-1) + \log a_{T_{j-1}A_j} \\ V_{j-1}^A(i-1) + \log a_{A_{j-1}A_j} \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}A_j} \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}A_j} \end{bmatrix}$$

$\times$ (Same for the other three states C, G, T),

$$V_j^I(i) = \log\frac{e_{I_j}(x_i)}{q_{x_i}} + \max\begin{bmatrix} V_j^C(i-1) + \log a_{C_jI_j} \\ V_j^G(i-1) + \log a_{G_jI_j} \\ V_j^T(i-1) + \log a_{T_jI_j} \\ V_j^A(i-1) + \log a_{A_jI_j} \\ V_j^I(i-1) + \log a_{I_jI_j} \\ V_j^D(i-1) + \log a_{D_jI_j} \end{bmatrix},$$

$$V_j^D(i) = \max\begin{bmatrix} V_{j-1}^C(i-1) + \log a_{C_{j-1}I_j} \\ V_{j-1}^G(i-1) + \log a_{G_{j-1}I_j} \\ V_{j-1}^T(i-1) + \log a_{T_{j-1}I_j} \\ V_{j-1}^A(i-1) + \log a_{A_{j-1}I_j} \\ V_{j-1}^I(i-1) + \log a_{I_{-1j}I_j} \\ V_{j-1}^D(i-1) + \log a_{I_{-1j}D_j} \end{bmatrix} \quad (8)$$

where $V_i^Y(i)$ is the log-odds score of the best path matching subsequence $x_{1\ldots i}$ to the sub-model up to state j, ending

with x_i being emitted by state Y_i, where Y in our model can be either A, C, G, T or I. On the other hand, $V_i^D(i)$ is the log-odds score of the best path ending in a silence state D.

UJP

Tomavic and Oakeley (17) introduced a correction to the PSWM using the UJP of the dependant bases divided by the background probability of the bases. Assessment of their method has been made using in implementation of the scoring function shown in their Equation 22.

Assessing interdependency in the binding site

We have measured the interdependency between the binding site positions using the mutual information (25) between each binding site position based on the Shannon entropies at each position given by the equation:

$$I(X;Y) = H(X)+H(Y) - H(X,Y) \quad (9)$$

where $I(X;Y)$ is the mutual information between position X and position Y and $H(X)$ is the entropy at position X and $H(X,Y)$ is the joint entropy between position X and position Y.

Evaluation of prediction methodologies

In our work we have compared the training-based TFBS prediction in prokaryotic binding sites between ungapped likelihood, UJP method, ungapped likelihood under positional background (ULPB), first-order HMM and PSWM.

For each method and each transcription factor, a leave-one-out cross-validation method has been used to obtain a score for each binding site in the training set: a model is built using all of the other binding sites for that transcription factor and that model is used to obtain a score for the binding site in question. The *P*-values of the binding sites were calculated by comparing the leave-one-out scores with the distribution of model scores obtained for the genome sequence of *E. coli K12 MG1655* using a full training-set model. The calculated *P*-values were then corrected for false discovery rate (FDR; 26) and used to draw the receiver operating characteristic (ROC) curves.

The ungapped (block) sequence alignments are used as suggested by RegulonDB (27) with no gaps in the sequences. We choose the orientation of the binding sites to be cis with the regulated genes, as given in RegulonDB. MSA for the first-order Markov model was carried out using clustalw (http://www.ebi.ac.uk/clustalw/).

We have compared the performance of ULPB against both UJP and PSWM for all the binding sites in the *E. coli K12 MG1655* for which we can obtain a training set of at least 20 sequences. We have also compared three binding sites in greater detail, including a comparison of the first-order HMMs, and in two cases, ChIP-on-chip data. These are the cAMP-receptor protein (CRP), LexA and ArcA. *E. coli* has been chosen because of the large number of experimentally verified binding sites sequences available and so provides the best data to test these ideas. The known binding site training sequences in this study have been obtained from RegulonDB.

CRP. It is one of the seven "global" transcription factors in *E. coli* (28), known to regulate >100 transcription units (29). CRP's activity is triggered by binding of the second messenger cAMP in response to glucose starvation and other stresses (29). CRP binding sites have proved to be particularly noisy as the computational searching for the consensus binding site can easily miss lots of known binding sites. CRP was chosen for its high promiscuity to the transcription factors.

LexA. It directly regulates ~30 *E. coli* transcription units involved in the 'SOS' response (30). Such transcription is induced in response to DNA damage. Under normal growth conditions, LexA binds to a specific 20 bp sequence within the promoter regions of these genes, repressing transcription by sterically occluding RNA polymerase (RNAP). LexA was chosen for its lower promiscuity to the transcription factors, which should exhibit better behavior than the CRP binding site.

ArcA. It is a global regulator that changes in relation to the expression of fermentation genes and represses the aerobic pathways when *E. coli* enter low oxygen growth conditions (31). ArcA was chosen for its different protein domain (CheY like) and a very low consensus of the binding site.

ChIP-on-chip Analysis

We have used a pre-prepared ChIP-on-chip analysis of LexA from Wade *et al.* (32) and CRP from Grainger *et al.* (33), and linked it with binding sites prediction score for various methods to, first, verify the prediction capability of the method and, second, show the linear correlation between the prediction scoring function and the binding site signal (see Supplementary Data).

Genome-wide linking

For each method and each binding site, we have scored the whole genome of *E. coli* K12 MG165 using the known binding sites from RegulonDB as a training set. Scored binding sites are linked with signal from the ChIP-on-chip analysis: for each probe on the array, the highest likelihood binding site within 1 KB is used. The binding sites linked with the signal are then tested against the known binding sites from RegulonDB by selecting a cut-off both for the likelihood and for the chip signal. The known binding sites detected above both cut-offs are considered as true positives; the other known binding sites detected below the cut-offs are considered false negatives; the unknown binding sites detected above the cut-offs are considered false positives; finally, the unknown binding sites detected below the cut-offs are considered true negatives. The cut-offs were optimized simultaneously for the best sensitivity and specificity to find thresholds that maximized their product.

RESULTS

Neighboring positions in binding sites show high levels of mutual information

We have identified base dependences in TFBSs using mutual information (25) between each base pair of the TFBS sequences. In all three binding sites analyzed (CRP, LexA and ArcA), there are high dependencies among the neighboring positions (Figure 2). The CRP binding sites show high mutual information particularly between positions 5, 6, 7, 8 and between positions 15, 16, 17, 18, 19. These sites also show longer range correlations between 6, 15, 17 and 19 and strong correlation between 7 and 16 and finally a correlation between position 8, 19 and 21. The LexA binding sites show higher correlations than the CRP binding site in most of the neighboring positions. There are also many distant correlations, for example, in positions 5, 6, 7, 8, 9 with the bases before 5 and position

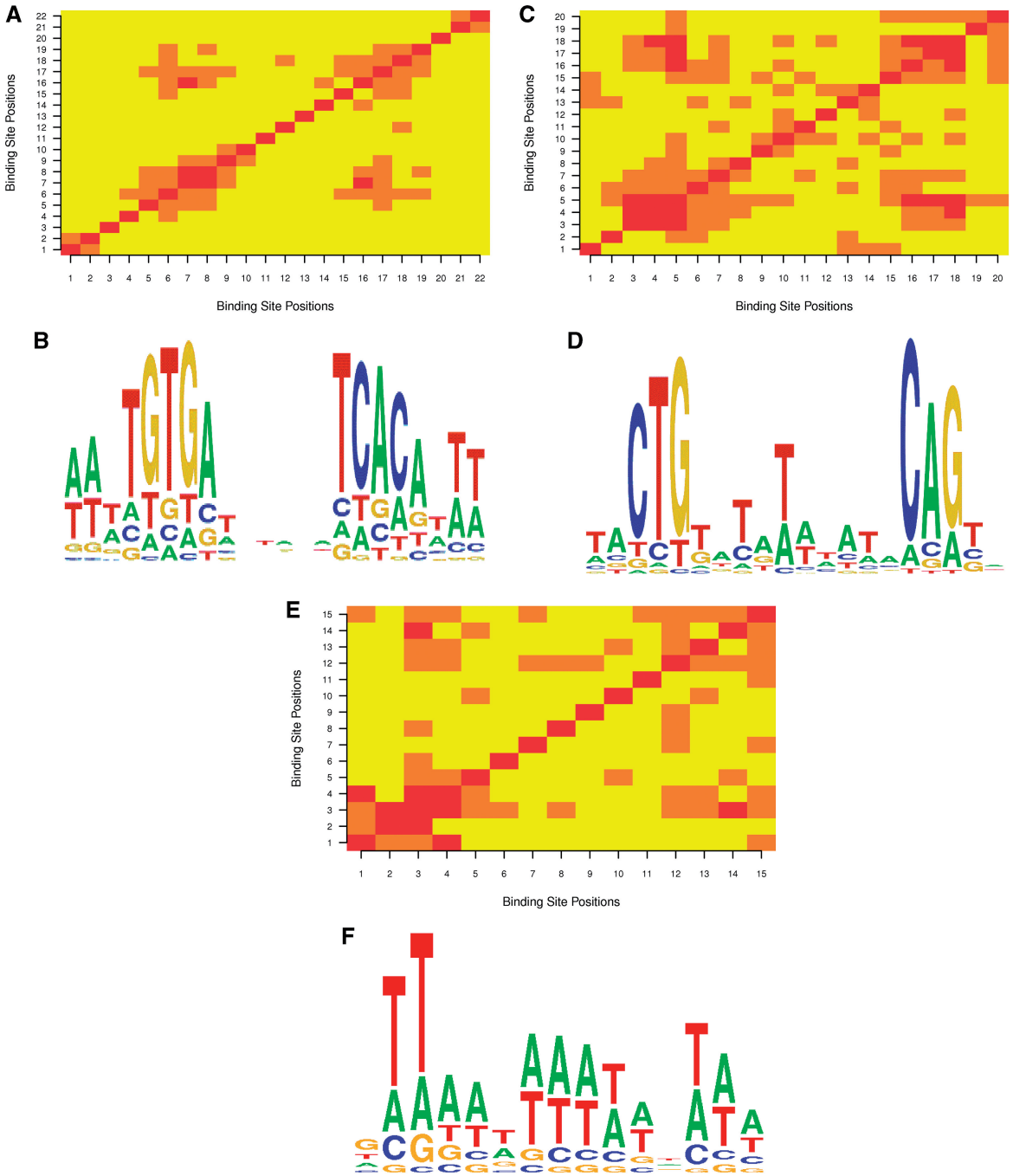

Figure 2. Heat maps and sequence logos (35) of the three binding sites under study showing mutual information between bases. Darker squares indicate higher mutual information. (**A**) shows the heat map of CRP, (**B**) Sequence logo for CRP, (**C**) shows the heat map of LexA, (**D**) Sequence logo for LexA, (**E**) shows the heat map of ArcA and (**F**) Sequence logo for ArcA. For all three genes, there are high levels of mutual information between many neighboring bases, as well as longer range interactions. Mutual information on the minor diagonal represents palindromic correlations.

5 with the bases 15 to 20. ArcA binding sites show multiple correlations in the distal and proximal five positions as well as some distant correlations between positions 3, 4 and 12, 13, 14 and 15. In all three cases, the central portion of the TFBS showed little mutual information between neighboring bases. There are also frequent occurrences of distant correlated mutations in palindromic positions. Many transcription factors, including CRP and LexA, bind as dimers (http://ecocyc.org/). Therefore, the associated TFBSs frequently consist of two similar anti-parallel sequences forming a separated, usually imperfect, palindrome. Thus correlations between the upstream and downstream portions of the TFBS are to be anticipated.

Predictions based on base interdependence outperform methods based only on position-specific information

We assessed both the ungapped models and the HMM models for the three transcription factors mentioned using training set of their known binding sites. The ROC curves (34) are shown for each binding site (Figure 3). For all three binding sites, the ULPB model shows a distinct advantage over other methodologies in predicting the binding sites.

The true discovery rate has been recorded for the binding sites tested against each method. The ULPB method shows a consistent improvement over position-specific methods and other neighboring-based methods, with area under curve 0.97 for CRP, 0.88 for ArcA and 0.98 for LexA. This is compared with the PSWM giving 0.92 for CRP, 0.82 for ArcA and 0.82 for LexA. Without positional background, the ungapped likelihood performs marginally worse, with 0.95 for CRP, 0.87 for ArcA and 0.98 for LexA. The first-order HMM shows a good performance for CRP with 0.96 and LexA with 0.98. On the other hand, first-order gapped HMM shows worst prediction for ArcA with 0.77 versus 0.82 for normal ungapped position-specific method (see Discussion). Table 1 demonstrates the area under ROC curve calculations for all of the methods. The ULPB model performs at least as well or better than all other methods for all three binding sites analyzed in detail. Analysis of all 22 binding sites demonstrates that the ULPB model performs better than PSWM in every case (Table 2), with very substantial improvements in some cases (e.g. FlhDC). ULPB substantially outperforms the UJP method of Tomovic and Oakeley in eight binding sites and is marginally better for further eight binding sites with the same performance for the other binding sites.

Combining likelihood with ChIP-on-chip signal improves prediction of known binding sites

In order to verify the predicted binding sites, a comparison has been made between computed likelihoods and ChIP-on-chip analyses of LexA (32) and CRP (33). All methods show reasonable correlations between likelihoods and ChIP-on-chip signal (see Supplementary Data). Likelihood scoring of binding sites can be further combined with ChIP-on-chip signal to improve prediction of known binding sites. The combination of ULPB and ChIP-on-chip predicts known binding sites with 73% sensitivity and 99% specificity (Figure 4). This is superior to other methods that have sensitivities of 69% (PSWM), 68% (ungapped) and 68% (first-order HMM) with similar specificities (Table 3 and Supplementary Data). For CRP, ULPB is able to predict known binding sites with higher sensitivity but with lower specificity.

ULPB platform: a web interface to the ULPB methodology

ULPB is a website giving public access to the algorithm described in this article. It predicts binding sites from a set

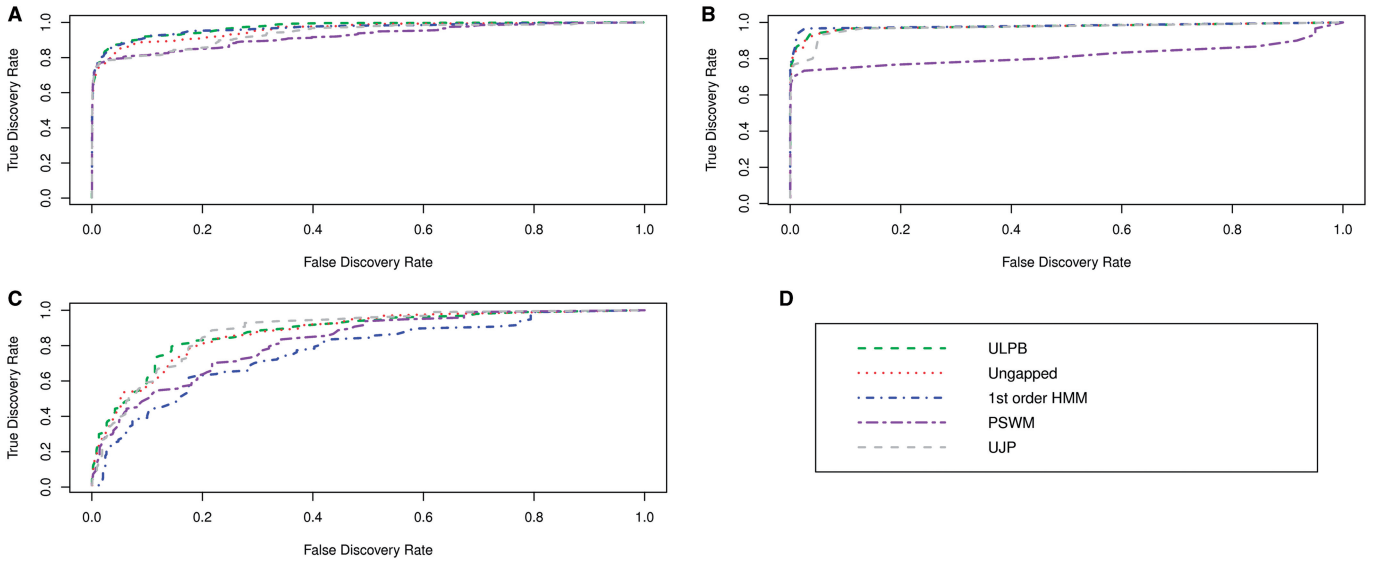

Figure 3. ROC curves for the binding sites being studied. (**A**) CRP, (**B**) LexA, (**C**) ArcA and (**D**) Figure legend. Each plot shows a comparison between Green: the ULPB, Blue: the gapped alignment scoring using Viterbi algorithm, Red: un-gapped alignment using the conditional probability, Purple: normal PSWM scoring and Grey: un-gapped joint probability. Observe that in all cases our novel ungapped method either outperforms or matches the level of all other methods.

Table 1. Area under ROC curves for all five methods applied to all three binding sites

Binding site	PSWM	First-order HMM	UJP	Ungapped model	ULPB
CRP	0.92	0.96	0.93	0.95	0.97
LexA	0.82	0.98	0.97	0.98	0.98
ArcA	0.82	0.77	0.88	0.87	0.88

Table 2. Area under ROC curves for two methods applied to all 22 regulators with at least 20 known binding sites in RegulonDB

Binding site	PSWM	UJP	ULPB
FlhDC	0.22	0.67	0.85
MetJ	0.28	0.4	0.67
AraC	0.35	0.58	0.83
OmpR	0.47	0.87	0.96
NarP	0.55	0.64	0.88
PhoP	0.62	0.97	0.97
GlpR	0.61	0.91	0.93
TyrR	0.58	0.82	0.84
NarL	0.74	0.87	0.98
LexA	0.76	0.93	0.98
NtrC	0.78	0.99	0.99
H-NS	0.54	0.54	0.65
ArgR	0.83	0.92	0.95
Lrp	0.67	0.72	0.75
SoxS	0.86	0.58	0.95
CpxR	0.78	0.84	0.84
ArcA	0.82	0.88	0.88
CRP	0.92	0.93	0.96
IHF	0.82	0.83	0.84
FNR	0.93	0.93	0.95
Fis	0.81	0.82	0.82
Fur	0.95	0.96	0.96

of search sequences based on a set of known binding sites sequences and using the ULPB method explained before. The website is integrated with xbase2 system (24) giving user access to searching >600 bacterial genomes (as of August 2009).

The website searches in three stages (Figure 5): the first stage is training, the second stage is the background scoring and third stage is for choosing the best cut-off for the binding site (default is 0.05 Q-value). A final option is motif filtering in which the returned predicted binding sites can be filtered by a user-supplied regular expression motif.

The binding site cut-off is selected with given *q*-value FDR cut-off under a set of background sequences generated from the search sequences given. The random set of sequences is generated after training a Markov chain of the transitions between the nucleotide types; in essence a Markov chain is constructed as in Figure 1.

The transition probabilities are captured from the search sequences. Starting with a random base {A, C, G, T}, these probabilities are then used to generate a sequence with the same length as the binding site.

The website is currently available on http://www.ulpb.bham.ac.uk

DISCUSSION

We have described a new methodology to score binding site likelihoods that uses interdependence between nucleotide bases and the positional background weights. We have shown that this provides a better scoring function compared to current position-specific methodologies and existing base-interdependent methods.

This method was tested in detail for three different *E. coli* transcription factors: CRP, LexA and ArcA, and against PSWM and UJP for a further 19 binding sites.

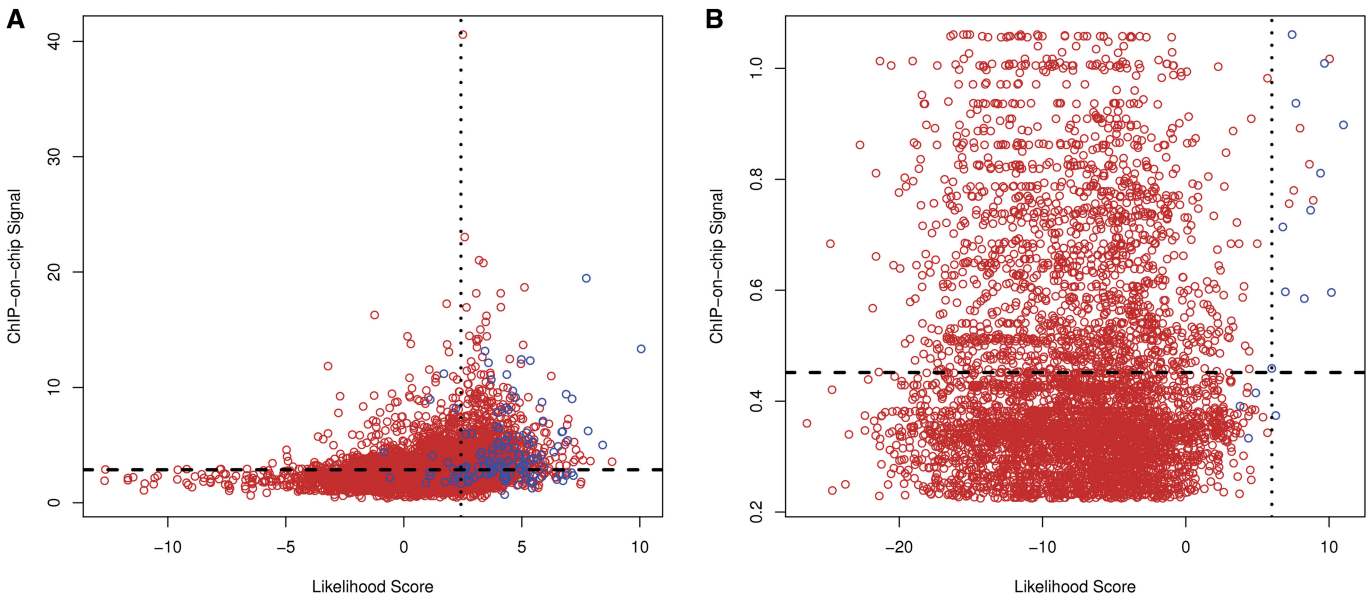

Figure 4. (**A**) ChIP-on-chip analysis of CRP linked with the whole genome showing probes corresponding to known binding sites as blue dots and other probes on the chip in brown. The horizontal line is shows the optimal signal cut-off and the vertical line shows the optimal likelihood cut-off. (**B**) ChIP-on-chip analysis of LexA linked with the whole genome; details as in (A).

Table 3. Sensitivity/specificity analysis of CRP and LexA linked with the ChIP-on-chip signal

Binding site	CRP		LexA	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
PSWM	58	86	69	99
First-order HMM	62	82	68	99
UJP	62	80	64	99
Ungapped model	57	86	68	99
ULPB	61	80	73	99

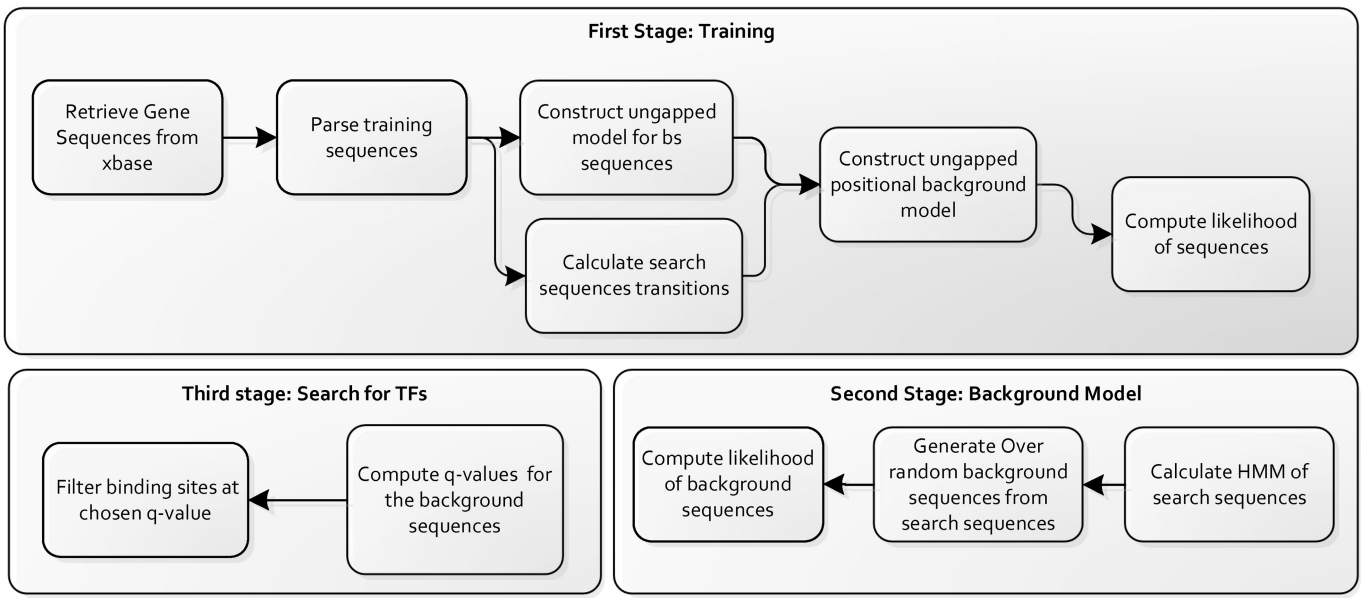

Figure 5. ULPB website passes through three stages in its process of the TFBS search. The first stage starts by computing the likelihood for the training sequences using ULPB. The second stage, a background model is generated from the search sequences and is used as a null hypothesis. The third stage determines the cut-off for the transcription factor likelihood as 5% of the background sequences, and then it scores the given search sequences and outputs the binding sites >5%.

E. coli was chosen as it is the prokaryote for which the most number of representative experimentally determined training sequences are available. CRP was chosen for its high promiscuity, ArcA was chosen for its low conservation and LexA was chosen for its high conservation. These binding sites were chosen as to represent most of the binding sites profiles. CRP and ArcA have shown a better performance on ULPB than the current position-specific, HMM and UJP methods. LexA on the other hand has shown a close performance between the methods.

The binding sites studied are all global regulators. It is difficult to apply training-based methodologies for TFBS prediction for transcription factors that regulate only a small number of genes because these methods need an appropriately sized training set to generate a reliable model. One approach to get round this could be to build a training set using known TFBSs for homologous transcription factors in closely related organisms.

The ULPB method was better than the first ungapped model since it gives higher weight for the binding site specific interdependencies versus the background interdependencies which increases the specificity of the method for the binding site against a certain set of search sequences. It has also shown improvement binding sites over the UJP method of Tomovic and Oakeley which uses joint probabilities.

The ungapped methods presented here generally proved to be a better scorer than the HMMs that include gaps that are representative of insertion and deletion events. Thus the interdependent effects are not as well captured by the evolutionary mutations included in a gapped MSA. In other words, the gapped alignment process actually disrupts the correlations between the bases forcing the HMM to select the best deletion or insertion or a nucleotide for the correlation, which introduces noise in the correlation profile. This effect is particularly apparent when comparing ArcA with LexA. The alignment introduces many more gaps in case of ArcA and almost no gaps with LexA (Figure 6), which could explain the difference between the blue curve (first-order HMM) and the green curve (ungapped) in Figure 3. Perhaps an alignment methodology that only allowed internal gaps would perform better.

The mutual information analysis also revealed that binding sites sometimes exhibit palindromic correlations.

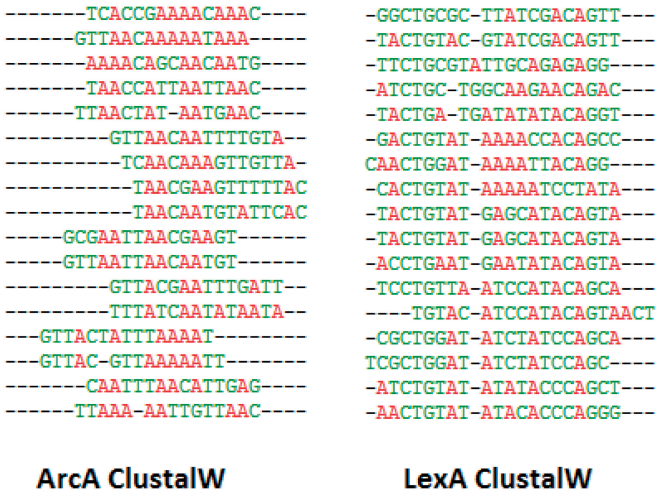

Figure 6. Clustalw MSAs of ArcA and LexA. The ArcA alignment has many gaps, especially at the start of the sequences. The LexA alignment has few gaps.

However, a model that included correlations with palindromic positions was less successful than the ungapped model presented (data not shown). It is possible that a model that uses a graph-theoretic tour of the mutual information matrix to capture long range and palindromic correlations could be more successful (21,22).

The combination of likelihood scoring and ChIP-on-chip or ChIP-seq analysis can be a powerful method for prediction of TFBSs. We have shown that for LexA, the ULPB method can help increase sensitivity of predictions without loss of specificity. The data for CRP were less conclusive; this is likely to be due to the high promiscuity of CRP for DNA increasing the noise in the CRP ChIP-on-chip data set and suggesting that complex chemical interactions contribute to the signal, beyond the consensus of the binding site alone. Thus in principle the combination of computational and ChIP techniques is potentially effective, but care needs to be taken over choice of transcription factor for analysis.

A possible extension of our work would be to use these methods to relate the sequence of the TFBS with its affinity. Such work would require a sizeable training set of measured affinities and could be useful in predicting the affinities of TFBSs for which no measurements are available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Steve Busby for his directions toward the choice of the binding sites; Joe Wade for help with the LexA data supplied; Nick Loman for help with the xBASE integration; and Jon Hobman and Selina Clayton for trialling the website.

FUNDING

Darwin Trust of Edinburgh (to R.A.S.). Funding for open access charge: University of Birmingham; University of Nottingham.

Conflict of interest statement. None declared.

REFERENCES

1. Leblanc,B. and Moss,T. (2001) DNase I footprinting. *Methods Mol. Biol.*, **148**, 31–38.
2. Aparicio,O., Geisberg,J.V., Sekinger,E., Yang,A., Moqtaderi,Z. and Struhl,K. (2005) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr. Protoc. Mol. Biol.*, **Chapter 21**, Unit 21, 23.
3. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science*, **316**, 1497–1502.
4. Hall,N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, **210**, 1518–1525.
5. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
6. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
7. Friberg,M., von Rohr,P. and Gonnet,G. (2005) Scoring functions for transcription factor binding site prediction. *BMC Bioinform.*, **6**, 84.
8. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
9. Ben-Gal,I., Shani,A., Gohr,A., Grau,J., Arviv,S., Shmilovici,A., Posch,S. and Grosse,I. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, **21**, 2657–2666.
10. Osada,R., Zaslavsky,E. and Singh,M. (2004) Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, **20**, 3516–3525.
11. Merkulova,T.I., Oshchepkov,D.Y., Ignatieva,E.V., Ananko,E.A., Levitsky,V.G., Vasiliev,G.V., Klimova,N.V., Merkulov,V.M. and Kolchanov,N.A. (2007) Bioinformatical and experimental approaches to investigation of transcription factor binding sites in vertebrate genes. *Biochemistry (Mosc)*, **72**, 1187–1193.
12. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK; New York.
13. Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.
14. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
15. Calladine,C.R. and Drew,H.R. (1986) Principles of sequence-dependent flexure of DNA. *J. Mol. Biol.*, **192**, 907–918.
16. Stormo,G.D., Schneider,T.D. and Gold,L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.*, **14**, 6661–6679.
17. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.
18. Zhang,M.Q. and Marr,T.G. (1993) A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, **9**, 499–509.
19. Agarwal,P. and Bafna,V. (1998) Detecting non-adjoining correlations within signals in DNA. In Istrail,S., Pevzner,P. and Waterman,M. (eds), *Second Annual Conference on Research*

in Computational Molecular Biology. The Association for Computing Machinery, New York, NY, pp. 2–7.
20. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
21. King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, **31**, e116.
22. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites. In Vingron,M., Istrail,S., Pevzner,P. and Waterman,M. (eds), *Proceedings of the Seventh annual International Conference on Computational Molecular Biology*. ACM Press, New York, NY, pp. 28–37.
23. Bulyk,M.L., Johnson,P.L. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
24. Choudhuri. (2004) Gene regulation and molecular toxicology. *Toxicol. Mechan. Met.*, **15**, 1–23.
25. Chouinard,J.-Y., Fortier,P. and Gulliver,T.A. (1996). SpringerLink (Online service). (1996) *Information theory and applications II 4th Canadian workshop, Lac Delage, Quâebec, Canada, May 28–30, 1995: Selected Papers*. Springer, Berlin, New York.
26. Hochberg,Y.B.Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B (Methodological).*, **57**, 289–300.
27. Gama-Castro,S., Jimenez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muniz-Rascado,L., Martinez-Flores,I., Salgado,H. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.
28. Martinez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.
29. Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.
30. Walker,G.C. (2000) Understanding the complexity of an organism's responses to DNA damage. *Cold Spring Harb. Symp. Quant. Biol.*, **65**, 1–10.
31. Nikel,P.I., Pettinari,M.J., Ramirez,M.C., Galvagno,M.A. and Mendez,B.S. (2008) *Escherichia coli* arcA mutants: metabolic profile characterization of microaerobic cultures using glycerol as a carbon source. *J. Mol. Microbiol. Biotechnol.*, **15**, 48–54.
32. Wade,J.T., Reppas,N.B., Church,G.M. and Struhl,K. (2005) Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites. *Genes Dev.*, **19**, 2619–2630.
33. Grainger,D.C., Hurd,D., Harrison,M., Holdstock,J. and Busby,S.J. (2005) Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc. Natl Acad. Sci. USA*, **102**, 17693–17698.
34. Zweig,M.H. and Campbell,G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, **39**, 561–577.
35. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.