Published online 5 April 2012

Nucleic Acids Research, 2012, Vol. 40, No. 14 ***e106***
doi:10.1093/nar/gks283

A flexible integrative approach based on random forest improves prediction of transcription factor binding sites

Bart Hooghe[1,2], Stefan Broos[1,2,*], Frans van Roy[1,2] and Pieter De Bleser[1,2,*]

[1]Department of Biomedical Molecular Biology, Ghent University, B-9052 Ghent, Belgium and [2]Department for Molecular Biomedical Research, VIB, B-9052 Ghent, Belgium

Received April 30, 2010; Revised and Accepted March 14, 2012

ABSTRACT

Transcription factor binding sites (TFBSs) are DNA sequences of 6–15 base pairs. Interaction of these TFBSs with transcription factors (TFs) is largely responsible for most spatiotemporal gene expression patterns. Here, we evaluate to what extent sequence-based prediction of TFBSs can be improved by taking into account the positional dependencies of nucleotides (NPDs) and the nucleotide sequence-dependent structure of DNA. We make use of the random forest algorithm to flexibly exploit both types of information. Results in this study show that both the structural method and the NPD method can be valuable for the prediction of TFBSs. Moreover, their predictive values seem to be complementary, even to the widely used position weight matrix (PWM) method. This led us to combine all three methods. Results obtained for five eukaryotic TFs with different DNA-binding domains show that our method improves classification accuracy for all five eukaryotic TFs compared with other approaches. Additionally, we contrast the results of seven smaller prokaryotic sets with high-quality data and show that with the use of high-quality data we can significantly improve prediction performance. Models developed in this study can be of great use for gaining insight into the mechanisms of TF binding.

INTRODUCTION

DNA-binding specificity of transcription factors (TFs) is traditionally viewed as consisting of a direct and an indirect readout component, and the proportion between them differs from one TF to another (1). The direct readout mechanism is well defined and involves recognition of specific DNA bases by amino acids. However, there is no deterministic recognition code for the interaction between DNA and protein sequences, essentially because of the influence of the three-dimensional (3D) structures of both macromolecules. The influence of the structure of the DNA-binding domain of the TF on the direct recognition code has been clearly shown for some TFs (2). If DNA-binding specificity were determined only by direct readout, then a probabilistic approach to TF–DNA recognition would suffice. The direct readout does not, however, fully explain the observed variety of sequence composition and binding affinity of binding sites for a specific TF (3). This is where the indirect readout mechanism comes in. Indirect readout is much less well defined but takes into consideration protein–DNA interactions that depend on base pairs that are not directly contacted by the protein. These protein–DNA interactions essentially reflect the influence of the structure and thermodynamic properties of the DNA before or upon binding by the TF. DNA is flexible and exhibits sequence-dependent deviations from the idealized B-DNA structure: the deviations arise from the stacking interactions of successive dinucleotides (4,5). These structural details have usually been neglected in the analysis of TF–DNA interactions: a probabilistic approach to direct readout is most commonly used as the sole component for prediction of transcription factor binding sites TFBSs, with varying degrees of success. Rohs *et al.* (6) recently emphasized the importance of the 3D structures of both macromolecules. Direct readout and indirect readout were renamed as base readout and shape readout, respectively. Base readout was subdivided according to either the major or the minor groove of the DNA, whereas shape readout was subdivided into global and local shape recognition. It was argued that individual TFs combine multiple readout mechanisms to achieve DNA-binding specificity.

*To whom correspondence should be addressed. Tel: +32 9 331 36 93; Fax: +32 9 331 36 09; Email: Stefan.Broos@dmbr.vib-ugent.be
Correspondence may also be addressed to Pieter De Bleser. Tel: +32 9 331 36 93; Fax: +32 9 331 36 09; E-mail: Pieter.DeBleser@dmbr.vib-ugent.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Methods for identifying TFBSs can be classified into two main groups on the basis of the type of data used to model the TF–DNA binding specificity. Sequence-based methods model the binding specificity from a collection of aligned sequences known to bind the TF *in vitro* or *in vivo*. Structure-based methods use information from available crystal structures of TF–DNA complexes [reviewed in Ref. (7)]. Most sequence-based methods treat DNA as a uniform static structure that is independent of the nucleotide sequence. For example, the widely used position weight matrix (PWM) method (8) takes into account only the nucleotide frequency at each position of the TFBS and assumes independence between those positions. The assumption that the nucleotides add to the binding affinity of TFs independently from each other is called the 'additivity' assumption. Based on theoretical concerns and a few experiments for some TFs (9–12), the correctness of this assumption and the quality of the approximation it yields have been discussed in the previous years (13–15). Recently, thanks to larger amounts of experimental data, it was shown that for most TFs, dependencies exist between nucleotide positions in their binding sites (16). This could be expected because it has been suggested that nucleotide positional dependencies observed within TFBSs arise from the structure and biophysical interactions of unbound and TF-bound DNA (15). Nucleotide positional dependencies are symptoms of shape readout rather than base readout. Nowadays, many sequence-based methods try to model nucleotide dependencies between positions, and thus they implicitly recognize the structural aspects of TF–DNA binding. They yield accuracy improvement over the classic PWM method for most TFs [e.g. Refs (17–20)]. A few publications present sequence-based methods that use sequence-dependent structural characteristics explicitly (21–28). Some of these methods, e.g. (25,28), report higher accuracies than those obtained by methods that model only nucleotide dependencies. Structure-based methods, by definition, take into account at least some structural characteristics of TF–DNA binding. Some of these methods are valuable for comparative modeling and they seem promising for TFBS prediction as well [e.g. (7,29)]. However, none of the structure-based methods have offered substantial improvement on the PWM method yet.

In this manuscript we present a sequence-based method that uses the random forest (RF) algorithm with features that cover either nucleotide positional dependencies or nucleotide sequence-dependent structural characteristics of the TFBS and its flanking sequences. We call the corresponding models the positional dependencies of nucleotides (NPD) model and the structural model. We also let our method combine both models and tried to integrate the PWM score in the combined model. The set of one-type models and combined models presented in this article should be seen as the products of our flexible integrative method, which can easily determine the most appropriate model to use. We measure the accuracy with which our models separate TFBSs from randomly selected genomic sequences, and we compare this measured value to the accuracy of the classic PWM method and the most recent alternative method, namely CRoSSeD (28).

Results are given for five eukaryotic TFs that bind differently to DNA: HIF1 (zipper-type group/Helix–Loop–Helix family), P53 (zinc-coordinating group/Loop–Sheet–Helix family), SP1 (zinc-coordinating group/BetaBetaAlpha-zinc finger family), STAT1 (Stat protein family) and TBP (Beta-sheet group/TATA box-binding family) (30). Our method was also used on seven prokaryotic data sets that were presented along with CRoSSeD (28) and a more recent Fis data set (31).

MATERIALS AND METHODS

Data

Positive sequences are those that are bound *in vivo* at least under some cellular conditions. They were extracted from various sources. Binding sites for HIF1, STAT1 and TBP were fetched from Pazar (32), for SP1 from TRANSFAC (licensed version 2008.4) (33), and for P53 from another paper (34). TBP binding sites were from human, mouse and rat. The binding sites for the other TFs were all human. When necessary, TFBSs were mapped back to genomic coordinates. PWMs available from TRANSFAC (licensed version 2008.4) (33) were used with the search algorithm MATCH (35) to align the fetched binding sites. These matrices were V$STAT1_01, V$SP1_Q2_01, V$TBP_01 and V$HIF1_Q3. The known TFBSs were positioned to the nearest TFBS predicted by the appropriate PWM using the TRANSFAC-given threshold values to minimize false negatives (minFN threshold values). These threshold values enable recognition of at least 90% of positive sequences, but come along with a high rate of false positives. We excluded the sequence if no predicted TFBS was found within 20 bp on either side of the position given by the database. The P53 binding sites from the paper were not re-aligned because they were already annotated in sufficient detail. We considered only P53 binding sites that were tagged as qualitative and gapless (34). In this way, our data sets of positives consisted of 55 binding sites for HIF1, 87 for P53, 243 for SP1, 209 binding sites for STAT1 and 88 for TBP. In order to assess the performance on prokaryotic data sets, we used binding sites for AraC (13 sites), ArcA (44 sites), Fis (135 sites), FlhDC (12 sites), IHF (70 sites), LexA (13 sites) and PurR (17 sites) from the CRoSSeD article (28). As an additional control for the prokaryotic data, we also used the large and qualitative ChIP-chip data set for Fis published by Cho *et al.* (31).

'Negative' or 'background sequences' are randomly selected from the human or *Escherichia coli* genome. We take 10 times as many negative sequences as the corresponding number of positives. We must provide enough negatives to ensure consistency of results, but not so many that the RF algorithm could suffer from an imbalance of the training data set, which would cause the focus to be too much on the classification accuracy of the majority class.

Structural characteristics

Structural characteristics used for this manuscript comprises characteristics calculated from scratch (see below for curvature and torsion calculations) and characteristics extracted from the literature. Most of these are correlated to some extent, but we let a feature selection procedure decide which characteristics and combinations thereof are most useful for identifying binding sites for each TF. Each DNA sequence-dependent structural characteristic is described by a list of all possible polynucleotides of a certain length, to which a numerical value describing the structural characteristic is assigned. For every characteristic, positions in a DNA sequence are scored by the value of the appropriate polynucleotide.

The calculation of sequence-dependent structural values requires an assumption of a certain 3D structure of the DNA. As we did not want to assume one specific DNA structural model, we implemented three different models: a model derived from protein-bound DNA (36), one from unbound DNA (23) and another from nucleosome-bound DNA (37,38). Each of these DNA structural models consists of values for all base-pair step parameters (roll, twist, tilt, rise, shift and slide) for each dinucleotide or trinucleotide. This enabled us to convert DNA sequences into 3D coordinates by using the rebuilding part of 3DNA (39), a program for analysis, rebuilding and visualization of 3D nucleic acid structures. For each of the DNA structural models, we did this conversion on 10 000 randomly generated sequences of 100 bp. From the resulting 3D coordinates, we then calculated the values of our structural characteristics. Values calculated for a specific structural characteristic but with coordinates coming from different DNA structural models were eventually treated as values for different structural characteristics. Curvature and torsion of the helix's axis were calculated from the coordinates of this axis only, each for the highest possible resolution. The formulas we used are as follows:

(i) Curvature: If a, b and c are three consecutive points on the helix's axis, then $\overrightarrow{U_A} = \overrightarrow{ab} \times \overrightarrow{ac}$ is orthogonal to the plane A formed by a, b and c. The curvature in b of the line containing a, b and c is given by the following equation:

$$C_b = \frac{2 \cdot \left|\overrightarrow{U_A}\right|}{\left|\overrightarrow{ab}\right| \cdot \left|\overrightarrow{bc}\right| \cdot \left|\overrightarrow{ac}\right|}$$

(ii) Torsion (dihedral angle): If a, b, c, d are four consecutive points on the helix's axis, then $\overrightarrow{U_A} = \overrightarrow{ab} \times \overrightarrow{ac}$ is orthogonal to the plane A formed by a, b and c, and $\overrightarrow{U_B} = \overrightarrow{bc} \times \overrightarrow{bd}$ is orthogonal to the plane B formed by b, c and d. Then the dihedral angle is given by the following equation:

$$T_{AB} = \cos^{-1} \frac{\left|\overrightarrow{U_A} \cdot \overrightarrow{U_B}\right|}{\left|\overrightarrow{U_A}\right| \cdot \left|\overrightarrow{U_B}\right|}$$

These calculations provide a value for every base position. However, this value is calculated with coordinates of more than just this one base (see equations above) and these coordinates are dependent on the identity of neighboring bases. We sought to determine an accurate relation between sequence and calculated structural values, and so we took the shortest length of polynucleotides for which the relative standard deviation on the corresponding mean structural value was $<1\%$. This polynucleotide length is 3, 4 or 5, depending on the characteristic and the DNA structural model. The calculated values of sequence-dependent structural characteristics (curvature and torsion of helix's axis) are available from the authors upon request. Other structural characteristics used in this manuscript were extracted from the 'literature' and comprise properties derived from either unbound or TF-bound DNA. They are all given as a value per dinucleotide, mostly with a considerably large standard deviation. The standard deviations, and their lack when expanding to polynucleotides longer than two bases, indicate that the structural characteristics of base-pair steps depend on the identity of neighboring nucleotides. Although we used higher nucleotide lengths having nearly no standard deviation on their mean value for the structural characteristics we calculated ourselves, the calculation was still based on the assumption of DNA structural models described by only dinucleotides or trinucleotides. The structure of a dinucleotide is known to be influenced by the identity of the neighboring nucleotides (27,40,41), and taking into account these next-nearest-neighbor effects might further improve the accuracy of the structural model. A description of the structural characteristics we used is given below.

'Curvature' and 'torsion' describe the DNA backbone in its highest resolution and thus provide at least a measure of bending. The characteristic we implemented, 'directed bending', does the same (42). Directed bending means the extent to which a dinucleotide tends to bend towards either the major or the minor groove when it is bound by a TF, and it is used as a measure of deformability of DNA. Values are determined on sequences bound by the TF CAP at sites where sequence dependence of bending is maximal (42). Pre-bending of free DNA (43) and TF-induced bending (44) have been recognized for more than a decade as structural motifs common to many TF–DNA complexes. 'Groove clash distance' and 'size' are both components of the clash function that was constructed to give a quantitative interpretation of the observed sequence dependence of TF–DNA interactions on DNA twist (45). A steric clash between exocyclic groups results from out-of-plane base-pair distortions. Its size is defined as the sum of the radii for the exocyclic groups interacting in the grooves. Clash distance is the distance between the centers of the interacting groups when they are in an 'idealized' conformation. Different geometries of the major and minor groove are taken into account and result in separate values per groove type (45). Groove shape is an interesting characteristic to explore because it was recently acknowledged that most TFs recognize the minor groove width upon specific binding (46). The value of groove width for

prediction of TFBSs was suggested by Liu *et al.* (23) in 2001. 'Minor groove opening' is a measure of the degree to which a base step is open in the minor groove, and hence it is related to the above-mentioned measure of groove clash size. The values are derived from high-resolution crystal structures of unbound DNA in BI conformation (23). 'Conformational tendency' is measured by the standardized Pearson residuals for the test of uniformity or homogeneity of the individual dinucleotide steps over different conformations, i.e. structural types of DNA (47). These values are derived from unbound DNA and represent the tendency of a dinucleotide to favor a specific DNA conformation. Uniformity of dinucleotides is tested between A-type, B-type and combined conformational families (A, B and A + B conformations) and within B-types of DNA (BI, BII, A/B, B/A, RESTB). RESTB is not assigned to any of the existing conformational families. We did not use the conformational tendencies of dinucleotides within A-forms of DNA because the dinucleotide AA/TT does not occur there (47). Almost one-third of dinucleotides from protein–DNA complexes adopt AI or AII conformations. This plasticity of DNA, which allows the conformation to change locally from the common B-form into an A-form, is one of the ways in which DNA achieves specificity in protein–DNA binding (44,48,49).

Random Forest algorithm

The RF algorithm (50) (http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm) is a tree-based machine-learning algorithm and is the engine of both our structural method and our NPD method. It is an ensemble classifier that consists of many individual decision trees (CARTs: classification and regression trees) and outputs the class that is predicted by the majority of those trees. Tree-based methods consist of non-parametric statistical approaches for regression and classification analyses. Classification trees are grown by recursively partitioning the observations into subgroups with a more homogeneous categorical response. At each node, the explanatory variable giving the most homogeneous subgroups is selected. For the CART tree learning algorithm, this selection is based on Gini impurity, which is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.

Tree-based methods can be very effective for selecting from large numbers of predictor variables, those that best explain the observations. They make no implicit assumptions about the form of underlying relationships between the predictor variables and the response, and so they might detect non-linear associations. The RF methodology forms an ensemble of unpruned classification or regression trees (CARTs) by bootstrapping samples of the training data and using random feature selection in the tree induction process. It generally exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5. The biggest disadvantage of RF is that its embedded feature selection procedure cannot handle large numbers of irrelevant features. For this reason, we performed a comprehensive filter feature selection and wrapper-based feature selection before the final model is trained (see next section). We used FastRandomForest (http://fast-random-forest.googlecode.com/), a parallelized implementation in Java. For further information, we refer to two publications that provide excellent explanations and examples on the use of RF for modeling dependencies among variables (51,52).

Building classification models

In the first stage of building a classification model, one model per characteristic is built. The structural method uses the above-mentioned structural characteristics, whereas the characteristics of the NPD method are represented by mononucleotides and dinucleotides. Hence, each sequence from the positive and negative set is converted to a series of structural vectors or is split up into mononucleotides or dinucleotides (Figure 1A and B).

We perform a comprehensive feature selection in order to obtain the final model. A first round of feature selection is performed in a purely statistical way to make a basic selection of positions where a difference exists between the values for the characteristic of the positives and those of the negatives (so-called filter feature selection) (Figure 1C). The statistical tests are applied with mild threshold values in order not to exclude too many features and to permit detection of their interactions by the RF algorithm later on. For the structural model, we consider values for all positions in the TFBS and for the 30 bases flanking it, as well as the mean value over all these positions, as features to be used in building the model. The Kolmogorov–Smirnov test at a false discovery rate threshold of 0.1 is used to determine the significance of differences between values at each position. The Wilcoxon rank test at a threshold of 0.05 is used to determine the significance of differences between values averaged over all 60 positions. For the NPD model, 30 mononucleotides flanking the TFBS start on both sides are considered. The basic selection of positions at which the mononucleotide distribution is different between positives and negatives is determined by the test for equality of proportions. More specifically, a position is selected when the sum of the logs of the *P*-values of proportion tests is significantly different from the background using a threshold of 0.1.

In the second round of feature selection, the preliminary model based on one characteristic is subjected to wrapper-based feature selection (Figure 1D). We repeatedly evaluate the accuracy of the model by cross-validation with the RF algorithm and remove features of the basic selection when this does not cause a significant decrease in accuracy (measured as either *F*-measure or AUC). AUC (area under the curve) represents the area under the receiver operating characteristic (ROC) curve, whereas *F*-measure is the weighted harmonic mean of precision and recall. This procedure of removing insignificant features is also called sequential backwards elimination

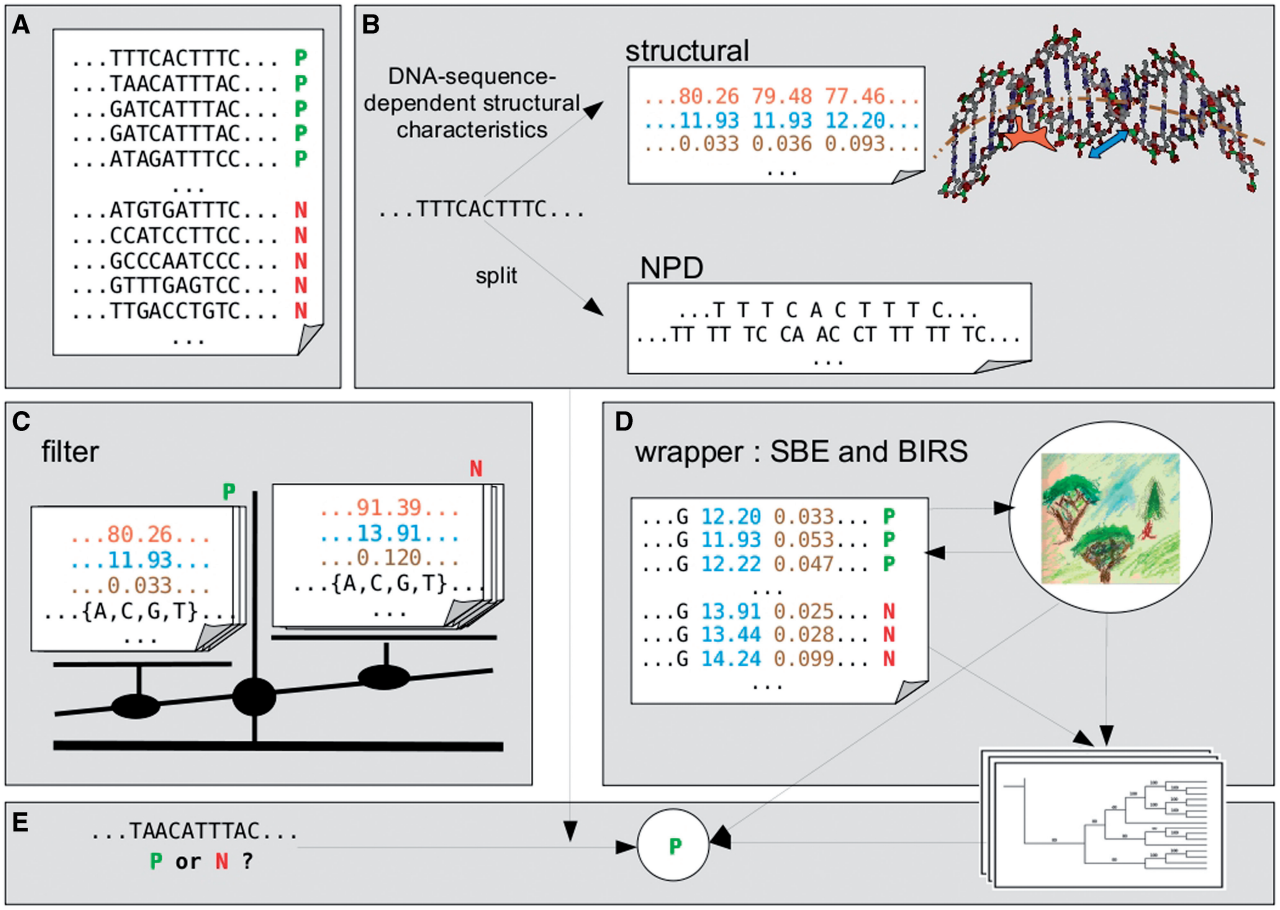

Figure 1. Overview of our approach: (**A**) The input from which models are built consists of the two classes of nucleotide sequences that the method should learn to separate. One class contains positive sequences (P, green) known to be bound *in vivo*; the other contains negative sequences (N, red) highly unlikely to be bound *in vivo*. (**B**) Each nucleotide sequence, from either class, is converted into multiple series of values; each series provides values for a specific DNA structural characteristic at all positions of the TFBS and its context (structural model), or simply consists of one base or two base parts of the sequence (NPD). (**C**) Basic selection of relevant features (i.e. positions) is made by statistical comparison of distributions of values for positive and negative sequences with mild thresholds. (**D**) Further selection is performed through wrapper-based feature selection, i.e. cross-validation performance evaluation with the RF algorithm. Per characteristic, redundant features are removed by sequential backwards elimination (SBE). Several models with one characteristic might be merged through BIRS. The final NPD model and final structural model can be merged into one integrative model. (**E**) The resulting model can be used by RF to predict the likelihood that a nucleotide sequence is a TFBS, after converting the sequence into series of the features contained in the model.

(SBE). It makes the model sparser, which permits better interpretation of the features it contains and which improves speed upon application.

At this stage, we end up with one model per characteristic. We rank all models according to their classification accuracy as determined by cross-validation (measured as AUC). Starting with the best performing one-characteristic model, we cumulatively merge it with lower-ranked models according to the best incremental ranked subset (BIRS) scheme (53); this implies the use of wrapper-based feature selection.

Combined models, i.e. models that contain characteristics from two or three different categories (NPD, structural or PWM score) are simply built by merging two or more models that are restricted to one category. The process of finding the combination that gives the best model can be easily automated by an extra round of wrapper-based feature selection.

When building PWMs for the eukaryotic sets, we automatically assigned their lengths by requiring that the start is on the assumed start position of the TFBSs and the end is characterized by three consecutive positions with an information content of at least 1.1. For the prokaryotic sets, it was necessary to use the entire sequence length for the PWM.

Evaluation of classification models

The evaluation of classification models is based on their prediction scores and provides an estimation of the accuracy of their classification. The prediction score of both the structural method and the NPD method is the RF confidence score, which is assigned to each sequence and indicates the certainty with which this sequence is predicted to belong to either the positive or the negative class. In the case of PWMs, we used the matrix similarity score (35). The evaluation of performance is visualized by

ROC curves and precision-recall curves. Each ROC and precision-recall curve shown is derived from a threshold-based average of 20 curves. Data for each of these 20 curves were obtained by training the model with a randomly taken subset of 80% of the data and testing that trained model on the remaining 20%. Principle component analysis was performed on the full models using the Weka 3 suite (54) and used to select a top five feature set for each TF (default parameters).

RESULTS

Based on the RF algorithm (50), we initially built two types of models. The so-called structural model uses one or more structural characteristics by employing their values at specific positions or their average value over all positions in the TFBS and its flanking sequences. The so-called NPD model accounts for positional dependencies at the nucleotide level, utilizing only nucleotide identities (mononucleotides and dinucleotides). The procedure of building and using these models is depicted in Figure 1 and explained in detail in the 'Materials and Methods' section. We start by discussing the classification accuracy of the classic PWM method, the structural method, the NPD method and combinations thereof, and compare our integrative method with a recent alternative method. This evaluation is performed on five high-quality eukaryotic data sets and eight prokaryotic data sets. Seven of these prokaryotic data sets are rather small and less well annotated. This led us to introduce a second, more qualitative Fis data set in order to assess the influence of data quality on the performance of the different methods. As an additional confirmation of the validity of the RF method, we evaluate the integrative TBP model on external data. Finally, we look at the selected features in each model and try to relate these features to what has been reported in the literature.

Classification accuracy

The ROC curve is a standard representation of the trade-off between false positive rate (FPR) and sensitivity. We use details of ROC curves to visualize the classification accuracy of the models. Regular ROC curves and their corresponding measure AUC cover the full range of FPRs from 0 to 1 and are thus of not much use for estimating the discriminatory power of a predictor of TFBSs (55). Genome-wide predictions performed with an FPR even as small as 0.01 are not really useful because they would return an overload of false positives, e.g. about 6 million for the human genome. Therefore, we focus on the part of the ROC curves, which corresponds to the lower, more relevant range of FPR. We also take our most integrative model as a reference model and for each model we list the FPR that corresponds to the true positive rate (TPR) that has an FPR of 0.01–0.1 for this reference model, corresponding to the bending point of the curves. Statistics of pair-wise comparisons of these FPRs are provided as well. We compare our models with each other and also compare their accuracy with the accuracy of our home-made high-quality PWMs and with the most recently proposed alternative method, CRoSSeD (28). The latter comparison will be discussed extensively in the next section.

For the eukaryotic transcription factors (Figure 2 and Supplementary Table S1), both structural and NPD models perform better than the PWM for four out of five TFs (HIF1, SP1, STAT1, TBP). Overall, the NPD model performs better than the structural model (four out of five cases). This is logical because the structural method almost exclusively captures the shape readout mechanisms of DNA-binding specificity. All base readout information gets lost upon conversion from a nucleotide sequence to vectors of structural characteristics. The NPD model, in contrast, is expected to capture base readout, as well as some portions of the shape readout that can be derived from nucleotide positional dependencies. Nevertheless, the structural models alone perform surprisingly well: they perform better than PWM in four out of five cases. For most eukaryotic TFs, merging the structural model with the NPD model leads to clear synergistic effects and achieves a classification accuracy that is superior to the accuracy of the separate models and PWM ('NPD_struct'). For three out of five eukaryotic transcription factors, inclusion of the PWM score even led to an additional improvement ('NPD_struct_PWM'). The RF strategy significantly improved upon the PWM method for all eukaryotic TFs (Supplementary Table S1).

For most prokaryotic models (Figure 3 and Supplementary Table S2), the NPD model and the structural model do not outperform the PWM. When considering the low-resolution prokaryotic data sets alone (Figure 3A–G), the structural or NPD model, or combinations thereof, perform better than the PWM model for only three out of seven TFs (ArcA,FlhDC and IHF). Combining the NPD model and the structural model leads to an improvement in five out of seven cases when compared with the individual models. Adding the PWM score did not result in an additional improvement, except for AraC. Compared with the other prokaryotic models, the high-quality Fis model performs exceptionally well (Figure 3H). This result clearly demonstrates the importance of using qualitative data when building classification models.

As an additional test, we also looked into precision-recall curves of the classification models for a growing number of background sequences (Supplementary Data S1). With this type of analysis, we tested the models for their ability to cope with a growing number of background sequences. For each TF we compared the combined RF model with the PWM for 10 different background sizes. We started with a 1:1 ratio and augmented the number of background sequences until we had a 1:10 ratio. Models that are less suited to cope with many background sequences show a sharper decline in the precision-recall curves when facing more negative sequences. The prokaryotic models gave mixed results. Again, the high-quality Fis model performs exceptionally better than the other prokaryotic models. The RF models of ArcA and IHF perform equally well as the PWM, whereas the rest of the TFs did not benefit from the more complex RF model. However, unlike the

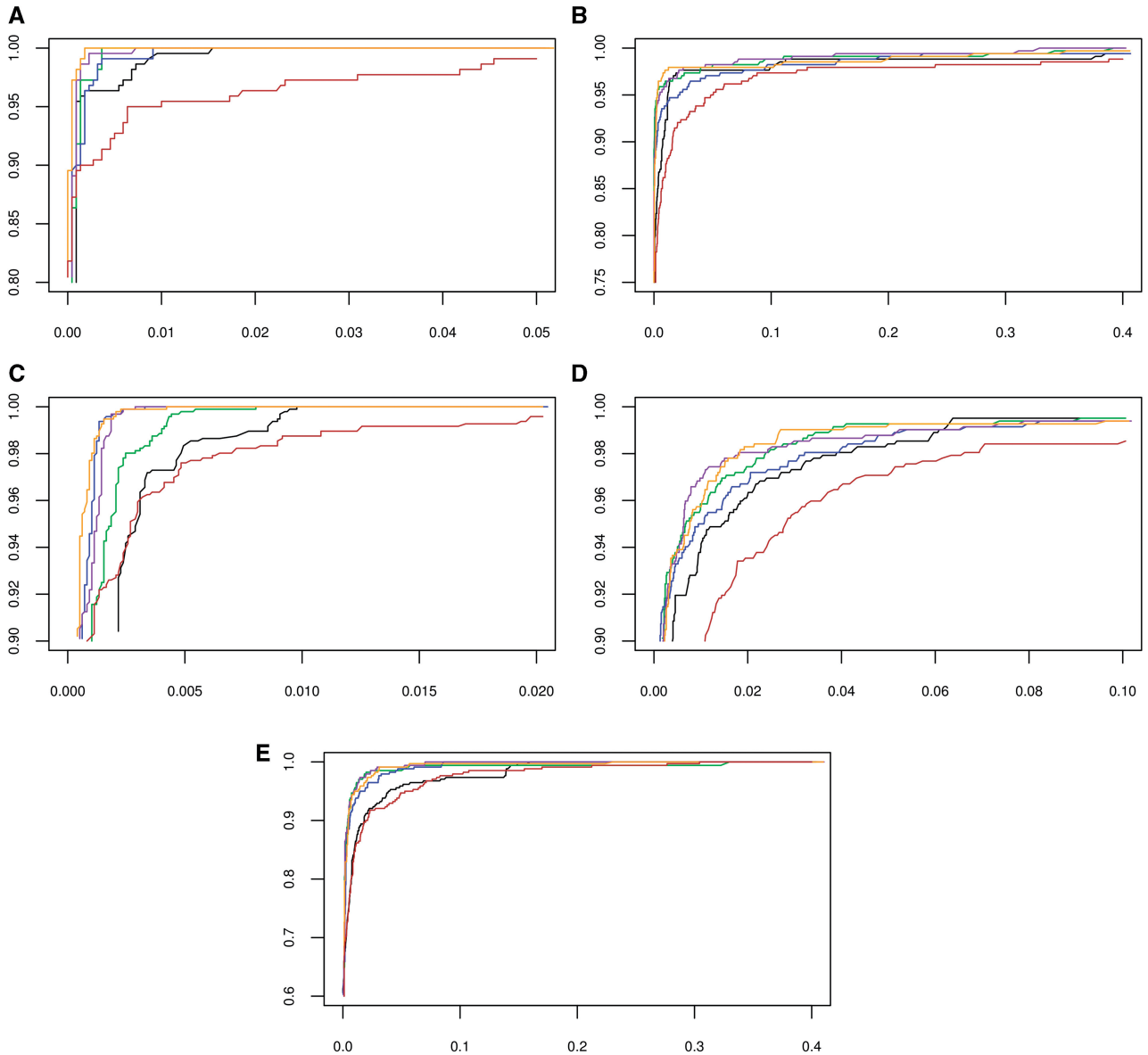

Figure 2. Accuracy of classification models in identifying TFBSs, as assessed for five eukaryotic TFs. Details of threshold-averaged ROC curves showing the trade-off between TPR (*Y*-axis) and FPR (*X*-axis); Classification models applied: PWM (black), NPD (green), struct (blue), NPD_struct (purple), NPD_struct_PWM (orange), CRoSSeD (brown). (**A–E**) ROC curves for various transcription factors: (A). HIF1 (B) P53; (C) SP1; (D) STAT1; (E) TBP.

prokaryotic models, the eukaryotic models gave consistent results. For all five eukaryotic TFs, the RF model turned out to be more robust against a growing number of background sequences compared with the simpler PWM model.

The difference in classification performance between the two Fis sets is striking (Figure 3H and Supplementary Table S2). The results indicate that with the high-quality Fis set, the RF model can improve upon the PWM method. In this case, NPD_struct_PWM is the best model and it is significantly better than all other models. It is clear that the overall classification accuracy of all the methods we compared is much better with the more reliable Fis data set. We speculate that lack of improvement for the RF models in the majority of prokaryotic sets is due to their relatively small sizes and poor quality of annotation, as is illustrated with this example.

Comparison with alternative sequence-based methods

A comprehensive overview of alternative sequence-based methods is given in Supplementary Data S2. Differences between our method and others includes accounting for the context of the TFBS, the use of several structural characteristics instead of just one, the use of structural values for specific positions rather than just the average value along the TFBS, the use of both structural characteristics and nucleotide positional dependencies, and the use of the RF algorithm. RF does not require any assumptions

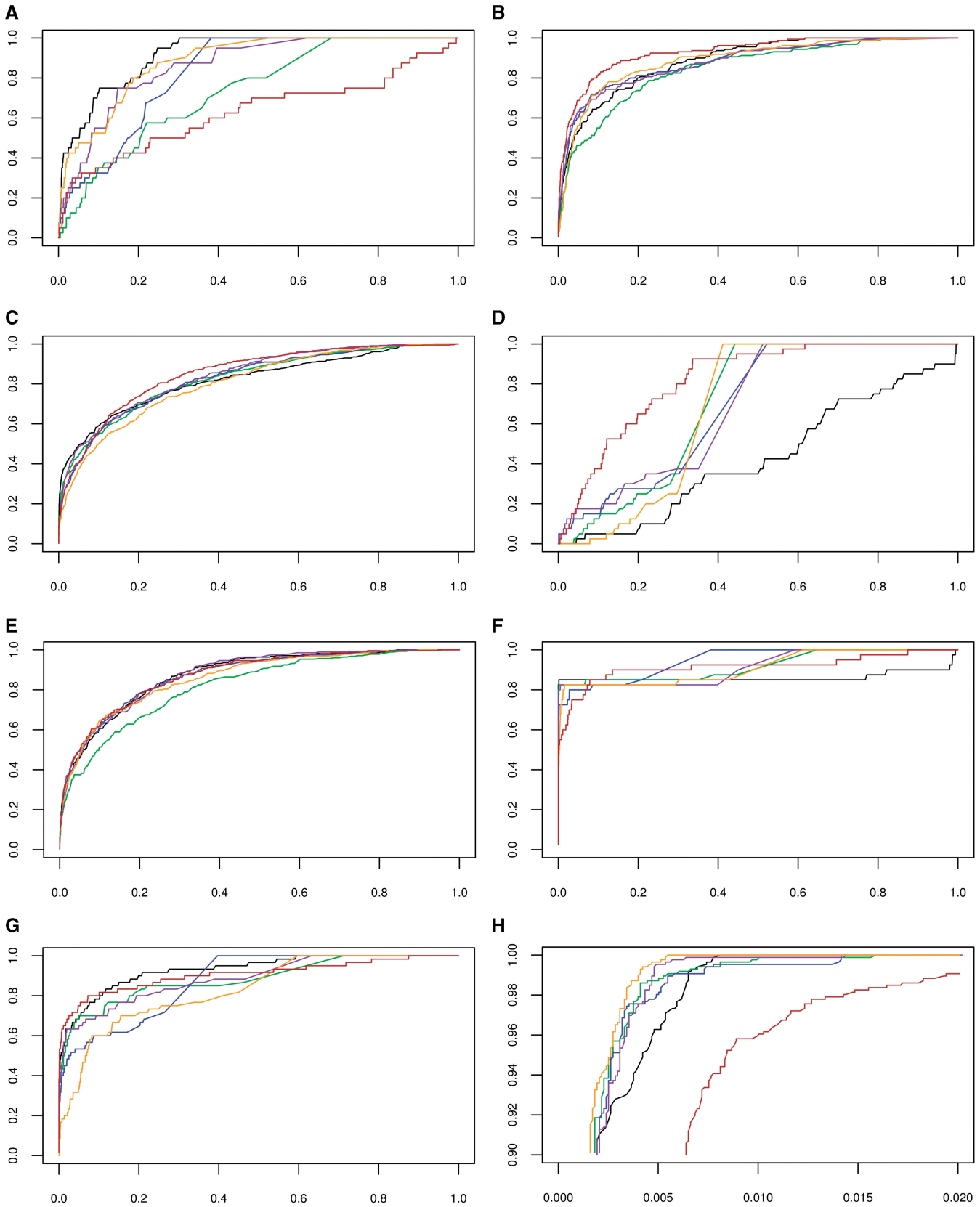

Figure 3. Accuracy of classification models in identifying TFBSs, as assessed for eight prokaryotic TFs. Threshold-averaged ROC curves showing the trade-off between TPR (*Y*-axis) and FPR (*X*-axis); Classification models applied: PWM (black), NPD (green), struct (blue), NPD_struct (purple), NPD_struct_PWM (orange), CRoSSeD (brown). (**A–H**) ROC curves for various transcription factors: (A) AraC; (B) ArcA; (C) Fis; (D) FlhDC; (E) IHF; (F) LexA; (G) PurR; (H) Fis [ChIP-chip set (31)].

about the form of underlying relationships between the predictor variables and the response. Hence, there is no need to assume independence or uniform contribution of multiple structural characteristics. Some other sequence-based methods use additional types of data to reduce the FPR of TFBS prediction, such as phylogenetic conservation (56), genome annotation [e.g. Refs (57,58)] or specific experimental results [e.g. Ref. (59)]. We only consider sequence-based methods not needing such additional information as methods comparable with ours. Some of these methods are SiteSleuth (27), promapper (25) and CRoSSeD (28). Each of them is based on a different classification algorithm, namely, support vector machine, Bayesian network and conditional random field, respectively. Furthermore, base readout and shape readout are captured in slightly different ways (e.g. other structural characteristics) and do not get equal chances due to arbitrary decisions. We conclude that with the exception of CRoSSeD (28), none of all previously presented methods have made clear comparisons to show how accurately their method identifies TFBSs compared with methods modeling dependencies between nucleotide positions, and that CRoSSeD is the current best performing alternative method. Here, we clearly show the worth of each of the 'pure approaches' (PWM, nucleotide positional dependencies, structural), and we show that integration of different approaches is beneficial to classification accuracy. We performed a quantitative comparison with the most recent alternative method, namely CRoSSeD (28). We compared our method with CRoSSeD both on the prokaryotic data set from the CRoSSeD article and on our eukaryotic data sets. The results on the eukaryotic data sets are depicted in Figure 2. For all eukaryotic TFs, CRoSSeD separates TFBSs from non-TFBSs less accurately than the PWM. Our integrative model ('NPD_struct' and 'NPD_struct_PWM') performs significantly better than CRoSSeD for all eukaryotic TFs.

The prokaryotic data sets that were used originally come from RegulonDB (60) and are remarkably different from the eukaryotic data sets we used. Most of the prokaryotic data sets show very little sequence conservation and only expose weak signals over a long distance [see Supplementary Data of Meysman *et al*. (28)]. The lack of strong nucleotide conservation in most prokaryotic data sets might have caused CRoSSeD to be developed with a different focus from our RF models. The different natures of the prokaryotic data sets are reflected by a much lower level of classification accuracy of the predictors and we were forced to list the FPR that corresponds to the TPR with an FPR of 0.05 or even 0.1 for the reference model 'NPD_struct_PWM', instead of the 0.01 used for the eukaryotic data sets (Supplementary Table S2). Our ROC curves and some conclusions differ from those shown in the paper presenting CRoSSeD (28). The different results must have been caused by differences in the evaluation setup. Many papers, including Meysman *et al*. (28), measure accuracy by the area under the ROC curve (AUC), but differences of its value might be irrelevant or even misleading, depending on the shapes of the ROC curves. Both CRoSSeD and our integrative method are among the best models in three out of seven cases (Figure 3A–G), but what is truly remarkable is that the PWM proves to be the best model in three out of seven cases when considering low FPRs only. We also compared our methods with the CRoSSeD method on the high-quality prokaryotic Fis set (Figure 3H). With this data set, the performance of all methods improves drastically. The RF method performs best, while the CRoSSeD method lags behind. These results make clear that data quality is an important determinant of model performance.

From both comparisons with CRoSSeD, we conclude that our approach performs better overall. The small prokaryotic data sets did not fully meet the requirements of our qualitative approach to evaluation of models, and hence conclusions should be made carefully.

Table 1. Performance of the TBP model on external ChIP-seq TBP data set (Mokry *et al*.), measured in ROC AUC

	PWM	RF model	CRoSSeD
ROC AUC	0.535	0.774	0.573

Evaluation of a model on external data

The seemingly small improvements in accuracy presented here may nevertheless make a huge difference when identifying TFBSs on large DNA sequences and genome-wide. Furthermore, it is interesting to evaluate models on data that do not originate from the same data set with which the models were built. In order to evaluate our method on external data, we tested the TBP model on an independent chIP-seq experiment for TBP (61). This is a very demanding test, since the models need to identify the TBP binding site in a wider peak region of the chIP-seq experiment. The same is then repeated for a background with the same length distribution. In Table 1, we compare the PWM method, our integrated model (containing structural and NPD characteristics) and the CRoSSeD tool in terms of ROC AUC for classification of sequences containing *in vivo* TBP binding sites and background sequences. Results clearly show that the PWM (AUC 0.535) and CRoSSeD (AUC 0.574) can barely discriminate between the TBP peaks and the background model, whereas our integrated model fulfills this task much better (AUC 0.774).

Features contained in the models

Supplementary Table S3 shows the features of the RF models. These features can reveal aspects of the DNA–TF binding mechanism. Even though the prokaryotic models do not perform that well in terms of classification, the selected features can tell us something about the binding mode of these TFs. All TFs have different models with different characteristics, representing their DNA-binding specificities. The structural characteristics are correlated to some extent (Supplementary Figure S1), but we let the feature selection procedures and the RF algorithm decide which features are most relevant for each TF. It should be noted that for each TF both

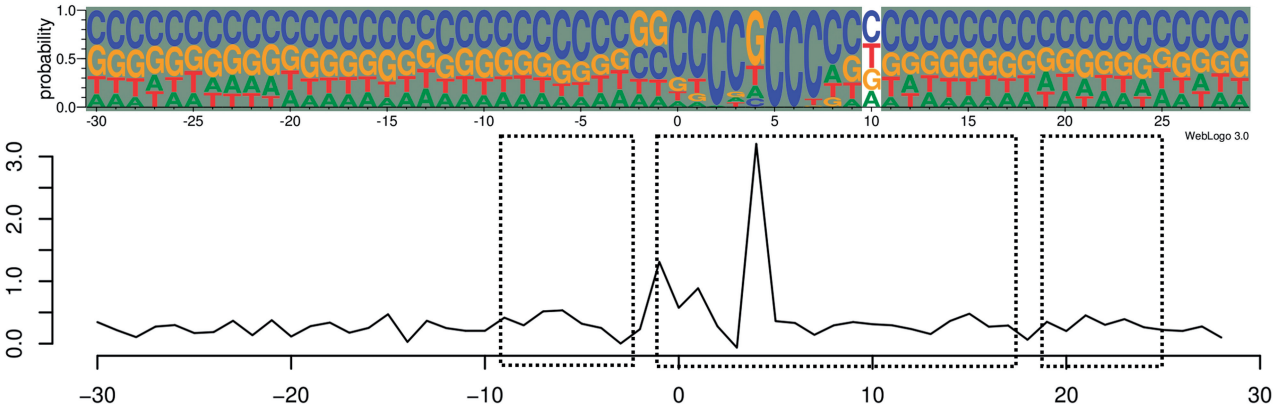

Figure 4. Visualization of our integrative model for SP1. Top: mononucleotide frequencies with the positions of the NPD model shown as shaded boxes. Bottom, average value of one of the structural characteristics contained in the structural model, namely conformational tendency restB; positions of the structural model are indicated by dotted-line boxes (*X*-axes indicate position relative to the aligned start of the SP1 binding sites).

the structural model and the NPD model include features at positions that precede the actual TFBS. Moreover, each model contains one or more mean values as feature which implies that the global structural *in vivo* context of the TFBS is an important feature next to more local shape readout mechanisms at or close to the binding site location. This global shape readout might reflect the general part of higher order protein–DNA interactions that determine binding specificity and functionality: the tendency of a nucleosome to bind the region in which the TFBS is embedded (6). It might thus be considered part of a so-called 'general binding preference' that was demonstrated to be important for improved prediction of TFBSs (57). A visualization of the SP1 model (Figure 4) clearly shows how the background genomic sequence in which SP1 binding sites are embedded is very similar to the consensus sequence of such sites. A PWM would thus predict many TFBSs, whereas the NPD model and structural model can look beyond position-independent nucleotide frequencies, each in its own way. In the next section, we will describe the most important features of each model, together with their biological relevance.

Biological relevance of the selected features

To assess the biological relevance of the selected features, we decided to do a principal component analysis (PCA) on the different TF models. For each model, we selected the top five principal components (Table 2), meaning the five most relevant features according to the PCA. We relate all of the selected features to what is known in the literature about structural protein–DNA complex formation. Unfortunately, for torsion-related features, we were unable to find explanations in the literature because this feature is not discussed in most protein–DNA reports. The PWM score is an important feature in most models and is considered the primary feature for direct readout. It should be noted that a strong deviation in the bending toward the major groove also means a deviation in the bending toward the minor groove. That is why we discuss these features as 'bending toward the major/minor groove'. The same goes for the conformational tendency of the DNA. We were able to explain most of the top features of each model, but unable to provide an explanation for the selected features for 'FlhDC'.

Although many prokaryotic classification models, in contrast to the eukaryotic models, did not result in any significant improvements over the simpler methods, the selected features and models can provide us with some valuable information about the binding mode of the protein. This information can be used to gain some insight even before any crystal structures are solved. In most prokaryotic models, the role of direct readout is very important. This is represented by the PWM score feature. This feature will not be discussed separately for every TF.

It is striking that for prokaryotic TFs the PWM score is the best feature in six out of eight models, whereas for eukaryotic TFs it is the best feature in only one out of five models. This can be explained by a recent systematic study on the differences between prokaryotic and eukaryotic TFBSs published by Wunderlich *et al.* (62), in which the authors calculated the average information content (IC) of both prokaryotic and eukaryotic TFBSs. They conclude that the average IC of a prokaryotic TFBS is 23 bits compared with 12.1 bits for eukaryotic TFBSs. This remarkable difference is mainly due to the shorter average length of the eukaryotic binding sites.

'AraC', a regulator of the araBAD operon in *E. coli*, binds as a dimer to the DNA (63). AraC proteins make all sequence-specific contacts in the major groove. Structural reports indicate that both monomers of the dimeric AraC proteins are separated by an AT-rich linker, resulting in an overall bend and a smaller overall minor groove clash size (64). This last feature is clearly reflected in the top five feature list of the AraC model.

In the 'ArcA' model, the groove width is a very important feature both as a positional feature and as a global mean feature. This is in agreement with the data on the OmpR/PhoB family of TFs, of which ArcA is a member (65,66). Just like clash size, width of both the major and the minor groove is an important feature in the winged helix–turn–helix (HTH) family of TFs. In this family of

Table 2. Results of the PCA analysis. For each TF model, we selected the five best features according to Weka PCA analysis

TF model	Feature	TF model	Feature
AraC	PWMmatrixscore_general minor_groove_clash_size_fullseqmean minor_groove_clash_size_p18 monont_p19 = G monont_p0 = A	**HIF1**	uniformity_A_fullseqmean dint_p5 = CG PWMmatrixscore_general dint_p6 = GT dint_p7 = TG
ArcA	PWMmatrixscore_general groovewidth_unboundLiu_fullseqmean groovewidth_unboundLiu_p0 groovewidth_unboundLiu_p1 groovewidth_unboundLiu_p-1	**P53**	uniformity_A_fullseqmean homogeneity_BI_fullseqmean homogeneity_RESTB_fullseqmean PWMmatrixscore_general homogeneity_RESTB_p2
Fis	PWMmatrixscore_general PWMcorescore_general uniformity_A_p-2 uniformity_A_fullseqmean uniformity_A_p-3	**SP1**	homogeneity_RESTB_fullseqmean PWMmatrixscore_general uniformity_AB_fullseqmean dint_p5 = CC dint_p6 = CC
IHF	bend_toward_major_groove_fullseqmean bend_toward_minor_groove_fullseqmean PWMmatrixscore_general bend_toward_major_groove_p-6 bend_toward_major_groove_p-7	**STAT1**	PWMmatrixscore_general dint_p13 = AA dint_p5 = TT dint_p12 = GA dint_p7 = TC
FlhDC	PWMmatrixscore_general monont_p-3 = C monont_p-20 = G monont_p-3 = T tors_1_nucleosome_p-7	**TBP**	bend_toward_major_groove_fullseqmean bend_toward_minor_groove_fullseqmean homogeneity_BII_fullseqmean bend_toward_minor_groove_p8 bend_toward_major_groove_p8
LexA	minor_groove_clash_distance_p-8 dint_p-8 = GC PWMmatrixscore_general minor_groove_clash_distance_p-7 minor_groove_clash_distance_p-9		
PurR	PWMmatrixscore_general PWMcorescore_general monont_p-5 = A monont_p-4 = A monont_p1 = T	**Fis ChIP-chip**	PWMmatrixscore_general monont_p14 = C bend_towards_minor_groove_p6 dint_p9 = TT monont_p0 = G

TFs, a helix is inserted in the major groove of the DNA, whereas the wings of the protein dimer are inserted in the minor groove (65).

'Fis' is known as one of the nucleoid-associated proteins (NAPs). Such proteins are responsible for the packing of the prokaryotic chromosome by bending and supercoiling of the DNA (67). For Fis, two models are available: one with a limited number of binding sites and one more trustworthy chIP-chip model, which we used as a quality control case. The smaller of the two models contains, among the direct readout features many features concerning the A/B-DNA tendency signifying the reported deviations from standard B-DNA (68). The top features of the chIP-chip model are a bit more diverse. Since Fis is one of the NAPs proteins, the appearance of the bending property in the list of PCA top features should come as no surprise. Other important features are both G/C mononucleotides on position 0 and +14. The presence of these features is very important because methylation of these positions on either strand is known to completely inhibit Fis binding (67). The location of these nucleotides is in agreement with the major groove contacts by Fis. The TT dinucleotide feature is also an important *in vivo* feature: it corresponds to the center of the AT-track that is responsible for the bending properties of the DNA in the binding site (31).

The top five components in the 'IHF' model consist mainly of features concerning DNA bending towards the major/minor groove. Since IHF is one of the most extreme DNA benders known, also called 'the master bender' (69,70), the inclusion and importance of the selected features should not be a surprise. This is also reflected in the RF model. The most important feature of this protein is the overall mean of the bend towards major/minor groove, making it one of the few prokaryotic models with a biophysical feature as a top feature, which is in agreement with the IHF's title as master bender.

For 'LexA', the most noticeable features are the minor groove clash size features between −7 and −9 (the linker region between two LexA half sites). This is also reported in the literature, where an unusually narrow minor groove and important clash interactions are observed in the linker region between two LexA half sites in order to fit into the network of interactions between the two half sites (71).

The selected GC dinucleotide feature is also of importance to the minor groove clash size: the occurrence of GC is disfavored because this dinucleotide has the largest minor groove clash size of all nucleotides. This is in agreement with earlier reports, which state that LexA has a preference for A/T-rich spacer regions (71,72).

In the model of the purine repressor (PurR), the top five features consist only of monomeric sequence features and PWM scores. This suggests that this model focuses on the direct readout of PurR binding.

For the 'HIF1' TF, three out of five top features are dinucleotide features. The dinucleotides together, one after the other, build the pattern 5′-CGTG-3′, known as the hypoxia-response element (HRE). This pattern is the most important determining factor of HIF1 binding and is fully conserved in every HIF1 binding site. These HREs are *cis*-regulatory DNA sequences for the specific binding to HIF1 and are necessary for transcription upon hypoxic conditions (73–75). The model was able to capture this sequence element very well.

For 'P53', the majority of important features concern the DNA conformation and the tendency to the A/B-DNA conformation. The DNA conformation is shown to be a very important determinant in the sequence-specific binding by P53. Although P53 binding sites are very degenerate, P53 can bind strongly to a wide range of binding sites. It has been suggested that a shift to a non-standard B-DNA conformation can drastically alter the binding capacity of P53 and that this conformational shift is responsible for the specific binding to the wide variety of P53 motifs (76).

'SP1' is known to unwind the DNA from 10.5 to 11.2 residues per turn, thereby greatly distorting the standard B-structure of the DNA toward a more A-DNA oriented structure and other deviant structures (77,78). Two out of five top features of the SP1 model confirm the importance of DNA conformational features in aiding the binding specificity of SP1 to the DNA, both of which are global features. The other top features are more sequence oriented. The two CC-dinucleotide features in the model are an indication of the cytosine enrichment in the canonical SP1 recognition element (CCCGCC). Furthermore, the importance of CC dinucleotides has been discussed by Zhu *et al.* (79) who found that methylation of the central CG dinucleotide did not impair SP1 binding, but methylation of the first CC dinucleotides significantly decreased SP1 binding specificity. This important feature of the specific binding of SP1 was correctly included as one of the top features in the RF model.

'STAT1', like all other STATs, shows a very strong preference for sequences containing two palindromic half-sites (TTC…GAA), leading to a dyad symmetry, to which the STAT1 dimer can bind (80). The inclusion of the dinucleotide features for AA, TT, GA and TC, together TTTC…GAAA, is the most specific variant of all STAT1 binding motifs according to an analysis made by Ehret *et al.* (81).

'TBP' is one of the most well known DNA benders (82,83) and it was shown that the unbound TATA box is already pre-bent (84). The properties of introducing a kink in the DNA are also well reflected in the model. When looking at the top five features, four out of five top features contain properties about DNA bending, confirming the tendency of TBP to bend the DNA.

DISCUSSION

It has been known for a few decades that the structure of DNA varies in a sequence-dependent manner (4,5). Some recent papers stressed the importance of sequence-dependent structural properties of DNA by showing that they are much less diverse than the nucleotide sequences, but at the same time they contain more information (85,86). That makes the structure space better suited than the nucleotide sequence space for seeking patterns (86–88). Several papers pointed specifically to the role of DNA shape in protein–DNA recognition (46,86,89,90). Rohs *et al.* (6) published a comprehensive review on this topic. In the past decade, only few proposed methods for TFBS identification explicitly took into account the nucleotide-sequence-dependent structural properties of DNA. However, many other methods implicitly capture some part of shape readout mechanisms of DNA-binding specificity when they model positional dependencies of nucleotides, and they tend to predict TFBSs more accurately than the widely used PWM.

For prokaryotes, the apparent lack of improvement for the more complex RF models can have several causes. The size of these data sets is relatively small, whereas complex models like the structural or NPD model might require bigger and better annotated data sets. The additional tests on the more qualitative Fis control set seem to confirm this hypothesis. A simpler method, like a PWM-based strategy, was developed for use with small data sets and apparently performs quite well on most prokaryotic data sets. An alternative, more biological explanation for the poor performance of our models on prokaryotes lies in the differences between prokaryotic and eukaryotic TFs. A systematic analysis of the differences in binding strategy between prokaryotic and eukaryotic binding sites revealed that prokaryotic binding sites tend to be longer and that they have more information content (62). In eukaryotes, the presence of the binding site alone is not enough and binding is often aided by signals in the flanking regions. Prokaryotes have few spurious binding sites, making the presence of one binding site alone a distinctive feature. This, in combination with the smaller and less qualitative set of binding sites, might lead to an overall decrease in performance of the more complex models and give the more simple PWM an advantage, as revealed by comparing the two Fis sets.

For eukaryotes, our results indicate that the inherent structural properties of DNA are involved in specific recognition by the TFs to an extent that depends on each TF, and that these properties can be used to refine predictions. Our results show that a purely structural model performs worse than a model capturing the positional dependencies of nucleotides most of the time. The latter type of model is represented in our comparison by our NPD model, which we believe models both base readout and a big portion of shape readout. The relative importance of the more simple

NPD characteristic consequently cannot be ignored when analyzing TFBS binding patterns in the eukaryotic models. We demonstrate, however, that structural properties contain information other than the nucleotide sequence, and that the use of this information can be used to further improve classification accuracy. We demonstrate that the PWM score that merely represents base readout in its most simple form, is sometimes complementary to the model combining the structural model and NPD model. Most importantly, we present an integrative approach that can easily combine two or three different approaches to establish the best possible prediction of TFBSs.

Further improvements of our purely structural model might be achieved by using higher resolution descriptions of structural characteristics and incorporation of additional ones, such as those available in the database for dinucleotide properties (91). Additionally, input for sequence-based methods is currently gathered in a way that favors the performance of detection methods using nucleotide identities. Sequences containing TFBSs are aligned by methods focusing on nucleotide conservation only, such as existing PWMs or multiple EM (expectation maximization) for motif elicitation (MEME) (92). It could be worthwhile to improve the alignment correction in a way that it takes into account structural vectors. This might even lead to a further improvement for the structural models.

Shape readout is thought to fine-tune binding affinity rather than determine the binding event (6). In this respect, the structural part of the combinatorial model might prove itself more important for discerning binding sites of TFs from the same TF family, as they have very similar or identical base readout mechanisms. Our method could also be useful for detecting binding sites of miRNAs because structure plays a dominant role in the RNA–RNA interaction (93).

Despite high-throughput experimental approaches to identification of TFBSs, improved *in silico* prediction of TFBSs is of great value. It allows more accurate identification of potential *in vivo* TFBSs on rapidly sequenced genomes and enhances our understanding of the TF binding processes. Our integrative method seems to be a good candidate for this purpose.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3, Supplementary Data sets 1, 2 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

We thank Dr Amin Bredan for careful linguistic editing and the four anonymous referees for their constructive comments, which greatly helped improve upon the original version of the manuscript. We also thank the ICT Department of Ghent University for partial support of this work.

FUNDING

Funding for open access charge: Flanders Institute for Biotechnology (VIB); Research Foundation Flanders (FWO); Agency for Innovation through Science and Technology in Flanders (IWT) [SB-091213 to S.B.].

Conflict of interest statement. None declared.

REFERENCES

1. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113–122.
2. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
3. Thayer,K.M. and Beveridge,D.L. (2002) Hidden Markov models from molecular dynamics simulations on DNA. *Proc. Natl Acad. Sci. USA*, **99**, 8642–8647.
4. Calladine,C.R. and Drew,H.R. (1986) Principles of sequence-dependent flexure of DNA. *J. Mol. Biol.*, **192**, 907–918.
5. Shakked,Z. and Rabinovich,D. (1986) The effect of the base sequence on the fine structure of the DNA double helix. *Prog. Biophys. Mol. Biol.*, **47**, 159–195.
6. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
7. Angarica,V.E., Perez,A.G., Vasconcelos,A.T., Collado-Vides,J. and Contreras-Moreira,B. (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, **9**, 436.
8. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
9. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
10. 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.
11. Liu,J. and Stormo,G.D. (2005) Quantitative analysis of EGR proteins binding to DNA: assessing additivity in both the binding site and the protein. *BMC Bioinformatics*, **6**, 176.
12. Liu,J. and Stormo,G.D. (2008) Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, **24**, 1850–1857.
13. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
14. O'Flanagan,R.A., Paillard,G., Lavery,R. and Sengupta,A.M. (2005) Non-additivity in protein-DNA binding. *Bioinformatics*, **21**, 2254–2263.
15. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.
16. Hu,M., Yu,J., Taylor,J.M., Chinnaiyan,A.M. and Qin,Z.S. (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **38**, 2154–2167.
17. Gershenzon,N.I., Stormo,G.D. and Ioshikhes,I.P. (2005) Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res.*, **33**, 2290–2301.
18. Marinescu,V.D., Kohane,I.S. and Riva,A. (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, **6**, 79.
19. Naughton,B.T., Fratkin,E., Batzoglou,S. and Brutlag,D.L. (2006) A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. *Nucleic Acids Res.*, **34**, 5730–5739.

20. Sharon,E., Lubliner,S. and Segal,E. (2008) A feature-based approach to modeling protein-DNA interactions. *PLoS Comput. Biol.*, **4**, e1000154.
21. Karas,H., Knuppel,R., Schulz,W., Sklenar,H. and Wingender,E. (1996) Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput Appl. Biosci.*, **12**, 441–446.
22. Ponomarenko,J.V., Ponomarenko,M.P., Frolov,A.S., Vorobyev,D.G., Overton,G.C. and Kolchanov,N.A. (1999) Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, **15**, 654–668.
23. Liu,R., Blackwell,T.W. and States,D.J. (2001) Conformational model for binding site recognition by the *E.coli* MetJ transcription factor. *Bioinformatics*, **17**, 622–633.
24. Burden,H.E. and Weng,Z. (2005) Identification of conserved structural features at sequentially degenerate locations in transcription factor binding sites. *Genome Inform.*, **16**, 49–58.
25. Pudimat,R., Schukat-Talamazzini,E.G. and Backofen,R. (2005) A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*, **21**, 3082–3088.
26. Gunewardena,S., Jeavons,P. and Zhang,Z. (2006) Enhancing the prediction of transcription factor binding sites by incorporating structural properties and nucleotide covariations. *J. Comput. Biol.*, **13**, 929–945.
27. Bauer,A.L., Hlavacek,W.S., Unkefer,P.J. and Mu,F. (2010) Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. *PLoS Comput. Biol.*, **6**, e1001007.
28. Meysman,P., Dang,T.H., Laukens,K., De Smet,R., Wu,Y., Marchal,K. and Engelen,K. (2011) Use of structural DNA properties for the prediction of transcription-factor binding sites in *Escherichia coli*. *Nucleic Acids Res.*, **39**, e6.
29. Morozov,A.V. and Siggia,E.D. (2007) Connecting protein structure with predictions of regulatory sites. *Proc. Natl Acad. Sci. USA*, **104**, 7068–7073.
30. Fulton,D.L., Sundararajan,S., Badis,G., Hughes,T.R., Wasserman,W.W., Roach,J.C. and Sladek,R. (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.
31. Cho,B.K., Knight,E.M., Barrett,C.L. and Palsson,B.O. (2008) Genome-wide analysis of Fis binding in *Escherichia coli* indicates a causative role for A-/AT-tracts. *Genome Res.*, **18**, 900–910.
32. Portales-Casamar,E., Kirov,S., Lim,J., Lithwick,S., Swanson,M.I., Ticoll,A., Snoddy,J. and Wasserman,W.W. (2007) PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol.*, **8**, R207.
33. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
34. Gowrisankar,S. and Jegga,A.G. (2009) Regression based predictor for p53 transactivation. *BMC Bioinformatics*, **10**, 215.
35. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
36. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
37. Satchwell,S.C., Drew,H.R. and Travers,A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.
38. Goodsell,D.S. and Dickerson,R.E. (1994) Bending and curvature calculations in B-DNA. *Nucleic Acids Res.*, **22**, 5497–5503.
39. Lu,X.J. and Olson,W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–1227.
40. Fujii,S., Kono,H., Takenaka,S., Go,N. and Sarai,A. (2007) Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.*, **35**, 6063–6074.
41. Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T. 3rd, Dixit,S., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2009) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299–313.
42. Gartenberg,M.R. and Crothers,D.M. (1988) DNA sequence determinants of CAP-induced bending and protein binding affinity. *Nature*, **333**, 824–829.
43. Parvin,J.D., McCormick,R.J., Sharp,P.A. and Fisher,D.E. (1995) Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature*, **373**, 724–727.
44. Dickerson,R.E. (1998) DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.*, **26**, 1906–1926.
45. Gorin,A.A., Zhurkin,V.B. and Olson,W.K. (1995) B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, **247**, 34–48.
46. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
47. Svozil,D., Kalina,J., Omelka,M. and Schneider,B. (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res.*, **36**, 3690–3706.
48. Spolar,R.S. and Record,M.T. Jr (1994) Coupling of local folding to site-specific binding of proteins to DNA. *Science*, **263**, 777–784.
49. Lu,X.J., Shakked,Z. and Olson,W.K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.
50. Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 28.
51. Lunetta,K.L., Hayward,L.B., Segal,J. and Van Eerdewegh,P. (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.*, **5**, 32.
52. Cordell,H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.
53. Ruiz,R., Jos,R.C. and Aguilar-Ruiz,J.S. (2006) Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recogn.*, **39**, 2383–2392.
54. Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I.H. (2009) The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, **11**, 10.
55. Medina-Rivera,A., Abreu-Goodger,C., Thomas-Chollier,M., Salgado,H., Collado-Vides,J. and van Helden,J. (2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.*, **39**, 808–824.
56. Zhang,Z. and Gerstein,M. (2003) Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J. Biol.*, **2**, 11.
57. Ernst,J., Plasterer,H.L., Simon,I. and Bar-Joseph,Z. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.
58. Narang,V., Mittal,A. and Sung,W.K. (2010) Localized motif discovery in gene regulatory sequences. *Bioinformatics*, **26**, 1152–1159.
59. Ramsey,S.A., Knijnenburg,T.A., Kennedy,K.A., Zak,D.E., Gilchrist,M., Gold,E.S., Johnson,C.D., Lampano,A.E., Litvak,V., Navarro,G. *et al.* (2010) Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, **26**, 2071–2075.
60. 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.
61. Mokry,M., Hatzis,P., de Bruijn,E., Koster,J., Versteeg,R., Schuijers,J., van de Wetering,M., Guryev,V., Clevers,H. and Cuppen,E. (2010) Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. *PLoS One*, **5**, e15092.
62. Wunderlich,Z. and Mirny,L.A. (2009) Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.*, **25**, 434–440.

63. Hendrickson,W. and Schleif,R. (1985) A dimer of AraC protein contacts three adjacent major groove regions of the araI DNA site. *Proc. Natl Acad. Sci. USA*, **82**, 3129–3133.
64. Lu,Y., Flaherty,C. and Hendrickson,W. (1992) Arac protein contacts asymmetric sites in the *Escherichia-coli* Arafgh promoter. *J. Biol. Chem.*, **267**, 24848–24857.
65. Martinez-Hackert,E. and Stock,A.M. (1997) Structural relationships in the OmpR family of winged-helix transcription factors. *J. Mol. Biol.*, **269**, 301–312.
66. Toro-Roman,A., Mack,T.R. and Stock,A.M. (2005) Structural analysis and solution studies of the activated regulatory domain of the response regulator ArcA: a symmetric dimer mediated by the alpha4-beta5-alpha5 face. *J. Mol. Biol.*, **349**, 11–26.
67. Pan,C.Q., Finkel,S.E., Cramton,S.E., Feng,J.A., Sigman,D.S. and Johnson,R.C. (1996) Variable structures of Fis-DNA complexes determined by flanking DNA-protein contacts. *J. Mol. Biol.*, **264**, 675–695.
68. Afflerbach,H., Schroder,O. and Wagner,R. (1999) Conformational changes of the upstream DNA mediated by H-NS and FIS regulate *E. coli* RrnB P1 promoter activity. *J. Mol. Biol.*, **286**, 339–353.
69. Travers,A. (1997) DNA-protein interactions: IHF–the master bender. *Curr. Biol.*, **7**, R252–R254.
70. Schneider,T.D. (2001) Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucleic Acids Res.*, **29**, 4881–4891.
71. Zhang,A.P., Pigli,Y.Z. and Rice,P.A. (2010) Structure of the LexA-DNA complex and implications for SOS box measurement. *Nature*, **466**, 883–886.
72. Lewis,L.K., Harlow,G.R., Gregg-Jolly,L.A. and Mount,D.W. (1994) Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in *Escherichia coli*. *J. Mol. Biol.*, **241**, 507–523.
73. Kajimura,S., Aida,K. and Duan,C. (2006) Understanding hypoxia-induced gene expression in early development: in vitro and in vivo analysis of hypoxia-inducible factor 1-regulated zebra fish insulin-like growth factor binding protein 1 gene expression. *Mol. Cell Biol.*, **26**, 1142–1155.
74. Michel,G., Minet,E., Ernest,I., Roland,I., Durant,F., Remacle,J. and Michiels,C. (2000) A model for the complex between the hypoxia-inducible factor-1 (HIF-1) and its consensus DNA sequence. *J. Biomol. Struct. Dyn.*, **18**, 169–179.
75. Camenisch,G., Stroka,D.M., Gassmann,M. and Wenger,R.H. (2001) Attenuation of HIF-1 DNA-binding activity limits hypoxia-inducible endothelin-1 expression. *Pflugers Arch.*, **443**, 240–249.
76. Kim,E., Albrechtsen,N. and Deppert,W. (1997) DNA-conformation is an important determinant of sequence-specific DNA binding by tumor suppressor p53. *Oncogene*, **15**, 857–869.
77. Shi,Y.G. and Berg,J.M. (1996) DNA unwinding induced by zinc finger protein binding. *Biochemistry*, **35**, 3845–3848.
78. Marco,E., Garcia-Nieto,R. and Gago,F. (2003) Assessment by molecular dynamics simulations of the structural determinants of DNA-binding specificity for transcription factor Sp1. *J. Mol. Biol.*, **328**, 9–32.
79. Zhu,W.G., Srinivasan,K., Dai,Z.Y., Duan,W.R., Druhan,L.J., Ding,H.M., Yee,L., Villalona-Calero,M.A., Plass,C. and Otterson,G.A. (2003) Methylation of adjacent CpG sites affects Sp1/Sp3 binding and activity in the p21(Cip1) promoter. *Mol. Cell. Biol.*, **23**, 4056–4065.
80. Chen,X.M., Vinkemeier,U., Zhao,Y.X., Jeruzalmi,D., Darnell,J.E. and Kuriyan,J. (1998) Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell*, **93**, 827–839.
81. Ehret,G.B., Reichenbach,P., Schindler,U., Horvath,C.M., Fritz,S., Nabholz,M. and Bucher,P. (2001) DNA binding specificity of different STAT proteins - Comparison of in vitro specificity with natural target sites. *J. Biol. Chem.*, **276**, 6675–6688.
82. Powell,R.M., Parkhurst,K.M. and Parkhurst,L.J. (2002) Comparison of TATA-binding protein recognition of a variant and consensus DNA promoters. *J. Biol. Chem.*, **277**, 7776–7784.
83. Juo,Z.S., Chiu,T.K., Leiberman,P.M., Baikalov,I., Berk,A.J. and Dickerson,R.E. (1996) How proteins recognize the TATA box. *J. Mol. Biol.*, **261**, 239–254.
84. Davis,N.A., Majee,S.S. and Kahn,J.D. (1999) TATA box DNA deformation with and without the TATA box-binding protein. *J. Mol. Biol.*, **291**, 249–265.
85. Gardiner,E.J., Hunter,C.A., Lu,X.J. and Willett,P. (2004) A structural similarity analysis of double-helical DNA. *J. Mol. Biol.*, **343**, 879–889.
86. Parker,S.C., Hansen,L., Abaan,H.O., Tullius,T.D. and Margulies,E.H. (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, **324**, 389–392.
87. Greenbaum,J.A., Pang,B. and Tullius,T.D. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res.*, **17**, 947–953.
88. Abeel,T., Saeys,Y., Bonnet,E., Rouze,P. and Van de Peer,Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.
89. Tullius,T. (2009) Structural biology: DNA binding shapes up. *Nature*, **461**, 1225–1226.
90. Rohs,R., West,S.M., Liu,P. and Honig,B. (2009) Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.*, **19**, 171–177.
91. Friedel,M., Nikolajewa,S., Suhnel,J. and Wilhelm,T. (2009) DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.*, **37**, D37–D40.
92. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
93. Long,D., Lee,R., Williams,P., Chan,C.Y., Ambros,V. and Ding,Y. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.