1438–1449 *Nucleic Acids Research, 2013, Vol. 41, No. 3* *Published online 24 December 2012*
doi:10.1093/nar/gks1301

The twilight zone of *cis* element alignments

Alvaro Sebastian[1,*] and Bruno Contreras-Moreira[1,2,*]

[1]Laboratory of Computational Biology, Department of Genetics and Plant Breeding, Estación Experimental de Aula Dei/CSIC, Av. Montañana 1005 and [2]Fundación ARAID, Paseo María Agustín 36, Zaragoza 50059, Spain

Received January 12, 2012; Revised November 9, 2012; Accepted November 14, 2012

ABSTRACT

Sequence alignment of proteins and nucleic acids is a routine task in bioinformatics. Although the comparison of complete peptides, genes or genomes can be undertaken with a great variety of tools, the alignment of short DNA sequences and motifs entails pitfalls that have not been fully addressed yet. Here we confront the structural superposition of transcription factors with the sequence alignment of their recognized *cis* elements. Our goals are (i) to test TFcompare (http://floresta.eead.csic.es/tfcompare), a structural alignment method for protein–DNA complexes; (ii) to benchmark the pairwise alignment of regulatory elements; (iii) to define the confidence limits and the twilight zone of such alignments and (iv) to evaluate the relevance of these thresholds with elements obtained experimentally. We find that the structure of *cis* elements and protein–DNA interfaces is significantly more conserved than their sequence and measures how this correlates with alignment errors when only sequence information is considered. Our results confirm that DNA motifs in the form of matrices produce better alignments than individual sequences. Finally, we report that empirical and theoretically derived twilight thresholds are useful for estimating the natural plasticity of regulatory sequences, and hence for filtering out unreliable alignments.

INTRODUCTION

Transcription factors (TFs) are proteins that bind to the promoters of genes, and thus regulate their expression by activating or repressing the transcription machinery. In consequence, the association of TFs to DNA-binding sites (DBSs) plays a major role in coordinating the cell. DBSs consist of typically short *cis*-regulatory elements that are located within intergenic regions with variable distances to neighbor transcription start sites. Any one TF can recognize a collection of similar DBSs, which can be grouped together to define a DNA motif. Motifs are most frequently represented as position-specific scoring matrices (PSSMs), which capture the occurrence of nucleotides in different positions of the DBSs (1).

Discovery, annotation and alignment of *cis* elements

Experimental methods to identify DBSs are challenging and are usually limited to determining *cis*-regulatory sites for one TF at a time. Among traditional techniques are electrophoretic mobility shift assay, chromatin immunoprecipitation (ChIP) and footprinting assays (2–4). Modern approaches such as protein-binding microarrays (PBMs), ChIP-chip and ChIP-Seq allow high-throughput and genome-wide discovery of DBSs (5–7).

There are many motif discovery tools available that produce PSSMs out of experimental DBSs (8–12). Such PSSMs are collected, curated and manually annotated in databases such as TRANSFAC (13), JASPAR (14) or RegulonDB (15). A different kind of method derives PSSMs from 3D structures of TF–DNA complexes stored in the Protein Data Bank (PDB) (16–21).

In addition, a great variety of computer programs have been developed in order to match pre-compiled motifs within genomic landscapes (12,22–26). These computational approaches effectively scan PSSMs along large genomic sequences in order to locate putative *cis*-regulatory elements, those with alignment scores above some arbitrary threshold.

The twilight zone of sequence alignments

In the context of protein sequences, the 'twilight zone' has been defined as the range of sequence length and sequence identity where there is a non-negligible probability that an alignment is wrong, taking structural similarity as a standard of truth (27,28). In general, algorithms for DBS discovery rely on alignments of nucleotide sequences and/or PSSMs and should also be expected to perform

*To whom correspondence should be addressed. Tel: +34 976716089; Fax: +34 976716145; Email: bcontreras@eead.csic.es
Correspondence may also be addressed to Alvaro Sebastian. Email: bioquimicas@yahoo.es

poorly in the equivalent twilight zone of DNA alignments. However, such a range of sequence length and identity remains to be defined for short nucleotide strings, as is the case of *cis*-regulatory elements. The reason for this might have been the small number of TF–DNA structural complexes deposited in the PDB, which has nonetheless increased steadily. It must be noted that Keich and Pevzner have already used the term 'twilight zone' for nucleotide sequences (29); however, their work focused on discovering motifs in genomic regions of variable length, and therefore was not primarily concerned with alignment quality, which is our main interest throughout this article.

The intrinsic limitations of pairwise alignments of short DNA strings

While aligning short DNA sequences such as DBSs, there are a few potential pitfalls that need to be overcome, such as (i) both the direct and the reverse complementary strand must be considered, while it might not be clear which orientation is biologically relevant; (ii) short nucleotide sequences are prone to yield non-relevant alignments; (iii) motifs or PSSMs of low information content usually produce poor-quality alignments and (iv) mismatches are common due to TF-binding plasticity. For these limitations, we believe it would be of great interest to characterize the twilight zone of short DNA alignments, in order to define thresholds that will set bounds on the quality of DBS alignments produced by any algorithm. As with protein alignments (26,27), here we take structural alignments of TF–DNA complexes as a standard of truth to decide whether a pairwise alignment of *cis* elements might be correct. In this framework, a correct alignment of a pair of DBSs will entail also a correct alignment of the TF domains that recognize them.

Article layout

First, we develop a method that superposes pairs of TF–DNA complexes in order to obtain the resulting structural alignments of *cis* elements recognized by homologous TFs. Second, we perform such superpositions within a non-redundant collection of TF–DNA complexes in order to discover root-mean-square deviation (RMSD) cutoffs that effectively classify pairs of regulatory sites with similar and dissimilar conformation. Note that this does not imply that similar structures recognize similar DNA sequences. Third, we calculate pairwise alignments of the *cis* elements (and corresponding PSSMs) of the former TF–DNA complexes and derive reliability thresholds, as in the work of Sander and Schneider (27), that define when similar DNA sequences are likely recognized by similar structures. Fourth, we propose a theoretical approach that produces cutoff values that approximate and validate the previous empirical thresholds. Finally, we analyze real-world motifs, some extracted from curated database TRANSFAC and others produced by PBM experiments (13,30), and measure the proportion of natural *cis* elements that are aligned beyond the twilight zone.

MATERIALS AND METHODS

Non-redundant set of 135 TF–DNA complexes

We retrieved all X-ray protein–DNA structures from the PDB (21) and their corresponding annotated interfaces from the 3D-footprint database (18) (June 2011 version). Complexes with single-stranded DNA, resolution >3.5 Å or with less than three nitrogen bases contacting protein residues were excluded. The selected structures were trimmed into protein DNA-binding domains (DBDs) and their corresponding DBSs. Domain boundaries were assigned with *hmmpfam* from software package HMMER v2.3.2 (31) and Pfam v.23 (32), with parameters –acc –cut_ga –cut_nc –cut_tc –E 10. DNA duplexes were also trimmed from the first to the last contacted nucleotide as annotated in 3D-footprint (http://floresta.eead.csic.es/3dfootprint/download/list_interface2dna.txt). Resulting DBDs were further filtered by (i) checking their 'TF activity' Gene Ontology annotation (33) and (ii) rejecting redundant domains over a 95% cutoff with CDHIT 4.0 (34). Eventually, a set of 135 TF–DNA complexes was obtained (listed in Supplementary Table S1), where 36 Pfam families are represented, emphasizing 26 homeodomain (PF00046) and 16 Zinc finger (PF00105) family members with domain lengths from 32 to 234 amino acids (domain description and statistics are supplied in Supplementary Table S2).

Dataset of 67 curated PSSMs and 873 DBSs

For 67 out of the selected 135 TFs, we were able to find high-quality PSSMs from curated databases, 61 from TRANSFAC v9.3 (13) and 6 from JASPAR v2009 (14). These matrices were manually trimmed to the same length of the *cis* elements captured in the corresponding TF–DNA complexes, to allow the comparison of sequence and structure-based pairwise alignments. Moreover, 873 DBSs used to derive 39 of the former 67 PSSMs were also retrieved from the TRANSFAC database (DBSs and PSSMs identifiers are listed in Supplementary Table S3).

Structural alignment of TF–DNA complexes

Each TF–DNA complex was structurally fitted to all the others in two steps. First, the program MAMMOTH (35) was used to perform the structural superposition of pairs of protein domains. Second, the resulting transformation matrices were applied to the coordinates of the DBSs in order to derive the equivalent *cis* element superpositions. RMSDs of superposed coordinates were calculated with β-carbon atoms (proteins) and with N9 (purines) and N1 (pyrimidines) atoms (DNA). Structural alignments were scored in terms of (i) the number of identical superposed nucleotides (StrIdent) and (ii) the sum of N9 and N1 atom pairs within 3.5 Å (StrScr). Out of 18 225 possible TF–DNA comparisons, we obtained 18 012 structural alignments after discarding self-alignments and errors. The full set of alignments can be accessed in Supplementary File 'alignment_data.csv'. A web application called TFcompare (http://floresta.eead.csic.es/tfcompare) has

been released to assist in the task of calculating structural pairwise alignments of TF–DNA complexes.

Sequence alignment of *cis* elements

Although many *cis* elements alignment and scoring methods have been reported in the literature, we focused on three main alignment strategies: (i) DNA sequence pairs were aligned using an ungapped Smith–Waterman algorithm (36) taking as scoring function the number of identical nucleotides (IdentScr); (ii) same as (i), but scoring identities with +1 and mismatches with −2, as in BLASTN (BlastScr) (37) and (iii) pairs of PSSMs were aligned as in (i) but taking as scoring function the sum of Pearson correlation coefficients of aligned columns (PearScr). Furthermore, PSSM alignments were also scored in terms of the −log(*E*-value), as calculated by STAMP (EvalScr) (38). Self-alignments and alignments shorter than three deoxynucleotides were discarded. The complete sequence alignment set is also included in Supplementary File 'alignment_data.csv'.

Twilight thresholds definition and linear fitting

Identity and similarity thresholds were calculated for each sequence alignment strategy by finding cutoff values that increasingly left 95, 90 and 75% of dissimilar TF–DNA complexes below the selected value. Cutoffs were then fitted to an equation of the form:

$$t(L) = a + b \cdot L$$

where L is the number of aligned nucleotides and the coefficients a and b are estimated by a standard linear model from the R software (39).

Receiver operating characteristic curves and predictive power of alignment methods

Receiver operating characteristic (ROC), originally employed to measure the accuracy of signal detection algorithms (40), have extensively been applied to classifiers of biological data (41). The area under a ROC curve (AUC) takes values in the range [0,1] and provides a numerical estimate of how accurate a classifier is; the higher AUC value, the higher the accuracy. Several ROC curves were calculated for the different alignment procedures (StrScr, StrIdent, IdentScr, BlastScr, PearScr and EvalScr), with the goal of measuring their performance in discerning true and false DBS alignments, which in turn correspond to similar and dissimilar pairs of TF–DNA complexes. Each ROC curve plots 'true-positive rate' (= sensitivity) on the *y*-axis versus 'false-positive rate' (= 1 − specificity) on the *x*-axis. The R library ROCR was used to perform these calculations (39,42).

Comparison between sequence and structural alignments of DBSs

Pairs of matched nucleotides in sequence alignments (IdentScr, BlastScr and PearScr) were compared with those in 'true' structural alignments of TF–DNA pairs (StrScr). Out of 1058 'true' structural alignments, 535 could be aligned by all sequence-based techniques and were further compared here. Sequence alignments with all identical positions were annotated as correct, otherwise as incorrect.

Theoretical estimation of twilight threshold values

Ten sets of 18 012 alignments of randomly generated DBSs and PSSMs were generated, conserving the distribution of sizes of the 137 trimmed *cis* elements captured in TF–DNA complexes. Artificial DBSs were built by random sampling of nucleotides with equal probability, whereas PSSMs were obtained by random shuffling of columns of TRANSFAC and JASPAR motifs. These artificial sequences and motifs were then aligned by applying the different algorithms explained earlier (IdentScr, BlastScr and PearScr). Average 95th percentile values of sequence identity and similarity values were calculated for each alignment length between 3 and 8, to be used as *in silico* twilight thresholds.

Alignment of PSSMs against their own individual DBSs

PSSMs and their component DBSs were trimmed to preserve only the core, removing the lower case flanks as labeled by 'convert-matrix' from the RSA-Tools suite (43). DBSs were then converted to binary PSSMs (columns with value of 1 in the appropriate nucleotide and 0 otherwise) and scored against their PSSMs using the PearScr algorithm. Several alignment sets of PSSMs and DBSs were tested: (i) the set of 39 TRANSFAC and JASPAR motifs of TFs with PDB structure versus their component DBSs; (ii) all TRANSFAC motifs versus annotated sites; (iii) all UNIPROBE motifs versus top 20 experimental sites; (iv) all UNIPROBE motifs versus top 100 sites and (v) all UNIPROBE motifs versus all their statistically significant sites. Alignment results for (ii) and (iv) can be checked in Supplementary Files 'transfac_alignments.csv' and 'uniprobe_alignments.csv'.

RESULTS

Structural alignment of TF–DNA complexes

The cornerstone of this work is the calculation of a large set of superpositions between pairs of TF–DNA complexes, which not only helps identify similar DBDs but also produces structure-based alignments of *cis* elements (DBSs). An analysis of superposed complexes, described in Figure 1A, unveils that similar DBDs in turn recognize DBSs with resembling geometries (the observed correlation coefficient was 0.6). Indeed, both the DBD and the DBS distributions show two clear peaks: a low RMSD region and a high RMSD region separated by a valley. These two peaks represent the concentration of similar and dissimilar structural alignments in low and high RMSD regions, respectively. The valley minimum is found to be 5 Å for DBDs (Figure 1B) and 3.5 Å for DBSs (Figure 1C). These two cutoff values were therefore chosen as standards of truth to distinguish between similar (true) and dissimilar (false) structural alignments.

When these standards are applied, there are four possible outcomes, summarized in Figure 2: (i) true

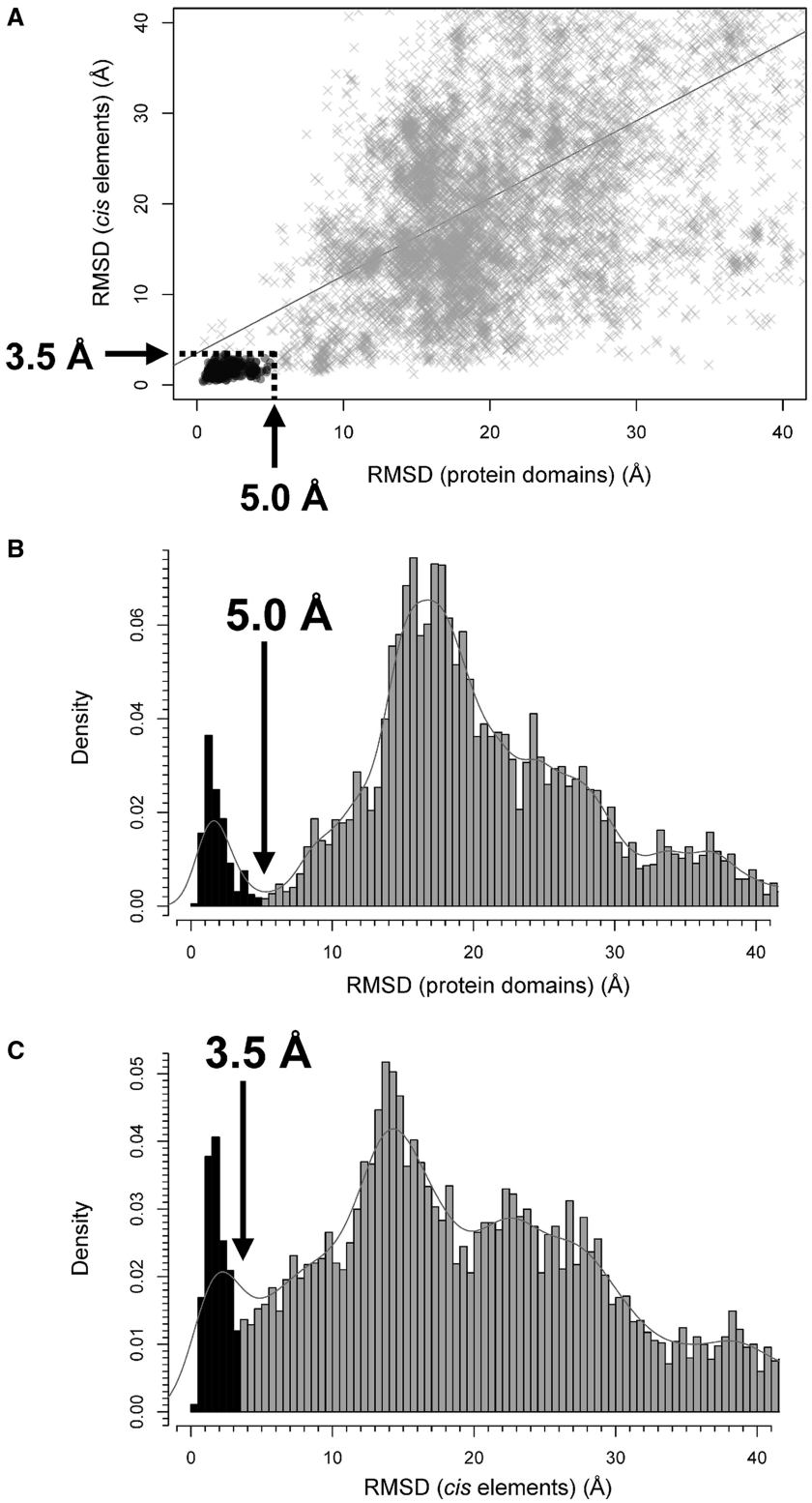
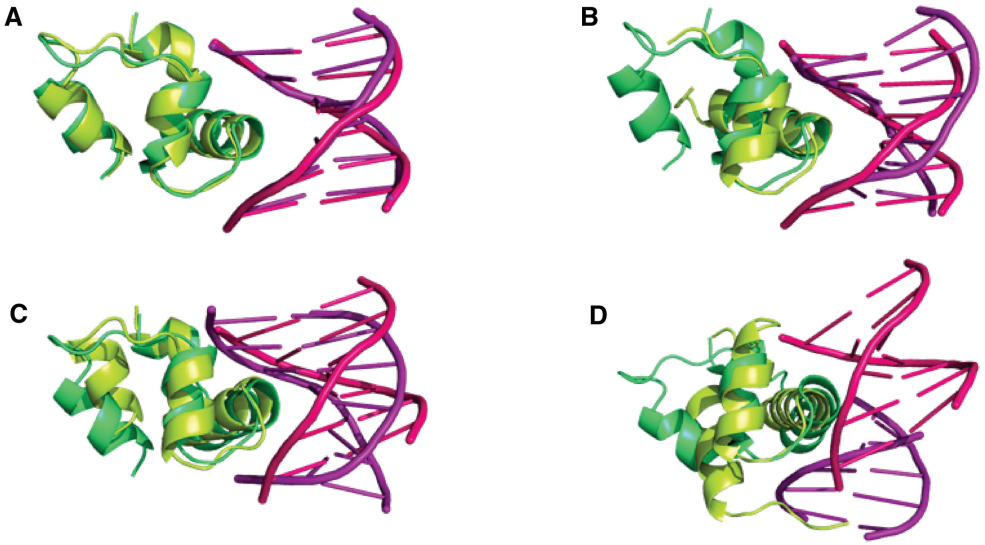


Figure 1. Analysis of 18 554 superpositions of transcription factors bound to their cognate *cis* elements. (**A**) Scatter plot of RMSD values of protein domain (DBD) and *cis* element (DBS) superpositions, with the resulting linear correlation ($y = 3.84 + 0.84x$; $R^2 = 0.364$). (**B**) Histogram of DBD RMSD values. (**C**) Histogram of DBS RMSD values. Pairs of structurally similar complexes are in black [circles in (A)], otherwise they are in gray [crosses in (A)]. Similar complexes are defined as pairs with DBD RMSD values < 5 Å and DBS RMSD < 3.5 Å.



	Query	Sbjct	Domain Query	Domain Sbjct	DBS RMSD	DBD RMSD	Classification	
A	1d5y_A1	1bl0_A1	PF00165	PF00165	1.32	1.02	True positive	SIMILAR
B	1d5y_A1	1cgp_A2	PF00165	PF00325	3.22	15.10	False positive	DISSIMILAR
C	1d5y_A1	1bl0_A2	PF00165	PF00165	5.77	3.57	False negative	
D	1d5y_A1	1b8i_B1	PF00165	PF00046	15.61	14.50	True negative	

Figure 2. Four possible outcomes of the superposition of a pair of TF–DNA complexes. (**A**) Both DBD and DBS align with RMSD values below thresholds (true positive). (**B**) Only DBS RMSD is under threshold (false positive). (**C**) Only DBD RMSD is below threshold (false negative). (**D**) both RMSD values are higher than thresholds. Case (A) is defined as a 'similar' pair of TF–DNA complexes (true alignment) and cases (B), (C) and (D) as 'dissimilar' pairs (false alignments). Protein domains are indicated as Pfam codes; RMSD values are given in Angstroms.

positives, similar complexes where $RMSD_{DBD}$ and $RMSD_{DBS}$ fall below the threshold; (ii) false positives, dissimilar proteins whose $RMSD_{DBS}$ falls below the threshold; (iii) true negatives, dissimilar complexes with both $RMSD_{DBD}$ and $RMSD_{DBS}$ above the thresholds and (iv) false negatives, pairs of similar proteins with $RMSD_{DBS}$ above the threshold. After applying these rules to our set of 18 012 superpositions, we find 1058 (5.9%) true structural alignments.

Figure 3 shows several examples of structural alignments of TF 3A01_A1 (PDB code 3A01, first protein domain of chain A) with other complexes. 3A01_A1 belongs to the homeodomain family, the most abundant in this study and one of the most thoroughly studied in the literature in terms of DNA-binding specificity (44,45). True superpositions in Figure 3A result in aligned *cis* elements with conserved positions. Instead, alignments of 3A01_A1 with other dissimilar complexes presented in Figure 3B (over RMSD thresholds) do not show conserved nucleotides. Figure 3C is an in-depth analysis of two homeodomains in which a structural superposition correctly pairs equivalent interface amino acid residues and equivalent *cis* element positions, while a sequence-based alignment favors achieving the highest score (in this case, identical nucleotides) at the cost of reversing the sequence strand orientation and subsequently failing to match equivalent interface amino acids.

As a summary, when TF–DNA complexes are dissimilar, structural superpositions produce random alignments with low associated scores that clearly indicate that their *cis* elements are not directly comparable. However, when two complexes have similar bound conformations, their structural superposition can be translated into a biologically meaningful alignment of regulatory elements, which matches interface residues, and that might not always be recapitulated by a standard sequence-based alignment. The web server TFcompare (http://floresta.eead.csic.es/tfcompare), presented in this article, allows the user to reproduce this kind of TF superpositions between structures deposited at the PDB.

Estimates of twilight thresholds for the alignment of *cis* elements

The previously defined standards of truth can be employed to assess the reliability of different approaches for aligning *cis* elements, which are usually short DNA segments (94% of elements recognized by single protein domains are between 3 and 8 nt long, see Supplementary Figure S1). Up to four sequence-based ungapped alignment techniques and scoring functions are evaluated here, which are summarized in Table 1 and further explained in 'Materials and Methods' section. Each of these strategies was benchmarked by plotting score distributions of similar and dissimilar structural superpositions versus alignment

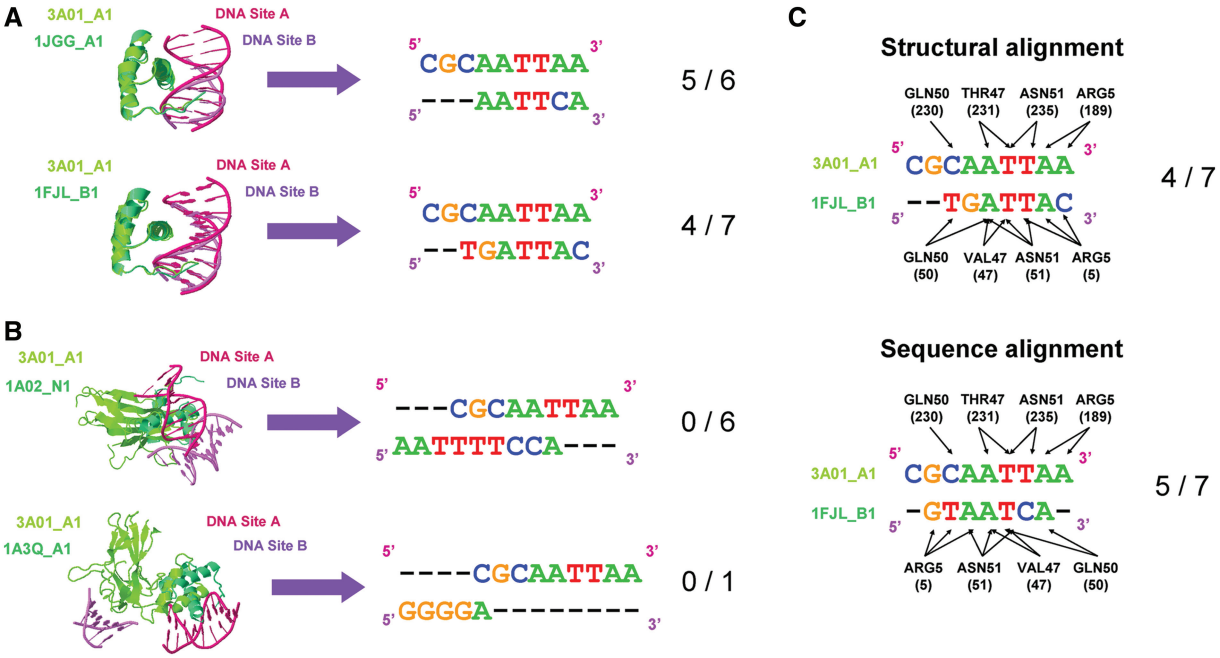


Figure 3. Examples of homeodomain 3A01_A1 superposed to other protein–DNA complexes. (**A**) True (similar) structural alignments showing the number of identical nucleotides over the aligned. (**B**) False (dissimilar) structural superpositions. (**C**) An in-depth comparison illustrating the differences between structural and sequence alignments of *cis* elements. Note that the structural alignment matches equivalent interface residues, numbered as in Noyes *et al.* (44), while the optimal sequence alignment implies taking the reverse complementary of the target sequence, breaking all interface similarities.

Table 1. Description of the different alignment and scoring methods tested in this work

Scoring method	Alignment type	Aligned data	Description
StrScr	Structural	PDB coordinates	Number of N9 and N1 nucleotide atom pairs within 3.5 Å
StrIdent	Structural	PDB coordinates	Number of identical superposed nucleotides
IdentScr	Sequence	Sequences	Number of identical aligned nucleotides
BlastScr	Sequence	Sequences	Sum of matches (+1) and mismatches (−2, default BLASTN scoring)
PearScr	Sequence	PSSMs	Sum of Pearson correlation coefficients of aligned PSSM columns
EvalScr	Sequence	PSSMs	Negative logarithm of the *E*-value calculated by STAMP

length. As true complex pairs were clearly outnumbered by unrelated pairs, we defined score cutoffs that minimized the chance of aligning *cis* elements of structurally dissimilar DBDs. Three percentages (95, 90 and 75%) of false superpositions were tested as twilight thresholds; the most restrictive (95%) is presented in Figure 4A (see numerical values in Supplementary Table S4 and Supplementary Figure S2). Thresholds in Figure 4A can be taken as confidence intervals for each alignment approach. Figure 4B plots the theoretical thresholds calculated after comparing and scoring random *cis* elements and PSSMs with the same collection of alignment strategies, as explained in 'Materials and Methods' section. As the numbers in Supplementary Table S5 confirm, the empirical and theoretical lines are significantly alike, with negligible root-mean-square errors with the exception of BlastScr. When theoretical estimates are compared with the corresponding empirical values, most of them have deviations of <0.1 units. Overall, these results suggest that both the structure-based benchmark and the independent theoretical simulations converge to similar twilight thresholds.

How are these twilight lines to be interpreted? For instance, for two aligned hexanucleotides, we will require a score of 5 with IdentScr (five identical bases) to call it a correct alignment; comparatively, a score of 6 would be required with BlastScr, while 4.3 would be enough for a PearScr alignment. In the extreme case of three aligned nucleotides, all strategies require a score of 3 to remove 95% of false alignments, with the exception of EvalScr, that instead requires an *E*-value < 0.10.

Since StrScr and RMSD are correlated functions, it could be anticipated that StrScr would yield the lowest twilight thresholds. However, it is surprising that for alignment lengths between 5 and 7, thresholds are unusually high. These overestimations can be explained as in this length range most of the superpositions in our dataset are similar, and consequently there are too few data with low scores to reliably calculate thresholds values.

BlastScr alignments yield thresholds with the highest values, which in most cases are equal to the number of aligned nucleotides, indicating that BlastScr is the most conservative approach when comparing *cis* elements.

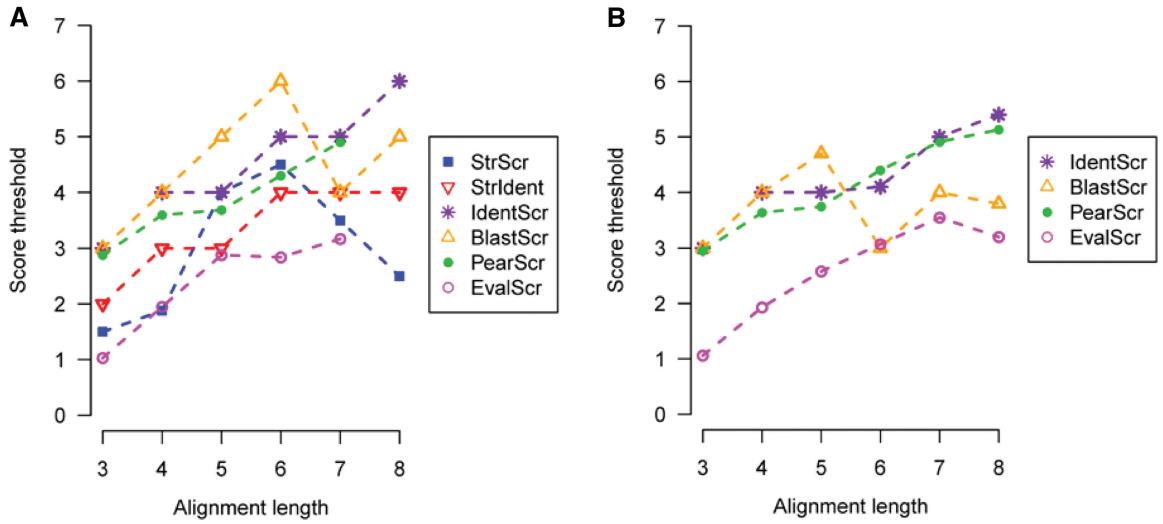


Figure 4. Twilight thresholds for different alignment techniques and scoring functions. (**A**) Experimental thresholds that leave 95% of false TF–DNA superpositions below the indicated cutoff values. (**B**) Theoretical thresholds that leave 95% of random DNA sequence alignments below the indicated cutoff values. An EvalScr of 3 corresponds to an *E*-value of 0.001. Not enough PSSMs of length = 8 were available for PearScr and EvalScr in (A).

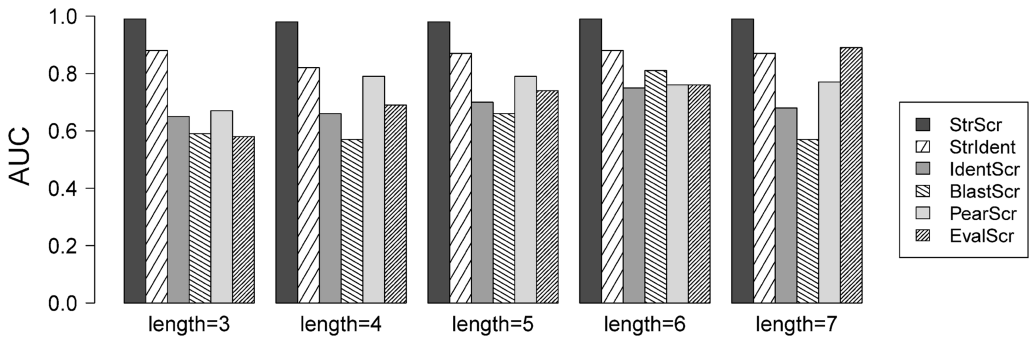


Figure 5. AUC for different alignment algorithms and scoring functions. Not enough data of length = 8 were available to draw ROC curves.

Nevertheless BlastScr thresholds in the range 5–7 follow a peak and valley behavior that is a consequence of the rather limited range of scores that this function produces. For example, alignments of length = 6 can only be assigned BlastScr scores 0, 3 or 6, whereas alignments of length = 7 are assigned values among −2, 1, 4 and 7.

StrIdent, IdentScr, PearScr and EvalScr follow a nearly linear tendency (regression coefficients are reported in Supplementary Table S4). StrIdent gives the lower values of these four techniques (excluding non-comparable EvalScr), confirming that structure-based alignment of *cis* elements is the most reliable approach in order to successfully filter out dissimilar complexes. Of course StrIdent values can only be obtained when 3D structures are available and therefore this hinders its practical application. IdentScr and PearScr display analogous behaviors and even similar intercepts and slopes in the linear regression, with the difference that IdenScr values are discrete and PearScr are continuous. IdentScr and PearScr twilight thresholds are approximately one unit higher than StrIdent ones, in agreement with the observation that structural alignments do not always match sequence ones (Figure 3C).

Evaluation of the reliability of *cis* element alignment strategies

ROC curves indicate that the most accurate function to evaluate *cis* element similarity is StrScr, reporting AUC in the range 0.98–1.00 (Figure 5, see also Supplementary Figure S3 and Supplementary Table S6).

StrIdent AUC values are in the range 0.82–0.88, not as higher as StrScr values because DBD structural similarity does not always imply *cis* element sequence similarity (as illustrated in Figure 3C). These observations further denote that the structure of *cis* elements is more conserved than their sequence. Nevertheless, the relatively high StrIdent AUC values suggest that sequence alignment guided by structural superposition is a reasonably accurate method in most cases.

The remaining sequence-based methods give lower AUC values, but there are differences in performance. PSSM strategies PearScr and EvalScr report the highest AUC values for most alignment lengths, so these methods

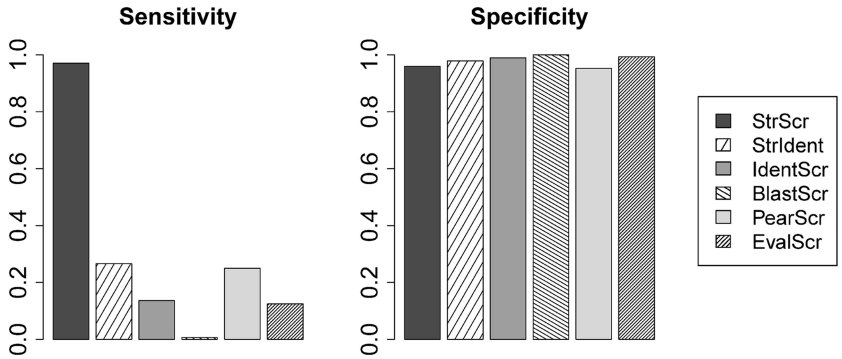


Figure 6. Sensitivity (true positive rate) and specificity (true negative rate) of different alignment algorithms and scoring functions when using the 95% twilight thresholds to recognize true and false *cis* element alignments.

should be preferred in order to distinguish true and false alignments. Both PSSM methods were expected to be more accurate than other sequence-based scores as they use matrices, richer than sequences in information content. It is also remarkable that EvalScr improves its accuracy as alignment length increases, again as expected, yielding an accuracy comparable to that of StrIdent for an alignment length of 7.

BlastScr appears to be the most erratic scoring function in this benchmark, as it produces the lowest AUC values on average. This behavior is a consequence of the limited range of scores that these functions produce for short DNA fragments, as already explained. However, this strategy performs remarkably well among alignments of six bases, for which there are many true and false instances in the training set.

Sensitivity and specificity of the different strategies were calculated upon enforcing the twilight thresholds to the alignment data. The results are summarized in Figure 6. Specificity was of course ~0.95 in all cases, as a consequence of using the 95% twilight thresholds. The sensitivity results, which estimate the fraction of true alignments recovered, are therefore more interesting. Apart from the maximum value of StrScr (0.97), all other approaches exhibited low sensitivity values, even StrIdent (0.27). This is another evidence suggesting that structure is much more conserved than sequence in TF–DNA complexes. This also means that as much as a third of structurally similar TF–DNA complexes can be successfully predicted by aligning their *cis* elements. Among sequence-based methods, only PearScr seems to achieve comparable sensitivity (0.25), which can probably be a consequence of the high information content of PSSMs. Overall, EvalScr is shown to be less sensitive than PearScr, even when both strategies were applied to exactly the same set of alignments. This averaged behavior does not challenge the previous observation that EvalScr improves as alignment length increases (Figure 5). Moreover, the analysis indicates that IdentScr should be preferred over BlastScr when aligning sequences instead of PSSMs.

The reported discrepancy between StrIdent and IdentScr seems to suggest that in many cases sequence alignments do not agree with structural alignments, as the example in Figure 3C. To obtain statistics about this kind of agreement, 535 true structure-based *cis* element alignments that could also be aligned by all sequence-based techniques were compared, as explained in 'Materials and Methods' section. The proportion of correct alignments produced by each method is reported in Supplementary Table S7, where results are ranked in terms of %sequence identity measured in the corresponding structural alignments. The comparison shows that when sequence identity is >80%, all methods produce >90% of correct alignments, as it would be expected for easy cases, but their performances notably differ below this point, with BlastScr displaying the worst accuracy on average. Both IdentScr and PearScr have similar performances across the board, with a significant drop of accuracy when %identity falls below 60. Taking all easy and hard cases together, we find that about half of aligned *cis* elements do not agree with the corresponding structural alignments.

A survey of twilight thresholds in two repositories of experimentally determined *cis* elements

In order to check the biological relevance of the calculated twilight thresholds, we analyzed PSSMs and individual DBSs from the curated, high-quality database TRANSFAC and from PBM experiments annotated in UNIPROBE. Theoretical thresholds were used to be able to evaluate alignments of up to 8 nt. In these experiments we used only the core regions of PSSMs and DBSs to avoid uncertainty, where the core is defined as a central region with high information content (see 'Materials and Methods' section for more details). It has been shown that the conservation of *cis* element individual positions is proportional to their number of protein contacts, so by dissecting the core regions we are expecting to effectively remove non-contacted nucleotides (46). Each individual core site was compared against its corresponding PSSM core (TRANSFAC alignments were implicit by positions defined in the original database) and the obtained score was compared with the twilight threshold, as illustrated in Figure 7. Results in Figure 8 suggest that a majority of DBSs produce scores over the cutoff values. This exercise reveals that the sensitivity of twilight thresholds when comparing sites recognized by the same protein is ~80% for TRANSFAC data, much higher than the previously

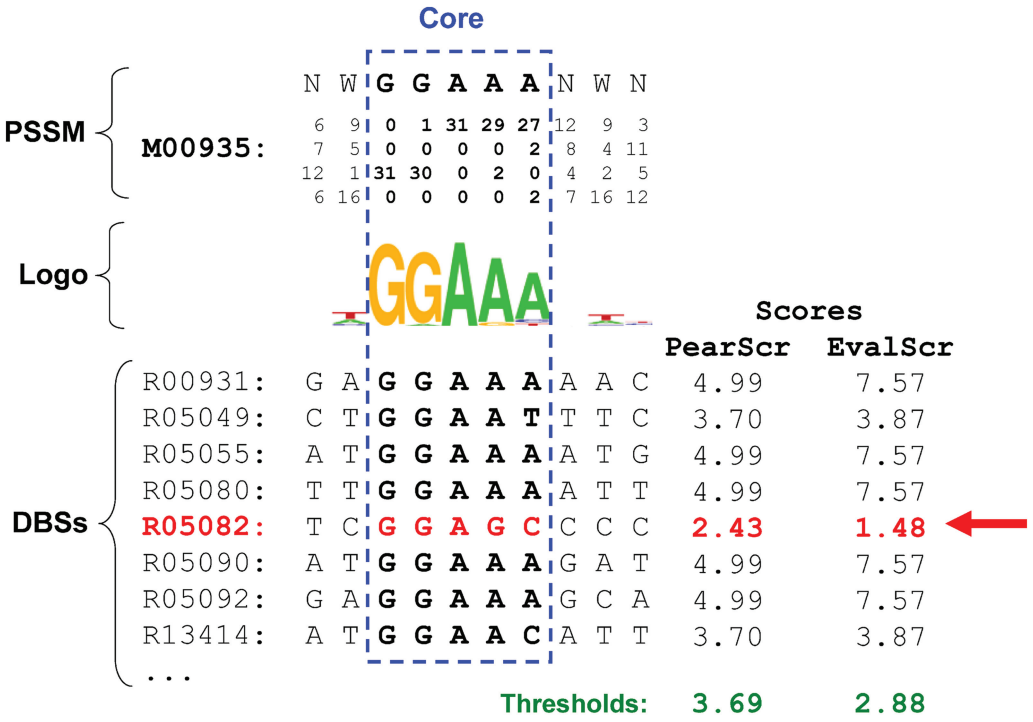


Figure 7. TRANSFAC *cis* elements for NF-AT transcription factor aligned to their respective PSSM cores (in bold), shown in the middle as a sequence logo. Numbers to the right are PearScr and EvalScr scores. Most aligned elements have scores above the respective twilight thresholds (3.69 for PearScr and 2.88 for EvalScr), while one site is clearly below (red color).

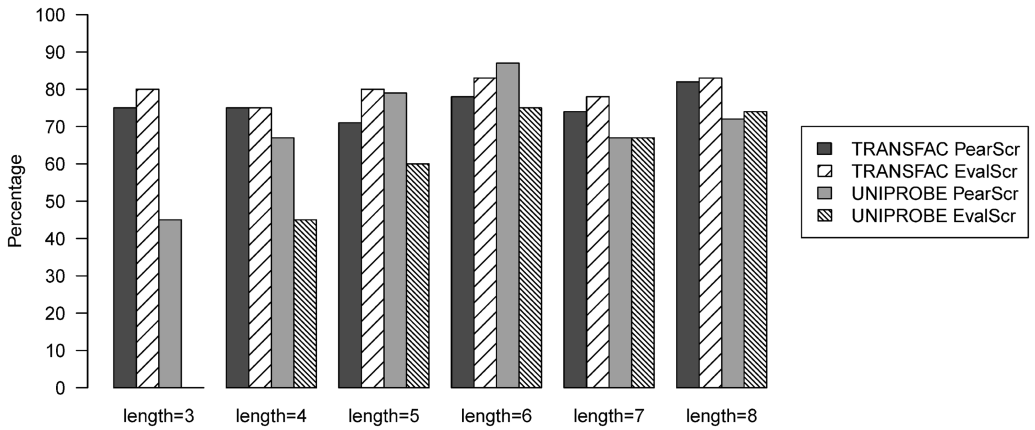


Figure 8. Percentage of individual *cis* elements with alignments scoring over the respective twilight thresholds in TRANSFAC (all annotated sites) and UNIPROBE (top 100 sites).

discussed figure of 25%. However, it is remarkable that a non-negligible number of elements diverge significantly with respect to their PSSMs. This is more noticeable among UNIPROBE alignments, likely because we arbitrarily considered the top 100 DBSs for each motif analyzed, which is a larger and more variable set than the average 21 sites per PSSM in TRANSFAC. These results reflect the natural binding plasticity of TFs, but we cannot rule out that some of these sites might have been incorrectly annotated as a consequence of experimental errors. When we align all microarray sites against the corresponding UNIPROBE PSSMs, it is observed that most of them have scores under the twilight thresholds, as would be expected, as most would not contribute to the resulting PSSM due to their low affinity. Supplementary Table S8 includes the results for other complementary tested datasets, as explained in 'Materials and Methods' section.

DISCUSSION

Alignment of genomic regulatory sequences is by nature a challenging problem, since it is not easy to separate real sequence signal from noise. Most efforts in this field have been focused on the comparison of promoter regions of co-regulated genes with the aim of finding

over-represented oligonucleotides, which might later be confirmed as relevant *cis* elements (10). A more recent view of the problem requires finding one or more over-represented motifs within a collection of sequenced immunoprecipitated chromatin fragments (6). Software tools devoted to this task can be termed as 'pattern discovery' computer programs. To our knowledge, it was in this context that the term 'twilight zone' was first applied to DNA sequences (29). In this work, we address a related problem that of comparing *cis* elements, an issue that nevertheless pattern discovery programs have to deal with. Indeed, a typical outcome of these tools is a PSSM that numerically captures the consensus of several putative *cis* elements found, which implicitly requires the alignment of such elements.

Inspired by previous work (27,28), here we took the set of TF–DNA complexes currently available in the PDB to derive structure-based thresholds that could be used for the evaluation of pairwise alignments of *cis* elements. Such a benchmark required the development of a structural alignment approach for TF–DNA complexes, which in summary consists of superposing the structures of a pair of protein domains and applying the obtained optimal transformation matrix to the bound DNA chains. This geometrical operation permits the direct comparison of both DNA duplexes, as done in previously published methods (17,47), and the derivation of the distributions shown in Figure 1, which ultimately defined the RMSD limits for calling 'true' and 'false' alignments of similar and dissimilar complexes, respectively (Figure 2). Taking into account these limits, the structural superposition of a pair of TF–DNA complexes can produce a biologically relevant alignment of their DNA sequences, which can subsequently be used as a reference to correct disputable sequence-based alignments (Figure 3).

By considering structural alignment as the 'gold standard' to compare different sequence alignment methods, we have calculated scoring thresholds (Figure 4A and Supplementary Table S4) for the sequence alignment techniques listed in Table 1. These thresholds can assist in removing a majority of non-reliable alignments and also have the inconvenience of losing many true alignments in the way. The differences in performance observed for StrIdent and StrScr to some extent support the statement made on proteins that 'high levels of sequence similarity or identity do not ascertain structural similarity' (27,28). Furthermore, as summarized in Supplementary Table S7, as much as 50% of the correct structural alignments can be retrieved by sequence-based techniques, the rest are erroneous alignments that maximize sequence identity but not necessarily structural similarity, as shown in Figure 3C. Despite these intrinsic limitations, we find that by fixing sequence-based thresholds the chance of aligning structurally dissimilar *cis* elements decreases in favor of matching similar ones, therefore improving alignment accuracy. For these reasons, twilight cutoffs were established to avoid aligning up to 95% of dissimilar DBSs. Our analyses indicate that the shorter the *cis* elements, the stricter the thresholds need to be, as expected and reported in earlier observations (48). In most cases, these empirical thresholds can be fitted to a linear model as function of sequence length, which potentially allows the calculation of thresholds for longer alignments (Supplementary Table S4). However, this extrapolation might not be necessary, since 94% of DBDs captured in the PDB recognize motifs between 3 and 8 nt. While these thresholds apply to single protein domains binding to DNA operator sites, many experimentally determined regulatory elements capture protein multimers bound. We believe that in most cases, these multimeric sites, often near-palindromic, can be split in individual components between 3 and 8 nt long, as exemplified in Supplementary Figure S4. However, we are also aware of complexes in the PDB in which two DBDs specifically contact the same nucleotides, and hence are not easily separated.

The number of complexes currently available in the PDB is the ultimate limit for the reliability of the empirical thresholds proposed in this work. For this reason, the results obtained with the *in silico* twilight cutoffs (Figure 4B) are very important, as they provide an independent assessment of their value. The observed agreement between structure-based and theoretical results validates our approach and paves the way for the calculation of theoretical thresholds beyond the complexes deposited in the PDB at any given time.

Among the sequence-based methods evaluated in this article, motif-based approaches outperform the rest, particularly as *cis* elements get longer. Indeed, PearScr and the derived EvalScr display better accuracy and lower thresholds than the other methods. The ROC curves for motif cores of length = 7 suggest that *E*-values should be the preferred scoring function, unless shorter DBSs are compared, in which case Pearson correlations should be preferred. It must be noted that other authors had already found these scoring functions to be superior in related work (22). In either case, when comparing regulatory elements of different DNA-binding proteins, the measured sensitivities reached 25% (for a fixed specificity of 95%). In contrast, after aligning experted-curated *cis* elements bound by the same protein, as done in our TRANSFAC analysis, the observed sensitivity was ~80%, suggesting that the twilight cutoffs are effectively able to recognize most cognate DBSs. These results suggest that twilight thresholds can facilitate the task of comparing a putative *cis* element to a collection of PSSMs (such as TRANSFAC, JASPAR or RegulonDB) in order to decide whether it might be recognized by a well-annotated TF. On average, such comparisons would successfully recognize four in five *cis* elements. Nevertheless, our results also indicate that one in five true elements, such as those annotated in TRANSFAC, would be incorrectly ruled out, revealing that PSSMs often fail to capture the specificities of regulatory sequences that drift away from the consensus. The UNIPROBE analysis reveals that dealing with high-throughput TF-binding data is not trouble-free, as there is no obvious rule for selecting the number of individual sites that will be eventually used to derive a PSSM. Microarray affinities can of course be used for this purpose. However, when affinity measurements are not available, the twilight values proposed in this article could assist in this task, despite their sensitivity

limitations. For instance, one could take the top fraction of ranked sites that would yield a similar sensitivity to that in TRANSFAC for core elements of the same length. More generally, the results of this work can be directly applied as a quality control mechanism for pattern discovery methods based on sequence alignments, contributing to the construction of quality-controlled PSSMs by reliably filtering out false alignments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–8, Supplementary Figures 1–4, and Supplementary Data 1–4.

ACKNOWLEDGEMENTS

The authors thank our colleagues from the STREG project and Carlos P. Cantalapiedra, Vladimir Espinosa and François Parcy for their comments on drafts of this article.

FUNDING

Funding for open access charge: Programa Euroinvestigación/Plant KBBE 2008 [EUI2008-03612].

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
2. Garner,M.M. and Revzin,A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res.*, **9**, 3047–3060.
3. O'Neill,L.P. and Turner,B.M. (1996) Immunoprecipitation of chromatin. *Methods Enzymol.*, **274**, 189–197.
4. Galas,D.J. and Schmitz,A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.
5. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
6. 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.
7. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.
8. Stormo,G.D. and Hartzell,G.W.,3rd. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
9. 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.
10. 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.
11. Frith,M.C., Hansen,U., Spouge,J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, **32**, 189–200.
12. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
13. 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.
14. Bryne,J.C., Valen,E., Tang,M.H., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–106.
15. 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.
16. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
17. Contreras-Moreira,B. and Collado-Vides,J. (2006) Comparative footprinting of DNA-binding proteins. *Bioinformatics*, **22**, e74–e80.
18. Contreras-Moreira,B. (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res.*, **38**, D91–D97.
19. Alamanova,D., Stegmaier,P. and Kel,A. (2010) Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC Bioinformatics*, **11**, 225.
20. 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.
21. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
22. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.
23. Frith,M.C., Fu,Y., Yu,L., Chen,J.F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.
24. Turatsinze,J.V., Thomas-Chollier,M., Defrance,M. and van Helden,J. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.*, **3**, 1578–1588.
25. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
26. Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.
27. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
28. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
29. Keich,U. and Pevzner,P.A. (2002) Finding motifs in the twilight zone. *Bioinformatics*, **18**, 1374–1381.
30. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
31. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
32. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
33. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

34. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
35. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
36. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
37. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
38. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
39. R Development Core Team. (2009) *R Foundation for Statistical Computing*. Austria, Vienna.
40. Green,D.M. and Swets,J.A. (1966) *Signal Detection Theory and Psychophysics*. New York, Wiley.
41. Sonego,P., Kocsor,A. and Pongor,S. (2008) ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief Bioinform.*, **9**, 198–209.
42. Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.
43. Thomas-Chollier,M., Sand,O., Turatsinze,J.V., Janky,R., Defrance,M., Vervisch,E., Brohee,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.
44. Noyes,M.B., Christensen,R.G., Wakabayashi,A., Stormo,G.D., Brodsky,M.H. and Wolfe,S.A. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–1289.
45. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Pena-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
46. Raviscioni,M., Gu,P., Sattar,M., Cooney,A.J. and Lichtarge,O. (2005) Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. *J. Mol. Biol.*, **350**, 402–415.
47. Siggers,T.W., Silkov,A. and Honig,B. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.*, **345**, 1027–1045.
48. Claverie,J.M. and Audic,S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12**, 431–439.