BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 20 2011, pages 2806–2811
doi:10.1093/bioinformatics/btr492

Sequence analysis

Advance Access publication September 4, 2011

SiTaR: a novel tool for transcription factor binding site prediction

Eugen Fazius[†], Vladimir Shelest[†] and Ekaterina Shelest[*]
Research Group Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, 07745 Jena, Germany

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Prediction of transcription factor binding sites (TFBSs) is crucial for promoter modeling and network inference. Quality of the predictions is spoiled by numerous false positives, which persist as the main problem for all presently available TFBS search methods.

Results: We suggest a novel approach, which is alternative to widely used position weight matrices (PWMs) and Hidden Markov Models. Each motif of the input set is used as a search template to scan a query sequence. Found motifs are assigned scores depending on the non-randomness of the motif's occurrence, the number of matching searching motifs and the number of mismatches. The non-randomness is estimated by comparison of observed numbers of matching motifs with those predicted to occur by chance. The latter can be calculated given the base compositions of the motif and the query sequence. The method does not require preliminary alignment of the input motifs, hence avoiding uncertainties introduced by the alignment procedure. In comparison with PWM-based tools, our method demonstrates higher precision by the same sensitivity and specificity. It also tends to outperform methods combining pattern and PWM search. Most important, it allows reducing the number of false positive predictions significantly.

Availability: The method is implemented in a tool called SiTaR (Site Tracking and Recognition) and is available at http://sbi.hki-jena.de/sitar/index.php.

Contact: ekaterina.shelest@hki-jena.de

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

Received on April 19, 2011; revised on August 16, 2011; accepted on August 19, 2011

1 INTRODUCTION

Despite the large number of available methods and tools for transcription factor binding sites (TFBSs) prediction, the problem of the high number of false positive (FP) predictions is not solved so far. The attempts to predict putative binding sites using the information about known TFBSs have a long history, with the main stream of PWM-based methods and deviations to applications of other sequence-based methods, such as Hidden Markov Models (HMMs) (e.g. Frith *et al.*, 2002), or less widely known approaches like, for example, intuitionistic fuzzy sets (IFS) theory (Garcia-Alcalde *et al.*, 2010). Further refinement of predictions can be done by including additional information, such as genetic and epigenetic factors in the query region, e.g. clade-specific evolutionary parameters, presence of nearby coding regions (Fu *et al.*, 2009) or phylogenetic relationships (e.g. Berezikov *et al.*, 2007; Hestand *et al.*, 2008) or/and genome-wide expression data (e.g. Siewert and Kechris, 2009). These ways are promising and have their field of application, and they also have a serious drawback: this additional information must be available, which is far from being always the case.

To find an optimal balance between high specialization provided by pattern-based methods and high sensitivity of the PWMs, several attempts have been made to combine these two approaches. These attempts were implemented in such tools as AliBaba2 (Grabe, 2002) and P-Match (Chekmenev *et al.*, 2005). AliBaba2 method is based on the examination of a TFBS context by application of an extension of the Berg and von Hippel model (Berg and von Hippel, 1987). It includes a step of aligning of known TFBSs with a query sequence with subsequent construction of a PWM matching each found motif. The P-Match algorithm uses individual TFBSs along with the corresponding PWMs. It searches for DNA subsequences matching one of the TFBSs from the set and calculates the matching score using the weight matrix. Both approaches, although they still include the obligatory step of the PWM construction, are a significant step forward in comparison with 'pure' PWM-based methods. Being methodologically interesting and promising, these tools are, unfortunately, hardly of use in practice: AliBaba2 is not maintained since 2002, and both tools are dependent on available PWM libraries, which are outdated in their public versions.

At the end of the day, PWMs remain the state of the art method. Consequently, the great majority of available tools are PWM based (see, for example, Bryne *et al.*, 2008; Cartharius *et al.*, 2005; Kel *et al.*, 2003) and differ mostly in the ways of calculating the cut-offs. Introduction of *e*-values, *P*-values and background modeling [e.g. in Matrix-scan of RSAT (Turatsinze *et al.*, 2008)] allow to reduce the number of FP to some extent. Other tools tend to use more mathematically sophisticated methods, such as variable order Bayesian trees and variable order Markov models [e.g. VOMBAT tool (Grau *et al.*, 2006)] without breakthrough results. We think, however, that the problem of motif prediction should not be approached by generalizations. We decided to develop a straightforward method based on direct comparison of input (searching) set motifs with the query sequence (see the Section 2). We exploit the idea of non-randomness of biologically meaningful sequence motifs, which can be detected by comparison of observed motifs' occurrences with those predicted to happen by chance. This simple mathematical approach provides high sensitivity and specificity of the method, allowing significant reduction of the FP predictions.

[*] To whom correspondence should be addressed.

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

We compared our method, implemented in a tool called SiTaR (Site Tracking and Recognition), with P-Match (http://biobase-international.com/, Chekmenev *et al.*, 2005) and a state of the art PWM-based tool Jaspar (http://jaspar.genereg.net/, Bryne *et al.*, 2008). SiTaR shows better performance false discovery rate (FDR) and F-measure, which is the weighted harmonic mean of precision and recall (see more details in Section 3.2).

In six independent cross-validation experiments for arbitrarily taken (TFs), we show that our method is not inclined to overfitting and can predict new motifs.

2 METHODS

2.1 SiTaR method

The method is based on the idea that we can calculate the number of motifs with given base composition and given number of mismatches in a random sequence (also with the given nucleotide composition). These predicted numbers can then be compared with the real occurrences of motifs in a query sequence.

Let the length of the random sequence be L, the length of the motif be l, the fractions of each nucleotide i in the motif being f_i and in the random sequence F_i. It is easy to show (see Supplementary Material) that the number of motifs μ with 0 mismatches can be calculated as:

$$M_{\mu,0}=(L-l+1)\prod_{i=1}^{4}F_i^{lf_i}, \tag{1}$$

With 1 mismatch:

$$M_{\mu,1}=l\cdot M_{\mu,0}\sum_{i=1}^{4}\left(f_i\frac{1-F_i}{F_i}\right), \tag{2}$$

and so on, see Supplementary Material for more details.

We can compare these numbers with the real motif occurrences. To count the real motif occurrences, the query sequence is scanned for motifs with a sliding window (Fig. 1, Step 1). Each motif of the input TFBS set [searching motif (SM)] is used as a search template with different number of mismatches (normally not more than one-third of the motif length; in reality it is up to four mismatches).

To calculate scores for the found motifs (FM), we first assign weights to all SMs. The weight for a searching motif μ with the number of mismatches x we define as:

$$W_{\mu,x}=\frac{M_{\mu,x}^{\text{count}}-M_{\mu,x}^{\text{pred}}}{\sqrt{M_{\mu,x}^{\text{pred}}}}, \tag{3}$$

where $M_{\mu,x}^{\text{count}}$ and $M_{\mu,x}^{\text{pred}}$ are measured and predicted occurrences, accordingly. The weight is, in fact, the overrepresentation of the counted motifs with the given number of mismatches over the calculated number of such motifs occurring by chance.

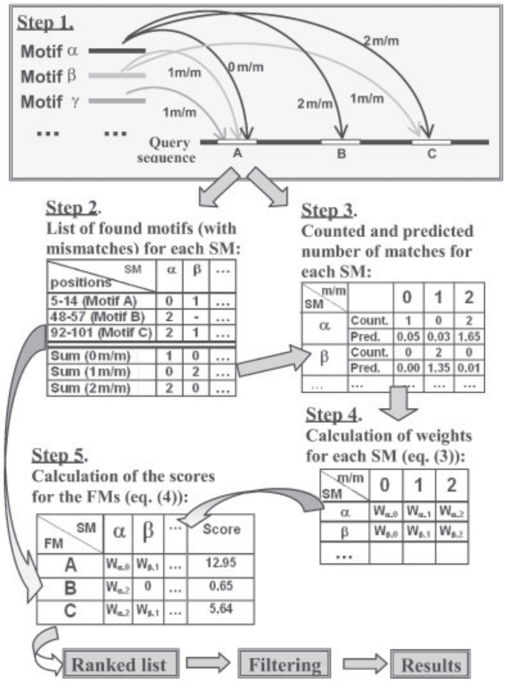
After the weights are assigned to each SM, the score for a found motif $\boldsymbol{A}$ is calculated as

$$S(A)=\max_{\mu}(W_{\mu,x(\mu)})\alpha\beta, \tag{4}$$

where $\max_{\mu}(W_{\mu,x(\mu)})$ is the maximal weight for all searching motifs, α is the portion of the SM set that have matched $\boldsymbol{A}$ with positive weight and β is the maximal match (i.e. the maximal portion of $\boldsymbol{A}$ symbols that are identical to SM symbols).

2.2 Filtering procedure

After the search is run (Fig. 1, Step 1) and the predictions for the putative motifs and their scores are made (Fig. 1, Steps 2–5), the list of the predicted motifs should be filtered to obtain the optimal results.


Step 1.
Motif α
Motif β
Motif γ
...
...
2m/m
0m/m
1m/m
1m/m
2m/m
1m/m
Query sequence
A
B
C
Step 2.
List of found motifs (with mismatches) for each SM:
Step 3.
Counted and predicted number of matches for each SM:
Step 4.
Calculation of weights for each SM (eq. (3)):
Step 5.
Calculation of the scores for the FMs (eq. (4)):
Ranked list
Filtering
Results

SM / positions	α	β	...
5-14 (Motif A)	0	1	...
48-57 (Motif B)	2	-	...
92-101 (Motif C)	2	1	...
Sum (0m/m)	1	0	...
Sum (1m/m)	0	2	...
Sum (2m/m)	2	0	...

m/m / SM		0	1	2
α	Count.	1	0	2
	Pred.	0.05	0.03	1.65
β	Count.	0	2	0
	Pred.	0.00	1.35	0.01
...	...	...	...	...

m/m / SM	0	1	2
α	$W_{\alpha,0}$	$W_{\alpha,1}$	$W_{\alpha,2}$
β	$W_{\beta,0}$	$W_{\beta,1}$	$W_{\beta,2}$
...			

SM / FM	α	β	...	Score
A	$W_{\alpha,0}$	$W_{\beta,1}$	...	12.95
B	$W_{\alpha,2}$	0	...	0.65
C	$W_{\alpha,2}$	$W_{\beta,1}$	...	5.64

Fig. 1. Main steps of the SiTaR algorithm. Step 1 (Search): each searching motif (SM) is used as a search template with different number of mismatches; Step 2: for each FM, we mark the number of mismatches between it and each searching motif; Step 3: for each SM, we write down the number of its counted (in the Step 2) occurrences with given number of mismatches and the corresponding predicted value of such occurrence by chance. Step 4: the weights for each SM are calculated according to formula (3). Step 5: calculation of the scores [formula (4)]. The ranked list of the scores is filtered according to the user's wishes.

To select the best hits, we provide scores for each found motif. Those with obviously low scores are filtered out automatically, but in many cases the decision about what to consider 'good' and 'poor' results depends on specific circumstances of a search, in particular on the reliability of the input TFBS set. For this occasion, we provide an opportunity to adjust the filtering threshold according to the desired number of true positive (TP) hits (estimated by the reidentification of true TFBSs inserted in random sequences, see Section 2.4) and acceptable number of FPs (estimated by the run on 'empty' random sequences). By default, the threshold is set to 100% of the TP with 1 allowed mismatch per site and the number of FP corresponding to this threshold is shown in the output. If these numbers are not satisfactory, the user can adjust them by: (i) selecting other number of mismatches; and (ii) selecting the optimal score. The search result is accompanied by a plot showing the dependency of true and FPs on the score value. The user-friendly design allows to 'move' the score threshold in the plot showing how it influences the both (TP and FP) parameters and to select the optimal combination.

2.3 Align or not align?

Principally, SiTaR does not require aligned sequences or even sequences of the same length. However, sometimes we encounter unreasonably long (up to 50 bp) TFBSs reported in literature and, consequently, in databases. [The normal TFBS length is 6–12 bp, in rare cases up to 18 bp. If a TFBS is too long (30–50 bp), it just means that the authors have not well defined the binding sites and reported approximate regions of the TFBS surroundings.] In such cases, to get searching motifs of reasonable length the reported sequences should be aligned with the other TFBSs of the set. If all TFBSs of the set are too long, we would recommend to reidentify the 'real' TFBSs within these sequences by looking for a common motif of 6–12 bp [using programs like MEME (Bailey and Elkan, 1994) or Gibbs Sampler (Thompson *et al.*, 2003)]. If the binding sites of the input set are well defined and have comparable length (allowing the same number of mismatches), they can be submitted to the program without further treatment.

We want to emphasize that it is not recommendable to look for motifs <6 bp in any case, because the probability of occurrence of such motifs by chance is too high. This is true for any kind of application independently of the searching method: SiTaR, PWMs or another.

2.4 Defining true and FPs

TFBS sets from Jaspar were used as the TP (input) sets in all experiments. In the test run on random sequences (Section 3.2.1), the motifs of the TP (input) set were hidden in a 10 kb random sequence and then reidentified. The results have been cleaned from redundancies (i.e. predictions of motifs in the same location. Location is considered the same if the predicted motifs overlap in more than half of the length). The ranked (from the highest to the lowest score) list of results was cut on the level corresponding to the desired number of TP (by default, 100%). All the predictions made besides hidden sites were considered as FPs.

SiTaR: the search was made with 0 mismatches, the threshold being adjusted if 100% of the TP (or the genuine TFBS, in the case of the test run on promoters, Section 3.2.2) has not been reidentified after the first run.

Jaspar: the search was made with a low threshold, which was then adjusted to get 100% of the TPs.

P-Match: P-Match uses the TRANSFAC matrices, which differ from those in Jaspar. To run P-Match on the exactly same sets of sequences as in the other considered tools, we created PWMs from the Jaspar TFBSs for each considered TF using Create Matrix option of the Professional version of Match^{TM}. The profiles in P-Match were applied with the lowest possible threshold (minFN), and then adjusted to get 100% of the TPs.

3 RESULTS

3.1 Validation of the method

To demonstrate that our models are not overfitted and can indeed predict new motifs, we conducted cross-validation experiments. For each of the selected TFs, the initial set of TFBSs was arbitrarily divided into two halves, one serving as a training set, the other being 'hidden' in a random sequence. The whole procedure was repeated 100 times. The idea was to check whether the approach allows to reidentify those genuine motifs, which were not used for the training.

We repeated the experiment for five arbitrarily chosen TFs from TRANSFAC database (Matys *et al.*, 2003) and one 'real-life' example taken from a paper reporting binding sites for Rlm1 TF (Jung and Levin, 1999) (Table 1). The average sensitivity and specificity by reidentification of the 'new' motifs are both 99.9%.

Table 1. Results of the cross-validation tests

TF name	No. of TFBSs	Re-identification	
		Sensitivity (%)	Specificity (%)
Srf	45	100	99.98
Mef2a	58	99.96	99.84
NFYA	114	100	99.99
Sox9	76	99.63	98.46
USF1	30	100	99.78
Rlm1	50	99.96	99.91
Average		99.93	99.66

Half of the TFBS set was hidden in a 10 000 bp long random sequence, the other half was used as training set. The analysis was repeated 100 times. The 're-identification' column shows the values of sensitivity and specificity for each TF averaged through 100 runs.

Table 2. FP predictions made by SiTaR, P-Match and Jaspar by reidentification of 100% and 90% of TP sites

TF	Jaspar ID	No. of sites	FP					
			100% recall			90% recall		
			SiTaR	P-Match	Jaspar	SiTaR	P-Match	Jaspar
ARID3a	MA0151.1	27	9	13	13	7	13	8
Arnt	MA0004.1	20	12	18	59	3	18	59
Ddit:Cebp	MA0019.1	37	0	0	47	0	0	22
Elf5	MA0136.1	44	2	3	372	2	3	32
Elk1	MA0028.1	29	0	2	256	0	2	123
Foxf2	MA0030.1	20	2	5	153	2	5	119
HNF4a	MA0114.1	66	0	0	NA	0	0	36
Mef2a	MA0052.1	58	0	1	29	0	1	1
MIZF	MA0131.1	20	0	0	184	0	0	5
NFATC	MA0152.1	25	13	20	83	11	20	58
NFIL3	MA0025.1	23	0	0	84	0	0	0
NFkB	MA0061.1	27	0	0	33	0	0	16
NFYA	MA0060.1	114	0	0	69	0	0	3
NHLH1	MA0048.1	54	0	0	7	0	0	3
NKx3-2	MA0122.1	24	2	15	335	2	9	170
Pax2	MA0067.1	30	8	20	278	6	20	158
Pbx	MA0070.1	18	0	0	68	0	0	1
RORA	MA0071.1	25	0	2	5	0	2	2
Sox5	MA0087.1	22	7	12	66	7	12	16
Sox9	MA0079.1	76	3	3	550	3	3	346
Srf	MA0083.1	45	0	0	NA	0	0	0
T	MA0009.1	20	0	0	4	0	0	0

NA, not achieved, 100% of TP could not be reidentified.

3.2 Comparison with other methods

We compared our method with Jaspar database searching tool (Bryne *et al.*, 2008) and with P-Match (Chekmenev *et al.*, 2005). Jaspar was selected as the most prominent PWM-based tool and P-Match is the representative of 'mixed' approaches, utilizing PWM information along with the alignments of individual TFBSs. The comparison was run for 22 TFs (Table 2), which were selected according to the following requirements: (i) the TF should be represented by a PWM in Jaspar; (ii) binding sites should have reasonable length (6–15 bp); and (iii) the number of sequences in the set should be not too small (>20). The sites used for the PWM construction were extracted from Jaspar and considered as the TP set.

3.2.1 Test run on random sequences with hidden true sites All motifs of the TP set were hidden in a 10 kb random sequence; any hit different from the hidden motif was considered as a FP prediction, since random sequences are not supposed to contain TPs. We considered the number of the FP predicted by each tool when recovering 100% and 90% of the TP (Table 2). For 100% sensitivity,

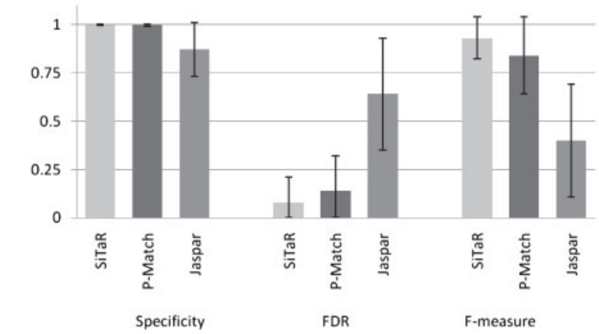


Fig. 2. Average specificity, FDR and $F_{0.5}$-measure of SiTaR, P-Match and Jaspar by 100% recall (see Supplementary Table S1 for more details).

the results of SiTaR and P-Match are comparable in two-thirds of cases, in the rest leaving the superiority to SiTaR. In comparison with Jaspar, SiTaR performs inevitably better for all considered TFs. For 90% sensitivity, Jaspar shows improved results, but still the superiority of SiTaR and P-Match is obvious.

To be able to quantitatively compare the methods, we calculate specificity [Spc = TN/(FP + TN), where TN stands for true negatives], false discovery rate [FDR = FP/(FP + TP)] and $F_{0,5}$-measure: $F_\beta = (1+\beta^2)\cdot\text{precision}\cdot\text{recall} / (\beta^2\cdot\text{precision}+\text{recall})$; in this case, β is taken equal to 0.5 to give more weight to precision than to recall. Average characteristics of each tool by 100% recall are shown in Figure 2 (see also Supplementary Material for more details). The specificity of SiTaR, P-Match and Jaspar are equally high for all tools and although there is a tendency of SiTaR and P-Match to outperform Jaspar, the difference is not significant. There is also no difference in the performance of SiTaR and P-Match in FDR and F-measure. However, these characteristics significantly differ for SiTaR and Jaspar, SiTaR showing apparently better results than Jaspar.

3.2.2 Test run on promoter sequences To demonstrate the performance of our tool on real biological sequences, we ran reprediction of known (experimentally proven) TFBSs in promoters where they belong. Information about genuine binding sites of the considered TFs was extracted from TRANSFAC Professional (version 2001.1). The sites were selected according to the following requirements: (i) there should be solid evidence of DNA binding; (ii) coordinates of the TFBS should be provided; and (iii) the corresponding promoter sequence should be available in TRANSFAC. In such way, we collected 44 genuine binding sites for 16 TFs in 38 promoters. The promoter sequences were cut to 2000 bp upstream transcription start sites (according to TRANSFAC). The search was made with the same input TFBS sets as in the previous experiment. The searching parameters were adjusted in such way that we could detect the real TFBSs. All predictions with the scores lower than the score of the genuine site were filtered out. The true TFBSs were successfully reidentified by SiTaR and Jaspar in all considered promoters (Table 3). P-Match, however, could not find the sites in nine promoters probably due to too high pre-defined cut-offs, which are not adjustable in the professional version of the tool (these cases are marked as 'not found', NF, in the Table 3).

Table 3. Results of the reprediction of genuine TFBSs in promoter sequences

TF name	Gene name	Site ID[a]	Coordinates[b]		Additional predicted sites		
					SiTaR	P-Match	Jaspar
ARNT	Cyp1a1	R00270	963	987	19	NF	34
Elk-1	EGR1	R14636	1564	1591	2	3	1
	Nr4a1	R16113	1732	1747	0	2	0
	SULT1A1	R28506	1854	1871	0	0	0
	TNF	R14625	1878	1890	26[c]	17[c]	43[c]
		R14624	1910	1922			
		R14623	1920	1929			
FoxF2	IL8	R05060	1071	1083	3	7	0
	NOS2	R05061	1205	1217	3	12	8
	Scgb1a1	R05040	1751	1765	4[c]	11[c]	17[c]
		R05041	1776	1790			
HNF4a	SERPINA1	R00114	1849	1884	0	0	0
	Transferrin	R01457	1924	1949	0	0	1
Mef2a	Ckm	R03585	843	860	0[c]	1[c]	1[c]
		R00244	947	976			
MIZF	RB1	R23825	1934	1956	0	0	0
NFATC	BACE1	R29890	1389	1409	0	10	1
	ITGB3	R24474	893	925	2	8	2
	IRF4	R24127	1582	1624	11	20	21
NFIL3	Abcb1a	R29073	1803	1834	7	NF	13
NFkB	IRF1	R29148	1715	1734	0	0	0
	IL8	R02909	1899	1908	0	1	0
NFYA	CDK1	R04344	1937	1958	0[c]	2[c]	1[c]
		R12441	1970	1990			
	Sox2	R22697	1930	1970	0	3	1
	Rrm2	R25977	1746	1783	2	3	3
Pax2	Tg	R08566	1920	1936	4	66	3
	TPO	R08567	1846	1868	6	16	44
	Gcg	R25062	1726	1771	2	21	7
PBX1	IL10	R25317	457	475	0	4	6
	NPY	R20786	300	318	1	NF	5
RORA1	Crygf	R03963	1789	1814	0	0	0
	NR1D1	R16582	1831	1853	9	NF	12
	RARB	R19784	1744	1755	1	NF	6
	Rbp1	R20171	840	850	2	0	0
SOX9	COL9A1	R23225	1563	1592	2	4	2
SRF	EGR1	R09707	1561	1581	2	4	4
	ACTA1	R00036	66	85	0	NF	0
	APOE	R00142	1664	1684	13	NF	20
USF1	Fgg	R00441	1918	1929	0	0	0
	APOE	R29834	1896	1915	2	NF	10
	APOC3	R13028	1900	1915	9	NF	20
		R13029	1929	1956			

NF, not found.

[a] TRANSFAC identifiers.

[b] Coordinates relative to the –2000 bp upstream the TSS.

[c] Threshold is shown for detection of all known TFBSs in the promoter.

In approximately 25% of promoters (9 of 38), the performance of the three tools was similarly successful: the number of predictions besides the known true sites was 0 or 1 (Table 3: R28506, R00114, R01457, R03585, R00244, R23825, R29148, R02909, R03963, R00441). However, promoter sequences are hard to interpret in terms of FP predictions, because we practically never can be completely sure in *non*-functionality of a predicted binding site. This kind of negative evidence is not published, and actually not investigated. On the other hand, our observations made for real TFBSs in mammalian promoters (based on the data from TRANSFAC) suggest that the number of occurrences of binding sites for the same TF in one promoter is rarely >10. Taking this into account, we can assume that all predictions that are <10 per promoter are equally good and can be considered not as false, but as potentially functional TFBSs. Under this assumption, we consider additional predictions as potentially true in 10 of 38 promoters (∼25%) (Table 3: R14636, R16113, R05060, R24474, R04344, R12441, R22697, R25977, R25317, R20171, R23225, R09707). In seven promoters, SiTaR is superior to P-Match, but the results of SiTaR and Jaspar do not differ (R05061, R29890, R08566, R25062, R20786, R19784, R00036). Only in one case of 38 considered SiTaR was outperformed by P-Match (Elk1 in TNFalpha promoter: one of the three known sites had a low score, hence the number of additional predictions grew to 26 opposed by 17 in P-Match, the number of Jaspar predictions being 43; we must admit, however, that the performance of all tools was not optimal in this promoter). SiTar was not outperformed by Jaspar in any of the considered promoters. Finally, in nine cases (23%) (R00270, R05040, R24127, R29073, R08567, R16582, R00142, R29834, R13028, R13029), SiTaR outperforms the both competitors.

All in all, SiTaR performed equally or better in 37 promoters of 38 considered.

4 DISCUSSION

SiTaR is a method for nucleotide composition-based detection of non-random matching motifs applied to prediction of TFBSs or other types of motifs. One of the evident advantages of the method is that it does not require equal length of the searching motifs, nor their alignment, nor a construction of a PWM nor any other modeling prior to the search. Any of those steps (aligning, trimming to equal length, etc.) leads to the loss of information (as any generalization) and hence introduces some degree of uncertainty in the final result. Thus, avoiding these steps we preserve as much information as there was in the initial set of TFBS sequences. It is notable that the demand for a high quality of the initial (searching) set of TFBS sequences is equally strong for any approach: ours or PWM based.

The other, quite important, consequence of the lack of the above-mentioned steps is saving of time and efforts. To use SiTaR, one needs to have just a set of known TFBS sequences; this can come directly from a publication or a database. To use any PWM-based tool, one needs to have the PWM. It either can be found in a matrix database, such as TRANSFAC or Jaspar, or has to be made *de novo* (from that set of known sequences, which we directly use in SiTaR). To do that, one has to (i) align the sequences; (ii) find a tool for PWM construction; and (iii) apply it. Of note, Jaspar and other prominent PWM-based tools like Matrix-scan or the public version of P-Match do not allow to construct own matrices.

This means that even when SiTaR demonstrates equal performance with a PWM-based tool, this result is achieved by lower expenses.

The experiment with reidentification of the TP set with three different methods showed general superiority of the tools that use aligning of individual motifs (SiTaR and P-Match). This conforms to the thesis that PWMs are too generalized. P-Match tries to overcome the PWM generalization by including the step of aligning of individual motifs of a PWM with subsets of the query sequence and applying the PWM characteristics when calculating the scores. This attempt significantly improves the prediction in terms of the number of FPs. However, P-Match remains dependent on PWMs, with all consequences discussed in the previous paragraphs. In SiTaR, we omit the generalization step completely.

The positive feature of PWMs is that they reflect conservation of particular nucleotides within the motif. This kind of information is not directly considered by our method, since we allow mismatches in any position of the motif. However, the theoretically allowed mismatches in the conserved positions are in practice corrected by matching of the ‘correct’ (true) motifs and accumulation of the correct matches. So the conservation of a certain base is in fact taken into account in an indirect way. To reinforce this effect, we put special emphasis on the multiple matching of the motifs when calculating the score. On the other hand, one should not underestimate the fact that TFBS sets can be very divergent, and some single true sites may strongly differ from the rest of the set (in fact, they can be completely unalignable). In a PWM (and, hence, in P-Match), such sites are in a way suppressed, so that their content is not reflected neither by the consensus sequence nor by the weights assigned to each position. On the opposite, our approach allows each motif to play its role independently of the rest of the set. So finally SiTaR is balancing between multiple and individual matching, allowing each of them to be equally successful if the found motif is recognized as non-random.

In terms of accuracy of predictions, by 100% of sensitivity SiTaR is superior to Jaspar when reidentifying TFBSs in random sequences (Fig. 2, Supplementary Table S1). The results of P-Match and SiTaR do not significantly differ, although also here there is a slight tendency of SiTaR to have lower FDR and higher $F_{0.5}$-measure (Supplementary Table S1).

Ideally, all sites presented in databases like TRANSFAC should be genuine, but we cannot exclude the possibility that one or two of them are actually false, simply because of the FP rate of the wet-lab experiments. For this reason, it not always makes sense to insist on the recovering of the whole TP set; in fact, in some TFBS sets we can observe just one sequence obviously not belonging to the rest of the set, ‘spoiling’ the whole picture. Taking into account such chances, we also considered reidentification of 90% of the TP. For this level of sensitivity, the difference in the performance of all considered tools is not anymore significant; however, the tendency of SiTaR to outperform Jaspar and P-Match remains the same (Supplementary Table S1).

In promoter sequences, the performance of SiTaR is comparable with that of Jaspar in 27 of 38 promoters (∼70%); in the rest 30%, however, SiTaR predicts less additional (potentially FP) sites. Since we do not know the real number of TP TFBSs in genuine promoters, we cannot estimate such characteristics as FDR, but we can judge about the sensitivity of the methods. In the considered examples, we can see that the sensitivity of Jaspar and SiTaR is equally high (100% of the real TFBSs are redetected), whereas the sensitivity of P-Match remains quastionable (we will return to this point in a few lines).

Our demands to the interpretation of the predictions in promoters can be formulated as following: (i) the true TFBS has to be recovered; (ii) there should not be too many TP sites (we allowed up to 10 additional TFBSs considering them as potentially TP; an excess over 10 was considered as potential FPs). Both criteria are met by SiTaR with success. It reidentifies all true TFBSs (of note, for this analysis we tried to select those true TFBSs that have not been used in the SM set); the number of potential FP predictions made by SiTaR is in average lower than that predicted by P-Match and Jaspar.

The other advantage of our tool is the possibility to interactively adjust the filtering threshold. The importance of this procedure is illustrated by the problems encountered by P-Match by redetection of TFBSs in real promoters. In SiTaR and Jaspar, the cut-offs are adjustable and this allows to reidentify all true TFBSs. In P-Match, on the opposite, there are only three fixed cut-offs and the lowest of them (MinFN, minimal false negatives) appeared to be still too high for the reidentification of the genuine TFBSs. As a result, the tool fails because of the problem that has nothing to do with the algorithm.

Normally, to change the cut-offs one has to return on the input page and change some input parameters. In some tools (like P-Match and Match^{TM} in TRANSFAC), a special procedure of creating a matrix profile is needed. In the others, like Jaspar, it is not clear what the score means, so the procedure of the threshold selection is 'blind'. Having had a severe experience with other tools, we endeavored to make the procedure of the threshold adjustment clear and simple. The user can watch simultaneously: the selected number of mismatches; the current score; the corresponding numbers of FP and TP. The interactive plot allows to see what will be the values of TP and FP if the score changes.

5 IMPLEMENTATION AND AVAILABILITY

The SiTaR tool is implemented in PHP. It is freely available at http://sbi.hki-jena.de/sitar/index.php.

ACKNOWLEDGEMENTS

We would like to thank Reinhard Guthke and Daniela Albrecht for critical reading and valuable comments on the article. We thank the first users of SiTaR for their trust, comments and suggestions for the interface design.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36.

Berezikov,E. *et al.* (2007) Exploring conservation of transcription factor binding sites with CONREAL. *Methods Mol. Biol*., **395**, 437–448.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol*., **193**, 723–750.

Bryne,J.C. *et al*. (2008) JASPAR, the open access database of transcription factor-bindng profiles: new content and tools in the 2008 update. *Nucleic Acids Res*., **36**, D102–D106.

Cartharius,K. *et al*. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.

Chekmenev,D.S. *et al.* (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.*, **33**, W432–W437.

Frith,M.C. *et al*. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*., **30**, 3214–3224.

Fu,W. *et al.* (2009) DISCOVER: a feature-based discriminative method for motif search in complex genomes. *Bioinformatics*, **25**, i321–i329.

Garcia-Alcalde,F. *et al.* (2010) An intuitionistic approach to scoring DNA sequences against transcription factor binding site motifs. *BMC Bioinformatics*, **11**, 551.

Grabe,N. (2002) AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol*., **2**, S1–S15.

Grau,J. *et al*. (2006) VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees. *Nucleic Acids Res*., **34**, W529–W533.

Hestand,M.S. *et al*. (2008) CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. *BMC Bioinformatics*, **9**, 495.

Jung,U.S. and Levin,D.E. (1999) Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway. *Mol. Microbiol.*, **34**, 1049–1057.

Kel,A.E. *et al*. (2003) MATCH A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.

Matys,V. *et al*. (2003) TRANSFAC® transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Siewert,E.A. and Kechris,K.J. (2009) Prediction of motifs based on a repeated-measures model for integrating cross-species sequence and expression data. *Stat. Appl. Genet. Mol. Biol*., **8**, Article 36.

Thompson,W. *et al.* (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res*., **31**, 3580–3585.

Turatsinze,J.V. *et al*. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc*., **3**, 1578–1588.