A promoter model is a combination of sequence elements modulating transcriptional regulation (Brazma et al., 2002). We could show that each real CE can be characterized by a ‘preferred’ distance or set of distances (Shelest, 2006). We have also shown that the approach of distance distributions can correctly identify the distance in real CEs.

2 METHODS

The underlying method of the distance distributions in TFBS pairs has been proposed by A. Konopka (Konopka and Smythers, 1987) and V. Makeev (Makeev et al., 2003), but in both cases the authors consider the distances only between neighboring binding sites, not taking into account all possible pairs as it is done in our approach, and do not allow a shift (see below). Briefly, we suppose that positions of the false positive TFBS are random, whereas the distances between the constituents of the functional pairs are not random. Let us consider a model system with uniform random distribution of points on a segment, where the points are representing the binding sites thus neglecting the extension of TFBSs. We consider a sequence of the length L, in which we find M_A sites of type A and M_B sites of type B, where A and B are being distributed randomly. It is evident that we then get M_A M_B pairs, the maximal number of different combinations of positions being L^2 (supposing that different sites can occupy the same place). We want to estimate what will be the distribution of the distances between sites A and B (A–B and B–A are considered as the same pair). Since the distances in genuine TFBS pairs are not rigid, we allow a slight shift of the sites, considering the pairs on some distance interval. Let \( f_{d, \delta} \) be the number of pairs on the distance interval from \( d \) to \( d + \delta \). It is easy to show that:

\[
\begin{align*}
\sum d \geq L - 1 & \quad f_{d, \delta} = \left( \frac{L}{2} + 1 \right) \frac{M_A M_B}{2} \delta, \quad 1 \leq d \leq L - 1 \\
\sum d = L - 1 & \quad f_{d, \delta} = \left( \frac{L}{2} + 1 \right) \frac{M_A M_B}{2} \delta, \quad 0 \leq d \leq L - 1
\end{align*}
\]

Note that factor \( M_A M_B / L^2 \) is the product of frequencies of TFBSs A and B, which were predicted in one sequence. In the case of a set of \( N \) sequences this factor must be changed to \( \sum_{k=1}^{N} \frac{M_A M_B}{L^2} \). To estimate the error of the predictions, we conducted computer simulations showing that the standard deviation is \( \sigma \approx \sqrt{f_{d, \delta}} \) (data not shown). The theoretically calculated distribution of distances between random sites will be called further on random distance distribution.

3 DistanceScan Overview

The workflow includes four main steps: (i) identification of pairs; (ii) measuring the number of pairs in a distance interval; (iii) calculation of random distance distributions for the sequence sets; and (iv) selection of the overrepresented peaks.
3.1 Identification of pairs
Presently DistanceScan is working with three kinds of inputs: (i) the results of a PWM-based TFBS prediction (MatchTM output; Kel et al., 2003); (ii) the results of motif prediction (Gibbs Sampler output; Thompson et al., 2003); and (iii) the results of two-step prediction of motifs and corresponding PWM search (MEME/FIMO output; Bailey and Elkan, 1994). DistanceScan considers all the positions of all predicted TFBSs or motifs. Further on, it examines all possible combinations of the positions, thus listing all possible pairs in the sequence. The distances are measured between the centers of the sites.

3.2 Counting the number of pairs at a certain distance in sets of sequences
The number of pairs \( f_{ob} \) in the set of sequences in a distance interval from \( d \) to \( d+\delta \) is directly counted for each \( d \) up to a maximum specified and for \( \delta \) up to a maximum specified. \( d \) and \( \delta \) are defined in special fields of the input form.

In some analyses, we do not expect that all sequences in the set contain the TFBS combinations (not all data are reliable, etc.). For such cases, the user has an opportunity to specify the minimal portion of sequences with pairs.

3.3 Calculation of random distance distributions for sets of sequences
The random distance distribution is calculated as described above, based on the lengths of sequences and frequencies of TFBSs in the considered set. Presently, the tool can work only with sequences of equal lengths.

3.4 Decision for the overrepresentation
To characterize the overrepresentation of a pair, we define a score as:

\[
S(d, \delta) = \frac{f_{ob} - f_{od}}{\sigma}.
\]

This score shows by how much the observed number of pairs is higher than the corresponding theoretical value, but it does not characterize the statistical significance of the peak. For this, we provide \( P \)-values assigned to each pair (and the corresponding score). The \( P \)-values are obtained by simulation with 1000 iterations. A larger number of iterations would slow down the counting. With 1000 iterations, the simulation results can slightly fluctuate between two repetitions of the same search. This is non-critical, because the fluctuations do not change the order of magnitude, so they cannot influence the decision about the significance. (In the \( P \)-value, not the value itself, but the order of magnitude is informative.)

4 VALIDATION
The effectiveness of the approach was first demonstrated by the results of a PWM-based TFBS prediction (MatchTM output; Kel et al., 2003); (ii) the results of motif prediction (Gibbs Sampler output; Thompson et al., 2003); and (iii) the results of two-step prediction of motifs and corresponding PWM search (MEME/FIMO output; Bailey and Elkan, 1994). DistanceScan considers all the positions of all predicted TFBSs or motifs. Further on, it examines all possible combinations of the positions, thus listing all possible pairs in the sequence. The distances are measured between the centers of the sites.

5 IMPLEMENTATION AND AVAILABILITY
The DistanceScan tool is implemented in R. It is freely available at https://www.omnifung.hki-jena.de/Rpad/Distance_Scan/index.htm. The source code can be adjusted to another input format on request.

REFERENCES


