BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 11 2010, pages 1460–1462
doi:10.1093/bioinformatics/btq132

Sequence analysis

Advance Access publication March 25, 2010

DistanceScan: a tool for promoter modeling

Vladimir Shelest, Daniela Albrecht and Ekaterina Shelest*

Research group Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, Beutenbergstr. 11a, 07745 Jena, Germany

Associate Editor: Limsoon Wong

ABSTRACT

Summary: The state of the art in promoter modeling for higher eukaryotes is predicting not single transcription factor binding sites (TFBSs), but their combinations. The new tool utilizes a previously developed method of distance distributions of TFBS pairs. We model the random distribution of distances and compare it with the distribution observed in the query sequences. Comparison of the profiles allows filtering out the 'noise' and retaining the potentially functional combinations. This approach has proved its usefulness as a filtering technique for the selection of TFBS pairs for promoter modeling and is now implemented as a tool in R. As an input, it can use the outputs of three different TFBS- and motif-predictive tools (Gibbs Sampler for motifs, Match^{TM} and MEME/FIMO for PWM-based search). The output is a list of predicted pairs on overrepresented distances with assigned scores, *P*-values and plots showing the distribution of pairs in the input sequences.

Availability: The tool is available at https://www.omnifung.hki-jena.de/Rpad/Distance_Scan/index.htm

Contact: ekaterina.shelest@hki.jena.de

Supplementary information: Supplementary Data are available at *Bioinformatics* online.

Received on November 17, 2009; revised on March 22, 2010; accepted on March 23, 2010

1 INTRODUCTION

A promoter model is a combination of sequence elements modulating transcription and characterizing promoters of a certain gene or group of genes (Bailey and Noble, 2003). It is widely accepted that in higher eukaryotes transcription factors (TFs) tend to cooperate and act in most cases in a synergistic manner and that it is this cooperation that confers the required specificity of transcriptional regulation (Brazma *et al.*, 1998; Fickett and Wasserman, 2000). Therefore, co-regulation of genes entails the existence of a characteristic combination of transcription factor binding sites (TFBSs) in their promoters. Such combination can be characterized not only by the set of TFBS, but also by the distances between the constituents: the distance between the binding sites should allow the cooperating factors to physically interact. We suppose that the false positive predictions of the TFBS are distributed randomly, whereas the distances in the functional TFBS combinations are non-random.

The principle possibility to use distances as a distinguishing feature was proven in the investigation of the behavior of real, experimentally characterized TFBS pairs that have been shown to bind cooperating TFs. Such composite elements (CEs) are combinations of two or more TFBSs, which provide synergistic action of the TFs, qualitatively different from a purely additive effect. The most comprehensive collection of composite elements can be found in the TRANSCompel database (Kel-Margoulis *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 E. Shelest (Shelest, 2006). Similar approaches have 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 $M_{\mathrm{A}}M_{\mathrm{B}}$ 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, δ. Let $f_{d,\delta}$ be the number of pairs on the distance interval from d to $d+\delta$. It is easy to show that:

$$\begin{cases} f_{d,\delta}=2(\delta+1)\left(L-d-\frac{\delta}{2}\right)\frac{M_{\mathrm{A}}M_{\mathrm{B}}}{L^2}, 1\le d\le L-1 \\ f_{0,\delta}=\left(2(\delta+1)\left(L-\frac{\delta}{2}\right)-L\right)\frac{M_{\mathrm{A}}M_{\mathrm{B}}}{L^2} \end{cases} \quad (1)$$

Note that factor $M_{\mathrm{A}}M_{\mathrm{B}}/L^2$ 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 $\sum_{i=1}^{N} M_{\mathrm{A},i}M_{\mathrm{B},i}/L^2$. 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.

*To whom correspondence should be addressed.

3.1 Identification of pairs

Presently DistanceScan is working with three kinds of inputs: (i) the results of a PWM-based TFBS prediction (Match$^{\text{TM}}$ 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 ($f^{\text{obs}}_{d,\delta}$) 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 δ up to a maximum specified. d and δ 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: $S=(f^{\text{obs}}_{d,\delta}-f_{d,\delta})/\sigma_{d,\delta}$. 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 correct re-identification of the real, experimentally proven CEs (Shelest, 2006). Several most abundant sets of the CEs from the TRANSCompel® Professional database, release 8.4, (http://www.biobase.de; Kel-Margoulis *et al*., 2002) were chosen as positive training sets: AP-1–NF-κB (13 sequences), AP-1–ETS (15 sequences), AP-1–NFAT (12 sequences), NF-κB–C/EBP (10 sequences), NF-κB–IRF (10 sequences), NF-κB–Stat (7 sequences), NF-κB–HMG I(Y) (9 sequences) and IRF–PU.1 (9 sequences). The sequences containing the corresponding CE were prolonged around the reported binding sites by 300 bp to either side from the center of the CE. DistanceScan correctly identified all distances in the CE as overrepresented. Several long-distance peaks registered by DistanceScan but not described in literature can be considered as false positive predictions (FP rate being up to 10%). (It is questionable if these pairs are false positives or they are not mentioned in literature because nobody checked the long-range pairs.) We did not make special efforts to get rid of the long-distance pairs, leaving this decision to the user.

To demonstrate the robustness of DistanceScan and to see the variability of its results in dependency of the input files (Gibbs Sampler, Match and MEME/FIMO outputs), we ran the analysis on the same set of promoter sequences used as the input for the three tools (see Supplementary Material). For the result to be more illustrative, we used the sequences containing one of the known CEs (NF-κB–IRF). It is obvious that the results of DistanceScan depend on the rate of true positive predictions of the motif-searching tools. Given that the true positive sites are correctly predicted, the results of DistanceScan do not strongly depend on the type of the input file. Although the exact width and the number of the peaks differ for different inputs, DistanceScan correctly reidentifies all but one CE in each case [8 CEs out of 9 for the Match input (89%), 8 out of 8 for MEME and 6 out of 7 for Gibbs (85%)]. (For more details, see Supplementary Material.) We have successfully applied the tool to the prediction of the functional combinations of TFBSs in modeling different pathways in higher eukaryotes and in fungi. We could not compare the DistanceScan tool with another tool, since to our knowledge there are no available tools specifically designed for searching TFBS combinations on specific distances.

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.

Funding: Deutsche Forschungsgemeinschaft (DFG) Priority Program 1160 'Colonisation and infection by human-pathogenic fungi' (to D.A.).

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*. AAAI Press, Melno Park, California, pp. 28–36.

Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **1**, 1–10.

Brazma,A. *et al*. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*., **8**, 1202–1215.

Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.

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

Kel-Margoulis,O.V. *et al*. (2002) TRANSCompel a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.

Konopka,A.K., and Smythers,G.W. (1987) DISTAN - a program which detects significant distances between short oligonucleotides. *Comput. Appl. Biosci.*, **3**, 193–201.

Makeev,V.J. *et al*. (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res.*, **31**, 6016–6026.

Shelest,E. (2006) Genetic network of antibacterial responses of eukaryotic cells. Bioinformatics analysis and modeling. PhD Thesis, Braunschweig Technical University, Braunschweig, Germany. Available at (http://www.bioinf.med.uni-goettingen.de/fileadmin/upload/publications/theses/Shelest.pdf).

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