Sequence analysis

ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors

Dmitri Papatsenko

Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA

Received on October 25, 2006; revised on January 12, 2007; accepted on February 6, 2007

Advance Access publication February 18, 2007

Associate Editor: Alex Bateman

ABSTRACT

ClusterDraw is a program aimed to identification of binding sites and binding-site clusters. Major difference of the ClusterDraw from existing tools is its ability to scan a wide range of parameter values and weigh statistical significance of all possible clusters, smaller than a selected size. The program produces graphs along with decorated FASTA files. ClusterDraw web server is available at the following URL: http://flydev.berkeley.edu/cgi-bin/cld/submit.cgi

Contact: dxp@berkeley.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Large number of programs have been developed to identify transcription regulatory regions in genomic sequences (Alkema et al., 2004; Berman et al., 2004; Frith et al., 2003; Markstein et al., 2002; Philippakis et al., 2005; Pierstorff et al., 2006; Rajewsky et al., 2002; Sinha et al., 2006; Sosinsky et al., 2003; Waleev et al., 2006). However, this important task still represents a challenge. One obstacle is the presence of large amount of non-functional binding-site matches (Papatsenko et al., 2002). Available binding motifs are imperfect and, often, thresholds in binding motif searches are not known. In addition, search for binding-site clusters may require size of the expected clusters or window size. This adds a second ambiguous parameter to the search. A statistical solution to cluster size problem was employed by A. Wagner in r-scan analysis (Wagner, 1997, 1998, 1999; Karlin and Brendel, 1992). ClusterDraw takes advantage of the r-scan algorithm, combined with an exhaustive search over a wide range of the binding site match P-values (Lifanov et al., 2003). The program calculates cluster significance from the sum $l$ (in bases) of $N - 1$ consecutive distances between all $N$ site matches present in a cluster and determines statistical significance for every possible cluster, smaller than a given size $l_{\text{max}}$. Among all overlapping clusters, the program selects those producing the best statistical scores. The described method is equivalent to a search for the best clusters in the parameter space defined by the motif match quality, size of the resolution window and position in a sequence.

2 ALGORITHM

2.1 Calculating motif match P-values

Calculation of cumulative match P-value for a word is based on the score $M$ calculated using position-weighted matrix (Prestidge and Stormo, 1993) (PWM, see Equation 1S, Supplementary Material). First, for a PWM given, the algorithm finds all possible words producing score higher or equal than the score $M$. Then, expected frequencies of all these words are calculated using standard approach (see Equation 2S, Supplementary Material). Sum, taken over all word frequencies for the words scoring higher or equal than $M$ is the cumulative match probability $P_M$ corresponding to the matrix score $M$ (see also Equations 2S–4S in Supplementary Material):

$$P_M = \sum_{j=1}^{R_M} \left( \prod_{k=1}^{n} q_{ji} \right)$$

In this formula, $R_M$ is the rank of PWM score $M$ among the PWM scores of all possible words for the matrix; $q_{ji}$ is the genome frequency of a character in the $i$th position of a word with the score rank $j$; $n$ is the total number of positions in the motif (matrix). $M$ to $P$ conversion tables are generated for each motif at the beginning of the program.

2.2 Calculating cluster significance

Cluster significance score $E$ is calculated from the cluster size $l$, the number of matches $N$, the match probability cutoff $P$ and the number of binding motifs in the search $T$ using binomial distribution (Wagner, 1997):

$$E(N, T, P, l) = -\log \left( 1 - \sum_{k=0}^{N-1} \binom{l}{k} (PT)^k (1 - PT)^{l-k} \right)$$

In the case of searches with multiple-binding motifs, for a binding motif $t$, a matrix score cutoff $M'$ corresponding to the selected match probability cutoff $P$ is calculated. Thus, for each motif, the identified matches will have the same match probability cutoff $P$. Given the number of binding motifs $T$, the probability cutoff is equal to $PT$ (see also Supplementary Material, Equations 5S–8S).
3 RESULTS AND DISCUSSION

ClusterDraw web server has a standard common gateway interface (CGI); current settings allow processing of up to 100 KB of sequence data. For a convenience of users, motifs can be entered as multiple alignments or position frequency matrices (PFMs). Basic interface provides only three options: minimal combination of binding motifs, cluster significance cutoff and background model/organism selection. Advanced interface provides options to control maximal cluster size, minimal match P-value, statistics and graphics. By default, ClusterDraw filters overlapping binding sites by finding local maxima; however, options are available to control this function and even extract overlapping sites/composite elements (Makeev et al., 2003; Walev et al., 2006). ClusterDraw output plot is shown in Figure 1A.

To validate performance of ClusterDraw, cluster significance profiles generated by ClusterDraw were compared to profiles generated by AHAB (Rajewsky et al., 2002). The AHAB was selected as one of the algorithms less sensitive to the window size and optimized to analyze the same type of data (i.e. fly enhancers), performance of AHAB versus other programs is available (Pierstorff et al., 2006). Results of the tests were quite striking, in most cases, cluster significance profiles produced by AHAB and ClusterDraw were highly correlated (see Fig. 1); predictive power of the both algorithms was similar as well. Differences were found in the ranks of the highest scores (see arrows in Fig. 1B). One can explain the agreement between the two different programs by the fact that they both perform exhaustive local searches. AHAB identifies the best out of all possible partitions for a given set of binding motifs in a window; ClusterDraw finds the best out of all possible overlapping clusters. The considered tests demonstrate efficiency of an exhaustive search strategy in detection of regulatory regions.

Performance tests for ClusterDraw and AHAB were also run on genomic sequences from mosquito Anopheles gambia, and honeybee Apis mellifera. These sequences contained Anopheles and Apis sim enhancers, recently identified in M. Levine lab (Zinzen et al., 2006). Given binding motifs presented in Drosophila sim enhancer, the both programs were able correctly predict sim enhancers in mosquito (see Fig. 1E) and honeybee (data not shown).

Absence of the window and the match score cutoff parameters in ClusterDraw, as well as correlation of the program predictions with other programs and experimental data provides new opportunities (Clyde et al., 2003; Ochoa-Espinosa et al., 2005) in the exploration of transcription regulatory regions.

ACKNOWLEDGEMENTS

Author thanks Mike Levine, who participated in algorithm improvement and provided data for testing. The work was supported by grant from Moore foundation to the Center of Integrated Genomics, University of California, Berkeley. Funding to pay the Open Access publication charges was provided by the Center for Integrative Genomics, University of California, Berkeley. The Center is supported by a grant from Moore Foundation.

Conflict of Interest: none declared.

REFERENCES


