DREiVe

Discovery of Regulatory Elements in Vertebrates

1. Intro.

The DREiVe algorithm integrates a traditional motif-discovery algorithm SPLASH with a local permutation-clustering algorithm SpeedyClust in order to identify transcription regulatory elements (e.g. enhancers and promoters) as evolutionarily conserved, order-independent clusters of short conserved motifs. This approach is based on the assumption that expression of orthologous genes is regulated by the similar set of transcription factors and therefore orthologous regulatory regions contain sets of conserved DNA motifs. DREiVe employs a non-alignment approach and therefore is able to identify rearrangements of binding sites within regulatory regions. Order of individual motifs within orthologous clusters as well as distances between them and their multiplicity can vary. DREiVe relies only on the evolutionary conservation of the binding sites for regulating transcription factors and does not require prior knowledge of these transcription factors or their binding sites.

Below we describe the DREiVe workflow as well as the selection of parameters for DREiVe analysis and interpretation of its outputs.

2. DREiVe Algorithm.

Sequence selection

Genomic sequences, RefSeq gene annotations as well as whole-genome sequence alignments were downloaded from UCSC Genome Browser. Human genome (hg19) is used as a reference sequence. The following species can be included into DREiVe analysis: horse (equCab2), elephant (loxAfr3), dog (canFam2), cow (bosTau4), rabbit (oryCun2), guineapig (cavPor3), mouse (mm9), rat (rn4), opossum (monDom5), platypus (taeGut1), chicken (galGal3), lizard (anoCar1) and frog (xenTro2). Selection of orthologous genomic sequences for DREiVe analysis is based on the chained and netted pairwise alignments with the reference genome. Sequence alignments are used by DREiVe only for retrieving orthologous sequences for the whole gene together with its

flanking intergenic regions. Much shorter regulatory regions within these sequences are identified by a non-alignment method (see below) in order to take into account permutations of binding sites within regulatory regions.

The following sequences are excluded from the analysis: (i) exons (the user can choose to include untranslated parts of exons), (ii) interspersed DNA repeat sequences (masked by Repeat Masker), (iii) low complexity regions: di- and tri- nucleotide repeats that are longer than the minimum conserved motif length (see below) with allowed differences from perfect repeats of 10% and (iv) sequences up to the minimum conserved motif length repeated at least three times. Sequences for each species are limited by 0.5 Mb. If this limit is exceeded then the species is removed from the analysis.

Discovery of conserved motifs

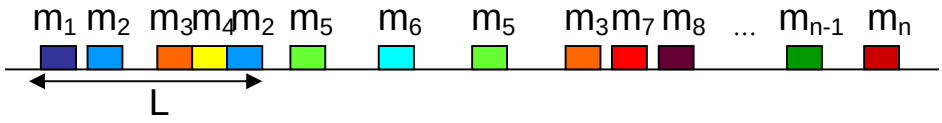
The SPLASH pattern discovery algorithm is used for the recognition of short conserved motifs that are present at least once in the minimal required number of the input orthologous sequences including reference species. These conserved motifs designate potential binding sites for transcription factors that regulate expression of orthologous genes through evolution. SPLASH discovers rigid motifs, represented as regular expressions, where positions conserved across sequences are shown by the corresponding symbol, and variable positions are represented by a ‘.’ character (e.g., AC.G.TTA.T). Conserved motifs can be located anywhere within the input sequences, including multiple times within the same sequence. The rate of false positive predictions at this step is high due to the short length of the conserved motifs and the massive amount of input sequences. Most false positive motifs are eliminated at the next step.

The following parameters determine the set of discovered motifs: (i) minimum number of sequences that must contain the motif, (ii) minimum number of matching nucleotides in the motif, (iii) motif density parameters: minimum number of matching nucleotides required over a given window length.

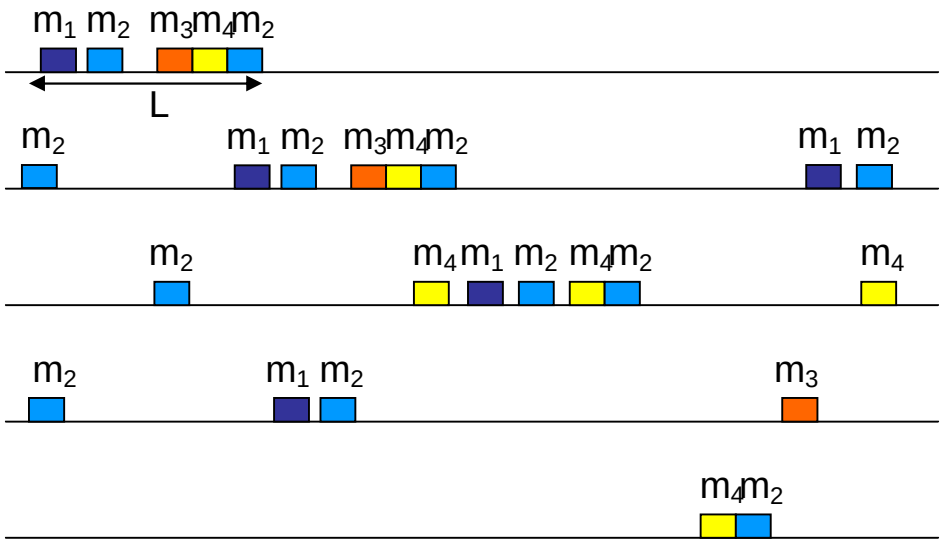
Discovery of motif clusters

Given a set of SPLASH conserved motifs our new clustering algorithm SpeedyClust identifies motif clusters that are conserved through the set of orthologous sequences. The order of individual motifs within orthologous clusters as well as the distances between them and their multiplicity can vary. Each motif in the cluster must be found in the minimal required number of orthologous clusters including the reference sequence. Orthologous clusters can be located anywhere in the input sequences.

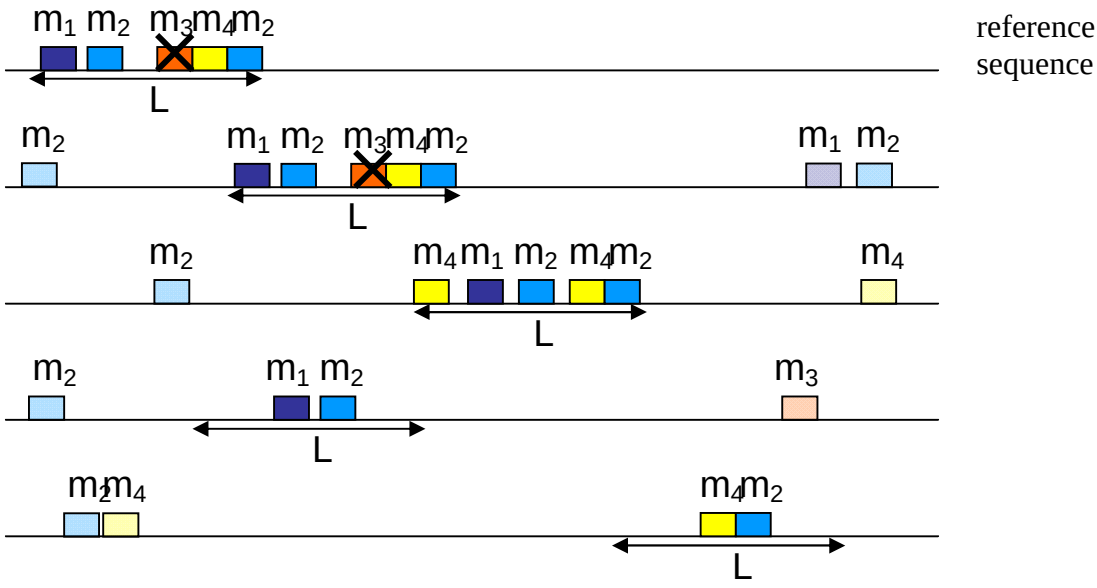
SpeedyClust output depends on two parameters: (i) maximum cluster length, *L* bp and (ii) minimum number of sequences that support each motif in the cluster, *j*. SpeedyClust analyses, in turn, each window of L bp in the reference sequence. The first window starts at the most 5' motif in the reference sequence and slides towards the 3' end making stops at each SPLASH motif. On the figure below the set of motifs {m_1, m_2, m_3, m_4} located on the reference sequence within the window of L bp forms the seed cluster.



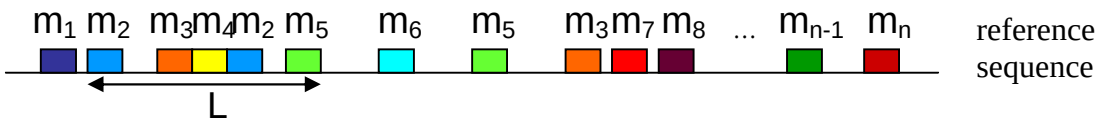
Next, all instances of motifs from the seed cluster in all non-reference sequences are highlighted.



Then the window of L bp that has the highest number of highlighted motifs for each non-reference sequence is selected. Motifs that are found in less than *j* orthologous clusters (j=3 in the example below) are eliminated from the seed cluster (m_3 in the example below).



Cluster {m_1,m_2,m_4} is saved and the SpeedyClust algorithm proceeds to the next window of L bp.



The resulting clusters are sorted by scores calculated as follows: first, each motif in the cluster is scored according to the number of orthologous clusters where it was found and the evolutionarily distances between corresponding species and the reference species (measured by the number of substitutions per fourfold degenerate site). Evolutionary distances for species whose sequences contain the motif are summarized and divided by the sum of evolutionarily distances for all species in the DREiVe analysis. Then each nucleotide in the cluster overlapping with the motif is assigned the same score as that motif. If a single nucleotide overlaps with multiple motifs, the highest score is assigned. Scores for each nucleotide in the cluster are summarized to give a cluster score. Only clusters (or sufficiently long single motifs) with scores above the cut-off of two minimal

motif lengths are reported. Motifs that did not contribute to selected clusters are eliminated. Redundant clusters are removed from the output but multiple clusters with overlapping coordinates in the reference genome can be reported.

Prediction of regulatory regions

Following cluster discovery and the elimination of random SPLASH motifs, each nucleotide in the reference sequence that overlaps with one of the saved motifs is assigned the same score as that motif. If a single nucleotide overlaps with multiple motifs, the highest score is assigned. Sequence conservation scores for each position in the reference sequence are calculated as the sum of nucleotide scores over the window of L bp which is centred at that position. In order to take into account the genomic regional variation in evolutionary rates we introduced a background correction: sequence conservation score for each position is normalized by the average sequence conservation score for the entire analysed region. Therefore all sequences with a sequence conservation score above 1 have better than average conservation. Next, nucleotide positions with a sequence conservation score above a predefined cut-off are selected. Motif clusters that overlap with the selected positions constitute DREiVe-predicted regulatory regions.

3. DREiVe Web Server.

Input.

To start DREiVe analysis the user types the search parameters into html form at http://dreive.cryst.bbk.ac.uk/search.html . DREiVe accepts gene names or RefSeq identifiers (NM_xxxxxx). It can analyze full genes, including introns as well as intergenic regions of predefined length. We suggest including in your analysis up to 50-100 kb of upstream and downstream intergenic sequences, which is expected to harbour regulatory regions for most human genes. Intergenic regions can also be limited by the neighbouring genes. Alternatively, the user can specify chromosome region from the human genome or download a fasta file that contains orthologous sequences for analysis. The user also selects a set of species from which orthologous sequences will be retrieved as well as a

minimal number of input sequences in which conserved motifs should be present. Other motif and cluster parameters include:

(i) density of motif: given the degenerate nature of transcription factor binding sites, a relatively loose density constraint of 6 exact matches for each 8bp window is recommended.

(ii) total minimal number of conserved positions in the motif: recommended range is 10-13 bp; several values can be tested in a single run.

(iii) maximal cluster length: recommended range is 300-600 bp (average length of human regulatory regions from TRRD database is 350 bp); several values can be tested in a single run.

(iv) cut-off for sequence conservation score: according to DREiVe analysis of human regulatory regions from TRRD database cut-off of 2 for sequence conservation score produces the most accurate predictions.

If the user wants to compare known promoters/enhancers with DREiVe-predicted regions of high sequence conservation, coordinates for known features can be entered in the input form and will appear on the graphical output. After submitting the query the user is redirected to the Web page that updates the status of the DREiVe job. The user can also provide an e-mail address and link to the Web page with the results will be sent to the user upon completion of the DREiVe job. DREiVe results will be saved on our server for 7 days.

Outputs.

DREiVe calculations can continue processing from several minutes to several hours depending on search parameters. DREiVe progress is reflected on the browser page that is updated every 10 seconds. A link to the set of input sequences appears after sequences have been retrieved. A message on completion of the SPLASH job appears once motif discovery is finished. Links to the final DREiVe output show SpeedyClust progress. DREiVe produces text and graphical outputs for each pair of parameters of minimal motif length and maximal cluster length (see http://dreive.cryst.bbk.ac.uk/dreive.manual/SIX1.full/main.html for example of the

output). Text output (http://dreive.cryst.bbk.ac.uk/dreive.manual/SIX1.full/300bp_10bp_clusters.seq.html) provides coordinates and sequences of DREiVe-predicted conserved regions for five different sequence conservation score cut-offs. Non-conserved sequences are masked and only conserved motifs together with flanking sequences of minimal motif length are shown. Each enhancer is linked (through "subset of overlapping clusters" link) to the page with information on motif clusters that overlaps with the predicted regulatory region. Cluster score, motif sequences and their coordinates for the reference genome and for all other species' sequences are shown for each cluster. Coordinates for motifs in non-reference genomes are positions of motifs in DREiVe-analyzed fasta file and are not the genomic coordinates (fasta file is accessible through the "Sequences" link on the main page e.g. http://dreive.cryst.bbk.ac.uk/dreive.manual/SIX1.full/fasta). Multiple occurrences of a single motif in the cluster are shown with multiple coordinates. Information is organized in the table and headings (species names) are linked to cluster sequences for corresponding species. Motifs in these sequences are highlighted with red. Therefore users can retrieve cluster sequences for any species that was included in the DREiVe analysis. Each enhancer can also be scanned against the non-redundant collection of binding profiles for vertebrate transcription factors (Jaspar CORE database). Matches with a profile score threshold of 80% are shown.

An example for graphical output for gene Six1 is shown at http://dreive.cryst.bbk.ac.uk/dreive.manual/SIX1.full/300bp_10bp_clusters.graph.html . The top line of the graph indicates the gene position on the human chromosome (hg19), the gene structure according to RefSeq annotation and the location of user-defined gene features (thick dark red line). The TSS is labeled with a dark blue flag and the genomic region between the TSS and the transcription termination site is highlighted with a thick blue line. Protein coding regions are shown as blue areas on the graph and 5'- and 3'-untranslated regions are shown in light blue. Low-complexity regions and interspersed repeats are shown with grey.

On the top panel, the local GC content is plotted using a window of 100 bp. Experimentally identified CTCF binding sites from 25 different cell lines are shown on the top panel as green rectangles. CTCF ChIP-seq binding data for human genome (hg19) from the ENCODE group in the University of Washington was downloaded from the UCSC Genome Browser. Only signal peaks that overlap in two replicates are shown. Area of the green rectangle is clickable and returns information about the corresponding cell line. Below is the list of 25 cell lines for which CTCF binding data is available:
A549 (lung carcinoma tissue, epithelium), AG04449 (buttock/thigh fibroblast, skin), AG04450 (fetal lung fibroblast, lung), AG09309 (adult toe fibroblast, skin), AG09319 (gum tissue fibroblasts, gingival), AG10803 (abdominal skin fibroblasts, skin), AoAF (aortic adventitial fibroblasts, heart), BE2_C (neuroblastoma, brain), GM12866 (B-lymphocyte, lymphoblastoid), Hasp (astrocytes, spinal cord), HAc (astrocytes, cerebellar), HBMEC (mammary epithelial cells, breast), HCM (cardiac myocytes, heart), HCPEpiC (choroid plexus epithelial cells, epithelium), HCT116 (colorectal carcinoma, colon), HEEpiC (esophageal epithelial cells, epithelium), HFF_MyC (foreskin fibroblast, foreskin), HMF (mammary fibroblasts, mammary), HPAF (pulmonary artery fibroblasts, blood), HPF (pulmonary fibroblasts, blood), HRPEpiC (retinal pigment epithelial cells, epithelium), HVMF (villous mesenchymal fibroblast cells, connective), MCF7 (mammary gland, adenocarcinoma, breast), NHDF_Neo (neonatal dermal fibroblasts, skin), RPTEC (renal proximal tubule epithelial cells, epithelium).

On the middle panel, the positions of the discovered conserved motifs along the human sequence are shown by vertical lines; the height of the line corresponds to the motif score (scale is on the left); grey horizontal line marks the cut-off for minimal motif length. Note that only motifs that contribute to the clusters are shown.

The bottom panel shows a graph of the conservation score along the human sequence (scale is on the left); grey horizontal line marks the sequence conservation score cut-off. DREiVe-predicted conserved regulatory regions are highlighted with a thick red line on the graph bottom. The area above the red line is clickable and linked to the page with information on motif clusters that overlap with predicted regulatory region (see above).

The user can choose to redraw the graph with a different sequence conservation score cut-off.

There is a fragment of DREiVe graphical output for gene SIX1 below. Five conserved regions were predicted by DREiVe upstream of SIX1 transcription start site (red lines on the bottom). One of them overlaps with experimentally identified enhancer (brown line on the top) and others are candidates for experimental validation.