Poster H-12

Regulatory motif discovery pipeline for co-regulated gene sets and chromatin-IP data: from motifs to functional instances.

Authors:
Laurence Ettwiller (*EMBL*)
Benedict Paten (*EBI*)
Mirana Ramialison (*EMBL*)
Ewan Birney (*EBI*)
Jochen Wittbrodt (*EMBL*)

Short Abstract: We present a new pipeline for discovering overrepresented motifs in pulled-down sequences from Chromatin-IP. The signal is first evaluated using random samples and motifs that constitute the signal are retrieved. These motifs are clustered and PWM(s) built. Positions of functional instances are evaluated using conservation across species.

Long Abstract:
Chromatin-IP experiments aim to find sequences bound by a particular transcription factor. With complete genomes available and advancements in microarray technology, large scale Chomatin-IP (Chip-Chip) can be conducted to obtain a genomewide picture of transcription factor binding locations, under the conditions defined by the experiment. Nevertheless, the resolution of this technology is rarely below hundreds of base pairs while the core-binding site for a transcription factor is typically around 6-8 base pairs. In order to precisely define the binding site and to identify commonly co-occurring sites (if the studied transcription factor acts in synergy with other transcription factor(s)), we present a computational method that first estimates if the sequences pulled by the ChIP experiment (sample) contain a signal in terms of over (or under)- representation of motifs compared to a background, and retrieve those motifs and their binding sites that constitute a signal.

The background definition used as reference needs to be defined but importantly no biases between the sample and the background should be found apart from the bias due to the regulatory motif over-representation. The background can therefore be all the sequences experimentally evaluated or randomly picked genomic sequences.

Because transcription factors can bind to many motifs and have variable information content and size, we evaluate a pattern alphabet of degenerate IUPAC characters to search for motifs of unlimited length but minimum occurrence number. To make this search feasible we employ a fast deterministic pruning method based upon a memory efficient suffix tree. Over-representation is calculated according to an approximate binomial model where a significance cut off is determined by a robust method of repeated background randomisation.

To decompose the set of overlapping and redundant putative motifs into a set of position

weight matrices we employ a greedy clique finding algorithm upon an undirected graph of motifs (the vertices) and their similarities (edges). Edges are created in an all-against-all procedure according to the degree of sequence overlap (which must be more than 70 percent) of motif instances in the sample sequences. In order to progressively resolve the different clusters (which constitute a final model), within sub-graphs, the most connected nodes and its directly connected nodes are removed from the graph. The operation is iterated until no cluster can be found. Clusters from the same initial subgraph constitute a family.

In order to locate the functional instances of the resulting position weight matrix (PWM,) pairwise alignments are done using orthologous sequences of the sample sequences using Blastz with relaxed parameters. Motifs that match the PWM(s) found are located in the resulting multiple alignment and each potential position are ranked according to the degree of conservation.

The complete analysis, including the alignments is summarized on a web interface that allows the user to navigate in a sequence or motif centric view.

The pipeline has been successfully benchmarked using available Chip-IP. We used a particularly well-studied example of transcription factor, the E2F family of transcription factor that binds to a well establish PWM. Using published Chip data on human [1] we were able to detect a strong signal that corresponds to the motif cluster matching the known E2F PWM. Using orthologous sequences in other vertebrates we were also able to show that the motif is overall more conserved in the sample than in randomly picked sequences suggesting an independent validation of the functionality of the newly found motifs. Furthermore, as most of the components of the cell cycle are conserved across all eukaryotes, the yeast orthologs of the human genes found to be regulated by E2F [1] were retrieved and the upstream sequences analysed by our pipeline. A signal can be detected that corresponds to the same E2F PWM found when analyzing the human sequence.

We are now successfully running the pipeline on a variety of novel experimental data and benchmarking the procedure to extend the analysis to microarray and syn-expression data.

[1] Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev. 2002 Jan 15;16(2):245-56.