Discovery of Primary, Cofactor, and Novel Transcription Factor Binding Site Motifs by Recursive, Thresholded Entropy Minimization

Ruipeng Lu[1], Eliseos Mucaki[2], and Peter Rogan(progan@uwo.ca)[1,2,3]

Departments of [1]Computer Science and [2]Biochemistry, University of Western Ontario, and [3]Cytognomix Inc., London ON.

Background:

We apply Shannon information theory to discover conserved motifs recognized by these transcription and cofactors in ChIP-Seq data from the Encyclopedia of DNA Elements. Motifs are built with Bipad, a C++ program [Nucl. Acids Res. 32.17 (2004): 4979-4991] that applies Monte Carlo-based entropy minimization to search multiple alignment space for homogeneous or bipartite models. These models can be used to determine the information contents (R_i) or binding affinity of functional binding sites and identify mutated sites.

Methods:

We built accurate information models for 168 transcription factors from unaligned sequences of ChIP-Seq fragments, biological and technical replicates, and from different cell lines. Resulting models were compared between replicates and with previously determined motifs. This process was then iterated to discover additional conserved sequence patterns in the same data. The original motif was masked, prior to derivation of a second or third model. Models consisting of low complexity, noise patterns common in intergenic regions were also thresholded to eliminate low read abundance ChIP-Seq peaks, and then reanalyzed. The quality control measures used to evaluate the accuracy of these models included: 1) determining the Euclidean distance between the current information weight matrix and previously published motifs, 2) evaluating the linearity of R_i vs binding energy to distinguish between correct and noisy, low complexity motifs, and 3) validation of predicted binding sites with experimentally proven sites in known target genes.

Results:

For 89 transcription factors, we successfully derived correct information models, and for 97 factors we discovered motifs of coregulatory cofactors. Information models for the same transcription factor were generally built on multiple replicates from different cell lines. Subtle differences were noted between models for the same TF derived from distinct tissue sources, presumably from differences in the composition of binding site targets in each tissue. Correct models (n=60) were distinguished from noise (n=25) based on R_i values of binding site substrates that are linearly related to binding energy using the F statistic. Further, we have so far validated models of 51 factors using two or more experimentally confirmed binding sites for each. For 133 factors, we determined normalized Euclidean distances between our derived information weight matrices and motifs reconstructed from oligonucleotide binding studies [Cell 158 (2014):1431-3]. The distances are <1 bit/nt for 75 factors, suggesting that our models and their motifs are nearly identical. For 18 factors, distances are 1-2 bits, indicating that the models differ at one or two positions. Distances were >2 bits/nt for 39 factors, where entropy minimization often revealed an *in cis* cofactor motif, rather than the experimentally derived motif. For example, entropy-derived motifs of CEBPB (n=10) concordant with experimentally derived binding sites differ by < 0.5 bit/nt, while cofactors of CEBPB (e.g. RUNX, IRF) and noise (poly-A rich) motifs exhibit much larger normalized distances (2.7-2.9 bits).

Conclusions:

Entropy minimization will find the most common motifs in a ChIP-Seq dataset, and is dependent on motif length and degree of conservation at that length. Our approach distinguishes correct binding site models from noisy, low complexity sequences that are coincidentally enriched. 132 cofactors were discovered for 97 of the 168 primary TFs. The unique distribution of some of these cofactor motifs suggests a potential association with tissue specific regulation by the primary TFs. This approach models not only strong binding sites (resembling consensus sequences), but also encompasses intermediate and weak sites. This should improve the accurate detection of mutations in known TF binding sites, to predict severity of these mutations, and to predict potential novel gene targets regulated by these TFs.