Significance score calculation

We describe how one can calculate a significance score for the special case of PRIORITY where the class parameter $\boldsymbol{c}$ is set in advance. In this situation, it is assumed that the structural class of the TF in question is known (see Section 3.2 in the paper).

We notice that under such an assumption, PRIORITY, like most other motif finding programs, reports the top scoring motif regardless of whether that set of sequences contains any motif at all. For instance, a motif finding program will find some motif even when the input sequences are generated from a random background model. Hence, if we are running PRIORITY on a set of n intergenic sequences, we need a way to assess the significance of the discovered motif in terms of n (and possibly the lengths of the sequences).

In order to find a significance measure, we run the algorithm five times (like we do on the real dataset), but on 50 randomly selected sequence sets of size n. These sequences are sampled randomly from the whole set of real intergenic sequences used by Harbison *et al*. as probes. We select the best score from the five runs for each of the 50 sets (just like we do when finding the best motif in an actual sequence set). We then fit a normal over the 50 points. Using the normal parameters, we calculate the significance of the learned motif from the actual set. Figure 1 shows a histogram of the scores obtained from 50 random sequence sets of cardinality 17, and the fitted normal curve.

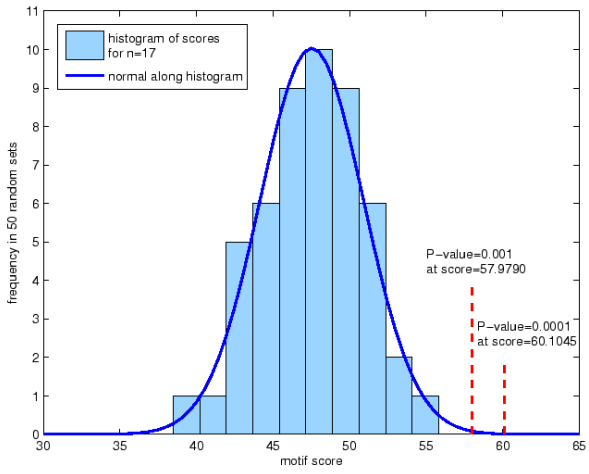

Figure 1: Histogram of the best scores among five runs for 50 collections of different random sequence sets of size 17. This is for the special case of PRIORITY with the inclusion of only the prior for basic leucine zipper TFs.

Harbison *et al*. have their own metrics to calculate the significance of a motif. Their metrics (AUROC and enrichment) are based on the number of bound and unbound sequences with a "match" to the motif. They also have a p-value calculation step similar to ours, taking random sets of sequences and generating a normal distribution over each of the scoring metrics. They use only those motifs in their final clustering algorithm which have a p-value less than 0.001 according to both criteria.

We believe that our method of scoring the motif using the joint posterior distribution (see equation 7) of the whole set of sequences is adequate to judge the significance of a motif. For example, the bZip protein Sko1 has 17 sequences in its probeset. PRIORITY with the single class prior finds a motif TACGTCAT very similar to the one with all three priors described in the main text of the paper. All other programs using no conservation information fail to find this motif (see Table 1 in main text). Only one program, CONVERGE—which uses conservation information—was reported to find it (see supplementary table). But this motif does not appear in Harbison *et al*.'s list of final post-processed motifs, possibly because of their criteria of significance. On the other hand, the motif found by PRIORITY has a score of 75.0849. As is evident from the dotted red lines in figure 1, this motif has a p-value of less than 0.0001 (lower than the p-value cut off of 0.001 used by Harbison *et al*.) according to our scoring system.

References

Harbison,C., *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome, *Nature*, 431:99–104.