Transcriptional regulation and control in eukaryotes and especially in multicellular organisms is one of the most intensively studied areas in the post genomic era. The rising importance lies in the direct impact on medical issues as cancer research, organ- and limb regeneration, ageing research as well as on developmental biology and molecular cell biology.

From experiments a number of transcription factor binding sites are known. These are often used in form of positional weight matrices. Diverse computational methods have been engineered to locate putative sites. Some of the sites however are non-functional and the interplay of the numerous transcription factors in fine tuning the cellular activities remains partly opaque.

Another obstacle seems to be, that the meaning of a factor can be context dependent. Though several extensions can be used to reduce noise: phylogenetic footprinting [3], information on gene expression data, co-regulation, tissue specificity and known composite modules [2]. To allow for complexity, the emphasis in searches is shifted to paired occurrence [1] and higher combinations of factor binding sites. [2]

In this context we have conducted an analysis using site predictions generated by EEL for promoters extracted from ENSEMBL database. We then combined the result with statistical approaches, mainly clustering and principal component analysis. The main aim is to explore the information lying solely within the predicted sites. The degree to which each feature vector contributes can be dissected, applying miscellaneous filters and other corrections to obtain optimal noise reduction. To this end we tested conversion of absolute site occurrences to binary representation, filtering for human matrices, homology filtering, data-normalization and several clustering methods on all of the manipulations as well as combinations of the approaches. Focus is put on the biological relevance of the derived models. This is ensured by using testsets recruited from literature.

References

