

Design for Efficient Motif Finding Algorithm

Amit Sinha

sinhaam@ececs.uc.edu
Department of ECECS
University of Cincinnati
Cincinnati, OH, 45221

Raj Bhatnagar

raj.bhatnagar@uc.edu
Department of ECECS
University of Cincinnati
Cincinnati, OH, 45221

Discovery of short repeated patterns (motif) is key to many problems in bioinformatics. The promoter region of genes are a common target for search for motifs (Transcription Factor binding sites). Since the transcription of different genes may be initiated by same transcription factors so the promoter sequences have common binding sites.

The search for motifs is difficult since motifs usually have a few mutations. So most approaches using exhaustive search consume too much time and resources. We propose a new algorithm which starts by collecting a set of candidate solutions and prunes away unnecessary patterns. Further, the candidate solutions are structured in a lattice which further leads to solutions much more quickly and uses less resources.

Another problem while searching for motifs is presence of a large number of false positives. Motifs are very short in length so the probability of a motif occurring merely by chance is very high. Since its not always possible to experimentally verify a binding site, we propose a algorithmic solution for rejecting noise. The transcription factors often bind one after the other in a serial order. So the binding sites are expected to occur in the same order in two or more promoter sequences. After identifying the common binding sites in a set of promoter sequences, only those patterns are likely to represent binding sites which occur in a similar order in the sequences.

Using a combination of faster search and rejection of false positives, our algorithm is able to find motifs efficiently with high accuracy.