1 Introduction

There are many algorithms that implement local motif search techniques but this paper propose a new approach to find globally optimal motif with relatively simple algorithm which search the motif space by branching from sample strings. They are called as PatternBranching and ProfileBranching algorithms.

2 Pattern Branching Algorithm

Discussing pattern branching algorithm ,let M be the motif and S be the sample containing n sequences with set S= S1,S2,S3. . . ..Sn and A0 be the occurrence of M with k mutations. We start with comparing A0 with each l-mer of each element of set S.We find min hamming distance d of A0 with S1 then S2 and so on. A score is then generated which is summation of all min distances calculated in previous steps for set S. A pattern set P is then generated which is called best nearest neighbour of A0.Suppose we have ATG as A0 and for k=1 the best nearest neighbours will be TTG,CTG,GTG,ACG,AGG,..etc which is total lCk.3k. Now we repeat the process to find score of each neighbour and choose the one having the minimum score as our next candidate A. This goes on till we find the actual globally optimal motif. Time Complexity of the algorithm is: O(n*l*N) where n is the no. of sequences, l is the length of kmers and N is length of each sequence Si.

3 Profile Branching Algorithm

In Pattern Branching algorithm, we make a profile matrix using the probabilities of A,T,G,C as (1/2,1/6,1/6,1/6) respectively.We now find entropy of A0(referred in pattern branching) and for each nearest neighbour later with k mutations.The mutated motif with maximum entropy value is chosen as A0.The process goes on and we find global optimal motif. Time complexity of this algorithm is : O(n*n * N*N) where n,N is same as used in pattern branching algorithm above.

4 Results

Now the results of this new approach used in paper is compared with other existing algorithms for implanted motif search and it is also tested on biological data whose motif is already known so that we can compare the efficiency of results for eg., pattern branching algorithm proposed here gives TGTGAAATAGATCACATTTT as output motif for E. coli CRP data where reference motif was TGTGANNNNGNTCACA .