Recognition of special pattern in the upstream DNA of Streptomyces

Streptomyces are soilbacteria and are responsible for the production of more than half of all known antibiotics. Up to now, two genomes of their species are sequenced and annotated, S. avermitilis and S. coelicolor. On the other hand, many aspects of the regulation of genes are still unknown. Finding new pattern with function of regulatory elements or promoters would contribute to a better understanding of genexpression.

The problems of finding new motifs in the upstream region of known genes are:

- unknown pattern
- not necessary preservation in front of different genes
- location of transcription start point can differ strongly

Two approaches:

1.) Find new pattern by looking for significant overrepresented words in all upstream regions.
2.) Find new pattern by comparing the upstream regions of “homological” genes.

At approach two, it turned out that a high conservation in gene often corresponds to high conservation in the upstream region. Better results can be achieved by taking genomes of mycobacterium into this comparison. But even in this case the conservation of the upstreams within Streptomyces and within Mycobacterium may trouble.

I’ve concentrated my work on the first approach.

Leaning on the essay of Bentley (et al.) “Bioinformatic identification of novel regulatory DNA sequence motifs in Streptomyces coelicolor” [1] and an essay of Saito (et al) “High-multiplicity of Chitinase genes in Streptomyces coelicolor” [2] we decided to look for special wordpairs and the space between them. In addition to [1] we permit greater distances between two words and gaps. Because of the gaps the method had to be changed decisively. Furthermore the search was restricted to those pairs, which are repetitions of the same or inverted word.
The following steps describe the method:

1.) Identify statistic overrepresented wordpairs in the upstream regions.
2.) Arrange suitable pattern via alignment of two position weight matrices.
3.) Transform the pattern in a Hidden Markov Model and find represents of the motifs with the Viterbi-algorithm.
4.) Confirm the motif via comparison of the respective genes of avermitilis and coelicolor.
    Compare the number of hits in the upstream regions to the hits in genes.

Some interesting wordpairs could be found. For example the pattern ‘TCTACagt-ctGTAGA’, ‘GTTTCaCGtGAAAC’ and ‘TTAGgTtAGgCTaACCTAA’ should be mentioned. The last one is a pattern, which had also been found by [1]. Very interesting about the second one is the appearance only in the middle of genome in S. avermitilis as well as in S. coelicolor.

Looking for similarities between motif pairs (with MEME or the Gibbs-Sampler) did not lead to good results. Moreover the wordblocks seam to be extended.
Maybe this can be attributed to preferring motifs with small distance and few gaps in the fourth step.