**CSCI 4314/5314 Project Proposal - Algorithms for Molecular Biology**

**Members:**

Sushma Colanukudhuru

Han Ngo

Christine Samson

Monte Anderson

Rasmi Lamichhane

Byron Bearden

**Introduction:**

Our application aims to implement Hidden Markov models for motif discovery. Hidden Markov models (HMMs) are probabilistic models that are well adapted to many tasks in bioinformatics, specially for predicting the occurrence of specific motifs in biological sequences. Hidden markov models have several advantages over pattern matching algorithms. They are statistical models that allow to represent a motif based on the probability to find a given nucleotide after another one.

**Advantages of HMM over other methods:**

* Using HMM for motif discovery is highly useful when searching for complex motifs.
* It makes it easier to find motif occurrences in divergent sequences.

**Objectives:**

Our application aims to efficiently explore all possible locations of Transcriptional Factor Binding site(TBF s) motifs in a set of sequences with high accuracy. It constructs an ensemble of k Hidden Markov Models(HMM) through local alignments of two sequences in the set and then progressively aligns HMMs in the ensemble to others in the set. The alignments with significant scores are in turn used to update the parameters of the k HMMs.

**Algorithm :**

It starts by locally aligning two arbitrary sequences in a sequence set using Smith-Waterman algorithms. A trace back is made to identify subsequences corresponding to the k largest alignment scores. ‘k’ is appropriately determined based on what works best for the given dataset. An initial ensemble of k profile HMMs will then be constructed from the k alignments. The profile HMM will contain two states Di and Mi for column i in the corresponding alignment. The deletion state Di does not emit any nucleotide and is used to describe the possible gaps in column i.Once the ensemble of HMM s is created,each of the remaining sequences in the database is aligned with the HMMs in the ensemble. The sequences with the best possible alignment scores are chosen to update the parameters of the HMM.

**Application to be benchmarked:**

**BioProspector**:

It uses Gibbs sampling strategy to examine the upstream region of genes in the same gene expression pattern group and looks for regulatory sequence motifs. BioProspector uses Markov background to model the base dependencies of non-motif bases. The parameters of the Markov background model is either estimated either from user-specified sequences or pre-computed from the whole genome sequences.

**Input and Output format used by BioProspector:**

BioProspector accepts FASTA format as input. The program searches for motifs from sequences a number of times and reports the number of top-scoring motifs specified by the user. A sample output by BioProspector looks as follows:

**Motif width (blk 1, blk 2); Gap [min gap, max gap]; Number of motif sites.**

Motif #1:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Width (15, 0); Gap [0, 0]; MotifScore 1.826; Sites 16

The Motif score is calculated as follows:

Motif score =( log(number of sites) \* (relative entropy of the motif) / motif width);

It also returns the starting positions of the motifs from both the forward and the reverse complements

>1 len 105 site #1 r 79  
TGTGAAAACGATCAA  
>1 len 105 site #2 f 20  
TGTGGCATCGGGCGA  
>2 len 105 site #1 r 73  
TGTGACGCCGTGCAA  
**Methods to generate dataset:**

Dataset of cyclic-AMP receptor proteins(CRP) is used. It consists of sequences each of length 150 bps. In addition, a simulated dataset will be generated, by replacing the nucleotide values at random positions in the original dataset and by inserting the motifs to be searched for at various positions in randomly generated sequences.

**Test cases to Test Algorithm Limits:**

Bioprospector when run on a dataset of cyclic-AMP proteins(CRP), identifies 18 of the 23 binding site motifs ,each of width 22.The starting positions of the TBFs returned by BioProspector deviate from the output generated by DNA foot printing methods. Our application aims to benchmark this parameter, i.e return the starting positions of the TBF binding sites motifs with better accuracy than BioProspector. Further BioProspector identifies only a single instance of the binding site in a particular sequence. Our application aims to identify multiple instances of the TBF binding site motifs in every sequence.

**Advantages of the Proposed application:**

* Multiple binding sites can be identified
* Sample ability can be increased, as the proposed method employs multiple HMMs

**Responsibilities and Milestones**

|  |  |  |  |
| --- | --- | --- | --- |
| *Week* | *Responsibility* | *Members* | *Milestone* |
| *4/9 - 4/15* | Generate the dataset.  Run BioProspector on the dataset. | Dataset - Monte, Christine  Run BioProspector on the dataset - Hannie, Sushma, Rasmi  Additional Research - Byron, Rasmi | Dataset generation is completed.  Coding of algorithm is done as much as possible. Performance of BioProspector on the generated dataset is analyzed. |
| *4/16 - 4/22*  *Note: Benchmark is due 4/19* | Start coding and run test cases. Perform comparative analysis with the baseline. | Coding and Implementation - Hannie, Sushma  Comparative analysis- Rasmi, Byron  Monte and Christine will help wherever the need be. | Coding is completed early in the week.  Comparative analysis of the performance of our application and BioProspector is done. |
| *4/23 - 4/29* | Test extensively on original and simulated datasets. Confirm if the benchmarking criteria are achieved. Implement visualisation if possible.  Start making the presentation | Benchmark testing-Sushma, Hannie  Visualisation-  Monte and christine  Presentation- All | Algorithm is completely coded and results are completely generated.  Visualization is implemented as much as possible.  Presentation is created and mock presentation is done. |
| *4/30 - 5/6*  *Note: Presentation is on 5/1 and 5/3* | Refine Algorithm if needed and implement visualisation | Algorithm Refinement- Anyone  Visualization - Monte, Christine  Presentation - All | Visualization is completed.  Presentation goes successfully, and we can deliver! |