Plan for Sharing Research Data and Information

This research will integrate research findings with educational and extension programs and activities at the Illinois Institute of Technology, other Institutions and Research Centers at large. This is one of the key goals in this proposal in support of NSF's goals to promote the discovery, integration, dissemination, and employment of new knowledge in service to society and achieve excellence in U.S. science, mathematics, engineering, and technology education at all levels. PIs of this proposal are actively engaged in various teaching and educational programs and are dedicated to providing diverse learning opportunities to students and general public with different educational backgrounds. We plan to integrate our research with different types of educational and extension programs. These activities will include seminar lectures in the Electrical and Computer Engineering, Biology, Computer Science, Math, and Bio-Medical Engineering departments; implement new research findings as teaching materials into the current core curriculum encourage participation of minority students in ECE, BME and Biology majors thru Research Projects; and develop joint educational program for high school students in the Chicago area. Furthermore, we plan to collaborate with other centers of Bioinformatics, Bioengineering and Research Institutes to foster education by applying engineering principles to cell biology, integrated with applied mathematics, computational science, bioengineering and medical sciences. Dissemination of information will also be thru participation in conferences, workshops and publication of Journal papers.

We are in the early stages of creating in our website an area that will include programs, databases (or database links), papers and relevant information to this research. This initiative is in its earlier stage but will be an important component of our work.

Currently you can find in this site:

1.- Publications relevant to this research

2.- Publication of the group

3.- MATLAB toolbox

Using MATLAB, we developed a toolbox to extract and manipulate the data required and put it in a format suitable to gene prediction and analysis. This toolbox is publically available through our research lab website. Using this toolbox, we extracted the following data from the NCBI for each tested genome: 1) the complete DNA sequence, 2) the exact locations of all known genes in the forward and reverse strands, 3) gene predictions obtained by GeneMark, 4) gene predictions obtained by Glimmer, and 5) the set of all possible open reading frames based on a pre-specified criteria. Based on this analysis, our program is able to classify the tested data into four different groups:

- Actual Translated Sequences (4,149 sequences): Open reading frames which GenBank indicates as sequences that translate into proteins,
- GeneMark Hypothetically Translated Sequences (695 sequences): Open reading frames which GeneMark indicates as genes but are actually not (GeneMark false positives),
- Glimmer Hypothetically Translated Sequences (2746 sequences): Open reading frames which Glimmer indicates as genes but are actually not (Glimmer false positives),

- Non-Translated Sequences (23,384 sequences): Open reading frames which do not appear on the list of Actually translated or hypothetically translated sequences. For this work, the open reading frame had to have: 1) A valid initiation codon; 2) A valid termination codon; 3) A sequence length greater than or equal to ninety-nine bases.

4.- MATLAB programs for TFBS and Motif discovery

The genome sequence datasets required for the analysis in our research were obtained from the National Center for Biotechnology Information (NCBI) and other open databases, such as the Tompa database. Using these available datasets, we extract and manipulate the data required and put it in MATLAB files, where the data is in format suitable to our analysis. These MATLAB data files are publically available through our research lab website; they include the following data from the NCBI and other databases for each tested genome: 1) the complete DNA sequence, 2) the exact locations of all known transcription factor binding sites in the forward and reverse strands, 3) performance evaluation algorithms based on pre-specified criteria. Based on this analysis, the tested data can be classified into four different groups:

- Synthetic data. The dominant nucleotide at each position in the motif appeared with the probability of 70%, while the remaining nucleotides appeared with a probability of 10%. Nucleotides in the Non-motif part are assigned a probability of 25% for each nucleotide.
- The cyclic-AMP receptor protein (CRP) binding sites. This dataset consists of 18 sequences from Escherichia coli. Each of the sequences is 105 nt long, and the dataset contains 23 motifs that had been experimentally determined.
- Drosophilo melanogaster Dscam introns in exon 6 cluster. the Dscam introns dataset consists of the 47 introns, which locate in the direct upstream of the 2nd up to the 48th alternative exon in the exon 6 cluster of the Dscam gene, and the motif is the reverse complement of the anchor sequence.
- Tompa database, The 52 datasets in Tompa database contain transcription factor binding sites drawn from mouse, human, fly and yeast. The binding sites are planted in three different types of background sequences: their original promoter sequence, the promoter sequence of a randomly chosen gene from the same genome and sequence generated using Markov chain of order 3. Four datasets without motifs were also added as negative controls.

The proposed discovery and detection algorithms consist of several m files including ‘IGSMC_init_dis.m’, ‘IGSMC_preinit.m’, ‘IGSMC_step_dis.m’, ‘empirical_distrib.m’, ‘nt2num.m’. To run the algorithms and see the performance on each dataset, open the m file with ‘Demo’ as prefix of the file name in the folder of dataset, choose the dataset to test and run the program.

The MATLAB programs for TFBS and Motif discovery are encrypted. To obtain the password for these source codes, please send a request and your contact information to email address: lhuang13@iit.edu .