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 NIH's goals to develop, maintain, and renew scientific human and physical resources that will assure the Nation's capability to prevent disease. 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, 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.- The first public domain program described below:

The complete prokaryotic genome sequences required for the analysis in our research were obtained from the National Center for Biotechnology Information (NCBI) [11]. Using MATLAB, we developed a toolbox to extract and manipulate the data required and put it in a format suitable to our 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, we program is able to classify the tested data into four different groups (See Figure 1):

- 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.

The sequence numbers given above for the four test groups are for the MG1655 E. coli genome. The following block diagrams give an illustrative description of the approach used to develop the proposed block code model.

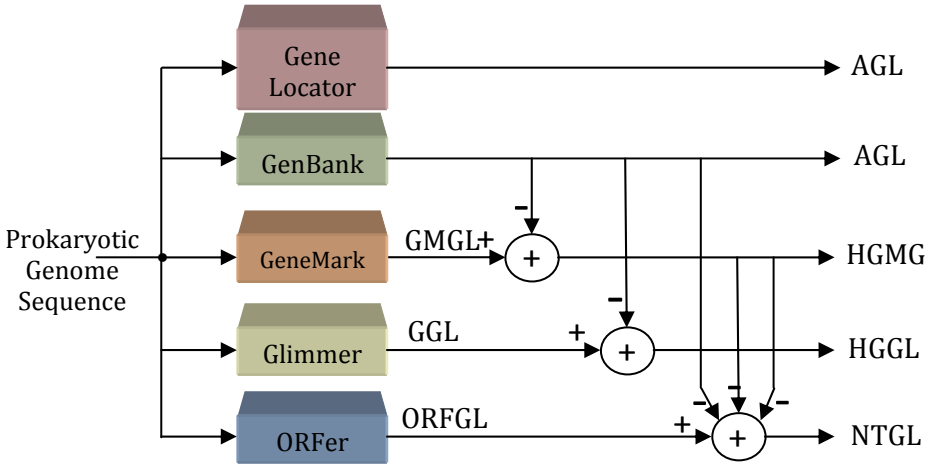

Figure 1. Schematic diagram of the output obtained by the developed programs

The notations used is Figure 1 are: AGL (Actual Gene Locations) corresponds to group 1, HGMGL (GeneMark Gene Locations) corresponds to group 2, HGGL (Hypothetical Glimmer Gene Locations) corresponds to group 3, and NTGL (Non-Translated Gene Locations) corresponds to group 4. The intermediate parameter GMGL (GeneMark Gene Locations) corresponds to the gene predictions obtained by GeneMark, GGL (Glimmer Gene Locations) corresponds to the gene predictions obtained by Glimmer, and ORFGL (Open Reading Frame Gene Locations) corresponds to the set of all possible genes that are greater than 99 bases long and start with a valid start codon and end up with a valid stop codon.

A more detailed explanation of our program and applications can be found in: www.ece.iit.edu/~biitcomm/public