CALIFORNIA STATE SCIENCE FAIR
2006 PROJECT SUMMARY

Name(s) **Hann-Shuin Yew**	**Project Number** **S0428**

Project Title

Investigation of Homopolymeric Runs in C. elegans Genome with Novel Model for Control Sequences

Abstract

Objectives/Goals

To create, test and optimize a new algorithm, the Guided Localized Model (GLoM), for simulating DNA that is more precise than currently available methods. It was hypothesized that GLoM would provide a closer approximation to the homopolymeric run frequencies in the C. elegans genome than current controls. Secondly, using a larger window size would create a more precise DNA simulation, with reduced overemphasis of repeated regions.

Methods/Materials

GLoM was developed in Java, a platform-independent programming language from Sun Microsystems. Processing and generating 97MB of data on a 1.3GHz processor with 512MB RAM and a 1000nt window size required 68 minutes.

The C. elegans chromosome sequence was downloaded from WormBase. Markov models were created from the Regulatory Sequence Analysis Tools (RSAT) of the Université Libre de Bruxelles, and skewed-random sequences - random controls with built-in nucleotide biases - were generated with Java. Five runs of each sequence were analyzed to minimize statistical fluctuations.

Homopolymeric run frequencies were then calculated and plotted with Logger Pro to provide a graphical comparison of the different DNA sequences.

Results

Of all the controls, GLoM provides the best fit to the original genome. However, it tends to overemphasize the nucleotide and codon biases in the original DNA.

Unexpectedly, changing the window size does not appreciably improve the precision of the algorithm, especially for the A/T runs. However, in the case of the C/G runs, expanding the window size creates a slightly closer fit to the original genome. Shrinking the window size also produces overly long (>40nt) runs.

Conclusions/Discussion

GLoM, the Guided Localized Model, is a fast, elegant method of generating a control sequence of any length. Compared with a standard Markov model, it gives a superior reflection of the homopolymeric run frequencies in C. elegans. Through this improved fit, GLoM demonstrates that the non-random character of the homopolymeric runs is probably due to localized and holistic mechanisms. In short, GLoM has a wide range of potential applications in the fields of computational and molecular biology.

Summary Statement

In this project, an algorithm that simulates DNA sequences more comparable to real DNA than those from currently used methods was created and tested, extending our understanding of the C. elegans nematode genome in the process.

Help Received

Many thanks to Dr. Andrew Fire of the Departments of Pathology and Genetics at the Stanford University Medical School for his invaluable mentoring.