



**CALIFORNIA STATE SCIENCE FAIR  
2012 PROJECT SUMMARY**

<b>Name(s)</b> <b>Larry Zhang</b>	<b>Project Number</b> <b>S0536</b>
<b>Project Title</b> <b>HMM Model Usage to Determine Gene-Specific CpG Ratios and Location of Genes</b>	
<p style="text-align: center;"><b>Abstract</b></p> <p><b>Objectives/Goals</b></p> <ol style="list-style-type: none"><li>1. To locate introns and exons within a sequence of DNA using dinucleotide frequencies that differ between the two</li><li>2. Calculate gene-specific CpG dinucleotide ratios between introns and exons and compare between related genes</li></ol> <p><b>Methods/Materials</b></p> <p>Lenovo Computer Sample Hidden Markov program DNA data from various databases (GenBank, fruitfly.org, etc.)</p> <p><b>Results</b></p> <p>The HMM was not able to successfully locate introns and exons on the DNA, though it did calculate specific CpG dinucleotide frequencies for each gene, which allowed me to calculate a CpG ratio. However, these CpG ratios ranged from 3-7, with no obvious or direct correlation between related genes such as BRCA1/BRCA2 within humans/mice, which both cause breast cancer, or OCA2 between humans, flies, and mice, which codes for eye color.</p> <p><b>Conclusions/Discussion</b></p> <p>The HMM program did not successfully locate introns and exons, probably due to the small size of training data to tell it the frequencies of dinucleotides between introns and exons. A successful training set would probably have around 200 genes, with introns and exons labeled. Additionally, this experiment used various genes from various locations in the genome of humans, mice, and flies, which brings in the complication that frequencies of dinucleotides simply are not constant between different regions of the genome. Additionally, the CpG ratios found had no real correlation between related genes, showing that dinucleotide frequencies (or more specifically, the CpG dinucleotide) have no relationship to the gene's function.</p>	
<b>Summary Statement</b> This project uses dinucleotide frequencies to calculate locations of introns vs exons and calculated unique frequencies for various species.	
<b>Help Received</b> Obtained sample HMM program from Dr. Li of Avid Academy	