Predicting the Function of Single Nucleotide Polymorphisms

Corey Harada

CS194 Honors Undergraduate Research Program

Introduction

A Single Nucleotide Polymorphism (SNP) refers to a point mutation in a DNA strand. This is the transposing of a single nucleotide into another base, for instance, a cytosine becoming thymine. They are a major source of genetic diversity. One major area of research involving SNPs is their role in disease.

SNPs can be found by sequencing the genome, and applying a few methods to detect them. However, learning the function of these SNPs is much more difficult. One difficulty in determining their function is that SNPs are often highly associated with each other. That is, a person who possesses a mutation at SNP A would have a higher chance of having a mutation at an associated SNP B. It is difficult, in this situation, to determine by experimental observation the effect of each SNP, because a given trait will be associated with multiple SNPs, and no easy way to determine which SNP is causal.

The goal of this project is to create a method to determine which SNPs are more likely to be responsible for phenotypic changes.

Approach

Our approach focuses entirely on SNPs located in the exons of coding regions of genes. While SNPs in other locations exist and can cause phenotypic changes, a SNP in the coding region, changing the resulting protein coded by that gene, has the most obvious potential function.

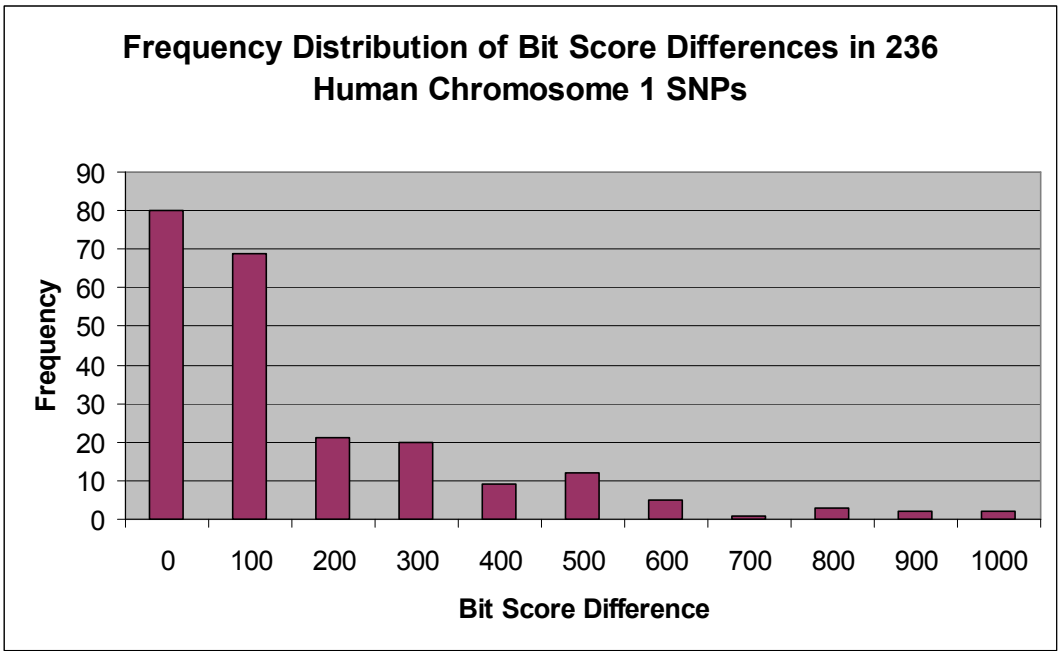
Our program will take an SNP, and locate its position in a gene. It will then translate the gene into two proteins, one with the mutation at the SNP, and one without. These proteins will be fed into the Basic Local Alignment and Search Tool, or BLAST, a program by the National Center for Biotechnology Information. BLAST will search it's protein databases for similar to the input. The idea is that proteins in humans are similar to those of proteins in other organisms. If a mutation causes the protein to become less well aligned with those of other organisms, then it is possible that the region the SNP is located in is highly important to the function of the protein, because changes there would be selected against. Thus, the region may be associated with some disease.

Detailed description of methodology / implementation

Our program uses Pygr, an open source bioinformatics extension for the Python language. We use it to retrieve the sequence of the human genome, based on position and chromosome. We obtain a list of SNPs, SNP positions, and alleles from the International Hapmap Project, a project that attempts to catalogue the genetic similarities and differences in humans. Synonymous SNPs, SNPs that cause no difference in the coding of amino acids and would thus result in the same protein, are filtered out. We then search for the gene containing each SNP, and then translate the gene into a protein, both with a mutation at the SNP and without.

Once we have our sequences, we call BLAST to analyze them, and parse the BLAST output. We calculate the difference in bit scores produced by BLAST. The bit score is essentially a measure of how well the sequence aligns with sequences in the BLAST databases. We sum the difference in bit scores over all aligned proteins between the two input proteins, and rank each SNP accordingly.

Experimental Results



Of these SNPs, 80 were equal to zero, i.e. there was no difference in bit score. This seems like it would be unlikely, as all of our inputs contained two different amino acid strands, and the chances that they would produce identical alignment scores seem like they would be lower. There are a few possible explanations for this. One is that the BLAST protein databases contain information on the diseased proteins for other species. In this case, the alignment scores would be exactly the same. Another is that the SNPs with a zero alignment score different code for non diseased phenotypes, which would not be selected against.

Conclusion

This algorithm has a few flaws. The SNPs that code for a phenotype that is not selected against, for instance, may produce a low bit score difference, despite in fact being a causal SNP for a specific trait. Additionally, the correctness and results of the algorithm depend entirely on the BLAST database. A good improvement on the algorithm might be to conduct a BLAST search on a specific set of databases, with proteins chosen such that it would produce the best results for our program.

Bibliography

Blast: Basic Local Alignment and Search Tool. June 6, 2008. National Center for Biotechnology Information. <http://blast.ncbi.nlm.nih.gov/>.

Hapmap Homepage. June 6, 2008. International Hapmap Project. <http://www.hapmap.org>.

UCLA Bioinformatics: Welcome to the Pygr Project! June 6, 2008. UCLA. <http://bioinfo.mbi.ucla.edu/pygr/>.
Acknowledgements

Professor Eleazar Eskin, for his help as my faculty advisor.
Professor Amit Sahai, for being in charge of the Honors Undergraduate Research Program.