Data Mining Project Proposal
Laura Mustavich
March 23, 2007

Problem Background

One of the largest data mining problems in the field of genetics is the identification of genes contributing to complex diseases. We are now finding that our traditional linkage and association analyses methods, which were highly successful in disentangling the genetic basis of simple Mendelian diseases, lack the power necessary to discover genes responsible for complex diseases. This is because so many genes often interact to cause a disease, that the affect of any one gene individually is so small, it is missed by traditional methods. It has now become apparent that if we are to discover genes which contribute to complex diseases, we must develop methods that incorporate the information given by epistasis, the interaction among genes. Several methods to do this have recently been developed. One such method is Multifactor Dimensionality Reduction (see sources below).

Actual Project Idea

Below is a list of my project ideas, in order of preference (#1 being most preferred):

1. Apply Multifactor Dimensionality Reduction to real world data. I would use one of the datasets below (whichever I can obtain first), and discuss my results.
2. Compare the results of Multifactor Dimensionality Reduction and a couple other epistasis methods, discussed in one of the papers below, on one of the datasets below.
3. Replicate the results of one of the papers below.

Background on Datasets

Dataset A: Alcoholism

Many loci have been associated with alcoholism, including markers in the ALDH and ADH genes, which function in alcohol metabolism. Taste receptor genes, such as those that confer the ability to taste the compound PTC, have been found to be associated with one's willingness to drink alcohol. People with certain genotypes can taste PTC, while people with other genotypes cannot, and for those that can taste PTC, it tastes bitter. In past studies, it has been shown that for people who can taste PTC, alcohol also tastes bitter to them, and so they drink less alcohol. Alcohol tastes sweeter to people who cannot taste PTC, and therefore, drink more of it. There is thus an association between a person's genotype at bitter taste receptor genes, and their willingness to drink alcohol. While one must be willing to drink large quantities of alcohol in order to become an alcoholic, no direct association between taste receptor genes and alcoholism has yet been established. Consequently, we hope to establish this association by using methods to detect possible epistasis among the taste receptor genes and other genes associated with alcoholism.

Dataset B: Taste Receptors and Cardiac Health

It is thought that taste receptor genes might be indirectly associated with cardiac health by influencing the foods that a person prefers, which in turn affects cardiac health. We hope to uncover the contributions of these taste receptor genes to cardiac health by using

methods to detect possible epistasis among these genes and food preference, which probably would be missed by other methods.

Multifactor Dimensionality Reduction Software

http://sourceforge.net/projects/mdr

Papers on Multifactor Dimensionality Reduction

The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases
http://content.karger.com/ProdukteDB/produkte.asp?Aktion=ShowPDF&ProduktNr=224250&Ausgabe=229632&ArtikelNr=73735

Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer
http://www.journals.uchicago.edu/AJHG/journal/issues/v69n1/012797/012797.text.html

Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions
http://bioinformatics.oxfordjournals.org/cgi/reprint/19/3/376

General Sources

Multifactor Dimensionality Reduction
http://en.wikipedia.org/wiki/Multifactor_dimensionality_reduction

Epistasis
http://en.wikipedia.org/wiki/Epistasis