Cooperative analysis of human diseases through genome-wide expression patterns and hierarchical clustering

David H. Wei
MB&B 452a Final Project
Genomics and Bioinformatics
Due: December 13, 2002

Introduction

The world today is drastically different than it was but a few years ago. An exponential increase in technology and computer power has opened science to vast new realms of knowledge and information. One of these realms is the microcosm of the human cell. In particular, the completed sequencing of the human genome has generated terabytes of information that is yet to be analyzed or even understood. Each human cell contains approximately 3 billion base pairs, which encode between 40,000 and 70,000 genes. Today, over one million expressed sequence tag (EST) sequences exist in public databases that represent nearly all of these human genes. In parallel with the scientific benefits from improved technology, medicine also has benefited from this growth in knowledge as well as from the concomitant improvements in science. Novel medical techniques are constantly being derived while older cures are made more efficient. One avenue through which this has occurred is the direct relation of the sequenced human genome to known specific diseases. For instance, now a growing amount of diseases have already been identified and related to particular genes. Diseases involving all different parts of the human body have been found to be caused by genetic abnormalities. The following list offers only a sampling of the diseases known: hemophilia, sickle cell anemia, obesity, cystic fibrosis, deafness, glaucoma, atherosclerosis, male-pattern baldness, asthma, Alzheimer, Parkinson, and a multitude of cancers. Up to this point however, only cancers have been the primary focus of genomic studies. Through expression profiling as well as classification of different forms of cancers, scientists have improved diagnostic techniques and therapeutic regimens.

In light of these new technologies and advances in medicine, this paper presents the following dualistic proposition: first, to increase focus on the application of genomic techniques on those diseases other than cancer; and secondly, to combine the aforementioned genomic technique of expression profiling with hierarchical clustering to form a novel cooperative strategy of disease analysis. Although it has been previously proposed that genomic studies should also be conducted for the other diseases, such as those of infectious nature, this paper differs in its combined consideration all genetic diseases as a whole and the integrative use of top-down hierarchical cluster analysis.

Expression Profiling and cDNA microarrays

Expression profiling utilizes cDNA microarrays in order to offer a better understanding of the overwhelming wealth of information that the sequenced human genome represents. Moreover, it allows the simultaneous measurement of mRNA expression for a large number of genes. The technique begins with the mRNA pool from two different cells, one from the control and the other from an altered variable cell. These pools of total mRNA extract are fluorescently labeled using oligonucleotide dT-primed reverse transcription by utilizing nucleotides tagged with either Cy3 (green) or Cy5 (red). These probes are then hybridized to a glass microarray slide that was previously prepared with a grid of PCR-amplified and robotically spotted cDNA clones (Figure 1). This process results in a slide with spots of variable ratios of intensity between green and red fluorescence. Such visible ratios provide an immediate tool for contrasting gene expression among a large number of genes. Previous studies of various forms of cancer involved generating whole-genome expression profiles in order to provide a basis for classifying new cases of cancer. Thus, each type of cancer generated a unique whole-genome expression profile that could serve as a comparison for a new unknown case of cancer and thereby help in diagnosing and categorizing the cancer. This would consequently improve the assignment of the appropriate type of therapy (maximizing efficacy and minimizing toxicity and costs) necessary for the corresponding diagnosis.

In the case above, expression profiling was simply a new route to the more effective and efficient classification of cancer. Classification was then used as a tool for improved diagnosis. However, the proposed strategy of this paper does not use expression profiling as a means of categorization, but on the other hand, uses it as an indirect means of comparing the general effect of one gene’s mutation on the entirety of the cell and the change in expression of the other genes across time.
lengths reflect the degree of similarity between the objects as evaluated by a pairwise similarity function. Also note that clustering methods are usually divided into the two general classes of supervised and unsupervised. Because of the lack of a priori knowledge of the complete range of expected gene expression patterns for any disease, unsupervised clustering approaches are favored.

Integration of time course expression data and hierarchical clustering

We now turn to the heart of the novelty in this combined approach of disease analysis. A few assumptions and preconditions must first be made. First, the genetic diseases studied must have previously identified non-lethal mutation(s) correlated with a single gene. Secondly, it must be assumed that a mathematical pairwise similarity function exists which may successfully describe the relationship between different expression data matrix. Thirdly, it must be given that the whole-genome expression data for each disease (as well as for a healthy cell) has been previously accurately measured and attained over a certain specified time course.

From this last assumption, we are then given a large number of unclustered time course gene expression data matrices--or many diseases and one healthy control condition (Figure 4)—that we may freely manipulate. We now create new matrices by comparing each unclustered time course expression data matrix for each disease with the control expression data matrix derived from the healthy cell. This is most similar to the perturbation matrix as used by Trey Ideker, et al. in analyzing systematic perturbations in metabolic networks. We now have a bank (Figure 5) of what we may call difference matrices that describe the change in the whole-genome expression data.

Figure 2. Example of cDNA data microarray. The vertical axis represents different genes, thus each row is a gene. The horizontal axis represents time course.

Figure 3. Expression data before and after random permutation with rows, columns, and both.

Figure 4. Database of expression data tables with multiples tables such as this for the many diseases. Difference matrices (each mutated gene), we may then hierarchically cluster the genes in a bottom-up fashion as described above. This strategy of multiple levels of expression data matrices may best be illustrated through the analogy of mathematical derivatives. Given the first expression data matrix which describes the current state of the whole-genome across time, we are essentially taking the first derivative as we create the difference matrix through comparing each disease with the control. As we find the similarity between the two difference matrices, we then fundamentally are considering the second derivatives.

Application of Novel Cooperative Strategy

Once these diseases are successfully clustered according to the amount they perturb the whole-genome human expression, new relationships between different diseases may be found. In relating different diseases, the possibilities are endless in the advantages that may be uncovered. Previously unknown connections between the ways different diseases affect different genes may indicate newfound and unorthodox

Figure 5. Database of difference tables that compare pairwise functional differences among the disease and the control.
possibilities of treating various diseases. For instance, the change in expression data over time for male-pattern baldness may be similar to the change in expression data over time for glaucoma, Alzheimer, deafness, etc. Commonalities in the changes levels of certain genes may also help develop alternative methods of therapy which may cure more than one disease simultaneously, or prevent one while curing another. Drug development may be targeted toward one disease without perturbation or incitement of another related one. This cooperative strategy relies heavily on the large amounts of expression data necessary for comparing different diseases. It also calls for uniform experiments of time course and does not account for possible lethal mutations, or multiple mutations in more than one gene. However, with the rapid increase of computer technology and exponential accumulation of scientific data, these problems may be solved in the near future. Certainly, if anything, we may conclude that the conception of such a novel strategy proves the genuine power inherent in genomics and bioinformatics—a power that we should not take lightly, but seriously consider in our endeavors as scientists.

References

1 http://news.bbc.co.uk/1/hi/sci/tech/1426702.stm