David Patariu
BIOC218: Computational Molecular Biology
Final Project
March 15, 2002

This paper presents a new method of analysis and scoring of gene expression data.

The intent of this new method is to overcome the difficulties/limitations associated with gene expression analysis due to the large amounts of data contained in gene expression profiles, and the difficulties associated with data analysis of large datasets.

By leveraging current sequence analysis methods used to compare and assess similarity between proteins or regions of proteins, a novel way to score and compare gene expression data was created.

The new process utilizes a binary hash of the gene expression data based on a comparison of normal versus unknown profiles to generate a score. Scores are compared to scores of known vs. normal for matches and identification of the cell type.

Details

Baseline profiles by cell type are generated from known gene expression profiles. The profile is then compared to the profile of a normal cell, and is scored. A score of one is given if the cell type profile exhibits increased expression of a particular gene relative to the normal cell. If expression is equal to or less than the expression of that of the normal cell, a score of zero is assigned. The result of this analysis will be a binary string. This binary string is then converted into a base ten number, which will represent the results of the comparison.

A profile for the unknown cell is then generated from microarray experimental data.

Expression levels are scored again against the normal profile with either a one or a zero. The unknown profile's binary number that was created is then converted into a base ten number, which would represent the unknown expression profile.

If the number that the question profile produced matches the number that the recognized profile produces, and then a match is achieved.

Methods

Microarray data from a recent publication (Garber et al, 2001) was used to test the new scoring algorithm. Seven genes were chosen after being noted as

having increased expression in both Small Cell Lung Carcinomas (SCLC) as well as Large Cell Lung Carcinomas (LCLC). Baseline profiles were calculated from the normal lung, SCLC and LCLC sample data and a score was generated for each. Then, the scoring system was tested, with both a known cell type, and then with data from and Adeno Carcinoma. A table with summary results is presented in this paper, and a Microsoft Excel worksheet is attached with supporting calculations.

Table 1

Sample Name	Gene Name								Sc
	Filler	HMGIY	FOSL1	PLAT	INSM1	SGNE1	QPCT	MYCL1	
SCLS Base Profile	1	0	0	0	1	1	1	1	14
LCLC Base Profile	1	1	1	1	0	1	1	0	24
SHU079_SCLC	1	0	0	0	1	1	1	1	14
SHAH026_Adneo	1	0	0	1	0	0	1	0	14
*Please see attached spreadsheet for supporting details									

Answers to Questions from Professor Brutlag:

1)Think critically about how the current method used to analyze data works?

Currently a large portion of microarray data is analyzed via a program called Cluster, which uses average link hierarchical clustering, with results displayed using Tree-view, a program that generates a dendrogram of the relationships. In general clustering works by using a dissimilarity measure between two groups.

When asking the questions is gene expression profile X equal to gene expression profile Y, using Cluster with Treeview would produces a dendrogram that may be difficult to interpret based on the complexity of the dendrogram. And positioning the unknown gene expression profile on the dendrogram may not does not answer the binary question of whether a gene expression profile matches or does not match a known profile.

Scoring techniques work to identify homologous regions in proteins. A modified scoring system would work equally well to identify cells that have the same gene expression profile.

- The proposed algorithm would not be as effective with protein homology mapping, since sequences can be of variable lengths and still have homologous regions. The nature of microarrays allows us to have

standard number of expressible genes, and this standard domain for profiles enables a comparison that yields a number for a particular profile.

2)What are its assumptions? Compare this to the assumptions of the new method.

The assumption of the Cluster/Treeview approach is that position in a visual dendrogram is adequate to answer the question of similarity of profiles. The new method assumes that the core library/profile is representative of the normal cell type, and that it has enough resolution so that there will not be duplicates.

3) What are its limitations? Compare this to the limitations of the new method.

The limitation of the Cluster/Treeview approach is that it is dependent on a visual representation of the data, and that “sameness” is not as clearly represented in this format.

The new method depends on the integrity of the base library of profiles, that they are representative of a particular cell type, and contain enough members to be statistically significant.

The new technique assumes that there is a large enough variability that no two sets of gene expression data from different profiles would produce the same final score, that all of the scores would be unique because expression is very unique. Currently, only scoring increased expression of genes exacerbates this problem by reducing the resolution of the analysis.

The scoring system of the new method lacks some resolution, since it is binary.

4) How could the new method be designed to eliminate an assumption or limitation of the old method?

To handle the problem of significance, the number of gene expression profiles used to generate the library profile could be made larger, or only include profiles that have a certain level of statistical certainty in them.

To mitigate the uniqueness issue, the threshold values that trigger a positive score could be adjusted on a per gene basis as more information is discovered about the actual expression of each gene in a profile, so that more items would be inclusive.

On the limitation of scoring of only increased levels of expression, another level of scoring could be added to the scoring table, representing under expressed genes, that would be scored on a zero or one scale, and added to the over

expressing score, creating a double long binary string that would produce a unique number.

5) How could it (gene expression analysis) be made to fit the biological situation better?

Current taxonomy and classification schemes attempt to make clear distinctions between entities, with the desired result being able to answer if a particular entity is or is not X. Despite the extreme level of complexity in gene expression, it has been shown in the literature that cell types can be distinguished via their gene expression. This new technique fits the ability to distinguish between cell types by generating a score based on gene expression that can be clearly mapped to a particular cell type.

6) The emphasis in computational molecular biology is on the computational. Describe how your solution meets these criteria.

The algorithm/methodology meets this requirement in that it eliminates subjective classification of cell types, and replaces it with a score that is generated/computed from the expression profile, and is a quantitative, not qualitative measure.

7) How can the current computational methods represent the biology better?

The currently used Cluster/Treeview tools could represent the biology better by not only producing a dendrogram of the results, but also including a similarity score like the one proposed in this paper, so that not only could one assess relationship among groups of cell lines, but also assess the sameness to a particular cell line.

Discussion:

Lung carcinomas have been the focus of gene expression profiling and provide a rich amount of data with which to hone classification techniques.

In working with microarray data, the size of the data became an issue, and limited the analysis that was possible. After importing the raw data from Small Cell Lung Carcinomas, Large Cell Lung Carcinomas, and Adeno Carcionomas into excel, the application was using well over 250mb or ram. It was extremely tedious to compare the expression of just one gene across datasets from three different carcinoma types, let alone ten. This resulted in choosing a subset of expressed genes instead of the 400+ genes noted in the literature to have notable expression patterns in lung carcinomas. Further analysis and testing with this new scoring approach should include this larger set of genes.

Ideally, a ROC curve would be created for this technique, proving the merit of this approach. Due to time constraints and the volume of the data that needed to be processed, it was not possible to generate this curve, but one could imagine creating tools that would automate the import, analysis, and comparison against known profiles, as well as importing unrelated gene expression data and observing if false positives are generated.

The process itself could also become just one step of a multi-step process of classification, where if the profile number generated did not match a known number, that an additive score that gave a rough approximation of similarity could be used as an adjunct to the absolute score and indicate what an expression profile is more similar too.

The original goal of this project is to answer the question:

>Could a tool be “trained” to recognize patterns in gene expression >that would be indicative of different kinds of cancers?

Although I would not say a tool was trained, a scoring technique was proposed that would lead to the same result of being able to definitively answer profiling questions.

Some of the secondary questions asked were:

>What are the best ways to score gene expression?
>How do different
>scoring algorithms impact the reliability/accuracy of results?
>Are some measurements more accurate than others?
Unfortunately, there was not enough time due to difficulty managing the data to perform an exhaustive analysis, but the initial results seem promising that this could be a very precise way to match gene expression profiles to cell types.

>Can work in pattern recognition for structure prediction/sequence
>matching be applied the area of gene expression analysis?
Yes, the algorithm was a modification of a scoring algorithm used to score protein sequences. The algorithm is more effective here because the domain of expression is bounded by what is tested on a chip, so that the final/maximum number or score is fixed. With a protein sequence, a mutation/insertion/deletion could change the sequence, making a binary hash of the sequence meaningless.

This methodology overcomes the limitation of being able to distinguish with certainty classes of (lung carcinomas) cells based on gene expression.

References:

1) Classification if human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, A. Bhattacharjee et al.,(2001) PNAS 98:13790-13795 http://www.pnas.org/cgi/content/full/98/24/13790
2) Diversity of gene expression in adenocarcinoma of the lung, M.E.Garber et al.,(2001) PNAS 98:13784-13789 http://www.pnas.org/cgi/content/full/98/24/13784
3) Cluster analysis and display of genome-wide expression patterns, M.B.Eisen et al.,(1998) PNAS 95:14863-14868
4) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implication, T. Sorlie et al.,(2001) PNAS 98:10869-10874
5) Molecular Portraits of Human Breast Tumors, C. M. Perou et al., (2000) Nature 408:747-752
6) Microarrays in primary breast cancer-lessons from chemotherapy studies, Lonning et al.,(2001) Endocrine-Related Cancer 8:259-263
7) Scanalyze User Manual, Michael Eisen