Pacific Symposium on Biocomputing 6:42-51 (2001)

MUTUAL INFORMATION ANALYSIS AS A TOOL TO ASSESS THE ROLE OF ANEUPLOIDY IN THE GENERATION OF CANCER-ASSOCIATED DIFFERENTIAL GENE EXPRESSION PATTERNS

GREGORY T. KLUS[@], ANDREW SONG[@], ARI SCHICK[@], MATTIAS WAHDE[%] and ZOLTAN SZALLASI[@]

[@]Department of Pharmacology, Uniformed Services University of the Health Sciences,Bethesda, MD, [%]Div. of Mechatronics,Chalmers University of Technology, Göteborg, Sweden

(reprint requests at *zszallas@mxc.usuhs.mil)*

Most human tumors are characterized by: (1) an aberrant set of chromosomes, a state termed aneuploidy; (2) an aberrant gene expression pattern; and (3) an aberrant phenotype of uncontrolled growth. One of the goals of cancer research is to establish causative relationships between these three important characteristics. In this paper we were searching for evidence that aneuploidy is a major cause of differential gene expression. We describe how mutual information analysis of cancer-associated gene expression patterns could be exploited to answer this question. In addition to providing general guidelines, we have applied the proposed analysis to a recently published breast cancer-associated gene expression matrix. The results derived from this particular data set provided preliminary evidence that mutual information analysis may become a useful tool to investigate the link between differential gene expression and aneuploidy.

Most human tumors display a set of well-defined aberrations at different levels of cellular biology and biochemistry. These include numeric chromosomal imbalance, termed aneuploidy[1], mutations in various genes, and an abnormal gene expression pattern[2]. One of the main aims of cancer biology is to find the causative relationship between these aberrations. Beyond scientific curiosity, understanding the link between these changes detected in tumors may have a profound impact on cancer therapy as well. If the abnormal gene expression patterns found in tumors were in fact a direct result of aneuploidy, then reversal of aneuploidy might be able to return tumor cells to a more normal gene expression pattern and phenotype, and therapies based on this approach should be investigated.

With the availability of data from the Human Genome Project specifying the various genes on each chromosome, it should now be rather straightforward to establish whether or not an extra chromosome or the loss of a chromosome is reflected in higher or lower expression levels of the genes present on that chromosome. For example, there are cases of pediatric acute lymphoblastic leukemia in which the sole karyotypic change is chromosome 5 trisomy[3]. In these cases the relative expression levels of the genes localized on this chromosome should be increased and this could be readily measured. However, the karyotype of most tumors is significantly more complex and the ploidy regulation of gene expression is likely superimposed by other regulatory mechanisms. Therefore, proving that differential gene expression patterns detected in cancer are generally induced by aneuploidy will probably involve a more complicated analysis of large-scale gene expression and karyotype databases.

The aim of the current paper is to describe a mutual information-based analytical framework for such an analysis, and to perform the first such analysis on a publicly available data set of breast cancer-associated gene-expression changes.

The causes of differential gene expression in cancer: Differential gene expression patterns in cancer result from the superimposition of the following three mechanisms:

1) Extra or missing chromosomes or chromosome regions (segmental aneuploidy). It is obvious that the often-detected complete loss of a given chromosomal region from a cell is reflected in the complete down-regulation of the genes present in that region. It is also well-known that increased copy number of a gene, called DNA amplification or the multiplication of a chromosomal region directly causes up-regulation of gene expression (see for example[4,5])

2) Many oncogenes act as transcription factors themselves or have a well-characterized direct effect on other downstream transcription factors. When these oncogenes (e.g., myc, src and ras) are overexpressed or mutated, they directly or indirectly change the expression level of several other genes[6].

3) The genetic network of a cell with a stable phenotype is self-consistent. In other words, the expression level of each gene is consistent with the expression level of its regulatory inputs. The very existence of cancer-associated differential gene expression proves that the genetic network of a given cell has several alternative stable states. These states are often called attractors in genetic network theory[7], and during malignant transformation the cell is induced to undergo attractor transition. It was also hypothesized, although never proved experimentally, that the cells can reach these alternative attractors after major perturbations of the genetic network, without the continued presence of oncogenes or aneuploidy. This idea is partially supported by the so called hit and run mechanism, when after malignant transformation the causative oncogene (e.g., ras) is lost but the cell still remains in its neoplastic state[8,9]. (It should also be noted that there are examples of reversible malignant transformation, when the cells revert to their non-malignant state after the overexpression of the causative oncogenes has been turned off[10,11].)

General analytical framework in order to establish aneuploidy as a major mechanism inducing cancer-associated gene expression patterns: If aneuploidy is its main driving mechanism, then differential gene expression in cancer will be induced as follows: First a group of genes will be up- or down-regulated due to chromosomal gain or loss. Then this aneuploidy induced gene expression pattern will be adjusted by the regulatory functions of the genetic network present in the cell, keeping the network consistent with the gene regulatory rules.

This hypothesis assumes that the genes present on the same chromosome or chromosome region will be often mis-regulated in the same tumor samples, showing a certain degree of co-regulation in gene expression measurements performed on a sufficiently large number of cancer samples. The level of co-regulation can be readily quantitated by simple means such as calculating the

Pearson correlation coefficient in continuous gene expression measurements[12]. In this paper, however, we propose to use mutual information instead of correlation coefficient (mutual information can be considered as a discretized form of the absolute value of correlation coefficients[13]) for two reasons. First the precision of massively parallel gene expression measurements is limited. Second, the degree of up- or down-regulation which can be expected to result from aneuploidy is not known. Thus, currently it is more informative to trinarize the data, classifying each gene as either unchanged or up- or down-regulated, rather than attempt to weight it with the ratios of mis-regulation. Trinerization can be readily performed after self-normalization of large-scale gene expression matrices as described by Chen et al[14].

Proposed analytical framework:

1. Take a cancer-associated gene expression matrix that was derived from a series of tumor samples of the same type (e.g. a set of primary mammary carcinomas) as population- and time-averaged gene expression data. Convert these data into a ternary matrix at an appropriate confidence level.
2. Calculate pair-wise mutual information for all gene pairs and create relevance networks of co-regulated genes with a mutual information level that is above the highest level detected in the gene expression matrix after randomization (i.e. above a threshold mutual information that can be still due to chance.)
3. Determine the chromosomal localization of the genes of the relevance network and compare it to the chromosomal distribution due to chance. This is determined by simulations assuming that co-regulated genes are randomly assigned to chromosomes.
4. If there are any relevance networks that show an unexpected clustering of genes located on the same chromosome, compare them to aberrations reported for that chromosome.

We will provide detailed description of the steps of this algorithm below, using a concrete breast cancer-associated gene expression matrix.

A **complete analysis** will require several complementary data sets:

1) A large body of gene expression measurements on a given type of cancer. The size of this data matrix is defined by the possible number of chromosome combinations or karyotypes associated with that type of cancer.

2) A catalog of the possible karyotypes of a given cancer. It is well established, that certain gains or losses of chromosomal regions or of whole chromosomes are frequently observed in a certain type of cancer, whereas others never occur[15]. The potential number of major karyotypes is an important reference point in this analysis: if there is a high number of potential configurations of aneuploidy then the number of required gene expression measurements will be proportionally higher.

3) The complete catalog of chromosomal localization of genes involved in the analysis, which will be soon available with the human genome project nearing completion.

A large number of studies on the karyotypes of cancer indicated, that certain chromosomal aberrations are often associated with a certain type of tumor, whereas others are never observed. (See for example[15]). This is also true for mammary tumors[15-18]. In this paper we were looking for relative enrichment of certain chromosomes in high mutual information relevance networks derived from a breast cancer associated gene expression matrix.

Mutual information analysis of a breast cancer-associated gene expression matrix: We have analyzed the breast cancer-associated gene expression matrix recently published by Perou *et al.*[2]. This publicly available data set contains cDNA microarray based relative expression levels of 5,584 genes for a number of both normal and neoplastic breast epithelial samples. For our analysis we have used only gene expression measurements derived from either breast cancer cell lines or primary breast tumors, 16 samples altogether. We have converted the continuous gene expression data into a ternary matrix, using a 2-fold up- or down-regulation as a threshold value. The ternary representation is justified by the current, relatively limited precision of massively parallel gene expression measurements and the fact that we have no estimates about the expected level of up- or down-regulation of gene expression induced by aneuploidy. The exact karyotype of these tumors have not been reported, but it is well known that most sporadic breast tumors have a chromosome set which is far from normal diploid[15-18]. The breast cancer cell lines included in the analysis are also known to have a highly aneuploidic karyotype[19].

In a recent technical paper[20] we have pointed out that the overall quantitative features of cancer-associated gene expression matrices show several consistent characteristics. Namely, the number of mis-regulated genes and the ratio of down-regulated versus up-regulated genes are not arbitrary but remain within a well-defined range for a given type of tumor. This data set had a high level of gene expression diversity. On average, 35% of all quantitated genes were mis-regulated in each sample. The high level of gene expression diversity was reflected in the high level of mutual information content of the data matrix even after randomization. It is also interesting to note, that the breast cancer samples examined here showed significantly more down-regulation than up-regulation of genes. In fact 13 out of 16 samples had more down- than up-regulated genes relative to normal, and in 10 out 16 samples the down-regulated genes outnumbered the up-regulated ones by 3 to 1.

Mutual information analysis: We have calculated mutual information for all possible gene pairs as described in Butte *et al.*[13] and Liang *et al.*[21] with appropriate modifications. For simplicity we kept the range of mutual information between 0 and 1 by using base 3 logarithm for the ternary data set. Therefore the entropy of the mis-regulation for a single gene was calculated as follows:

$$H(A) = -\sum_{i=1}^{3} p(x_i)\log_3(p(x_i)) \quad (1)$$

where $p(x_i)$ is the frequency based probability that gene A will take the value of x_i (i=1,..3) out of the three possible states of 0 (no change), 1 (up-regulation) or -1 (down-regulation). The mutual information for gene pairs A and B is defined as

(2) $MI(A,B) = H(A) + H(B) - H(A,B)$

Randomization of the data matrix: We needed to establish a threshold mutual information level (recently termed and abbreviated as TMI by Butte et al.[13]) above which we considered two genes being co-regulated. Random distribution of 1's, 0's and -1's in a matrix will lead to a certain level of background MI distribution. This is routinely assessed by randomizing the gene expression matrix and then recalculating the pair-wise MI for all gene pairs. We have performed permutative randomization on the gene expression matrix as described in Wahde and Szallasi[22]. This will randomize 1's, 0's and -1's within each row and will retain the average number of mis-regulated genes in the data matrix. The high number of mis-regulated genes of this data matrix predicted a high level of background MI level. Indeed, as demonstrated on Figure 1, after randomization there were several gene-pairs with a pair-wise MI level of up to 0.75. Therefore we have set TMI at this level.

Mutual information analysis, matrix randomization and graphic representation was implemented in Borland Delphi 3. The computation time for calculating the pair-wise mutual information for the complete 5584x16 matrix is about 3 min.

Calculating the chance chromosomal distribution of relevance networks: In an ideal case to prove the involvement of aneuploidy in differential gene expression patterns, one would expect fully connected relevance networks with high mutual information content where all or most genes are localized on the same chromosome. However, these ideal clusters will be "diluted" by the superimposed effect of gene co-regulation and by the fact that certain chromosomal aberrations occur together with higher frequency. On the other hand, if differential gene expression is driven by gene co-regulation with no ploidy effect at all, then one would expect that the genes present in high mutual information clusters, if they exist at all, would be nearly randomly distributed among all chromosomes. This latter assumption has formed the null hypothesis of our statistical analysis. We determined the likely distribution of chromosomal assignments within each relevance network assuming that those genes are randomly localized on chromosomes. Since the exact number of genes on each chromosome has not been determined yet (with the exception of Chr. 21 and 22), we have assumed that the number of genes/chromosome is proportional to the size of the chromosomes measured in megabases. (These data can down-loaded from the web site of National Center for Biotechnology Information at www.ncbi.nlm.nih.gov/.) Human chromosomes vary in size between 263 Mb (Chr. 1) and 47.7 Mb (Chr. 22). Therefore, we assumed that a gene in a relevance network will be assigned with

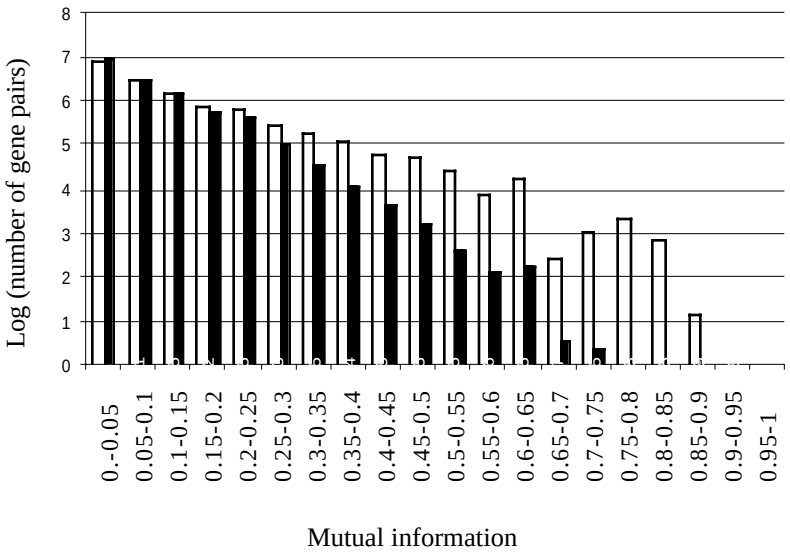

Figure 1 The distribution of mutual information amongst all possible gene-pairs for the actual data set (open columns) and for the average of ten randomized data sets (filled columns). The randomization of the gene expression matrix and the calculation of pair-wise mutual information for all possible gene pairs were performed as described in the text.

about 5-fold higher probability to e.g. Chr. 1 than to Chr. 22. In other words, we assumed that in the absence of ploidy regualation the probability that a given gene is present in a given relevance network will be proportional to the size of the chromosome on which the gene is localized. This assumption will become more accurate as more information becomes available from the human genome project.

We have implemented the following simulation in Matlab: we set the simulated cluster size, i.e. number of genes, to a given detected relevance network of high mutual information (see table 1). Then we have randomly assigned the genes of that cluster to chromosomes in such a way that the probability of assignment was proportional to the size of the chromosome. Finally, we have calculated how frequently we have seen a chromosomal distribution similar to the one observed in the relevance networks derived from the original data set. For each relevance network we ran 1000 simulations and determined whether at 99% confidence level the detected chromosomal distribution is due to chance.

Summary of findings: We have identified 65 relevance networks at a TMI level of 0.75. The majority of these were small clusters, namely 35 gene pairs and 16 gene triplets. None of the gene triplets were localized on the same chromosome. Preliminary analysis suggested, that it is likely (>10% chance) that two genes in a

relevance network of three genes will be localized on the same chromosome. Therefore, further examination of these small clusters was not informative. We have identified 14 relevance networks with more than 3 genes. The chromosomal localization of each gene was determined by a sequence-based BLAST search against the human genome data-base maintained by NCBI. (Available at http://www.ncbi.nlm.nih.gov/genome/seq). This has ensured that the chromosomal localization of the actual gene probes were determined even if a given microarray probe carried the wrong gene identification. The chromosomal distribution of the genes of these networks is listed in Table 1. All relevance networks were fully connected at a MI>0.75 level. 13 out of the 14 relevance networks showed chromosomal distributions that could be caused by chance (at 99% confidence level) assuming the random chromosome assignment described above.

Relevance network #3, however, displayed significant "enrichment" of genes originating on three chromosomes. This relevance cluster of 13 genes contained four genes from chromosome 17, three genes from Chr. 1, and two genes from Chr. 11, and the remaining four genes were from different chromosomes. This distribution of chromosomal assignment is unlikely due to chance at a 99% confidence level. It is well documented that chromosomes 1, 11 and 17 belong to the group of chromosomes that show numerical aberration with the highest frequency in breast cancer[15-18]. These chromosomes often show numerical changes together[15-18]. It is also known that loss of heterozygosity involving these chromomes is frequently detected in these tumors, and these chromosomes are more often lost than gained in breast cancer[15-18]. These data showed excellent correlation with the fact that the mis-regulation of genes involved in this relevance network represented mainly down-regulation. (The genes present in this relevance network were down-regulated in 8 tumors, up-regulated in one tumor and unchanged in 7 samples.) In this case, the relevance network gave a very good indication of the abnormal behavior of chromosomes associated with it.

Discussion: In this paper we have introduced mutual information analysis as a tool to establish a causative link between aneuploidy and differential gene expression in cancer. The limited sample number of the available gene expression data in breast cancer and the lack of a comprehensive database of karyotypes has obviously limited our analytical efforts at the moment. Nevertheless, in one case our analysis turned up a large relevance network of high mutual information in which the genes' chromosomal assignment was non-random. Furthermore, the three chromosomes highly represented in this relevance network (Chr. 1, 11 and 17) have been reported to show coordinated numerical aberrations in breast cancer[15-18]. These chromosomes are often lost which corresponds well with the frequent coordinated down-regulation of these genes in the breast cancer associated gene expression matrix examined.

The fact that only one out of fourteen relevance networks showed signs of involvement of aneuploidy suggests that chromosomal aberrations may play a limited role in the differential gene expression detected in breast tumors. However, the relevance network with non-random chromosomal assignment provide a

preliminary proof of principal and suggest a wider application of mutual information for this type of analysis.

Abbreviations: Chr.; Chromosome, TMI: threshold mutual information, MI: mutual information

Acknowledgment: The opinions and assertions contained herein are the private opinions of the authors and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences or the U.S. Department of Defense.

Relevance Network	Chromosomes represented by 1 gene	2 genes	3 genes	4 genes
#1 17 genes (2 unknown)	1,2,3,4,5,8, 12,13,15,16	17	10	
#2 15 genes (1 unknown)	1,2,7,8,14, X	2,9,18,19		
#3 13 genes	2,4,5,10	11	1	17
#4 11 genes	4,6,13,14, 19,20,X	2,17		
#5 10 genes	2,3,4,6,10, 12	5,15		
#6 10 genes (2 unknown)	2,11	3,4,10		
#7 9 genes (1 unknown)	1,2,3,6,8, 10,11,19			
#8 8 genes	3,9,10,15, 17,19	1		
#9 7 genes (2 unknown)	1,3,5,16,19			
#10 6 genes (1 unknown)	15	6,12		
#11 5 genes (1 unknown)	1,2,9,12			
#12 5 genes	15	3,12		
#13 5 genes	1,4,7,21,22			
#14 4 genes	1,2,5,7			

Table 1. List of chromosomal assignments of genes present in relevance networks with high mutual information and with more than 3 genes. See further details in the text.

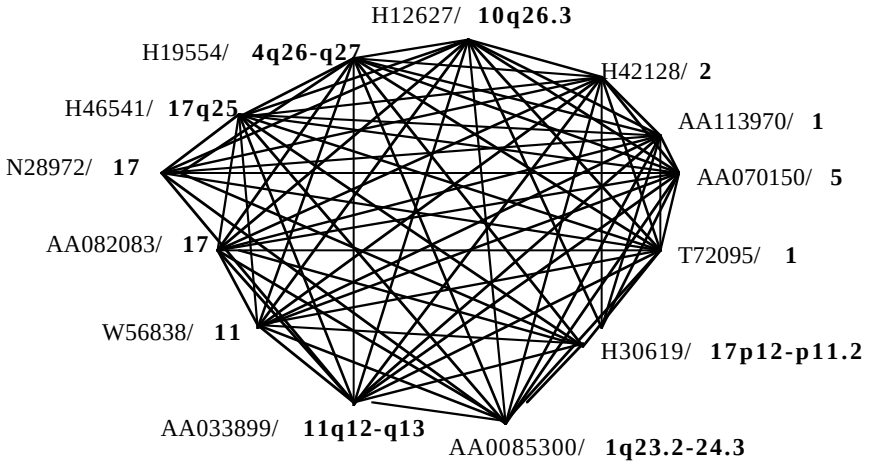

Figure 2. Relevance network #3. The gene accession number and the corresponding chromosomal localization (in bold letters) is listed for each gene.

References:

1. Sen, S. Aneuploidy and cancer. *Curr Opin Oncol* 12:82-88 (2000)
2. Perou, C.M., *et al.* Distinctive gene expression patterns inhuman mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A* 96:9212-9217. (1999)
3. Sandoval, C., *et al.* Trisomy 5 as a sole cytogenetic abnormality in pediatric acute lymphoblastic leukemia. *Cancer Genet Cytogenet* 118:69-71 (2000)
4. Menard,S,, *et al.*Role of HER2 gene overexpression in breast carcinoma. J *Cell Physiol* 182:150-62 (2000)
5. Galitski, T., *et al.* Ploidy regulation of gene expression. *Science* 285:251--254. (1999)
6. Coller,, H.A., *et al.* Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc Natl Acad Sci U S A* 97:3260-3265 (2000)
7. Kauffman, S. The origins of order. *Oxford: Oxford University Press*. (1993).
8. Lau, C.C., *et al.* Plasmid-induced "hit-and-run" tumorigenesis in Chinese hamster embryo fibroblast (CHEF) cells. *Proc Natl Acad Sci U S A* 82:2839-2843 (1985)
9. Plattner, R, *et al.* Loss of oncogenic ras expression does not correlate with loss of tumorigenicity in human cells. *Proc Natl Acad Sci U S A* 93:6665-6670 (1996)
10. Felsher, D.W., and Bishop, J.M. Reversible tumorigenesis by MYC in hematopoietic lineages. *Mol Cell* :199-207 (1999)

11. Baasner S., *et al.*Reversible tumorigenesis in mice by conditional expression of the HER2/c-erbB2 receptor tyrosine kinase. Oncogene 13:901-911 (1996)
12. Eisen, M.B., *et al.* Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-14868 (1998)
13. Butte, A.J. and Kohane, I.S. Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. *Pacific Symposium on Biocomputing* 5:415-426 (2000).
14. Chen, Y., Dougherty, E.R., and Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images.*J. Biomed.Optics* 2:364--374. (1997)
15. Mertens, F., *et al.* Chromosomal imbalance maps of malignant solid tumors: a cytogenetic survey of 3185 neoplasms. *Cancer Res.*57:2765-2780. (1997)
16. Botti C, *et al.* Incidence of chromosomes 1 and 17 aneusomy in breast cancer and adjacent tissue: an interphase cytogenetic study. *J Am Coll Surg* 190:530-539 (2000)
17. Marinhom A.F., Botelho, M., and Schmitt F.C. Evaluation of numerical abnormalities of chromosomes 1 and 17 in proliferative epithelial breast lesions using fluorescence in situ hybridization. *Pathol. Res. Pract.* 196:227-233 (2000)
18. Ferti-Passantonopoulou A.D., and Panani A.D. Common cytogenetic findings in primary breast cancer. *Cancer Genet Cytogenet* 27:289-298. (1987)
19. Whang-Peng, J., *et al.* Cytogenetic studies of human breast cancer lines: MCF-7 and derived variant sublines. *J Natl Cancer Inst* 71:687-695 (1983)
20. Klus, G.T., *et al.* Use of overall quantitative features of cDNA microarray measurements in cancer research. (at: http://www.usuhs.mil/pha/faculty/zoltan.shtml) (2000)
21. Liang, S., Fuhrman, S. and Somogyi, R. REVEAL, A General Reverse Engineering Algorithm for Inference of GeneticNetwork Architectures *Pacific Symposium on Biocomputing* 3:18-29 (1998).
22 Wahde, M. and Szallasi, Z. Generative model based analysis of cancer associated gene expression matrices. *Proceedings of the First International Conference on Systems Biology*. (in press, 2001)