

Joint Analysis of DNA Copy Numbers and Gene Expression Levels

Doron Lipson¹, Amir Ben-Dor², Elinor Dehan³, and Zohar Yakhini²

¹ Computer Science Dept., Technion, Israel
dlipson@cs.technion.ac.il, corresponding author

² Agilent Laboratories, Palo Alto, CA
{[amir_ben-dor](mailto:amir_ben-dor@agilent.com), [zohar_yakhini](mailto:zohar_yakhini@agilent.com)}@agilent.com

³ Functional Genomics Unit, Sheba Medical Center, Israel
ed486@endeavor.med.nyu.edu

Abstract. Genomic instabilities, amplifications, deletions and translocations are often observed in tumor cells. In the process of cancer pathogenesis cells acquire multiple genomic alterations, some of which drive the process by triggering overexpression of oncogenes and by silencing tumor suppressors and DNA repair genes. We present data analysis methods designed to study the overall transcriptional effects of DNA copy number alterations. Alterations can be measured using several techniques including microarray based hybridization assays. The data have unique properties due to the strong dependence between measurement values in close genomic loci. To account for this dependence in studying the correlation of DNA copy number to expression levels we develop versions of standard correlation methods that apply to genomic regions and methods for assessing the statistical significance of the observed results. In joint DNA copy number and expression data we define significantly altered submatrices as submatrices where a statistically significant correlation of DNA copy number to expression is observed. We develop heuristic approaches to identify these structures in data matrices. We apply all methods to several datasets, highlighting results that can not be obtained by direct approaches or without using the regional view.

1 Introduction

Alterations in DNA copy number (DCN) are characteristic of many cancer types and are thought to drive some cancer pathogenesis processes. Alterations include large chromosomal gains and losses as well as smaller scale amplifications and deletions. Genomic instability triggers the overexpression or activation of oncogenes and the silencing of tumor suppressors and DNA repair genes. Alterations in DCN have been initially measured using local fluorescence in situ hybridization-based techniques. These evolved to a genome wide technique called Comparative Genomic Hybridization (CGH, see [8]), now commonly used for the identification of chromosomal alterations in cancer [10, 2]. In this genome-wide cytogenetic method differentially labelled tumor and normal DNA

are co-hybridized to normal metaphases. Fluorescence level ratios between the two labels allow the detection of chromosomal amplifications and deletions. This method has, however, a limited resolution (10-20 Mbp) which makes it impossible to predict the borders of chromosomal changes or to identify changes in copy numbers of single genes and small genomic regions. In a more advanced method termed *array CGH* (aCGH) the tumor and normal DNA are co-hybridized to a microarray of thousands of genomic clones of BAC, cDNA or oligonucleotide probes ([13, 11, 6]). The use of aCGH allows the determination of changes in DCN of relatively small chromosomal regions. When using oligonucleotide arrays the resolution can, in theory, be finer than single genes.

The development of high resolution mapping of DCN alterations and the progress of expression profiling technologies enable the study of the effects of chromosomal alterations on cellular processes and how these are mediated through altered expression of genes residing in altered regions. By measuring DNA copy numbers and mRNA expression levels on the same set of samples we gain access to the relationship of copy number alterations to how they are manifested in altering expression profiles. In [12] the authors used (metaphase slides) CGH to identify large scale amplifications in 23 metastatic colon cancer samples and performed expression profiling on the same samples. They observed some correlation between DCN and expression levels but on overall effect of alterations on transcription. In [14] an opposite observation is reported, for breast cancer samples. That is: a strong global correlation between copy number changes and expression level variation is observed. Similarly, Hyman et al [7] studied copy number alterations in 14 breast cancer cell lines and identified 270 genes with expression levels that are systematically attributable to gene amplification. The statistics used by both latter studies is based on simulations and takes into account single gene correlations but not local regional effects. Recently, Linn et al studied expression patterns and genome alterations in DFSP and discovered common 17q and 22q amplifications that are associated with elevated expression of resident genes [9].

Our purpose in this paper is to provide algorithmic and statistical methods to rigorously support data analysis designed to improve our understanding of copy number to transcription relationships (specifically in aCGH data). Regions of high correlation are potentially related to the tumor pathogenesis. More specifically, genes affected by changes in DCN potentially play a role in driving tumor differentiation. Note that the correlation between expression data vectors and their corresponding (same gene) DCN data vectors should behave completely random if all the variation in the DCN vector arises due to experimental errors. We are therefore mostly interested in detection of statistically significant correlations. These might not show up when low resolution and global data analysis approaches are employed. For example, low penetrance (not all cells in the sample) and low prevalence (not all samples in the study) alterations might effect expression below the 2-fold mark and only in some of the samples, but in a significant manner when a genomic region is considered. In addition – the detection of regions that manifest a significant correlation can aid in detecting actual low

penetrance alterations in high resolution even if the DCN data, alone, do not support such discoveries.

Throughout the paper we use C and E to denote the DNA copy number and gene expression data matrices, where the (i, j) -th entry of each matrix represent the data for the i th gene in the j th sample. We abbreviate “DNA copy number” and “gene expression” as DCN and GE respectively.

In Section 2 we describe methods to quantify the correlation observed for any pair of rows $C(i, \cdot)$ and $E(i, \cdot)$. In Section 3 we extend these to account for the regional character of DCN alterations. In Section 4 we discuss submatrices of affected samples and genes. Conclusions are discussed in Section 5.

Datasets We demonstrate our methods on two joint DCN-GE breast cancer datasets. The first, described in [14] is of 6,095 genes across 41 samples (4 cell lines, 37 primary breast tumors). The second, from [7] is of 13,824 genes across 14 cell-line samples. All datasets are measured on cDNA microarrays.

2 Correlation Scoring Methods

Consider a single gene g and let $u = u_g$ and $v = v_g$ denote the corresponding DCN and GE data vectors of g . Let n denote the number of samples (length of u and v). In this section we present several approaches for scoring g by looking for dependencies between u and v .

2.1 Pearson Product Moment Correlation

The most common measure of the dependence between two vectors u, v is the Pearson correlation coefficient:

$$r(u, v) = \frac{\sum(u - \bar{u})(v - \bar{v})}{\sqrt{\sum(u - \bar{u})^2} \sqrt{\sum(v - \bar{v})^2}}. \quad (1)$$

r measures the degree to which u and v maintain a *linear* relationship. It may therefore be less suitable when the DCN and GE values follow some non-linear relationship. Nonetheless, previous large-scale DCN-GE comparative studies [14] used Pearson correlation as a sole scoring method to evaluate dependence.

2.2 Separating-Cross Correlation

A different methodology for comparing gene copy measurements with gene expression levels, such as the one described in [7] utilizes user chosen thresholds for classifying DCN measurements as deleted or amplified and for classifying GE measurements as under-expressed or over-expressed. Such methods do not rely on any assumption of linearity or on the value of the mean; however, they are somewhat dependent on the specific choice of thresholds. The separating-crosses scores we now introduce are a generalized approach to threshold-based analysis of the dependence between two vectors.

We can view the two vectors u and v as n points (u_i, v_i) in the plain. An axis parallel cross $t = t_{x,y}$, centered at (x, y) , partitions the plain into four quadrants, denoted by A_t, B_t, C_t , and D_t (See figure 2). We denote by a_t the number of points (u_i, v_i) that belong to the quadrant A_t . The other quadrants counts b_t, c_t and d_t are defined similarly. Clearly, $a_t + b_t + c_t + d_t = n$.

Roughly speaking, u and v are correlated if there exist a cross t such that both a and d are large (compared with b and c). More generally, assume we are given a function of the quadrants counts (such a function is called a *cross-function*), $f(a, b, c, d)$. We are interested in the maximal obtainable value of f :

$$F(u, v) = \max_t \{f(a_t, b_t, c_t, d_t)\}. \quad (2)$$

The function F is called a *separating cross score function*.

Let π denote the ranks of the samples with respect to the vector u . That is, $u(\pi^{-1}(1)) < \dots < u(\pi^{-1}(n))$. For example, for $u = (2, 1.5, 9, 0.4)$, $\pi = (3, 2, 4, 1)$. Similarly, we denote by τ the samples permutation induced by v . Since cross-functions (and thus score functions) depend only on quadrants counts and not on the actual locations of the points, we have $F(u, v) = F(\pi, \tau)$. Thus, for every function $f(\pi, \tau, t)$, we can compute $F(\pi, \tau)$ by examining $(n-1)^2$ possible crosses. We describe one cross score function, the *Maximal Diagonal Product* (MDP). Consider the separating-cross-function $DP(\pi, \tau, t) = a_t \cdot d_t$, which we call *Diagonal Product* (DP), and the corresponding score function:

$$MDP(\pi, \tau) = \max_t \{DP(\pi, \tau, t)\}. \quad (3)$$

A useful attribute of the MDP score is that it provides a distinction between samples that contribute to the maximum score (points within A_t and D_t) and those that do not (points within B_t and C_t). We make use of this attribute in identifying affected samples in Section 4. The combinatorial nature of this score allows rigorous calculation of its statistical properties, the discussion of which is beyond the scope of this paper.

3 Regional Analysis

Some cancer related alterations in genomic DNA have direct effect on mRNA levels, possibly leading to downstream functional deficiencies. These alterations are most likely *localized* in one or more of the following aspects:

1. The alteration in genomic DNA is limited to certain chromosomal segments.
2. The expression of all genes within a specific genomic segment may not be effected to the same extent.
3. Not all samples contain identical or similar genomic alterations.
4. Within specific samples, alterations occur with varying levels of penetrance.

Previous work on DCN-GE expression relationship consider only correlation between the gene-expression levels of single genes and their respective DNA copy

number measurements. Pollack et al [14] study the global behavior of DCN-GE correlation and show that the distribution of Pearson correlation values between DCN and GE differs from the expected distribution. They report 54 amplified genes with moderately or highly elevated expression levels. Hyman et al [7] also demonstrate global single gene correlations and identify 270 genes with expression levels significantly influenced by changes in DCN .

CGH based studies show that chromosomal alterations frequently apply to long stretches of the genome that may span a large number of genes. The expression pattern of a gene that is affected by such an aberration is expected to correlate not only with the copy number of its own coding DNA but also with the DCN measurements of neighboring genes. We therefore expect that analysis that takes regional effects into account to yield better results that might offset the negative effects of noise in the data or low penetrance. Both [14] and [7] did not account for such regional considerations. In this section we suggest a framework for considering local correlation between genomic alteration and variance in gene expression levels, that accounts for regional effects.

Given a gene g_i we define its k -neighborhood as the continuous sequence of genes indexed by $\Gamma_k(i) = (i - k, \dots, i + k)$. A straightforward approach to quantifying the correlation of the gene's GE vector $E(i, \cdot)$ with the DCN vectors in its neighborhood $\Gamma_k(i)$ is by calculating the average correlation of $E(i, \cdot)$ to each of the respective DCN vectors:

$$r(i, \Gamma_k(i)) = \frac{1}{2k + 1} \sum_{j=i-k}^{i+k} r(i, j), \quad (4)$$

where $r(i, j)$ is any correlation measure between the vectors $E(i, \cdot)$ and $C(j, \cdot)$.

Alternative approaches to regional correlation include the correlation of $E(i, \cdot)$ to the vector of (weighted or uniform) average DCN in $\Gamma_k(i)$, or the product of the p-values of the respective correlations.

Permuted Data When performing analyses that take gene order into account we compare results to a null model that assumes neighboring genes are independent of each other. To this end, we also perform our analysis on gene-permuted matrices E' and C' where the same permutation was applied to the rows of both matrices. We expect regional effect results to be dependent on the original chromosomal order of the genes.

Computing p-Values To identify regions where DCN and GE correlate beyond the extent expected for the consistent DCN values we perform the simulation analysis outlined below. The general idea is to evaluate a locus dependent p-value for chromosomal regions. Correlations in regions where a very consistent DCN measurements are observed need to cross much higher thresholds to be significant since distributions expected at random in such regions have larger variation (weaker smoothing effect of averaging because of the consistent DCN values). Consider a neighborhood $\Gamma_k(i)$ and fix L , the size of simulation applied. We randomly draw $L-1$ expression vectors (rows of E) indexed by i_1, i_2, \dots, i_{L-1} ,

and for each compute $r_l = r(i, \Gamma_k(i))$, the correlation of the random expression vector to the neighborhood $\Gamma_k(i)$. To the correlation $r^* = r(i, \Gamma_k(i))$, actually observed at i , we assign its rank ρ amongst $r_1, r_2, \dots, r_{L-1}, r^*$, a number between 1 and L . The p-value for the region correlation observed at i is $pV(i) = \rho/L$.

3.1 Results

We applied the above locus dependent p-value calculations to investigate copy number to expression correlations in [14]. Figure 1 depicts the cumulative distribution of $pV(i)$, where i ranges over all genes in the dataset. As expected, randomly permuting the dataset yields a straight line that can be used as reference (curve E), while significant single gene correlations (i.e. $r(i, i)$, curve C) are overabundant at all p-values. Significant correlations are even more abundant when computed for neighborhoods of size $k = 2$ and $k = 10$ (curves B and A, respectively). Note that these results depend on both the chromosomal order (as the gene-permuted data yields a lower abundance of significant correlation scores than single gene correlations, curve D) and on direct DCN to GE correlations (due to the method of calculating $pV(i)$). The region-dependent $pV(i)$ scores enable the identification of loci where the gene expression levels significantly correlate with the DCN measurements with greater statistical confidence. For illustration, consider a threshold of $pV(i) \leq 0.001$. A random dataset of 6000 genes is expected to contain 6 genes with this score whereas single gene correlations yield 164 such genes (FDR = 3.7%). Averaged correlation against $\Gamma_2(i)$ yields 214 significant loci (FDR = 2.8%) and working with $\Gamma_{10}(i)$ yields 289 significant loci (FDR = 2.1%). Thus, using region-based analysis delivers almost 80% more loci where DCN-GE correlation may be identified with high confidence. Furthermore, additional regions of correlation are thus detected (details in the supplement [1]).

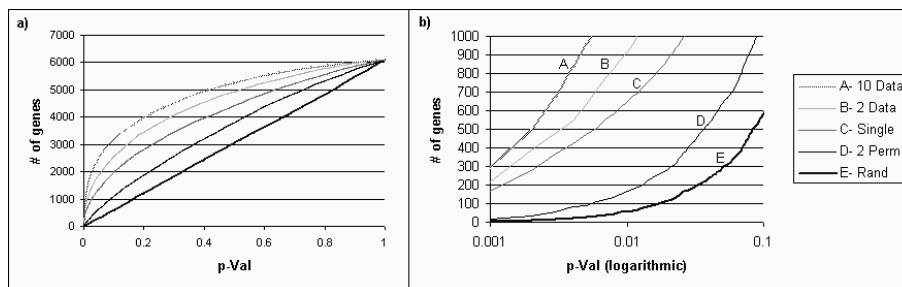


Fig. 1. a) Cumulative distribution of values of $pV(i)$ (p-value of the average correlation $r(i, \Gamma(i))$), for neighborhoods of size $k = 10$ (A), $k = 2$ (B), single genes (C), and for gene-permuted data with a neighborhood of size $k = 2$ (D). Randomly permuted data yields a straight line, as expected (E). b) Zoom of lower values of the same curves at logarithmic scale. Note the different values attained at $pV(i) = 10^{-3}$.

4 Genomic-Continuous Submatrices

In the previous section we mentioned that genomic alterations are often localized to a subset of the samples as well as to a specific chromosomal segment. In this section we discuss the task of detecting both the genomic segment in which an aberration has occurred, the affected samples and the transcriptional effect of the aberration. To this end we define a model of significantly altered genomic-continuous submatrices and present two algorithmic approaches for detecting them. We then apply the suggested methods to the two breast cancer datasets.

4.1 Definition

For a given pair of DCN and GE matrices C, E over an ordered set of genes G and a set of samples S we define a *genomic-continuous submatrix* (GCSM) as $M = G' \times S'$ where $G' \subset G$ is a continuous segment of genes and $S' \subseteq S$ is a subset of samples, and its *complement submatrix* $\bar{M} = G' \times \{S - S'\}$. Let $C(M)$ and $E(M)$ denote the projections of the matrices C and E on the subsets G' and S' (the DCN and GE *submatrices* corresponding to M).

Under our biological model, a genomic alteration in a given chromosomal segment and a given sample should affect most of the DCN measurements in this segment but only some of the respective GE measurements (since changes in expression depend on other factors that determine regulation). Informally, we say that a given GCSM M is *significantly amplified* if:

- Most DNA copy values in the set $C(M)$ are positive.
- Some genes $g_i \in G'$ have higher expression values $\{E(i, j) : s_j \in S'\}$ compared to $\{E(i, j) : s_j \notin S'\}$.

More formally, let us define a score that reflects the degree to which M is significantly amplified. First, we define a score $F(M; C)$ that reflects overabundance of positive values in $C(M)$ in comparison to $C(\bar{M})$ using the hypergeometric distribution. Let $N = |C(M \cup \bar{M})|$ and $n = |C(M)|$. Let K and k be the number of positive values in $C(M \cup \bar{M})$ and $C(M)$, respectively. Given N, n, K the hypergeometric probability of finding k or more positive values in $C(M)$ is:

$$F(M; C) = HG(N, K, n, k) = \sum_{i=k}^N \frac{\binom{n}{i} \binom{N-n}{K-i}}{\binom{N}{K}}. \quad (5)$$

Similarly, we define a score $F(M; E)$ that reflects the overabundance of genes in G' that are significantly differentially expressed, comparing S' and $S - S'$, in the correct direction (higher in S' than in $S - S'$). A TNoM (Threshold Number of Misclassifications) score may be assigned to each gene according to its performance as a S' versus $S - S'$ classifier [5, 3]. Rigorous p-values can be calculated for TNoM. If the probability, for a single gene, of obtaining a score of s or better under the null model is $p(s)$ then the number of genes with scores s or better, amongst the $|G'|$ genes examined, is Binomial($n, p(s)$) distributed.

Let $n(s)$ denote the number of genes with such scores actually observed in the data. Let $\sigma(s)$ be the tail probability of the Binomial($n, p(s)$) distribution, at $n(s)$. $F(M; E)$ is then defined to be $\max_{0 \leq s \leq |S'|} -\log(\sigma(s))$. For a more detailed description of differential expression overabundance please see [4].

Under the null model, DCN and GE vectors are completely uncorrelated. A total score for an amplification in M is:

$$F(M; C, E) = -[\log_{10} F(M; C) + \log_{10} F(M; E)] \quad (6)$$

The above discussion addresses amplifications only. However, any deletion in a subset S' is equivalent, under F , to an amplification in $S - S'$.

4.2 Algorithmic Approach

Locating a partition of samples that maximizes TNoM overabundance for a given set of genes is by itself a difficult task that has been approached by heuristic methods [4]. The problem of locating a partition that maximizes a combined hypergeometric and TNoM overabundance score is clearly at least as hard, and consequently we resort to heuristic approaches for locating significantly altered GCSMs. Note that due to the fact that we are looking for continuous segments only, all possible segments may be enumerated in $O(n^2)$, where n is the number of genes. The difficulty remains in determining which subset S' maximizes $F((G' \times S'); C, E)$ for a given segment G' . We suggest two algorithmic approaches:

Max-Hypergeometric Algorithm As the definition of the score of a GCSM M is composed of two parts, a reasonable heuristic approach to locating high-scoring GCSMs is to select the sample partitions that maximize one part of the score – the hypergeometric score – for each possible segment, and for these partitions to calculate the combined score. For a given segment G' the calculation of $\max_{S' \subseteq S} [-\log(F((G' \times S'); C)]$ may be performed in $(O(|S|))$ time by ordering the samples according to decreasing number of positive entries, as described in the following psuedo code:

Algorithm 1 MHA - Max-Hypergeometric Algorithm

Input: C, E, t - a significance threshold, l - maximum segment length.

Output: A list of high scoring GCSMs, L .

for all segments $G' \subset G$ of length $\leq l$ **do**

For each sample $s_i \in S$ let $p_i = \#$ of positive entries in $C(G', s_i)$.

Order the samples s.t. $p_{\pi(1)} \geq \dots \geq p_{\pi(|S|)}$.

$maxScore = \max_{1 \leq i < |S|} F((G', \{S_{\pi(1)}, \dots, S_{\pi(i)}\}); C, E)$.

if $maxScore > t$ **then**

Add $M = (G', S')$ to L .

Consistent Correlation Algorithm One shortcoming of MHA is that it depends on a sufficiently strong pattern in the DCN measurements alone in order

to detect high-scoring GCSMs. However, in Section 3 we argued that in some cases significant correlation between DCN and GE patterns is indicative of a chromosomal aberration even when the DCN signal *per se* is weak. The second algorithmic approach relies on DCN -GE correlations for locating candidate partitions S' . To this end, we make use of a helpful attribute of the MDP correlation score. Recall from Section 2.2 that for a given gene g_i the score $\text{MDP}(i)$ defines a cross-threshold t for which the product $A_t \cdot D_t$ is maximized (see Figure 2). It is therefore straightforward to separate the samples that contribute to the score $\text{MDP}(i)$ – those within A_t or D_t – from those that do not (within B_t or C_t). Taking into consideration the chromosomal neighborhood of g_i we can increase our confidence that g_i 's expression level in a specific sample is affected by the aberration. Consider, for example, a sample s that falls in D_t when computing $\text{MDP}(i, j)$ for $E(i)$ and all relevant $C(j)$. The probability of such an event occurring at random decreases exponentially with k .

For a gene g_i and a sample $s \in S$ we define the *sample MDP score* of s as:

$$\text{SMDP}(s, i) = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} \{ [1_{s \in A_t(i,j)} \text{MDP}(i, j)] - [1_{s \in D_t(i,j)} \text{MDP}(i, j)] \},$$

where $A_t(i, j)$ and $D_t(i, j)$ are the sets of samples that fall into quadrants A_t and D_t for the threshold t that attains the MDP for $E(i)$ and $C(j)$. Note that $-\text{MDP}(i, \Gamma_k(i)) \leq \text{SMDP}(s, i) \leq \text{MDP}(i, \Gamma_k(i))$ and extrema are attained if s falls in either A_t or D_t in all of the crosses.

The above method allows us to rank the samples $s_i \in S$ according to increasing odds that they have been affected by an amplification. As before, this ranking suggests $O(|S|)$ partitions to be evaluated. In practice, we may opt to run the algorithm on a filtered set of genes $\tilde{G} \subset G$ that pass some minimal regional correlation threshold, in accordance with the statistics mentioned in Section 3.

Algorithm 2 CCA - Consistent Correlation Algorithm

Input: C, E, \tilde{G} - a subset of genes, k - the neighborhood size, t - a significance threshold, l - maximum segment length.

Output: A list of high scoring GCSMs, L .

for all genes $g_i \in \tilde{G}$ **do**

For each sample $s_j \in S$ calculate $p_i = \text{SMDP}(s_j, i)$.

Order the samples s.t. $p_{\pi(1)} \geq \dots \geq p_{\pi(|S|)}$.

for all segments $G' \subset G$ of length $\leq l$ s.t. $g_i \in G'$ **do**

$\text{maxScore} = \max_{1 \leq i < |S|} F((G', \{S_{\pi(1)}, \dots, S_{\pi(i)}\}); C, E)$.

if $\text{maxScore} > t$ **then**

Add $M = (G', S')$ to L .

Analysis Note that the two algorithms are appropriate for two different types of high-scoring GCSMs. MHA is optimal when $F(M; C)$ is a dominant factor of the total score, i.e. when the DCN measurements alone point to a chromosomal

aberration. CCA is appropriate when there is a strong correlation between $E(M)$ and $C(M)$ suggesting that both $F(M; C)$ and $F(M; E)$ have a significant part in the total score. In the latter case we expect that a chromosomal alteration has a significant effect on transcriptional activity. A third extremal case, for which neither algorithm is appropriate, is the case in which $F(M; E)$ alone is a dominant factor of the total score. A biological interpretation of this event is co-regulation of neighboring genes, unlinked to chromosomal aberration (as in an operon). This type of effect is not in the scope of this study and may be overlooked by both algorithms.

4.3 Results

We applied both algorithms to detect high-scoring GCSMs in the joint DCN-GE breast tumor data of Pollack et al [14], using a threshold of $F(M; C, E) > 30$. A considerable number of significant GCSMs was detected both by MHA and CCA spanning genomic segments of 5-52 loci, or 0.25-4Mbp. The two algorithms locate similar but not identical high-scoring GCSMs. MHA produced no GCSMs with scores above the given threshold for either randomly permuted data or the same dataset where the genomic order was randomly permuted, verifying that the high scores attained were due to regional phenomena (see Figure 3). A distribution with intermediate values is obtained when MHA is run on a the dataset where only the genomic order of the expression vectors was randomly permuted, suggesting that the DCN matrix on its own is accountable only for a fraction of the high scores obtained for the complete data. CCA does not produce any results on random data since, by definition, it initiates a search only when the regional correlation score is statistically significant. Genomic aberrations were located in various chromosomal segments, including 1p, 1q, 3p, 8p, 8q, 13q, 17q and 20q reported in [14] and in several additional locations. These include a GCSM with $F(M) = 60.3$ in 17q (28-32Mbp), and an a GCSM with $F(M) = 50.2$ in 11q (69-73Mbp). Genes affected by altered GCSMs include TP53, FGFR1 and ERBB2, known to be involved in breast cancer (see supplement [1] for a complete list of high-scoring GCSMs, and some affected genes). Similar results were obtained for the joint DCN-GE breast cancer cell-line dataset of Hyman et al [7]. Our method validates the novel alterations in 9p13 (GCSM score 25.5) and in 17q21.3 (GCSM score 37). The same aberrations were located in a subset of the samples of [14], with GCSM scores of 32.5 and 53 respectively. Figure 4 depicts a significant alteration in 17q11 that is significantly associated with altered expression levels of 5 resident genes (data from [14]).

5 Conclusion

The advanced stage of expression profiling technologies, coupled with continued development of aCGH and other technologies for measuring DCN alterations enable a more comprehensive study of the molecular profile of cancer. Specifically, these technology advances enable the generation of joint data sets where

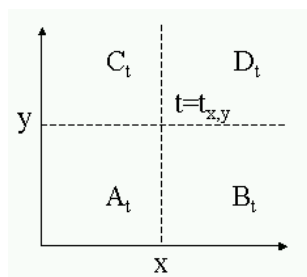


Fig. 2. Notation for separating crosses

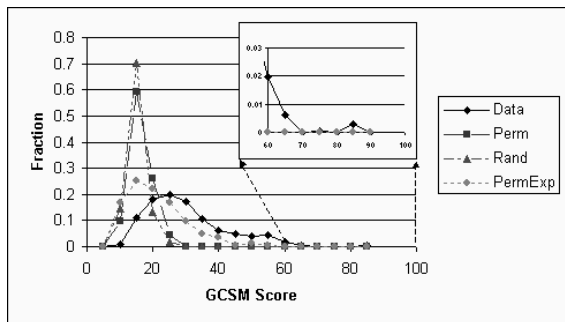


Fig. 3. Distribution of GCSM scores obtained by MHA on data of [14] (Data), on randomly permuted data (Rand), and on randomly permuted gene order (both DCN and GE - Perm; only GE - PermExp).

DCN and GE values are measured on the same set of samples and genes. Using these data to understand cancer pathogenesis mechanisms related to alterations and their transcriptional effects will require the development of adequate data analysis methodology and tools. In this paper we provide a first statistical and algorithmic treatment of this methodology.

We describe methods for assessing the correlation of DCN to GE for single genes, taking into account regional effects. We also develop a statistical and algorithmic framework for identifying altered regions with a correlated transcriptional effect. Applying our methods to published breast cancer data we provide strong evidence of the transcriptional effects of altered DCN and identify chromosomal segments and sample subsets where these effects are more pronounced. Specifically, we identify regions where a statistically significant alteration affects more than one resident gene.

In future work we intend to improve the algorithmics and to further develop rigorous statistics for biologically meaningful structures in DCN-GE data.

References

1. Supplement data available at <http://bioinfo.cs.technion.ac.il/cghexp/>.
2. B.R. Balsara and J.R. Testa. Chromosomal imbalances in human lung cancer. *Oncogene*, 21(45):6877–83, 2002.
3. A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. In *Proceedings of RECOMB*, pages 54–64, 2000.
4. A. Ben-Dor, N. Friedman, and Z. Yakhini. Class discovery in gene expression data. In *Proceedings of RECOMB*, pages 31–8, 2001.
5. M. Bittner et al.. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–40, 2000.
6. I. Hedenfalk et al.. Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *PNAS*, 100(5):2532–7, 2003.

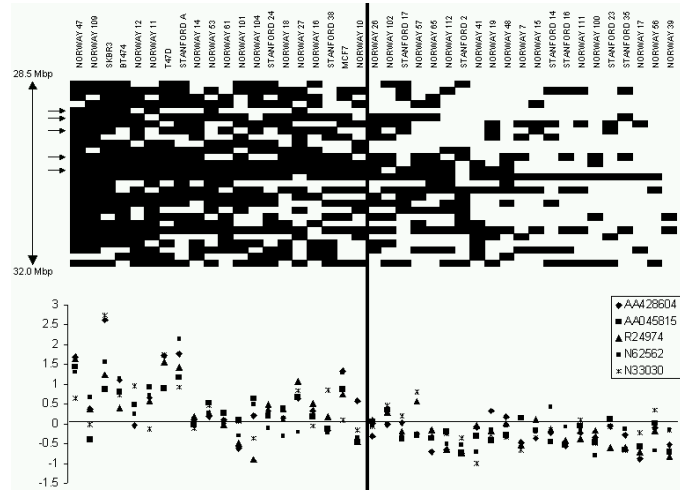


Fig. 4. A significantly amplified GCSM identified in 17q11 in breast tumor data [14] (score of 60.3). Each column represents a sample. Samples are partitioned into two subsets (S' and $S - S'$) by a vertical line. Top panel depicts the DCN submatrix ($C(M)$) for the respective genomic segment (28.5–32.0 Mb), where dark entries represent positive values. Lower panel depicts GE profiles of 5 resident genes (respective rows in $E(M)$) that are significantly differentially expressed between S' and $S - S'$. The positions of these genes in the genomic segment are indicated by arrows in the top panel.

7. E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringner, G. Sauter, O. Monni, A. Elkahoulou, O.P. Kallioniemi, and A. Kallioniemi. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, 62:6240–5, 2002.
8. O.P. Kallioniemi, A. Kallioniemi, D. Sudar, D. Rutovitz, J. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. *Semin Cancer Biol*, 4(1):41–46, 1993.
9. S.C. Linn et al.. Gene expression patterns and gene copy number changes in DFSP. *Amer J of Pathology*, 163(6):2383–2395, 2003.
10. F. Mertens, B. Johansson, M. Hoglund, and F. Mitelman. Chromosomal imbalance maps of malignant solid tumors: a cytogenetic survey of 3185 neoplasms. *Cancer Research*, 57(13):2765–80, 1997.
11. D. Pinkel et al.. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Gen*, 20(2):207–211, 1998.
12. P. Platzer et al.. Silence of chromosomal amplifications in colon cancer. *Cancer Research*, 62(4):1134–8, 2002.
13. J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, and P.O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1):41–6, 1999.
14. J.R. Pollack et al.. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*, 99(20):12963–8, 2002.