10th International IFAC Symposium on
Computer Applications in Biotechnology
Preprints Vol.1, June 4-6, 2007, Cancún, Mexico

A COMPUTATIONAL PROCEDURE FOR THE INTEGRATIVE ANALYSIS OF GENOMIC DATA AT THE SINGLE SAMPLE LEVEL

M. Zampieri[1], R. Spinelli[2], I. Cifola[2], C. Peano[2], D. Basso[1], F. Rocco[3], S. Ferrero[4], E. Fasoli[4], P. Mocarelli[5], C. Battaglia[2], S. Bicciato[1]

[1]Dept. of Chemical Engineering Processes, University of Padova
via Marzolo, 9 – Padova Italy
[2] Interdepartmental Center for Bio-molecular Studies and Industrial Applications, University of Milano
via Fantoli 16/15 – Milano Italy
[3] Urology Unit, University of Milan
via Commenda 15, – Milano Italy
[4]Anatomy Pathology Unit, San Paolo Hospital, University of Milan
via Rudini 8 – Milano Italy
[5]Dept. of Experimental, Environmental Medicine and Medical Biotechnologies, University of Milano-Bicocca, Monza, Italy

Abstract: The integrative analysis of DNA copy number levels and transcriptional profiles, in context of the physical location of genes in a genome, still represents a challenge in the bioinformatics arena. A computational framework based on locally adaptive statistical procedures (Locally Adaptive Statistical Procedure, LAP and Global Smoothing Copy Number, GLSCN) for the identification of imbalanced chromosomal regions in single samples is described. The application of LAP and GLSCN to the integrative analysis of clear cell renal carcinoma patients allowed identifying chromosomal regions that are directly involved in known and novel chromosomal aberrations characteristic of tumors.

Keywords: Genomic data, copy number, gene expression, microarray, kernel smoothing methods, statistical tests

1. INTRODUCTION

Disclosures from the genome sequencing projects are inducing molecular biologists to adopt a novel, systemic approach termed systems biology. Systems biology elevates the study of biological systems from the single entity level to higher hierarchies, such as entire genomic regions, groups of co-expressed genes, functional modules, and networks of interactions. In this context, high-throughput genomic data represents a fundamental discovery tool to understand and reconstruct biological mechanisms and regulatory networks. The massive and rapid accumulation of structural and functional information has required the development of computational frameworks able to turn genomic data into accurate and robust biological hypotheses about the genetic and epigenetic mechanisms regulating the transcriptional machinery (Beer and Tavazoie, 2004). Moreover, recent studies on the relationships between gene structure and gene function in eukaryotic genomes showed how groups of physically contiguous genes are characterized by similar, coordinated transcriptional profiles (Caron *et al*., 2003; Versteeg *et al*., 2004) and suggested a relationship between genomic structural abnormalities and expression imbalances (under- or over-expression). In particular, Caron et al (2001) illustrated how whole chromosome views reveal a higher order organization of the genome, as there is a strong clustering of expressed genes with most chromosomes presenting large regions of highly transcribed genes, called RIDGEs (regions of increased gene expression), interspersed with regions where gene expression is low. Moreover, the pioneering study by Garraway and colleagues (Garraway *et al*., 2005) illustrated how the combination of pre-existing gene expression profiles with genome-wide copy number (CN) data can lead to the identification of novel lineage-specific

oncogenes associated with copy number gain in tumor specimens. Only recently, however, have single studies reported the simultaneous generation of genome-wide maps of copy number alterations (CNAs) and transcriptional activity to study the global effects of chromosomal instability on gene expression (Tsafrir *et al*., 2006; Kotliarov *et al*., 2006). Indeed, genomic instability in human samples can be monitored using microarray-based techniques, in particular high-density single nucleotide polymorphism (SNP)-mapping arrays (Bignell *et al*., 2004). These oligonucleotide arrays permit the simultaneous genotyping of more than 100,000 SNPs and thus provide information on loss of heterozygosity (LOH) and chromosomal alterations with a detection limit reaching 20 kb. More importantly, when copy number profiles of chromosomal instability are confronted with transcriptional data in various tumor samples, a clear impact of DNA copy number change on gene expression can be observed.

Given these experimental evidences, the integration of high-throughput genomic and transcriptional data with gene structural information (i.e., chromosomal localization) represents a major challenge for bioinformatics and computational biology. Indeed, an integrated approach would allow deciphering how the structural organization of genomes influences its functional utilization, identifying how transcription factors regulate gene expression through target genes, and discover novel cancer biomarkers.

Several computational approaches have been adopted to identify chromosomal regions of increased or decreased expression from transcriptional data (Beer and Tavazoie, 2004; Bignell *et al*., 2004; Crawley and Furge, 2002; Toedling *et al*., 2004; Levin *et al*., 2005; Callegaro *et al*., 2006) and to quantify the degree of cooperativity between the number of copies of a gene and its expression level (Garraway *et al*., 2005; Tsafrir *et al*., 2006; Kotliarov *et al*., 2006; Cifola *et al*., 2006). All these methods order the data according to the chromosomal location, smooth transcriptional and copy number signal (e.g., using windows of fixed length or containing a pre-selected number of genes or with a variable bandwidth) and finally calculate correlation coefficients and significance between CN and mRNA expression. In particular, Callegaro *et al*. (2006) and Cifola *et al*. (2006) presented an integrated framework based on two non-parametric model-free bioinformatics tools to identify genomic regions characterized by concomitant alterations in copy number and in regional transcriptional activity. Both Locally Adaptive Statistical Procedure (LAP, Callegaro *et al*., 2006) and Global Smoothing Copy Number (GLSCN) account for variations in gene distance and density and are based on the computation of a standard statistic as a measure of the difference in genomic and gene expression patterns between groups of samples. Once calculated, the statistics are sorted, on each chromosome, according to the chromosomal coordinate (in base pairs) of the corresponding gene. For each chromosome, the statistic is locally smoothed using non-parametric estimation of regression function over the positional coordinate. Chromosomal regions with CN alterations and transcriptional imbalances are identified using a permutation procedure. In particular, gene positions are randomly shuffled and the randomly generated statistics are smoothed to generate the null smoothed distribution. This empirical null distribution is finally used to estimate the q-value measure of significance.

Although effective, LAP and GLSCN are limited to the differential analysis of populations of samples, thus precluding the identification of genomic anomalies affecting single patients. Thus, the purpose of this work is to present a computational framework based on LAP and GLSCN for the identification of genomic regions characterized by concomitant alterations in copy number (CN) and in regional transcriptional activity in single tumor samples.

2. METHODS

In its original version, LAP calculates a statistic for ranking probes in order of strength of the evidence for differential expression in two or more populations. Specifically, given a matrix **X** of normalized expression levels x_{ij} for gene i in sample j ($i = 1, 2, ..., G$; $j = 1, 2, ..., n$) and **Y** a response vector y_j ($j = 1, 2, ..., n$) for n samples, the statistic d_i can be defined as the ratio of change in gene expression r_i to the standard deviation in the data set s_i for each probe set i:

$$d_i = \frac{r_i}{s_i + s_0} \tag{1}$$

where the quantities r_i and s_i assume different formulations in different experimental designs (e.g., two- and multi-class problems, paired data, quantitative responses, time course experiments, survival analyses) and the estimates of gene-specific variance over repeated measurements are stabilized by a fudge factor s_0 (see Tusher *et al*., 2001 and SAM technical manual for details).

Considering the analysis of a single patient j from a population of m tumor samples with normalized expression level x_i^j for gene i and a populations of n normal specimens with average gene expression $\bar{x}_i^{norm}$, the statistic d_i is defined as:

$$d_i^j = \frac{x_i^j - \bar{x}_i^{norm}}{s_i + s_0} \tag{2}$$

where standard deviation s_i for each probe set i is estimated using all tumor and normal samples:

$$s_i = \left\{ a \left[\sum_{j=1}^{m} \left(x_i^j - \bar{x}_i^{tum} \right)^2 + \sum_{k=1}^{n} \left(x_i^k - \bar{x}_i^{norm} \right)^2 \right] \right\}^{1/2} \tag{3}$$

$$a = \frac{m+n}{m \cdot n} \cdot \frac{1}{m+n-2}$$

Similarly to LAP, Global Smoothing Copy Number (GLSCN) analyzes differential copy number values for individual mapping arrays probes (SNPs) in two populations. For the analysis at the single patient level, the statistic of GLSCN has been modify to subject CN data of a tumor sample *j* to a hypothesis test, in which the null and alternative hypotheses are formulated respectively as:

$$H0: CN_i^j = median\left(CN_{allSNPs}^j\right)$$
$$H1: CN_i^j \neq median\left(CN_{allSNPs}^j\right) \quad (4)$$

where CN_i^j is the copy number value of SNP *i* in sample *j*, the median $CN_{allSNPs}^j$ is calculated over all SNP probes in sample *j*, and variance is assumed to be constant.

Once calculated, expression and CN statistics are converted into smoothed scores using a kernel regression estimator with fixed or automatically adapted local plug-in bandwidth. As described in (Toedling J *et al*., 2005; Callegaro *et al.*, 2006; Cifola *et al*., 2006) smoothing of the statistic can be formally stated as a non-parametric regression problem where the score is to be estimated over the chromosomal coordinate. Non-parametric regression problems can be approached using various methods, as kernel smoothing, orthogonal series, spline functions or wavelets. A critical issue in selecting the regression strategy is represented by the procedure for adapting the smoothing parameters. Indeed, the smoothing parameters, e.g the bandwidth, can be adapted globally or locally (Herrmann, 1997). Both LAP and GLSCN use the *lokern* function adapted from the Gasser-Müller type estimator (Herrmann, 1997; Gasser and Müller, 1979) for smoothing the statistic scores. However, in the case of the gene expression data, given the heterogeneous distances and densities of RNA probes on the chromosomes, the optimal bandwidth is estimated iteratively minimizing the asymptotic mean squared error. Instead, in the case of CNAs, a fixed bandwidth (e.g., of 1 Mb) can be chosen in consideration of the relatively homogeneous distribution of SNP probes along the chromosomes.

Finally, chromosomal regions with transcriptional imbalances and smoothed CN scores significantly different from the median CN value are identified using a permutation procedure under the assumption that each gene has a unique neighborhood and that the corresponding smoothed statistic is not comparable with any statistic smoothed in other regions of the genome. The G statistic scores are first randomly assigned to G chromosomal locations through permutations and then, for each permutation, smoothed over the chromosomal coordinate. Thus, observed and null statistics are smoothed and compared exactly over the same region, taking into account variations in the gene distances and in gene density. The permutation process, over B random assignments, allows defining the null smoothed statistic for gene/SNP *i*. The significance of the differentially expressed genes, i.e., the p-value p_i for gene/SNP *i*, is computed as the probability that the random null statistic exceeds the observed statistic over B permutations. This p-value has the peculiarity to be local since the observed smoothed statistic is compared only with null statistics smoothed on the same neighborhood of chromosomal position *i*. Indeed, during the permutation process, the chromosomal position is conserved while the statistics are randomly shuffled. Once the distribution of empirical p-values has been generated, the q-value is used to identify differentially expressed chromosomal regions. Q-values allow quantifying significance in light of thousands of simultaneous tests.

3. RESULTS

In the context of a research project focused on the identification of clinical biomarkers for renal cell carcinoma (RCC), the single sample versions of LAP and GLSCN have been applied to the analysis of 6 paired normal/tumor samples of human clear cell renal carcinoma (ccRCC). Using Affymetrix high-density oligonucleotide microarray technology, a transcriptional profiling (on GeneChip Human Genome U133 Plus 2.0 arrays) and a genome-wide SNP-mapping of CNAs (on GeneChip Human Mapping 100K SNP arrays) were performed. Raw signal intensities were converted to expression values using the robust multi-array average (RMA) procedure and to CN values using Affymterix Copy Number Analysis Tool (CNAT, v3.0). To assign probe sets to genes, the 47,401 HG-U133 Plus 2.0 probe sets were annotated to obtain Entrez Gene IDs and chromosomal positions, using the *annotate* package of Bioconductor for the R environment (http://www.bioconductor.org). This re-annotation step, in addition to the filtering out of probe sets without a unique chromosomal position and those referring to the X and Y chromosomes, resulted in the selection of 16,473 unique gene IDs for further studies. CN data for SNPs without a unique chromosomal position as well as for SNPs on the X chromosome were filtered out (the 100K arrays do not contain SNPs on Y chromosome) and the resulting dataset comprised a total of 112,990 SNPs. Re-annotated gene expression and copy number data of any single patient were analyzed with the single sample versions of LAP and GLSCN, respectively. Specifically, the 16,473 expression values have been converted into statistic d_i using Eq. (2) and (3). The statistic scores have been further smoothed over the chromosomal coordinates, using the lokern function. The smoothed scores have been then randomly assigned to 16,473 gene loci over 100,000 permutations and smoothed over the chromosomal coordinate. Similarly, the 112,990 CN values have been converted into statistic scores using the hypotheses of Eq. (4) and the statistics smoothed over the chromosomal coordinates, using the lokern function. The smoothed scores have been then randomly assigned to 112,990 chromosomal positions over 100,000 permutations and smoothed over the chromosomal coordinate.

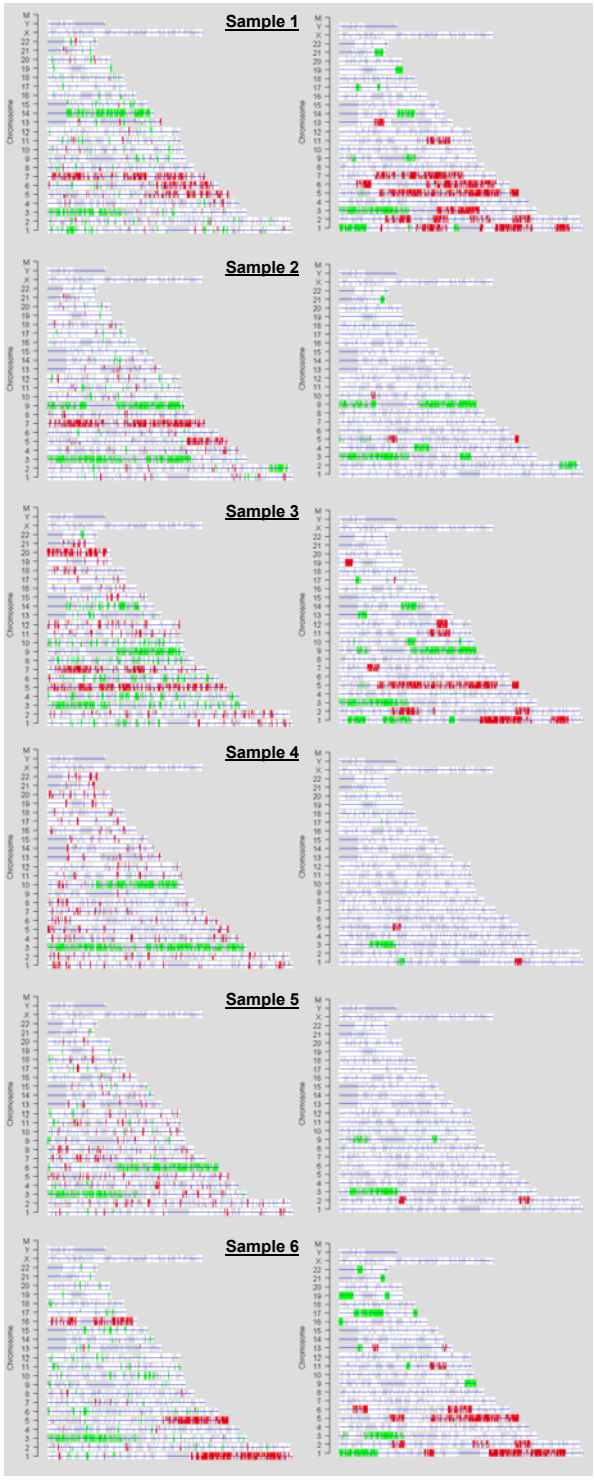

Fig. 1. Single sample whole genome plots of the chromosomal regions with CNAs, (left panel), and gene expression imbalances, (right panel), at a q-value=0 a q-value<0.05 for CN and gene expression data, respectively. The white bars indicate locations and orientations of all probe sets in the microarray, the red perpendicular lines represent the exact chromosomal locations and orientations of genes with CN gain or up-regulated, and the green lines the location of probes with CN loss or down-regulated.

Finally, to integrate gene expression and CN data, p-values have been computed as the probability that the random null statistic exceeded the observed statistic over the permutations for the 16,473 unique gene loci used in the LAP analysis. Once the distribution of empirical p-values had been generated, q-value was used to identify chromosomal regions affected by transcriptional imbalances and CNAs. The single sample analysis with LAP and GLSCN generated the high-resolution genomic maps of Figure 1 (setting q<0.05 and q=0 for transcriptional activity and CNAa analysis, respectively).

To evaluate the relationship between CNA and transcriptional activity, the relative statistic scores for 16,473 well annotated chromosomal positions (genes) were categorized into three classes: *increased* (gain of CN or up-regulated), *unchanged* and *decreased* (loss of CN or down-regulated). At q=0 for CNA analysis and q<0.05 for transcriptional analysis, the concordance of categorization in the various samples is reported in Table 1.

Table 1 Percentage concordance between number of genes with transcriptional and CN alterations in *increased*, *unchanged* and *decreased* categories

Sample #	Concordance %		
	Increased	*Unchanged*	*Decreased*
1	56	79	36
2	1	97	67
3	31	82	44
4	0.5	98	16
5	0.6	98	27
6	63	81	25

4. Discussion and Conclusions

A novel mathematical and statistical framework to combine microarray profiles of transcriptional activity and copy number alterations (CNAs) at genome level has been developed and applied to study single tumor samples. The integrative analysis of genomic and transcriptional data using locally adaptive single-sample statistical procedures allowed identifying sample-specific, as well as global, associations between DNA copy number changes and regional gene expression levels. In particular, a novel finding is the concomitant loss of sequences of the short arm of chromosome 3 and of the long arm of chromosome 9 and the gain of chromosome 5. These abnormalities in copy number are highly correlated with the down-regulation of the transcriptional activity of genes located in chromosomes 3 and 9 and with the up-regulation of transcripts from chromosome 5.

To our knowledge, this is the first computational platform able to directly combine, on a single sample base, SNP-based CN data and transcriptional profiles at the level of gene loci for 16,473 unique genes using Affymetrix microarrays. The identified chromosomal areas, presenting concomitant alterations in genomic and transcriptomic profiles, could be tumor-specific regions containing candidate clinical biomarkers or patient-specific abnormalities to be related to the disease etiology or outcome.

ACKNOWLEDGEMENT

This work was supported by grants from the Italian Ministry of University and Research (MIUR-FIRB RBNE01HCKF1 and RBNE01TZZ8, COFIN 2005069853, and ONCOSUISSE Collaborative Cancer Research Project OCS 01517–02–2004).

REFERENCES

Beer MA, Tavazoie S. (2004). Predicting gene expression from sequence. *Cell*, 117, 185-98

Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR. (2004). High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res*, 14(2), 287-295

Callegaro A, Basso D, Bicciato S. (2006). A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions. *Bioinformatics*, 22(21), 2658-66

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science,* 291, 1289-92

Cifola I, Zampieri M, Peano C, Spinelli R, Beltrame L, Basso D, Fasoli E, Ferrero S, Bosari S, Bicciato S, Battaglia C. (2006). Correlation between copy number alterations and transcriptional activity using high-density microarray technology in a metastatic renal carcinoma cell line. *BMC Genomics*, submitted

Crawley JJ, Furge KA. (2002). Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. *Genome Biology*, 3, RESEARCH0075

Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, Beroukhim R, Milner DA, Granter SR, Du J. (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, 436(7047), 117-122

Gasser T, Müller HG. (1979). Kernel estimation of regression functions. Smoothing Techniques for Curve Estimation. Lecture Notes in Math. 757, 23-68. Springer, New

Herrmann E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Graphical and Computational Statistics*, 6, 35-54

Kotliarov Y, Steed ME, Christopher N, Walling J, Su Q, Center A, Heiss J, Rosenblum M, Mikkelsen T, Zenklusen JC. (2006). High-resolution Global Genomic Survey of 178 Gliomas Reveals Novel Regions of Copy Number Alteration and Allelic Imbalances. *Cancer Res*, 66(19), 9428-9436

Levin AM, Ghosh D, Cho KR, Kardia SL. (2005). A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics*, 21(12), 2867-74

Storey JD, Tibshirani R. (2003). Statistical significance for genome-wide experiments. *Proc Natl Acad Sci U S A*, 100(16), 9440–9445

Toedling J, Schmeier S, Heinigm M, Georgi B, Roepcke S. (2005). MACAT—microarray chromosome analysis tool. *Bioinformatics*, 21(9), 2112-3

Tsafrir D, Bacolod M, Selvanayagam Z, Tsafrir I, Shia J, Zeng Z, Liu H, Krier C, Stengel RF, Barany F. (2006). Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res*, 66(4), 2129-2137

Tusher VG, Tibshirani R, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9), 5116-5121

Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH. (2003). The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res*, 13, 1998-2004