Exploratory Analysis of Cancer SAGE Data

Ricardo Martinez[1], Richard Christen[2], Claude Pasquier[2], Nicolas Pasquier[1]

[1] Laboratoire I3S (CNRS UMR-6070), Université de Nice – Sophia-Antipolis, 2000 route des lucioles, Les Algorithmes, 06903 Sophia-Antipolis, France; {rmartine,pasquier}@i3s.unice.fr

[2] Laboratoire de Biologie Virtuelle (CNRS UMR-6543), Université de Nice – Sophia-Antipolis, Centre de Biochimie, Parc Valrose, 06108 Nice cedex 2, France; {christen,claude.pasquier}@unice.fr

Abstract. Using several analyse techniques for the hierarchical clustering of a SAGE expression dataset of 822 tags from 74 tissue samples (normal and cancer) we show that cleaning the dataset (tags and experiments) is critical and that attribution of a tag to a gene is not easy. Comparison of cancers from various tissues is a difficult task as tissue samples cluster according to tissue origin and not as cancer or normal.

1 Introduction

The SAGE method is based on the sequencing of concatemers of short (14 basepairs; recently 17 bp) sequence tags that originate from the 3'-nearest cutting site of a restriction enzyme) to estimate transcripts abundance [VZVK95], to estimate the expression level of eukaryotic transcripts without prior knowledge of their sequences and is more sensitive than the EST method [SZL+04], but requires knowledge of the complete genome. The advantage of the SAGE method is to perform a random sampling of transcripts in a particular tissue, with little sequencing effort.
The dataset proposed for analysis comprises several difficulties:

1. PCR and sequencing may produce a number of errors. A single error may lead to non recognition of a transcript or wrong attribution. Some tags may be present in more than one gene. Finally, since restriction enzymes may not cut with 100 % efficiency, some tags may be wrong.
2. Tissue samples originate from two different sources (i.e. bulk or cell line) that may influence gene expression. Cancerous tissue are usually provided after surgery, a "cancer" sample may contain more healthy tissue than cancer, leading to a "wrong" identification.
3. Analyzes using DNA chips concluded that cancer cells are more alike normal cells of the same tissue than cancer cells from a different tissue: there are many more tissue-specific genes than genes involved in cancers [RSE+00,SRW+00]. Thus, trying to classify in two classes, normal versus cancer, in order to identify specific tags can be difficult. Also, cancers may have different origins (deregulation of oncogenes versus breakdowns of chromosomes for example) searching for two classes only may be problematic.
4. Interpretations. Even after removal of tags that do not show any significant change among samples, many tags remain to be classified. One may then use tools such as THEA [PGJC04] to automatically annotate clusters or nodes from a classification tree with statistically significant information extracted from for example GeneOntology, if each tag is linked to a gene.

Tissue	Cancer bulk	Cancer cell line	Normal bulk	Normal cell line	Total
Brain	8	7	5	1	21
Breast	6	3	2	0	11
Colon	2	4	2	0	8
Kidney	0	2	0	0	2
Ovary	3	4	0	2	9
Pancreas	0	3	2	2	7
Prostate	3	6	2	0	11
Peritoneum	0	0	1	0	1
Skin	1	0	0	1	2
Vessel	0	0	0	2	2
Total	23	29	14	8	74

Table 1. Repartition of conditions by cell state and source: cancer (C), normal (N), bulk (Bu), cell line (Ce).

The main goal of our analysis was to investigate the influence of cleaning the dataset. We propose to validate removal of spurious tags or experiments and therefore increase the signal. In an exploratory analysis we used the small dataset. This paper focuses on the following steps: i) Pruning of non-significant tags; ii) data normalization; iii) selection of differentially expressed genes; iv) deletion of outlier biological conditions; v) classification of biological conditions.

2 Tags selection

Tags are often annotated based on the SAGE Genie principles [BOG+02] and linked to a series of expression data (often EST sequences), a step that is difficult to automate. It is often difficult to understand and appreciate the methods used for tag attribution, we therefore developed specific tools. First, every human ENST sequence was downloaded from Ensembl. Tags present in transcripts of a single gene were labelled as good (436) attributed corresponding ENSG numbers[3]. Tags present in transcripts originating from several genes were labeled as bad (219) and removed from further analysis.
Next, all EMBL human sequences (including ESTs) were downloaded to search non attributed tags (167). Every sequence recognized was blasted for ENSG attribution. This step led to a further 80 tags attributed to a ENSG number. Reasons for tag non attribution are likely to be: i) location in a region not yet identified as a gene; ii) location in the mitochondrial genome (very few protein coding genes), which was not taken into account; iii) tag resulting from the partial digestion of a transcript, and therefore not located in the 3-most domain.
At this point we had clearly less tags linked to genes than if we had used a tool such as SAGE Genie or other tools. But the first tag of the list was linked to a mitochondrial sequence by SAGE Genie, while at the Global Gene Expression Group project it mapped to Unigene Hs.476965 (G1/S transition control protein-binding protein IEF-8502)[4]. The SAGE Genie linked this tag to a sequence of accession number BE874599. Blast of this sequence provided a hit on the mitochondrial human genome, but at a position that was identified as '16S ribosomal sequence'. Such sequence has no polyA tail of any sort, and does not contain a repeat of A anywhere in the sequence.

[3] `http://www.ensembl.org/`
[4] `http://sciencepark.mdanderson.org/ggeg`

At this step we are rather confident that every data resulting from large scale analysis using web based tools, should be critically assessed either using two different public tools or ad-hoc scripts and databases[5].

3 Algorithms and methods

We used the Significance Analysis of Microarrays (SAM)[6] method to select differentially expressed genes. SAM computes a statistic di for every gene i, measuring the strength of the relationship between gene expression and the response variable (cancer bulk, cancer cell line, normal bulk and normal cell line). The cutoff for significance "Delta" was fixed at 0.21 implying a False Discovery Rate of 5 %.
It is critical to take into account condition variations and in particular *outliers* that introduce noise in the classification [LMV04]. Thus, we developed a methodology for finding outliers using Principal Component Analysis (PCA) and hierarchical clustering methods:

1. Using PCA as an exploratory tool to determine the optimal number of clusters.
2. Applying hierarchical clustering algorithms to identify outliers and remove them.
3. Applying again PCA analysis to verify that variability level is not decreased when each of these conditions is removed.
4. Cluster to verify that the clustering was improved.

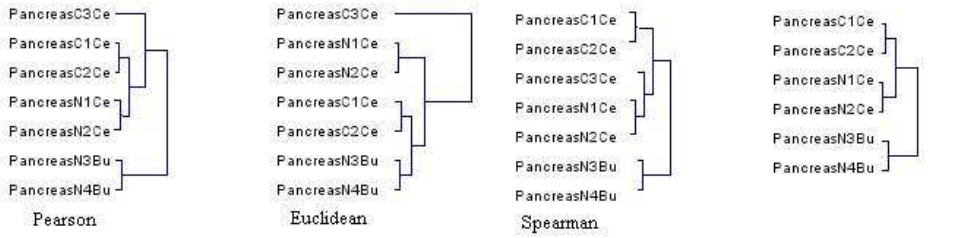
We tested 5 algorithms (K Means, Fanny, Partial Least Squares, Unweighted Pair Groups Method Average (UPGMA) and DIvisive ANAlysis (DIANA)) and 5 measures of distance (Euclidean, Pearson, Manhattan, Spearman and Tau) according to 3 different consistency measures (average proportion of non-overlap, average distance between clusters and average distance between cluster means) [DD03]. We selected UPGMA and DIANA algorithms and Pearson, Euclidean and Spearman distances that are the most efficient with this dataset.

4 Experimental results

4.1 Biological condition selection

The 7 pancreas conditions are distributed in 3 classes: cancer cell line (C1Ce, C2Ce and C3Ce), normal cell line (N1Ce and N2Ce) and normal bulk (N3Bu and N4Bu), as is shown by the first 3 PCA components that explain 98.59 % of the total variance. The hierarchical trees obtained for the different distance measures are shown in figure 1(a). Trees obtained with the UPGMA and the DIANA algorithms are similar.
When using the Pearson and Euclidean distance measures, condition PancreasC3Ce is placed in an isolated cluster, and when the Spearman measure is used it is associated with normal conditions. Removing this condition, the first 3 components explain 99.03 % of the total variance and the result of clustering is shown in figure 1(b).
Using a similar process for other tissues, the 16 outlier conditions obtained are listed in table 2. These results confirm the natural division of conditions in three classes corresponding to the first 3 components of PCA analysis. Furthermore, in all experiments we can see that the variance explained by the first 3 components is always improved, up to 4.31 % for Ovary conditions, when outlier conditions are removed.

[5] `http://www.ncbi.nlm.nih.gov/`
[6] `http://otl.stanford.edu/industry/resources/sam.html`



(a) All conditions (b) Without outliers

Fig. 1. Hierarchical clustering of the pancreas conditions.

Organ/Tissue	PCA (first 3 components)	Outliers	PCA without Outliers (first 3 components)
Brain	98.46 %	{N4Ce,C1Bu,C14Bu, C5Bu, C9Ce}	99.02 %
Breast	95.57 %	{C6Bu}	97.38 %
Colon	98.56 %	{}	98.56 %
Ovary	93.60 %	{N1Ce, N2Ce, C4Ce, C6Bu}	97.91 %
Prostate	98.02 %	{N1Bu,C7Bu,C9Bu, C8Ce, C1Ce}	98.70 %
Pancreas	98.59 %	{C3Ce}	99.03 %

Table 2. PCA analysis of conditions by tissues.

For each tissue, we systematically observe three classes: cancer, bulk and cell line, with bulk and cell line clearly separated (see figure 1(b)). This observation therefore confirms previous analyzes that showed cell source to be of crucial influence on gene expression.

4.2 Hierarchical clustering of biological conditions

We applied PCA analysis and found that the first 6 components explain 98.22 % of the variance corresponding to the 6 tissue clusters. Comparing these results with PCA analysis on the initial dataset showed that gene and condition selections have eliminated data noise.

Then, we applied the UPGMA and DIANA algorithms to the cleaned dataset and the tree obtained by consensus for both algorithms, and for the Pearson and the Spearman distances, is shown in figure 2. For the Euclidean distance, the distribution is similar but branches to the leaves are longer.

Comparing clustering trees obtained with the initial dataset (not shown) and figure 2 clearly showed that the selection process improved data quality since length of terminal branches were considerably reduced. We can observe a first degree classification by tissue that is accurate for Pancreas, Brain, Breast, Colon and Prostate tissues, but mixes Ovary tissue conditions with other tissue conditions. We can also see a clear second degree classification, among conditions of the same tissue, by cell source: bulk and cell line. Among Pancreas, Breast, Brain and Colon condition clusters, we can observe a third degree classification by state: cancer and normal.

In conclusion, clustering clearly separates cell sources, corroborating previous results on SAGE and DNA chips data [NSS01,RSE+00]. We can conclude that there are important differences between bulk and cell line conditions that should not be ignored. We believe that when conducting studies for finding "interesting gene cancer knowledge" involving multiple tissues SAGE libraries, the study must be first oriented toward a decomposition of the conditions by tissues and then by cell sources to finally focus the analysis on cell states.

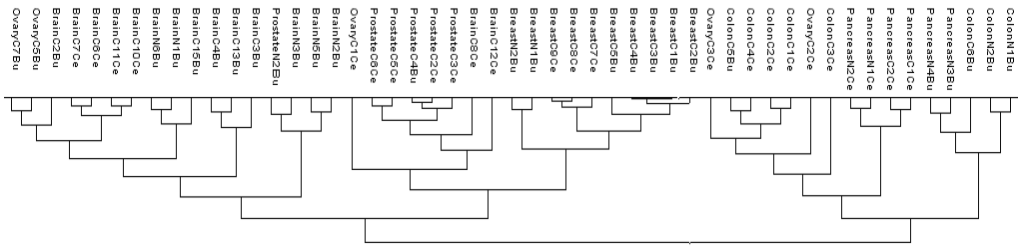


Fig. 2. Hierarchical clustering of conditions.

Eventually, we applied the C5.0 unsupervised classification method to produce classification rules of biological conditions by tissue, cell state and cell type. Three different class attributes characterizing each condition were created: tissue type (Pancreas, Ovary, Brain, Prostate and Breast), cell source (bulk or cell line) and cell state (cancer and normal). Boosting and cross validation options were activated. The numbers of rules with maximal accuracy generated for each class decomposition of conditions are shown in table 3.

Class	Number of rules	Max accuracy
Bulk	5	100 %
Cell line	5	100 %
Cancer	1	80 %
Normal	3	80 %
All 6 tissues types	1	60 %

Table 3. Rules by class and their maximal accuracy.

Using the cell source classification, 5 exacts rules, i.e. with perfect accuracy, were generated. For the cell state classification, only 1 and 3 rules respectively, all with with only 80 % of accuracy, were generated. Considering tissue classification, only 1 rule with 60 % accuracy was generated. This result is logical since there are 6 different tissues, thus disturbing the classification, and cells from different tissues but originating from cell lines tend to become more similar from the tag expression levels viewpoint. These results confirm that in the small cleaned dataset, there is an intrinsic division of conditions by cell source that is more natural than by cell state.

5 Conclusion

Most SAGE studies made use tags of 14 bp. However, a recent study showed the clear advantage of using a tag of 15 bp [DBB+05]. Even longer tags will be better. Recently, the SAGE protocol was enhanced with a new tagging enzyme (MmeI), which produces 21-22 bases tags [SSR+02], allowing direct mapping to the transcripts [VC04]. When numerous tags are available removing tags present only once, that may result from errors, is possible. Sequence errors have little effect on the quantification of moderately expressed genes but not for rare transcripts. About 6.7 % of Long SAGE ditags will have acquired mutations prior to ligation, cloning and sequencing [VC04], arguing for a robust tag attribution to a transcript.
Only reliably annotated tags can be included in the final analysis [SSL+04]. Annotation of SAGE tags to genes and their corresponding Unigene cluster numbers revealed that on average only 30 % of all tags (including less abundant tags) could be reliably annotated based on the SAGE Genie principles [BOG+02]. Annotation improved to about 70 %

for tags with intermediate to abundant expression levels. Remaining tags either could not reliably be associated with a gene (e.g. annotated to unclustered ESTs) or were not present in a single gene.
In conclusion, algorithms used to analyze SAGE data have a strong influence on results [DBB+05] and using a single computer program and a single source of sequence data (annotations) would result in a weaker analysis. We have also shown incoherence of results between different public web tools, and an obvious error of gene attribution for the first tag at least. Removing outlier experiments also decreases noise and increases reliability of clustering. Finally, we saw that searching for classification rules identifying normal and cancerous tissues among tissues of different origins is difficult as rules of maximal accuracy discriminate tissue origins. Using several datasets containing each one numerous samples from the same tissue could improve the results.

References

[BOG+02] Boon K., Osorio E.C., Greenhut S.F., Schaefer C.F., Shoemaker J., Polyak K., Morin P.J., Buetow K.H., Strausberg R.L., De Souza S.J. and Riggins G.J. An anatomy of normal and malignant gene expression. *Natl. Acad. Sci. USA*, 99(17):11287-11292, 2002.

[DD03] Datta Su. and Datta So. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459-466, 2003.

[DBB+05] Dinel S., Bolduc C., Belleau P., Boivin A., Yoshioka M., Calvo E., Piedboeuf B., Snyder E.E., Labrie F. and St-Amand J. Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. *Nucleic Acids Res.*, 33(3):e26, 2005.

[LMV04] Loguinov A.V., Mian I.S. and Vulpe C.D. Exploratory differential gene expression analysis in microarray experiments with no or limited replication. *Gen. Bio.*, 5(3):R18, 2004.

[NSS01] Ng R.T., SanderJ. and Sleumer M.C. Hierarchical cluster analysis of SAGE data for cancer profiling. *Proc. BIOKDD conf.*, pp 65-72, 2001.

[PGJC04] Pasquier C., Girardot F., Jevardat de Fombelle K. and Christen R. THEA: ontology-driven analysis of microarray data *Bioinformatics*, 20(16):2636-2643, 2004.

[RSE+00] Ross D.T., Scherf U., Eisen M.B., Perou C.M., Rees C., Spellman P., Iyer V., Jeffrey S.S., Van de Rijn M., Waltham M., Pergamenschikov A., Lee J.C., Lashkari D., Shalon D., Myers T.G., Weinstein J.N., Botstein D. and Brown P.O. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, 24(3):227-35, 2000.

[SSR+02] Saha S., Sparks A.B., Rago C., Akmaev V., Wang C.J., Vogelstein B., Kinzler K.W. and Velculescu V.E. Using the transcriptome to annotate the genome. *Nat. Biotechnol.*, 20(5):508-512, 2002.

[SRW+00] Scherf U., Ross D.T., Waltham M., Smith L.H., Lee J.K., Tanabe L., Kohn K.W., Reinhold W.C., Myers T.G., Andrews D.T., Scudiero D.A., Eisen M.B., Sausville E.A., Pommier Y., Botstein D., Brown P.O. and Weinstein J.N. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, 24(3):236-44, 2000.

[SSL+04] Shippy R., Sendera T.J., Lockner R., Palaniappan C., Kaysser-Kranich T., Watts G. and Alsobrook J. Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics*, 5(1):61, 2004.

[SZL+04] Sun M., Zhou G., Lee S., Chen J., Shi R.Z. and Wang S.M. SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics*, 5(1):1, 2004.

[VZVK95] Velculescu V.E., Zhang L., Vogelstein B. and Kinzler K.W. Serial Analysis of Gene Expression. *Science*, 270:484-487, 1995.

[VC04] Viatcheslav R.A. and Clarence J.W. Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics*, 20(8):1254-1263, 2004.