Integrating and analyzing big next generation sequencing data of cancer

Emanuel Weitschek*†, Fabio Cumbo†‡, Eleonora Cappelli‡, and Giovanni Felici†

National Research Council of Italy

UNINETTUNO

Motivation: Due to the great advances of Next Generation Sequencing (**NGS**) techniques, bioinformaticians are faced with large amounts of genomic and clinical data, which are growing exponentially. A striking example is The Cancer Genome Atlas (**TCGA**), whose aim is to provide a comprehensive archive of biomedical data about tumors. In this work, we integrate **RNA-seq** and **DNA-methylation** experiments extracted from TCGA and perform a supervised classification analysis, to distinguish successfully the tumoral samples from the normal ones and to extract reliable rule-based classification models that contain salient features (i.e., genes and methylated sites). These features, which are related to the investigated tumor, can be studied by domain experts in order to obtain new knowledge about cancer. Finally, our proposed integration and analysis method can be adopted with success for further studies on different data sources and NGS experiments.

Materials and Methods: I. *RNA-seq and DNA-methylation integration technique:* n samples are collected each one with its m features and their class labels (conditions), e.g., normal - tumoral. Every sample i is represented by the vector $f_i = (f_{i1}, f_{i2}, \cdots, f_{im}, f_{ic})$, where $f_{ij} \in R$, i = 1, … ,n, j = 1, … , m and $f_{ic} \in$ {normal, tumoral}. The data matrix is obtained by the vectors $f_1, f_2, \cdots, f_n$, where the rows represent the samples and the columns the features.
In the **RNA-seq data matrix** (Table I) the features are genes and each cell stores the RSEM (*RNA-seq by Expectation Maximixation)* value of each gene expression measure. For the **DNA-methylation data matrix** (Table II) we consider a **new measure** that refers to the *genes* of the methylated sites and define it as the **sum of the beta values of the methylated sites that are present in the gene**. We integrate the data by joining the matrices (Table III) on common sample IDs an by considering the RNA-seq RSEM value and the new DNA-methylation measure on the genes.
II. *Supervised data analysis:* we apply supervised classification methods, i.e., the tree-based **C4.5**, the rule-based **RIPPER** and **Camur** (Classifier with Alternative and MUltiple Rule-based models) classifiers, which permit to identify the features (genes) that are related to the particular cancer under study.

Table I. EXAMPLE OF THE BREAST CANCER RNA-SEQ DATA MATRIX EXTRACTED FROM TCGA

SampleID	G1rna	G2rna	···	GMrna	Class
S1	1.53	4.31	···	0.27	Tumoral
S2	0.35	5.36	···	0.65	Normal
S3	1.11	11.49	···	0.42	Tumoral
S4	2.02	13.10	···	0.50	Tumoral
···	···	···	···	···	···
SN	2.75	4.04	···	0.48	Tumoral

Table II. EXAMPLE OF THE BREAST CANCER DNA-METHYLATION DATA MATRIX EXTRACTED FROM TCGA

SampleID	G1dmet	G2dmet	···	GMdmet	Class
S1	1.23	3.21	···	6.23	Tumoral
S2	3.54	5.10	···	5.76	Normal
S3	0.93	9.32	···	3.53	Tumoral
S4	4.25	7.34	···	7.61	Tumoral
···	···	···	···	···	···
SN	3.76	6.15	···	5.59	Tumoral

Table III. EXAMPLE OF THE BREAST CANCER INTEGRATED DATA MATRIX

SID	G1rna	G2rna	···	G1dmet	G2dmet	···	Class
S1	1.53	4.31	···	1.23	3.21	···	Tumoral
S2	0.35	5.36	···	3.54	5.10	···	Normal
S3	1.11	11.49	···	0.93	9.32	···	Tumoral
S4	2.02	13.10	···	4.25	7.34	···	Tumoral
···	···	···	···	···	···	···	···
SN	2.75	4.04	···	3.76	6.15	···	Tumoral

Results: We test our integration method and the supervised classification algorithms on RNA-seq and DNA-methylation experimental data extracted from TCGA and related to the Breast (**BRCA**), and the Kidney renal papillary cell carcinoma (**KIRP**) cancers. A compact overview of the data sets is provided in Table IV (where RNA stands for RNA-seq, DNA for DNA-methylation, and INT for the integrated data sets). In Table V, we report the classification performances in terms of accuracy (percentage of correct classified instances) of the two methods tested in 10-fold cross-validation. It is worth noting that all classification tasks are solved with very promising results (accuracy > 97%).

BREAST CANCER:
TMEM220rna<167.58 AND
HTR1Bdmet>1.92
AND SNORA63dmet < 0.45
OR DSTrna < 22424.14

```
FIGF   – 44 MMP11 – 20
SPRY2  – 37 COL10A1 – 14
SCN3A  – 25 DS      – 14
PAMR1  – 20 ARHGAP2 – 12
```

Table V. CLASSIFICATION PERFORMANCES (ACCURACY IN %)

Cancer	Experiment	C4.5	RIPPER
BRCA	RNA	98.5	98.1
BRCA	DNA	97.9	96.1
BRCA	INT	97.3	98.1
KIRP	RNA	99.3	98.9
KIRP	DNA	-	-
KIRP	INT	98.2	98.2

Cancer	exp	samples	tumoral	normal	features	[MB]
BRCA	RNA	1218	1104	114	20486	199
BRCA	DNA	859	769	90	20234	246
BRCA	INT	835	758	77	40718	375
KIRP	RNA	273	241	32	20485	44
KIRP	DNA	10	10	0	20221	4
KIRP	INT	273	241	32	40703	58

Conclusions: In this work, we presented an application of Next Generation Sequencing data integration and supervised analysis. We defined a new gene-wide measure for DNA-methylation that permits the analysis with supervised classification methods and the integration with RNA-seq data.

Contacts:
E. Weitschek
emanuel@iasi.cnr.it
E. Cappelli
eleonora.cappelli@uniroma3.it