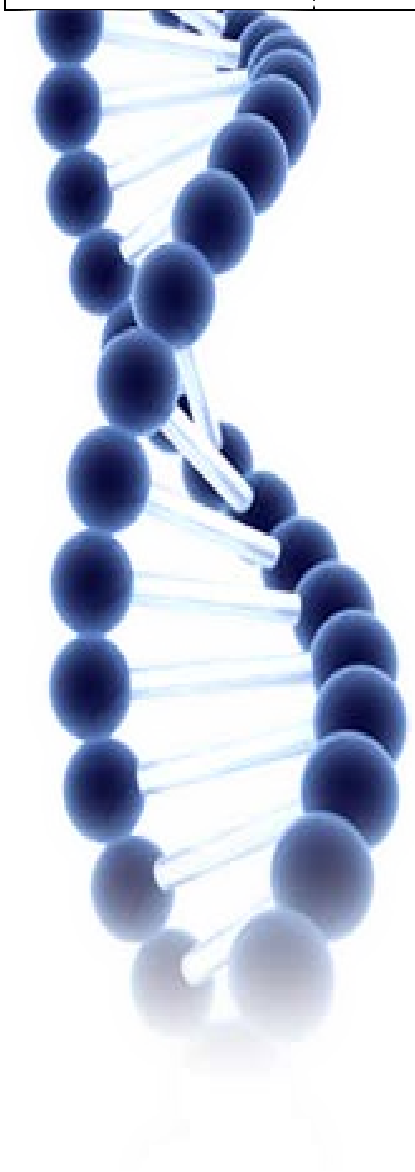


Stomach Cancer Gene Visualization

"Information Visualization"
Final Project

Autori:	*Calcaterra Stefano* *Mione Federico*	Data:	*26/07/2014*
Versione	*0.1*	Data:	*26/07/2014*

Indice

Indice delle figure

1 TCGA

Acronimo di The Cancer Genome Atlas, è un'opera globale e coordinata per accelerare la comprensione del cancro attraverso l'applicazione di tecnologie di analisi del genoma, compreso il sequenziamento su larga scala del genoma umano.

The Cancer Genome Atlas *Understanding genomics to improve cancer care*

Figura 1: Logo TCGA

Ha come obbiettivo principale il miglioramento della nostra capacità di diagnosticare, curare e prevenire il cancro.

Offre una piattaforma gratuita che consente di ricercare e scaricare al link https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp, rispettando la privacy dei pazienti, interi dataset contenenti informazioni cliniche utili per analisi genomiche.

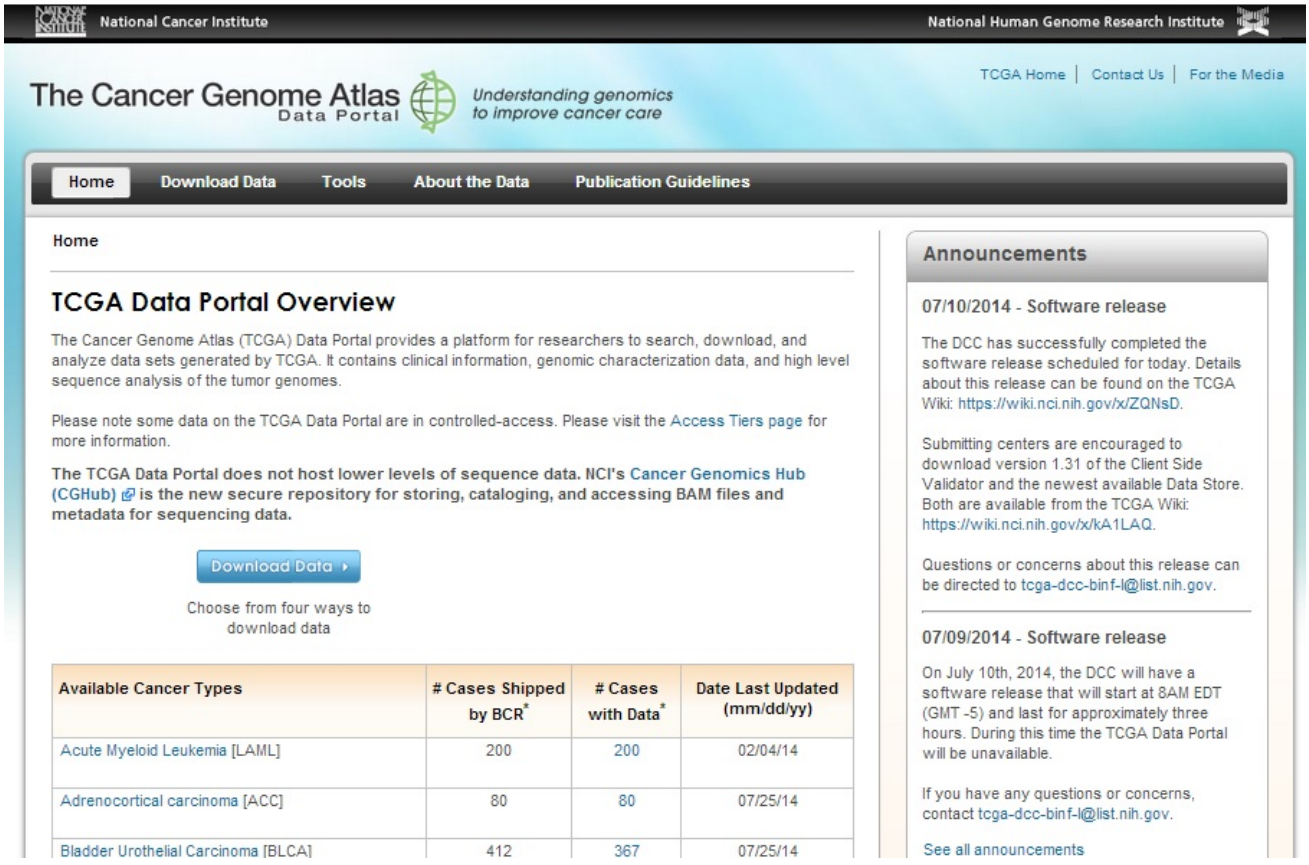


Figura 2: TCGA Data Portal Overview

1.1 Descrizione Dataset

Per lo svolgimento del progetto è stato preso in considerazione il dataset del TCGA relativo all'adenocarcinoma allo stomaco, costituito da circa 271 file di tipo gene.quantification.txt (dimensione totale = 330 MB).

Ogni file esaminato si riferisce ad uno specifico paziente ed è stato generato applicando la tecnica del RNA-Sequencing, una metodologia per l'analisi del trascrittoma (ovvero l'insieme dei trascritti di RNA che caratterizzano un dato stadio di sviluppo di una cellula). Ciascun file contiene quindi, la lista dei geni del paziente (in media ~24000) e per ognuno di essi il valore del RPKM (Reads Per Kilobase per Million mapped reads), oltre ad altre informazioni.

Il valore RPKM rappresenta una misura dell'espressione genica ricavata con metodi di normalizzazione applicati nella tecnica di sequenziamento del RNA.

Figura 3: a sinistra viene mostrata una porzione del dataset STAD contenente file gene.quantification.txt. A destra un file gene.quantification.txt con l'elenco dei geni e i valori di RPKM ad essi associati.

Tutti i nomi dei file presenti nei dataset si uniformano ad un particolare standard (barcode); è stato quindi possibile estrarre da essi il codice identificativo del paziente e la codifica della classe di appartenenza.

La figura che segue mostra il barcode per la nomenclatura dei file.

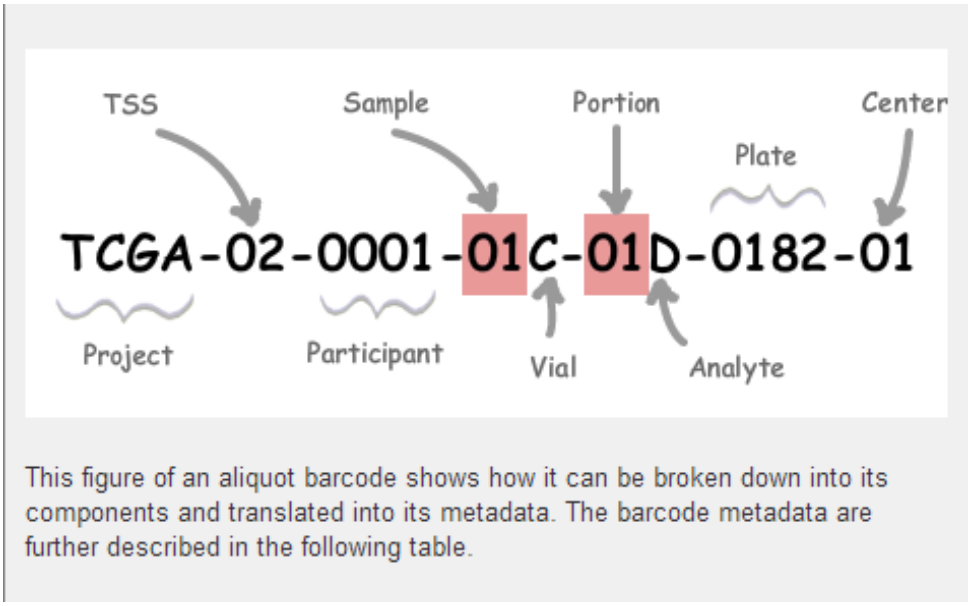


Figura 4: barcode

Come si nota, il barcode è una composizione di una collezione di identificatori. Ai fini del progetto sono stati presi in considerazione:

- Il Participant, identificativo del paziente
- Il Sample, identificativo del tipo di campione esaminato.
  Può assumere diversi valori:
  - da 01 a 09, identificano il tipo Tumoral
  - da 10 a 19, identificano il tipo Normal
  - da 20 a 29, identificano il tipo Control
- Il Vial, identificativo dell'ordine del campione in una sequenza di campioni.
  Assume valori compresi tra A e Z.

2 Estrazione dei dati

Per estrarre le informazioni di interesse si è proceduto alla creazione di un Job Map-Reduce utilizzando il framework Apache Hadoop.

2.1 Fase di Map

Nella **fase di Map**, a partire dal nome del file di ogni paziente, vengono estratti il Participant, il Sample e il Vial che, insieme, andranno a costituire le "key" del mapper.

Per ogni file vengono inoltre estratti i nomi dei geni e il valore del RPKM ad essi associato che, combinati insieme attraverso un separatore (il carattere "$"), andranno a rappresentare i "value" del mapper.

Da questa prima fase si ottengono quindi coppie chiave – valore del tipo:

[Participant(SampleVial), gene$RPKM]

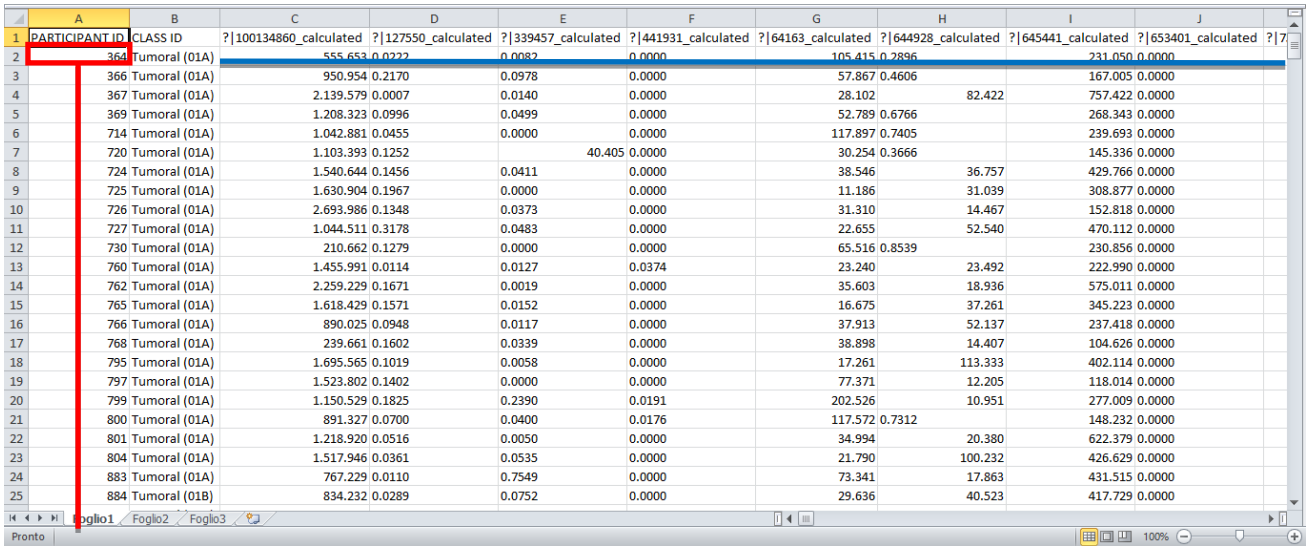
2.2 Fase di Reduce

La **fase di Reduce** prende in input l'output del mapper raggruppato per chiave, ovvero coppie chiave-valore nelle quali la chiave è costituita dalla stringa *Participant(sampleVial)* e il valore è una lista di *gene$RPKM* associati alla chiave.

In questa fase inoltre viene stabilita la forma che avrà l'output del job sia nel File-System di Hadoop che in un file con estensione csv.

2.3 Output generato

E' stato previsto un solo tipo di output in forma matriciale contenente il nome dei geni sulle colonne e l'identificativo dei partecipanti nelle righe.



PARTICIPANT ID	CLASS ID	?|100134860_calculated	?|127550_calculated	?|339457_calculated	?|441931_calculated	?|64163_calculated	?|644928_calculated	?|645441_calculated	?|653401_calculated
364	Tumoral (01A)	555.653	0.0222	0.0082	0.0000	105.415	0.2896	231.050	0.0000
366	Tumoral (01A)	950.954	0.2170	0.0978	0.0000	57.867	0.4606	167.005	0.0000
367	Tumoral (01A)	2.139.579	0.0007	0.0140	0.0000	28.102	82.422	757.422	0.0000
369	Tumoral (01A)	1.208.323	0.0996	0.0499	0.0000	52.789	0.6766	268.343	0.0000
714	Tumoral (01A)	1.042.881	0.0455	0.0000	0.0000	117.897	0.7405	239.693	0.0000
720	Tumoral (01A)	1.103.393	0.1252	40.405	0.0000	30.254	0.3666	145.336	0.0000
724	Tumoral (01A)	1.540.644	0.1456	0.0411	0.0000	38.546	36.757	429.766	0.0000
725	Tumoral (01A)	1.630.904	0.1967	0.0000	0.0000	11.186	31.039	308.877	0.0000
726	Tumoral (01A)	2.693.986	0.1348	0.0373	0.0000	31.310	14.467	152.818	0.0000
727	Tumoral (01A)	1.044.511	0.3178	0.0483	0.0000	22.655	52.540	470.112	0.0000
730	Tumoral (01A)	210.662	0.1279	0.0000	0.0000	65.516	0.8539	230.856	0.0000
760	Tumoral (01A)	1.455.991	0.0114	0.0127	0.0374	23.240	23.492	222.990	0.0000
762	Tumoral (01A)	2.259.229	0.1671	0.0019	0.0000	35.603	18.936	575.011	0.0000
765	Tumoral (01A)	1.618.429	0.1571	0.0152	0.0000	16.675	37.261	345.223	0.0000
766	Tumoral (01A)	890.025	0.0948	0.0117	0.0000	37.913	52.137	237.418	0.0000
768	Tumoral (01A)	239.661	0.1602	0.0339	0.0000	38.898	14.407	104.626	0.0000
795	Tumoral (01A)	1.695.565	0.1019	0.0058	0.0000	17.261	113.333	402.114	0.0000
797	Tumoral (01A)	1.523.802	0.1402	0.0000	0.0000	77.371	12.205	118.014	0.0000
799	Tumoral (01A)	1.150.529	0.1825	0.2390	0.0191	202.526	10.951	277.009	0.0000
800	Tumoral (01A)	891.327	0.0700	0.0400	0.0176	117.572	0.7312	148.232	0.0000
801	Tumoral (01A)	1.218.920	0.0516	0.0050	0.0000	34.994	20.380	622.379	0.0000
804	Tumoral (01A)	1.517.946	0.0361	0.0535	0.0000	21.790	100.232	426.629	0.0000
883	Tumoral (01A)	767.229	0.0110	0.7549	0.0000	73.341	17.863	431.515	0.0000
884	Tumoral (01B)	834.232	0.0289	0.0752	0.0000	29.636	40.523	417.729	0.0000

Figura 5: output- matrice originale

Come evidenziato in Figura 5, l'output sui file csv per la matrice originale ha una struttura del tipo: PARTICIPANT ID | CLASS ID | GENE-1|…|GENE-n|CLASS.

- PARTICIPANT ID: è l'identificativo del paziente (= Participant nel barcode).
- CLASS ID: rappresenta la classe di appartenenza del paziente codificata (Tumoral, Normal, Control) e il Sample e il Vial ad esso associati.
- GENE-1|…|GENE-n: identificatori delle colonne sono i nomi dei geni, per ciascun gene e ciascun paziente si ha, associato, il valore RPKM.
- CLASS: rappresenta la classe di appartenenza del paziente (Tumoral, Normal, Control), è una codifica del Sample.

Gli attributi CLASS e CLASS ID sono stati introdotti nelle matrici ai fini dell'analisi dei dati. Il CLASS ID, come accennato, è costituito dal Sample e dalla relativa codifica (Es: *Tumoral (01A)*) e viene preso in considerazione per analisi di maggior dettaglio. L'attributo CLASS invece è una forma di ridondanza della codifica del Sample ed è stato introdotto per generalizzare il campo di analisi.

2.4 *Dettagli di sviluppo*

Come precedentemente accennato, la generazione dell'output sul file csv avviene direttamente nella fase di Reduce, andando ad inserire l'intestazione della matrice ("intestazione") e i relativi dati ("stringa").

La figura 6 mostra la porzione di codice per la creazione dell'output.

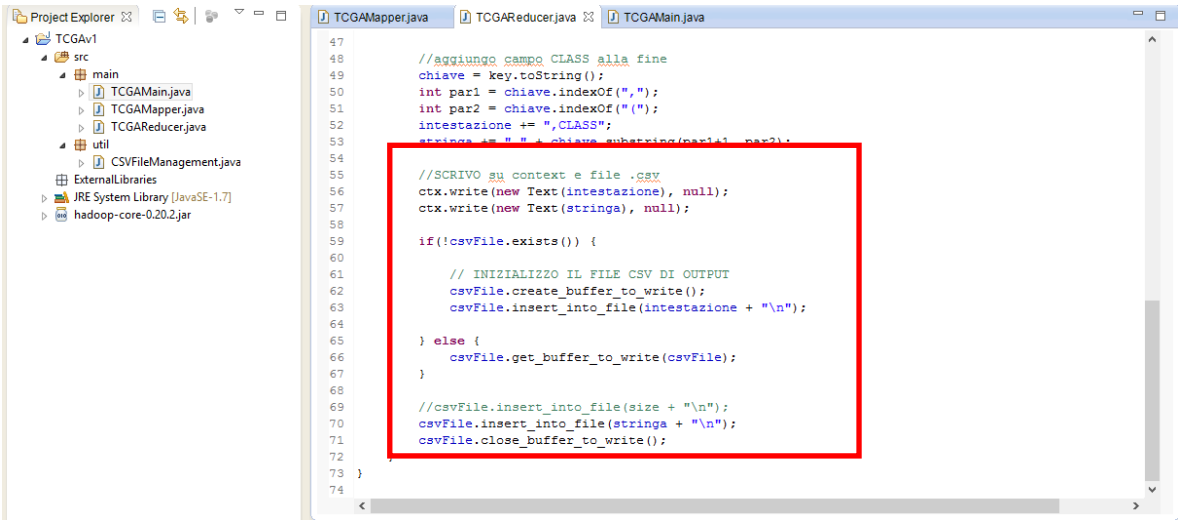


Figura 6: porzione di codice relativo alla generazione dell'output

3 Analisi dei Dati

L' analisi dei dati è stata condotta utilizzando il software WEKA, acronimo di "Waikato Environment for Knowledge Analysis", che fornisce un insieme di algoritmi per l'apprendimento automatico e l'estrazione di conoscenza dai dati (attività di machine learning e data mining).

3.1 Classificazione

La classificazione ha come obbiettivo l'estrazione di modelli che descrivono classi di dati per predire valori categorici o continui. La costruzione del modello viene generalmente fatta a partire da un insieme predeterminato di classi o concetti (Training set).

Per effettuare la classificazione con Weka sono stati utilizzati alberi decisionali "Trees" (in questo caso l'algoritmo J48), utilizzando come *Test Options* la *Cross-validation*. In questo modo i record vengono suddivisi in un determinato numero di folds (nel nostro caso 10) e ogni fold, a turno, funge da validation set per le altre folds che costituiscono il Training set. Alla fine dell'analisi vengono calcolati diversi valori tra i quali il *mean square error* o *errore quadratico medio* che misura la discrepanza quadratica media fra i valori dei dati osservati ed i valori dei dati stimati, il numero e la percentuale delle istanze classificate correttamente e non e l'*accuracy* espressa in termini di *Precision*, *Recall* e *F-Measure* .

L'output della classificazione, a meno di componenti opzionali, è così costituito:

- **Run information**: informazioni relative al programma di apprendimento, al nome della relazione, al numero di istanze e di attributi e alle modalità di prova che sono state coinvolte nel processo.
- **Classifier model (full training set)**: il testo del modello di classificazione che è stato prodotto sull'intero training set.
- **Summary**: un elenco di statistiche che riassume come il classificatore è stato in grado di predire la classe delle istanze del Test set, tra le quali l' errore quadratico medio e la correttezza delle istanze classificate.
- **Detailed Accuracy By Class**: informazioni dettagliate sulla precisione di ogni classe di predizione (Precision, Recall, FMeasure).
- **Confusion Matrix**: mostra il numero di casi che sono stati assegnati a ciascuna classe. Sulla sua diagonale principale si individuano i valori classificati correttamente, mentre nelle restanti celle si individuano gli errori di predizione.

Di seguito si riportano come esempio porzioni di output ottenute applicando "Tree" al dataset STAD.

In blu vengono messe in risalto le varie componenti dell'output della classificazione, in rosso i risultati delle analisi.

```
10  === Classifier model (full training set) ===
11
12  J48 pruned tree
13  ------------------
14
15  ESM1|11082_calculated <= 0.2558
16  |   RNFT2|84900_calculated <= 0.4895: Normal  (34.0/1.0)
17  |   RNFT2|84900_calculated > 0.4895: Tumoral  (5.0)
18  ESM1|11082_calculated > 0.2558: Tumoral  (232.0)
19
20  Number of Leaves  :     3
21
22  Size of the tree :  5
23
24
25  Time taken to build model: 2.66 seconds
26
27  === Stratified cross-validation ===
28  === Summary ===
29
30  Correctly Classified Instances         256               94.4649 %
31  Incorrectly Classified Instances        15                5.5351 %
32  Kappa statistic                          0.751
33  Mean absolute error                      0.0567
34  Root mean squared error                  0.2323
35  Relative absolute error                 26.1994 %
36  Root relative squared error             70.9967 %
```

Figura 7: Porzione di output dell'analisi ottenuta con l'algoritmo J48

Nel riquadro evidenziato in rosso (Figura 7) si nota il testo del modello di classificazione prodotto sull'intero training set, nel Summary sono riportate invece il numero di istanze classificate correttamente e non e l'errore quadratico medio.

L'alta percentuale di istanze classificate correttamente dimostra la validità e la bontà del modello di classificazione ottenuto mediante l'applicazione dell'algoritmo J48.

Nella figura che segue vengono messe in risalto l'accuracy del modello di classificazione per le due classi del dataset (Tumoral e Normal) e la relativa matrice di confusione.

Attraverso la diagonale principale della matrice di confusione possiamo ricavare il numero delle istanze classificate correttamente.

Figura 5: Porzione di output dell'analisi ottenuta con l'algoritmo JRip

Analizzando i valori di Precision, Recall e F-Measure abbiamo la conferma della bontà del modello di classificazione; tutti e tre i valori sono infatti prossimi all'unità.

```
38
39  === Detailed Accuracy By Class ===
40
41                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
42                   0.962     0.182       0.974     0.962       0.968      0.894    Tumoral
43                   0.818     0.038       0.75      0.818       0.783      0.894    Normal
44  Weighted Avg.    0.945     0.164       0.947     0.945       0.946      0.894
45
46  === Confusion Matrix ===
47
48     a   b   <-- classified as
49   229   9 |   a = Tumoral
50     6  27 |   b = Normal
51
```

Figura 8: Porzione di output dell'analisi ottenuta con l'algoritmo J48

Nella figura 9 viene messa in evidenza la sezione dell'output **Run Information**, in essa abbiamo indicazioni dell'algoritmo di classificazione applicato (in questo caso J48), del numero di istanze e del numero di attributi analizzati, del *Test Option* (cross-validation) e del numero di *fold* scelte.

```
1  === Run information ===
2
3  Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
4  Relation:     output_stomach_completo-weka.filters.unsupervised.attribute.Remove-R2
5  Instances:    271
6  Attributes:   29701
7  [list of attributes omitted]
8  Test mode:10-fold cross-validation
9
```

Figura 9: Porzione di output dell'analisi ottenuta con l'algoritmo J48

```
=== Summary ===

Correctly Classified Instances         256               94.4649 %
Incorrectly Classified Instances        15                5.5351 %
Kappa statistic                          0.751
Mean absolute error                      0.0567
Root mean squared error                  0.2323
Relative absolute error                 26.1994 %
Root relative squared error             70.9967 %
Total Number of Instances              271

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.962     0.182      0.974     0.962     0.968      0.894    Tumoral
                 0.818     0.038      0.75      0.818     0.783      0.894    Normal
Weighted Avg.    0.945     0.164      0.947     0.945     0.946      0.894

=== Confusion Matrix ===

   a   b   <-- classified as
 229   9 |   a = Tumoral
   6  27 |   b = Normal
```

Normal text file — length : 1583 lines

Figura 10: Porzione di output dell'analisi ottenuta con l'algoritmo Tree J48

La figura 10 mostra la corrispondenza tra il numero di istanze classificate correttamente riportato nel Summary e lo stesso valore ricavabile dalla diagonale principale della matrice di confusione. In questo caso il modello di classificazione è stato ricavato dall'applicazione dell'algoritmo J48.

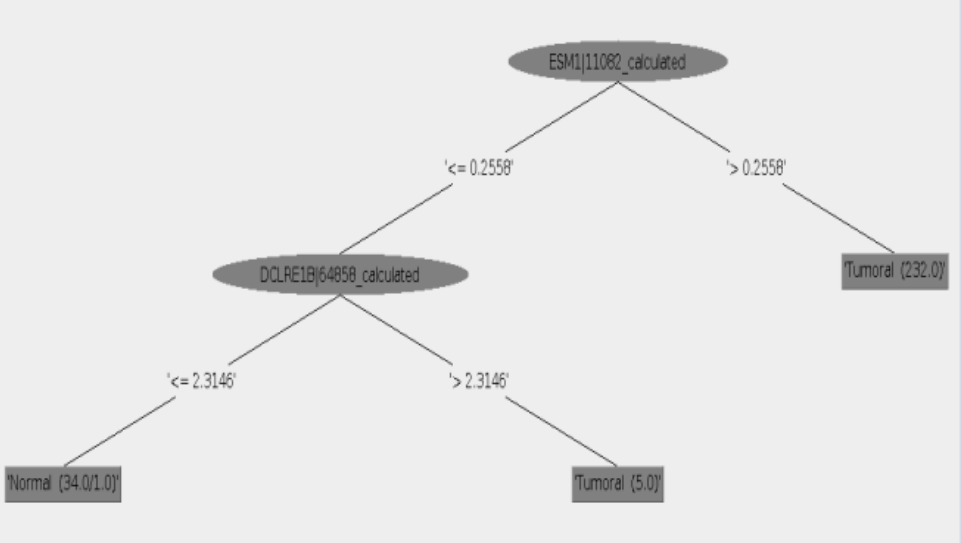


Figura 11: Visualizzazione dell'albero decisionale

Nella figura 11 si ha la visualizzazione dell'albero decisionale relativo al testo del modello di classificazione di figura 7.

3.2 Clustering

Le tecniche di clustering si applicano per suddividere un insieme di istanze in gruppi che riflettano qualche meccanismo o caratteristica naturale del dominio di appartenenza delle istanze stesse. Queste proprietà fanno sì che delle istanze siano accomunate da una "somiglianza" più forte rispetto agli altri dati della collezione.

Lo scopo di un algoritmo di clustering è quello di suddividere un insieme di dati in gruppi che siano quanto più possibile coerenti fra loro e allo stesso tempo diversi l'uno dall'altro (l'alta similarità intra-cluster e la bassa similarità inter-cluster).

Il clustering rappresenta la forma più comune di apprendimento non supervisionato (nessun uso di esperti umani per assegnare le istanze alle classi). L'input chiave di un algoritmo di questo tipo è dato dalla misura della distanza che viene utilizzata per suddividere le istanze in gruppi.

Per effettuare l'analisi di cluster è stato applicato su ogni dataset l'algoritmo K-Means, il più importante algoritmo di *Flat clustering* (crea un insieme di cluster piatto, senza una struttura gerarchica che metta in relazione i cluster l'un l'altro). Obbiettivo di K-Means è quello di minimizzare il valor medio del quadrato della distanza euclidea dei documenti dal centro del cluster a cui sono stati assegnati.

Il centro di un cluster è definito come la media di tutti i documenti presenti nel cluster (centroide).

Per la valutazione dei cluster, come accaduto per la classificazione, è stato utilizzato l'attributo CLASS.

Di seguito si riporta una porzione della schermata di output del clustering sul dataset STAD-Stomach_cancer e le relative visualizzazioni dei cluster.

All'interno della sezione Run Information vengono evidenziati, in rosso, l'algoritmo di clustering applicato e il numero di istanze e di attributi valutati.

Nella sezione KMeans si trovano informazioni relative al numero di iterazioni dell'algoritmo e alla somma dell'errore quadratico, seguite dalla lista degli attributi analizzati, dal numero e dal valore delle istanze totali e delle istanze per ciascun cluster creato (in questo caso 2, uno per le istanze di tipo Normal, l'altro per le istanze di tipo Tumoral, come messo in evidenza nella figura 13).

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     output_stomach_completo_v2-weka.filters.unsupervised.attribute.Remove-R1-2
Instances:    271
Attributes:   29700
[list of attributes omitted]
Test mode:Classes to clusters evaluation on training data
=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 27
Within cluster sum of squared errors: 125923.62440227614
Missing values globally replaced with mean/mode

Cluster centroids:
                                                 Cluster#
Attribute                                Full Data       0          1
                                           (271)      (61)       (210)
=============================================================================
5S_rRNA|?|125of139_calculated                   0         0          0
5S_rRNA|?|126of139_calculated                   0         0          0
5S_rRNA|?|127of139_calculated              0.0349    0.0368     0.0344
5S_rRNA|?|128of139_calculated              0.0066    0.0024     0.0079
5S_rRNA|?|129of139_calculated              0.0028    0.0044     0.0024
```

Figura 12: Porzione di output dell'analisi ottenuta con l'algoritmo Simple K-Means

```
29719  VCY|9084_calculated                                    0        0        0
29720  XKRY|9082|1of2_calculated                             0        0        0
29721  XKRY|9082|2of2_calculated                             0        0        0
29722  ZFY|7544_calculated                                1.2257   1.7024   1.0872
29723
29724
29725
29726
29727  Time taken to build model (full training data) : 47.02 seconds
29728
29729  === Model and evaluation on training set ===
29730
29731  Clustered Instances
29732
29733  0       61 ( 23%)
29734  1      210 ( 77%)
29735
29736
29737  Class attribute: CLASS
29738  Classes to Clusters:
29739
29740    0   1  <-- assigned to cluster
29741   52 186 | Tumoral
29742    9  24 | Normal
29743
29744  Cluster 0 <-- Normal
29745  Cluster 1 <-- Tumoral
29746
29747  Incorrectly clustered instances :   76.0      28.0443 %
```

Figura 13: Porzione di output dell'analisi ottenuta con l'algoritmo Simple K-Means

Questa seconda porzione di output oltre ad indicare il numero e la percentuale di istanze per cluster, il *class attribute* e la tipologia di istanze appartenenti a ciascun cluster, mette in evidenza anche il numero e la percentuale delle istanze clusterizzate in modo errato, in questo caso circa il 28% (76/210) un valore non molto trascurabile, indice di un modello di clustering non proprio perfetto e accurato come lo erano stati i modelli di classificazione.

Figura 12: Visualizzazione dei cluster distinti in base al colore

4 Visualizzazione dei dati

La visualizzazione dei dati è stata condotta utilizzando D3.js (Data-Driven Documents) una libreria Java-Script che utilizza i dati per guidare la creazione e il controllo delle forme dinamiche e interattive che vengono eseguite su di un web browser. In altre parole questa libreria fornisce funzionalità per la generazione di documenti HTML il cui contenuto è dinamicamente determinato dai dati.

Gli studi e i grafici generati sono consultabili all'indirizzo http://infoviscancer.github.io/ suddivisi nelle seguenti due macro-categorie che saranno approfondite rispettivamente nei paragrafi 4.1 e 4.2:

- Classificazione
- Clustering

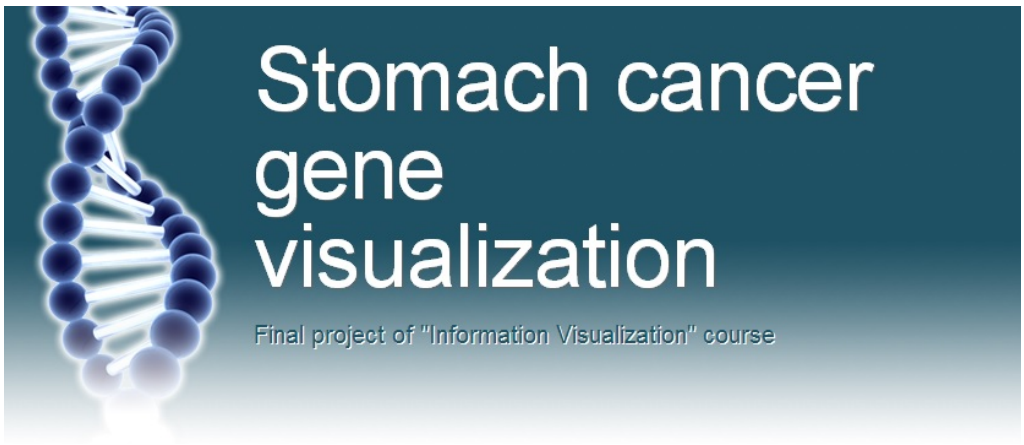


Figura 14: sito web del progetto

4.1 Classificazione

Nella macro-categoria "Classificazione" (visibile all'indirizzo http://infoviscancer.github.io/classification.html) attraverso la creazione di due grafici vengono visualizzati i modelli di classificazione ricavati durante la fase di Cross-Validation e la distribuzione dei pazienti in funzione del valore dell'RPKM dei geni coinvolti nell'albero di classificazione calcolato con WEKA (algoritmo J48).

4.1.1 Alberi di classificazione

4.1.1.1 Descrizione

Questo diagramma mostra per ogni folds della Cross-Validation il relativo albero di classificazione calcolato, sia in forma testuale che grafica.

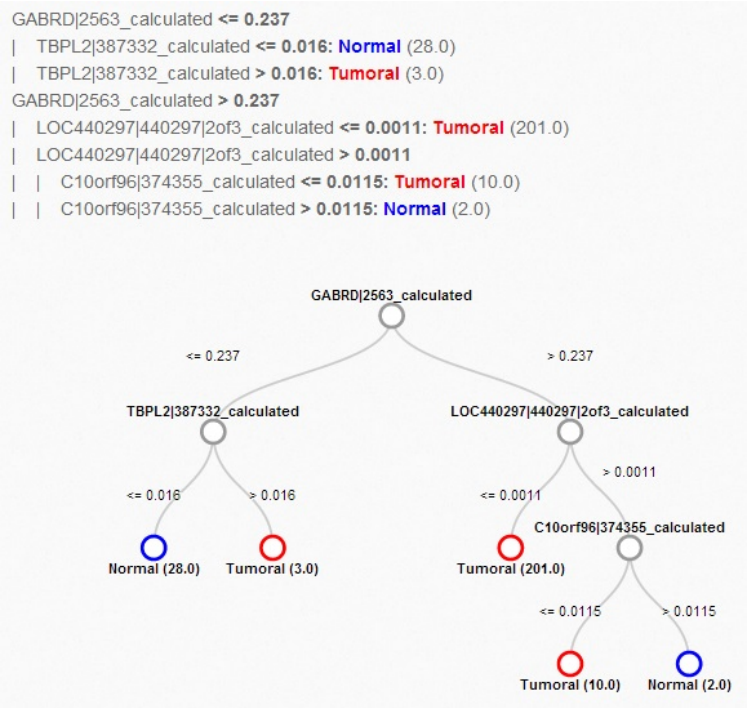


Figura 15: albero di classificazione

Passando il mouse sopra un nodo non foglia dell'albero vengono mostrate le seguenti informazioni aggiuntive ricavate dal sito http://rest.genenames.org/ tramite interrogazione automatica:

- **Symbol**: simbolo ufficiale del gene approvato dall' HGNC, si tratta di una forma abbreviata del nome del gene. I simboli sono approvati in conformità alle linee guida per la nomenclatura del gene umano;
- **Hgnc_id**: identificativo univoco del gene per l' HGNC;
- **Name**: il nome del gene completo approvato dall' HGNC;
- **Locus type**: classe genetica di apparteneza del gene;
- **Alias symbol**: simboli e nomi alternativi che sono stati utilizzati per riferirsi al gene. I sinonimi possono essere dati dalla letteratura, da altri database o possono essere aggiunti per rappresentare l'appartenenza ad una famiglia genica.
- **Location**: indica la posizione citogenetica del gene o della regione sul cromosoma.

Inoltre cliccando sul nome di un gene è possibile scaricare l'intero set di informazioni in formato json.

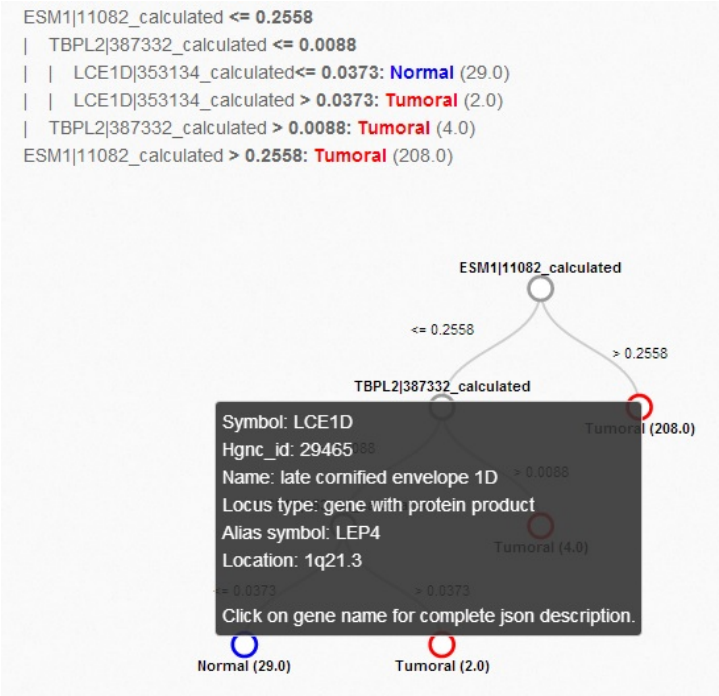


Figura 16: dati aggiuntivi albero di classificazione

Per mezzo di uno slider è possibile passare da un albero di classificazione ad un altro.

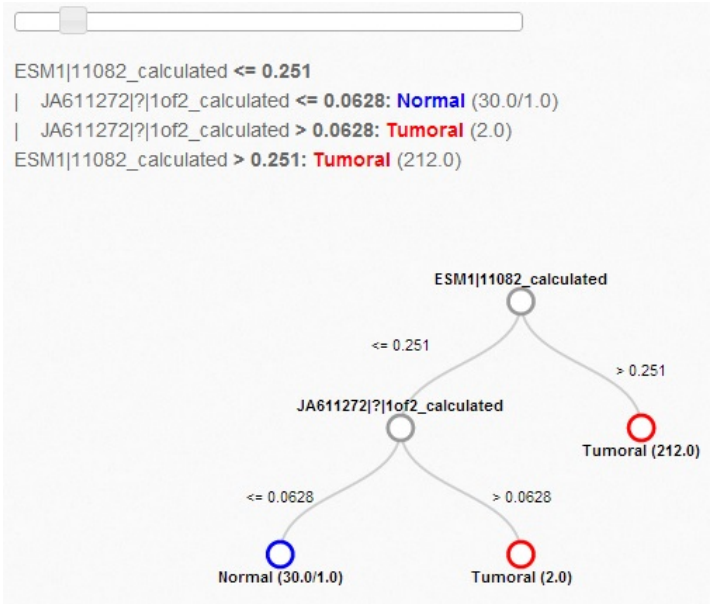


Figura 17: slider di selezione dell'albero di classificazione

4.1.1.2 Dati di input

Per l'input di questa visualizzazione sono state prese in considerazioni le varie regole di output, riscritte in formato json al fine di essere visualizzate come albero attraverso la librearia d3js.

4.1.2 Scatter Plot

4.1.2.1 Descrizione

Il grafico mostra la distribuzione dei pazienti in funzione del valore di RPKM dei geni coinvolti nell'albero di classificazione calcolato con WEKA attraverso l'algoritmo J48.

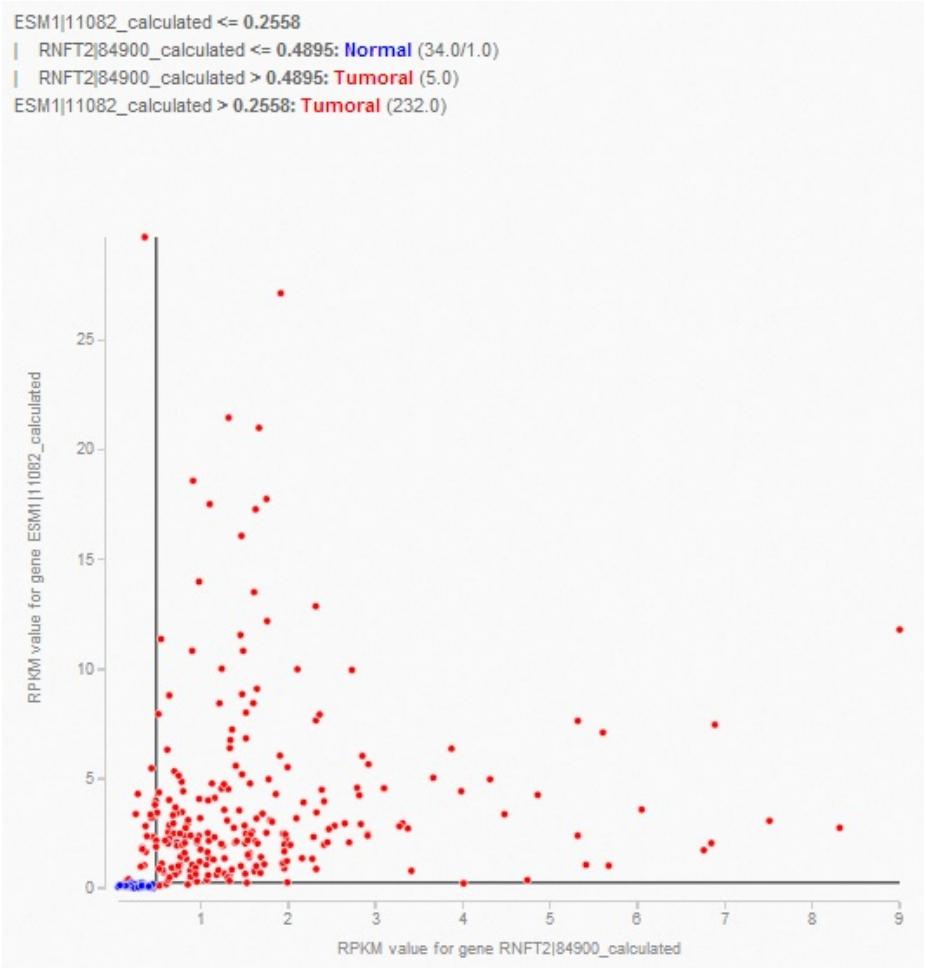


Figura 18: scatter plot

La figura 18 mostra l'output prodotto dalla classificazione, in esso compaiono solo due geni che vengono utilizzati per distinguere i pazienti nelle classi Tumoral e Normal. I due geni sono presi come riferimento e vanno a costituire gli assi del diagramma, su di questi viene riportato il valore di RPKM.

Ogni paziente viene quindi collocato sul grafico in base al valore di RPKM assunto dai geni presi come riferimento dall'algoritmo e colorato in rosso se appartenente alla classe Tumoral o in blu se appartenente alla classe Normal.

Gli assi visualizzati con una linea più marcata rappresentano i limiti calcolati dall'algoritmo per classificare le istanze in Normal e Tumoral. Ad esempio la linea orizzontale (y=0,2558) indica

che tutti i pazienti con un valore di RPKM superiore a 0,2558 per il gene *"ESM1|11082_calculated"* appartengono sicuramente alla classe Tumoral; mentre quelli con un valore inferiore appartengono alla classe Normal solo se l'RPKM del gene *"RNFT2|84900_calculated"* è minore o uguale a 0.4895 (a sinistra della linea verticale x=0,4895) altrimenti appartengono alla classe Tumoral.

Passando il mouse sopra un punto (un paziente) vengono mostrate informazioni aggiuntive relative a:

- identificativo del paziente
- classe di appartenenza
- valori di RPKM dei due geni presi come riferimento dall'algoritmo

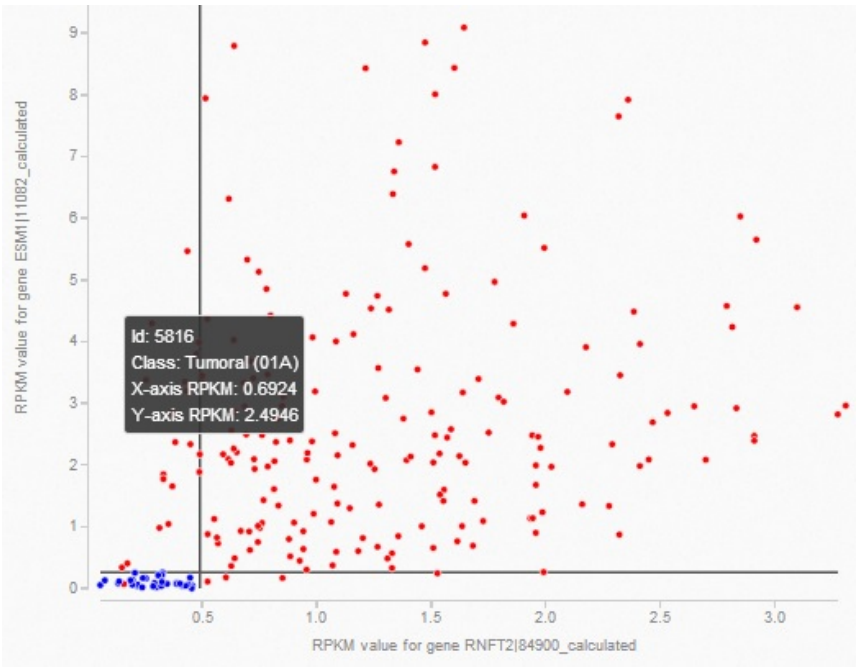


Figura 19: informazioni aggiuntive scatter plot

Selezionando con il mouse una porzione del grafico è possibile inoltre effettuare uno zoom.

Infine attraverso l'uso di due sliders, uno per l'asse x e l'altro per l'asse y, è possibile modificare i geni di riferimento. La modifica avviene visualizzando il gene con valore di correlazione di Pearson più alta verso il gene di riferimento considerato dall'algoritmo.

La correlazione di Pearson è stata calcolata attraverso il software "MALA", immettendo come input una versione modificata della matrice inversa dell'originale (avuta come output dal processo di map-reduce) e selezionando i due geni della regola presa in esame. Sono stati poi estratti i primi 10 geni in ordine decrescente di valore di correlazione.

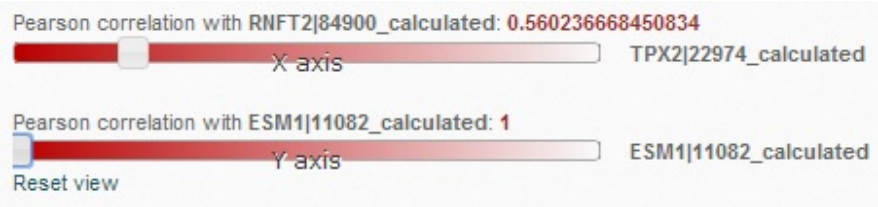


Figura 20: sliders scatter plot

Inoltre cliccando sul nome dei geni, visualizzati sopra e al lato degli sliders, è possibile scaricare l'intero set di informazioni in formato json ricavate dal sito http://rest.genenames.org/.

4.1.2.2 Dati di input

I dati di input sono costituiti dalla matrice ottenuta a seguito dell'esecuzione del job di Map-Reduce (vedi paragrafo 2.3) , seguito dalla selezione dei geni presi in considerazione.

4.2 Clustering

Nella macro-categoria "Clustering" (visibile all'indirizzo http://infoviscancer.github.io/clustering.html) attraverso la creazione di quattro grafici vengono evidenziate per ogni gene le differenze, sulla base dell'RPKM, tra la classi "Tumoral" e "Normal".

I grafici utilizzati sono:

- diagramma a bolle (bubble chart)
- chord diagram
- slopegraphs
- grafico a barre

4.2.1 Diagramma a bolle (bubble chart)

4.2.1.1 Descrizione

Questo diagramma mostra, per ogni gene, il valore dell'output dell'operazione di clustering. La visualizzazione segue una struttura a spirale collocando il gene con valore medio più alto al centro del diagramma.

Il valore di ogni gene è proporzionale alla superficie della bolla.

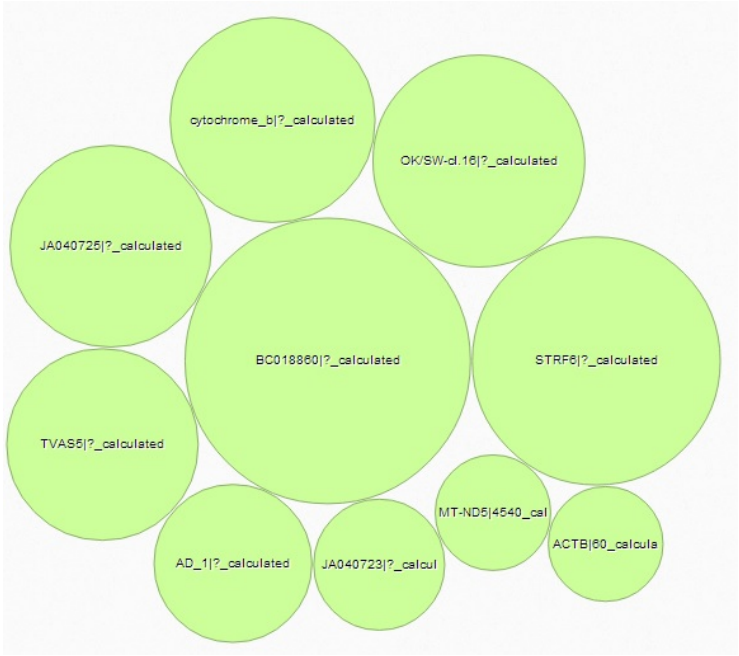


Figura 21: diagramm a bolle

Passando il mouse sopra una bolla vengono mostrate le informazioni relative al gene selezionato.

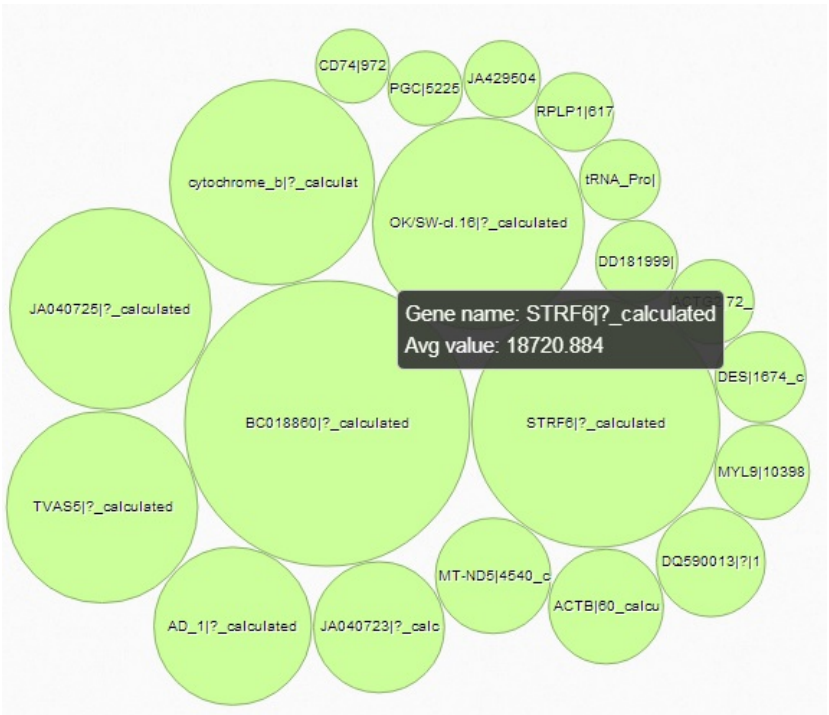


Figura 22: dati aggiuntivi diagramma a bolle

Per mezzo di due sliders è possibile modificare rispettivamente il numero di geni visualizzati (10, 20, 30, 40, 50, 60, 70, 80, 90, 100) e la classe (Normal o Tumoral).

Figura 23: sliders di selezione numero geni e classe

Di seguito viene mostrato il risultato della variazione del numero di geni visualizzati e della classe, quest'ultima distinta in base al colore.

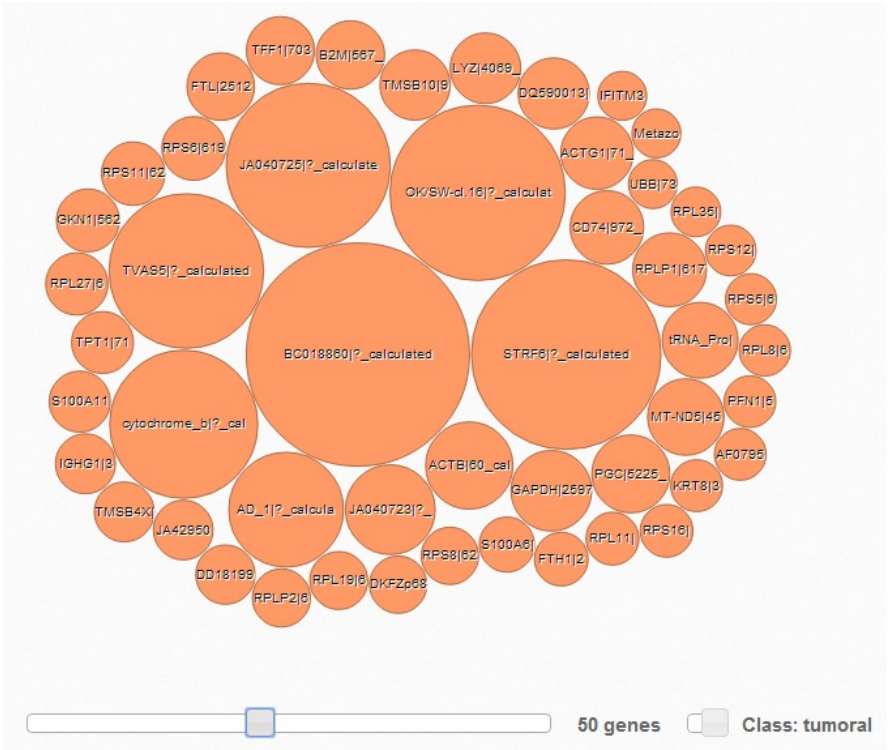


Figura 24: Diagramma a bolle - visualizzazione di 50 geni della classe tumoral

4.2.1.2 Dati di input

I dati di input sono costituiti da 10 file json per la classe Normal e altrettanti per la classe Tumoral ciascuno contenente i nomi dei geni da visualizzare e il relativo valore di output del clustering ordinati in ordine decrescente.

4.2.2 Chord diagram

4.2.2.1 Descrizione

Il diagramma mostra l'interazione tra geni normali e tumorali ordinati in ordine decrescente dell'output del clustering.

La distinzione delle classi Tumoral e Normal si ottiene in base al colore dei semicerchi che limitano il grafico, rispettivamente nero per i primi e bianco per i secondi.

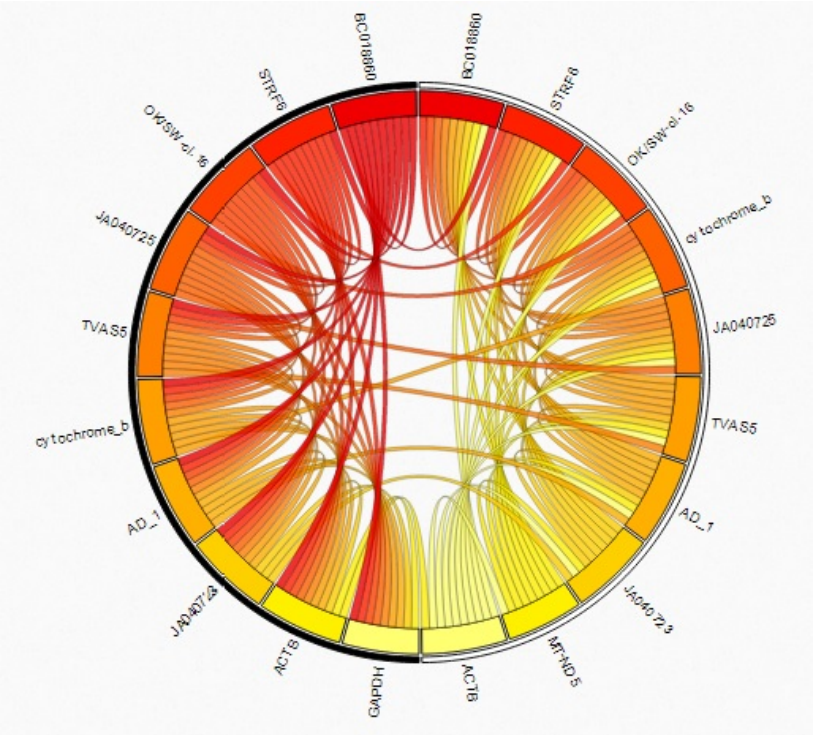


Figura 25: chord diagram

Si noti che ogni gene ha una interazione con tutti gli altri geni della sua classe. Inoltre, se il gene è presente anche nell'altra classe, viene visualizzato un collegamento (quindi un'interazione) tra il gene in classe normale e in classe tumorale.

Per mezzo di due sliders è possibile impostare il numero di geni da visualizzare (10, 20, 30, 40 o 50) ed eliminare le sole relazioni intra-classe mostrando le altre.

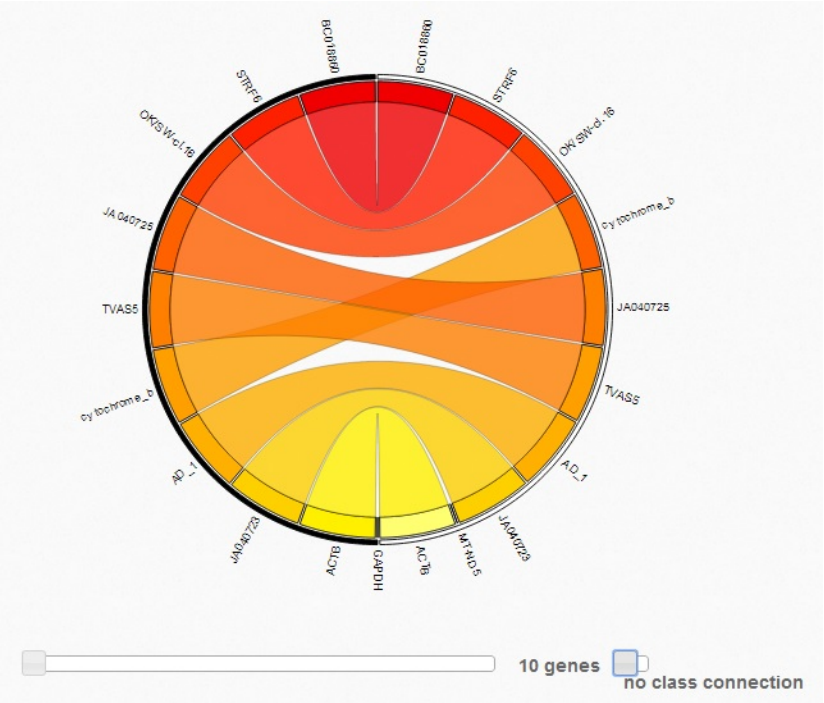


Figura 26: chord diagram - interazioni inter-classe

4.2.2.2 Dati di input

I dati di input sono rappresentati da due gruppi di 5 file csv, ogni file contiene un numero di geni variabile (10, 20, 30, 40 o 50) in base al livello di dettaglio che si vuole visualizzare. Inoltre si ha un gruppo per ogni tipo di visualizzazione da effettuare (con connessioni intra-classe o meno).

4.2.3 Slopegraphs

4.2.3.1 Descrizione

Il diagramma mostra l'interazione tra geni normali e tumorali ordinati in ordine decrescente dell'output del clustering.

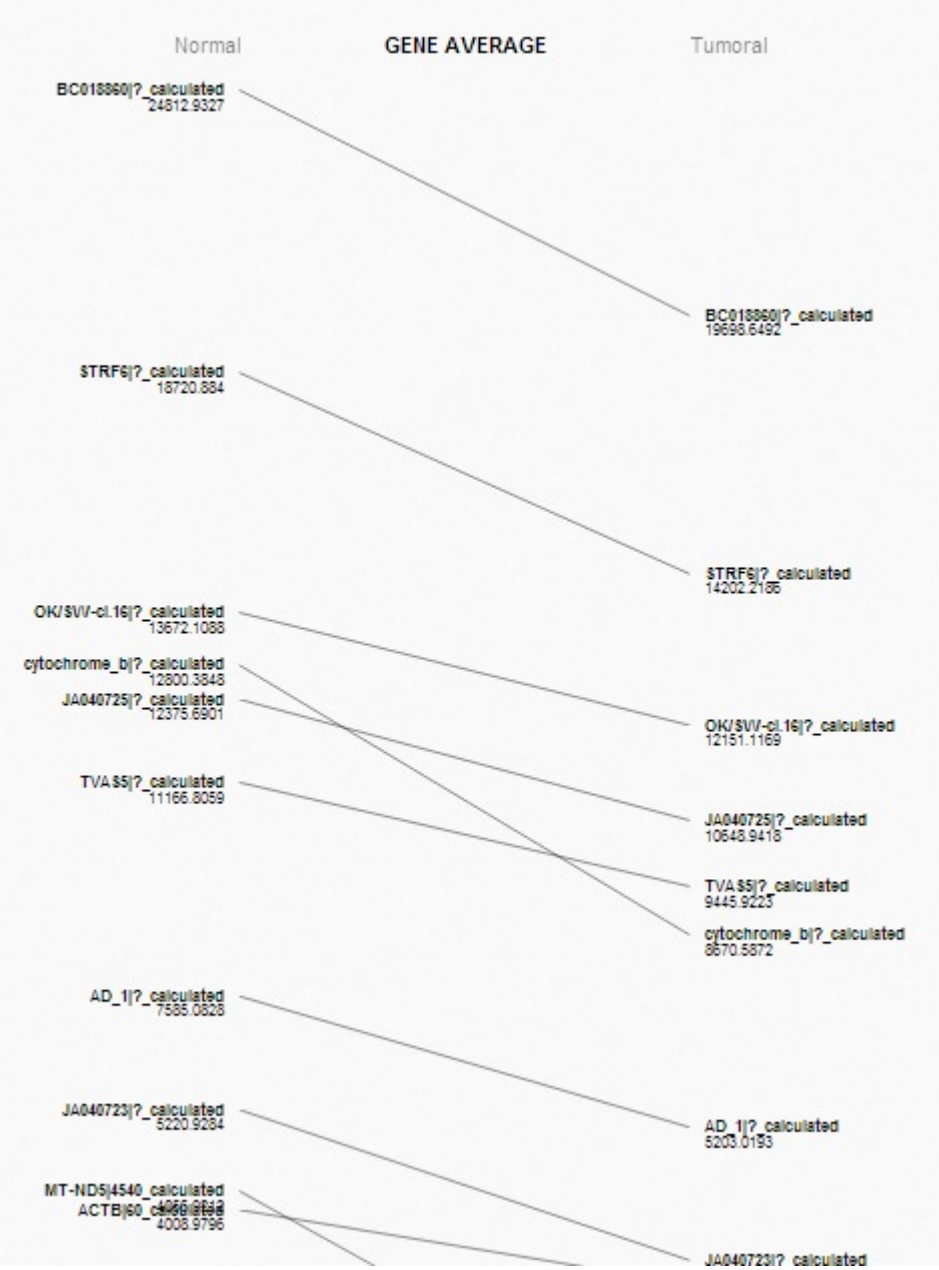


Figura 27: diagramma slopegraphs

Questo diagramma mostra in modo esplicito la corrispondenza tra i primi N geni (dove N è selezionabile tramite uno slider) della classe Normal e i primi N geni della classe Tumoral in una scala discendente. Se un gene non trova corrispondenza nei primi N geni dell'altra classe viene comunque visualizzato con valore medio pari a zero nella classe in cui non ha trovato corrispondenza.

Passando il mouse sopra il nome del gene la linea che collega quest'ultimo con il suo corrispondente viene evidenziata in rosso e appare la classe a cui appartiene, il nome del gene e il valore associato dall'output del clustering.

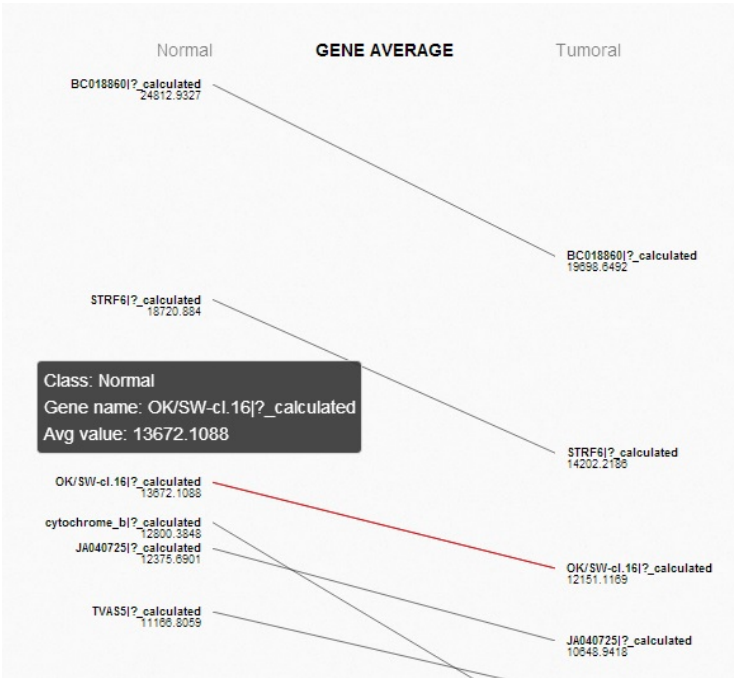


Figura 28: dati aggiuntivi slopegraphs sul gene

Passando, invece, il mouse sopra una riga questa viene evidenziata in rosso e vengono anche mostrati sia i valori medi della classe normal che quelli della classe tumoral.

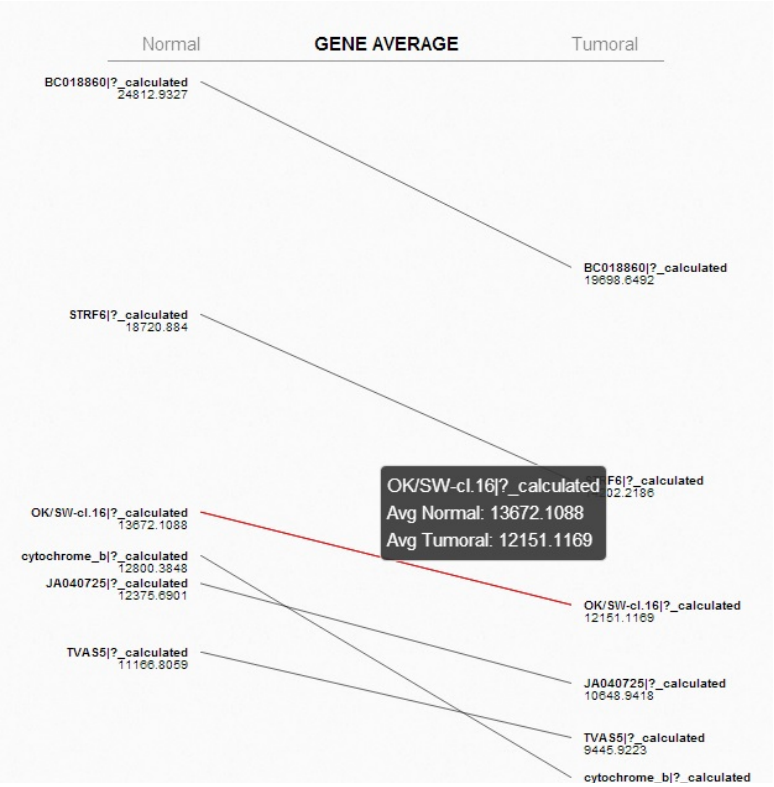


Figura 29: dati aggiuntivi slopegraphs sul collegamento

Per mezzo di due sliders si è ottenuta rispettivamente la possibilità di impostare la dimensione della lista ordinata di geni (10, 50 o 100) e di ingrandire il diagramma per una più agevole consultazione.

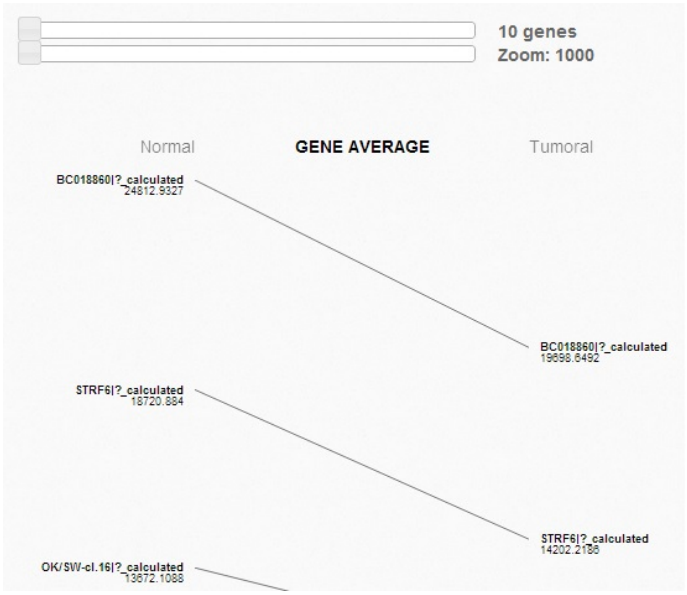


Figura 30: sliders di selezione numero geni e zoom

4.2.3.2 Dati di input

I dati di input sono costituiti da 3 files json, uno per ciascun numero (10, 50 o 100) di geni da visualizzare. Ogni file è composto dai nomi dei primi N geni con valore di output del clustering più alto seguiti dal relativo valore, sia per la classe Normal che per la classe Tumoral.


```
var all_data_10 = {
"0": {
"BC018860|?_calculated": 24812.9327,
"STRF6|?_calculated": 18720.884,
"OK/SW-cl.16|?_calculated": 13672.1088,
"cytochrome_b|?_calculated": 12800.3848,
"JA040725|?_calculated": 12375.6901,
"TVAS5|?_calculated": 11166.8059,
"AD_1|?_calculated": 7585.0828,
"JA040723|?_calculated": 5220.9284,
"MT-ND5|4540_calculated": 4056.9212,
"ACTB|60_calculated": 4008.9796
},
"1": {
"BC018860|?_calculated": 19698.6492,
"STRF6|?_calculated": 14202.2186,
"OK/SW-cl.16|?_calculated": 12151.1169,
"JA040725|?_calculated": 10648.9418,
"TVAS5|?_calculated": 9445.9223,
"cytochrome_b|?_calculated": 8670.5872,
"AD_1|?_calculated": 5203.0193,
"JA040723|?_calculated": 3183.6099,
"ACTB|60_calculated": 3012.0267,
"GAPDH|2597_calculated": 2569.6558
}
};
```


Figura 31: file json con 10 geni

4.2.4 Grafico a barre

4.2.4.1 Descrizione

Il diagramma a barre mostra, per ogni gene, la differenza del valore di output del clustering tra la classe Normal (asse y superiore) e la classe Tumoral (asse y inferiore) evidenziando l'appartenenza gene-cromosoma attraverso una gamma di colori.

Per realizzare questa visualizzazione è stato necessario scrivere un programma Java per arricchire i dati di output del clustering di Weka con il valore del cromosoma e dello specifico locus a cui ogni gene appartiene.

Per mezzo di due sliders, uno orizzontale e uno verticale, è possibile effettuare rispettivamente uno zoom sull'asse delle ascisse o sull'asse delle ordinate.

Figura 32: grafico a barre

Passando il mouse sopra una barra vengono mostrate le seguenti informazioni aggiuntive, alcune ricavate dal sito http://rest.genenames.org/ tramite interrogazione automatica:

- **Gene name**: nome del gene;
- **Avg RPKM value**: valore dell'output del clustering;
- **Chromosome**: cromosoma del gene;
- **Locus**: posizione del gene all'interno del cromosoma.

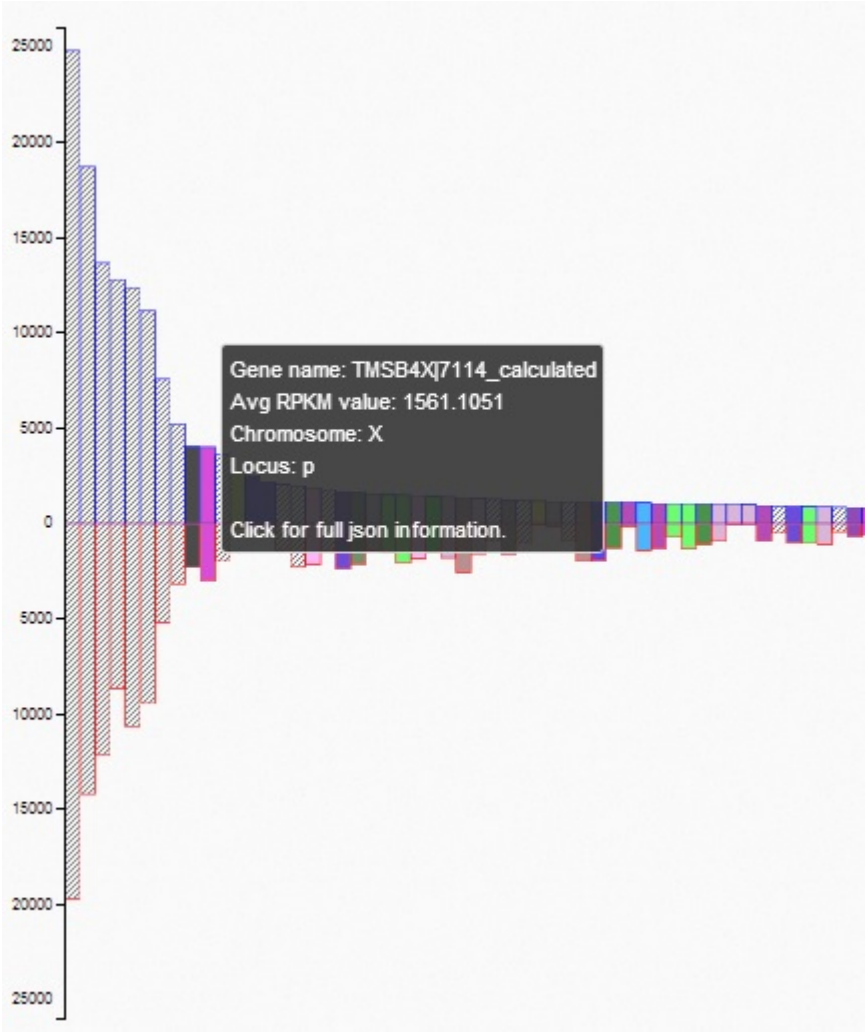


Figura 33: informazioni aggiuntiva grafico a barre

Tramite il menù mostrato nella figura seguente è possibile effettuare le operazioni:

- **NORMAL 1000 AVG DESCENDANT**: ordinare in modo decrescente il valore di output del clustering dei primi 1000 geni appartenenti alla classe Normal;
- **TUMORAL 1000 AVG DESCENDANT**: ordinare in modo decrescente il valore di output del clustering dei primi 1000 geni appartenenti alla classe Tumoral;
- **NORMAL FULL AVG DESCENDANT**: ordinare in modo decrescente il valore di output del clustering di tutti i geni appartenenti alla classe Normal;
- **TUMORAL FULL AVG DESCENDANT**: ordinare in modo decrescente il valore di output del clustering di tutti i geni appartenenti alla classe Tumoral.

Figura 34: menù grafico a barre

E' inoltre presente una legenda che permette, in base al colore, di comprendere a quale cromosoma appartenga un determinato gene. Per mezzo di un click su di un cromosoma è possibile applicare un filtro alla visualizzazione andando a mostrare solo i geni relativi al cromosoma in questione.

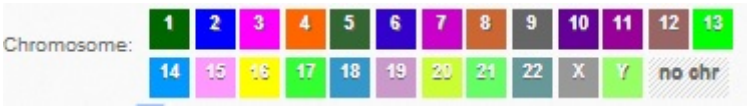


Figura 35: legenda cromosoma

4.2.4.2 Dati di input

I dati di input sono costituiti da 4 files, uno per ciascuna delle 4 differenti modalità di visualizzazione selezionabili dal menù. Ogni file è composto da:

- nome del gene
- valore dell'output del clustering normal
- valore dell'output del clustering tumoral
- cromosoma di appartenenza
- posizione citogenetica del gene (regione sul cromosoma)
- locus di appartenenza all'interno del cromosoma

i geni sono elencati in ordine decrescente del valore di output del clustering.

```
gene,rpkm_normal,rpkm_tumoral,chromosome,p_or_q,locus
BC018860|?_calculated,24812.9327,19698.6492,,,
STRF6|?_calculated,18720.884,14202.2186,,,
OK/SW-cl.16|?_calculated,13672.1088,12151.1169,,,
cytochrome_b|?_calculated,12800.3848,8670.5872,,,
JA040725|?_calculated,12375.6901,10648.9418,,,
TVAS5|?_calculated,11166.8059,9445.9223,,,
AD_1|?_calculated,7585.0828,5203.0193,,,
JA040723|?_calculated,5220.9284,3183.6099,,,
MT-ND5|4540_calculated,4056.9212,2310.2934,mitochondria,,
ACTB|60_calculated,4008.9796,3012.0267,7,p,
DQ590013|?|1of2_calculated,3605.2425,2001.1237,,,
MYL9|10398_calculated,2739.3368,299.1312,20,q,
DES|1674_calculated,2507.3146,186.3599,2,q,
ACTG2|72_calculated,2169.1035,136.9224,2,p,
DD181999|?_calculated,2050.1554,1388.8544,,,
tRNA_Pro|?|1of20_calculated,1974.7049,2275.3102,,,
RPLP1|6176_calculated,1871.9509,2157.3051,15,q,
JA429504|?|1of4_calculated,1788.4225,1404.5668,,,
PGC|5225_calculated,1682.9445,2401.6369,6,p,
CD74|972_calculated,1651.0728,2140.5576,5,q,
TMSB4X|7114_calculated,1561.1051,1418.0327,X,p,
RPL27|6155_calculated,1520.909,1528.7798,17,q,
ACTG1|71_calculated,1491.839,2075.4266,17,q,
B2M|567_calculated,1475.2502,1849.2798,15,q,21-
TPT1|7178_calculated,1405.2412,1513.0518,13,q,
FTL|2512_calculated,1382.7791,1800.5877,19,q,
GAPDH|2597_calculated,1375.8034,2569.6558,12,p,
RPS11|6205_calculated,1319.1545,1591.5651,19,q,
Metazoa_SRP|?|35of109_calculated,1276.8148,944.4363,,,
RPS6|6194_calculated,1209.05,1602.913,9,p,
AF079515|?_calculated,1193.7485,1057.0295,,,
MYH11|4629_calculated,1181.7274,65.4523,16,p,
FLNA|2316_calculated,1162.9335,157.6572,X,q,
DQ582201|?_calculated,1150.18,937.7625,,,
LYZ|4069_calculated,1145.4643,1971.4133,12,q,
TMSB10|9168_calculated,1142.2601,1949.8254,2,p,
```

Figura 36: file csv di input per grafico a barre

Per rendere più veloce l'utilizzo del filtro in base ai cromosomi sono stati creati, attraverso un'applicazione java appositamente sviluppata, un numero di file csv pari al numero dei cromosomi. Ogni file contiene esclusivamente geni relativi ad un determinato cromosoma.

5 Sviluppi Futuri

Per rendere maggiormente completo il progetto dal punto di vista dell'analisi dei dati si potrebbe pensare di espanderlo utilizzando ulteriori algoritmi di classificazione e clustering.

Si potrebbero inoltre comparare i modelli di analisi prodotti da Weka, con modelli prodotti da altri software e mettendo in evidenza graficamente similitudini e differenze, validità e precisione.

6 Riferimenti

- *TCGA The Cancer Genome Atlas* - https://wiki.nci.nih.gov/display/TCGA/The+Cancer+Genome+Atlas

- *WEKA: Machine Learning Algorithms in java* – Dipartimento di Informatica e Sistemistica Antonio Ruberti, Sapienza Università di Roma

- *Machine Learning with WEKA* – Eibe Frank, Department of Computer Science, University of Waikato, New Zeland

- *Gene expression profile analysis: normalization, clustering and classification* - Emanuel Weitschek, Giulia Fiscon, Giovanni Felici, and Paola Bertolazzi, Department of Engineering Roma Tre University, Rome, Italy; Institute of Systems Analysis and Computer Science National Research Council, Rome, Italy; Department of Computer, Control and Management Engineering, Sapienza University, Rome, Italy

- *TCGA and GenData, The Cancer Genome Atlas* - Emanuel Weitschek, Giulia Fiscon, Fabio Cumbo

- *MALA: A microarray clustering and classification software*