**A Study on Gene expression in non-tumoral liver tissue and recurrence-free survival in hepatitis C virus-positive hepatocellular carcinoma**

**Report Submitted as part of the Masters Research Project for**

**BINF 798**

**By**

**Monica Devi Manam**

**Dr. Iosif Vaisman**

**Acknowledgement :**

**I would like to thank Dr.Iosif Vaisman for guiding me throughout the project.**

**Abstract :**

The aim of this study was to understand gene expression signatures of hepatocellular carcinoma (HCC) recurrence in subjects with hepatitis C virus (HCV) infection. Traditional clinico-pathological endpoints are recognized as weak predictors of RFS. It has been suggested that gene expression profiling of HCC and non-tumoral liver tissue may improve prediction of RFS, aid in understanding of the underlying liver disease, and guide individualized patient management. Gene Dataset was taken from Geo database. Gene expression profiling data was used to determine the molecular signature of HCV-associated HCC and to develop a predictor of RFS using machine learning using various algorithms.

**Introduction :**

Hepatocellular carcinoma (HCC) is the fifth most common cancer worldwide by annual incidence and the third leading cause of cancer death . Wide geographic variation in age-adjusted incidence and death rates is well recognized. Most alarming is the fact that age-adjusted incidence and death rates for cancer of the liver and intrahepatic bile duct show a statistically significant increasing trend in the past three decades in the USA and many other countries, even though other major cancers are on a decline . This rise is being attributed, at least in part, to an increase in incidence of hepatitis C virus (HCV) infections and non-alcoholic steatohepatitis, pathological states that are also growing in the US population. The resistance of HCC to existing treatments and the lack of biomarkers for early detection make it one of the deadliest cancers. Surgical resection, liver transplantation, and ablation by radiofrequency or ethanol injection are now conventional therapies at early disease stages. Even with these options, survival at 5 years is poor and ranges between 50% and 70%

One of the key reasons for poor long-term survival in HCC is high incidence of recurrence, a complication that cannot be prevented effectively by new and existing therapies [[7](http://www.molecular-cancer.com/content/9/1/74#B7),[8](http://www.molecular-cancer.com/content/9/1/74#B8)]. Many clinico-pathological features, such as tumor size, number of tumors in liver, capsule state, cell differentiation, venous invasion, and the extent of intrahepatic spreading are commonly used in clinical diagnosis as predictive risk factors for HCC recurrence and health prognosis in patients[[9](http://www.molecular-cancer.com/content/9/1/74#B9),[10](http://www.molecular-cancer.com/content/9/1/74#B10)]. However, the prospective utility of these attributes for predicting recurrence-free survival (RFS) may be more limited as HCC is being diagnosed at earlier stages.

Several research groups have performed gene expression profiling of both tumor and nontumoral specimens and identified gene signatures of recurrence-free [[11](http://www.molecular-cancer.com/content/9/1/74#B11)], or overall survival [[12](http://www.molecular-cancer.com/content/9/1/74#B12)] in HCC patients who undergo tumor resection. These studies carry an exciting potential to open the field to more selective chemoprevention as a follow-up to surgical interventions since patients with greatest risk of death or recurrence can be potentially identified [[12](http://www.molecular-cancer.com/content/9/1/74#B12)] and/or individualized therapies may be devised based on the molecular profiles of poor-prognosis markers of HCC

we report the molecular network signatures of HCV-associated HCC, as well as the outcome of the analysis of the predictive value of gene expression in tumor- and nontumoral tissue-derived samples.

**Materials & Methods:**

Global gene expression (25,073 genes) data of the samples from HCC (black circle), surgical margin derived nontumoral samples (white circle), and control samples (nontumoral samples from patients with metastatic liver tumors; black triangle) was visualized using principal components analysis. Raw microarray data was archived in Gene Expression Omnibus (GEO) database was used to retrieve "GSE17856" dataset and it is available to the public. Obtained from 47 subjects with HCV-associated HCC who underwent complete removal of the tumor from the Icahn School of Medicine at Mount Sinai.

Machine Learning Algorithms Used:

**SMO :** Sequential minimal optimization is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines

SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers \alpha_i, the smallest possible problem involves two such multipliers. Then, for any two multipliers \alpha_1 and \alpha_2, the constraints are reduced to:

0 \leq \alpha_1, \alpha_2 \leq C,

y_1 \alpha_1 + y_2 \alpha_2 = k,

and this reduced problem can be solved analytically: one needs to find a minimum of a one-dimensional quadratic function. k is the sum over the rest of terms in the equality constraint, which is fixed in each iteration.

**J48: C4.5** is an algorithm used to generate a decision tree, J48 is the java implementation of it.

This algorithm has a few base cases.

All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

**Random Forests:** are an ensemble learning method for classification :operate by constructing a multitude of decision trees. at training time and outputting the class that is the [mode](http://en.wikipedia.org/wiki/Mode_(statistics)) of the classes output by individual trees.

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way. The following technique was described in Breiman's original paper[[1]](http://en.wikipedia.org/wiki/Random_forest#cite_note-breiman2001-1) and is implemented in the [R](http://en.wikipedia.org/wiki/R_(programming_language)) package randomForest

**Naive Bayes** :classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions.

In simple terms, a naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a [supervised learning](http://en.wikipedia.org/wiki/Supervised_learning) setting. In many practical applications, parameter estimation for naive Bayes models uses the method of [maximum likelihood](http://en.wikipedia.org/wiki/Maximum_likelihood); in other words, one can work with the naive Bayes model without accepting [Bayesian probability](http://en.wikipedia.org/wiki/Bayesian_probability) or using any Bayesian methods.

**List Of abbreviations :**

HCC: Hepatocellular Carcinoma; HCV: Hepatitis C Virus; FDR: False Discovery Rate; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; RFS: Recurrence-Free Survival; SAFE: Significance Analysis of Function and Expression.

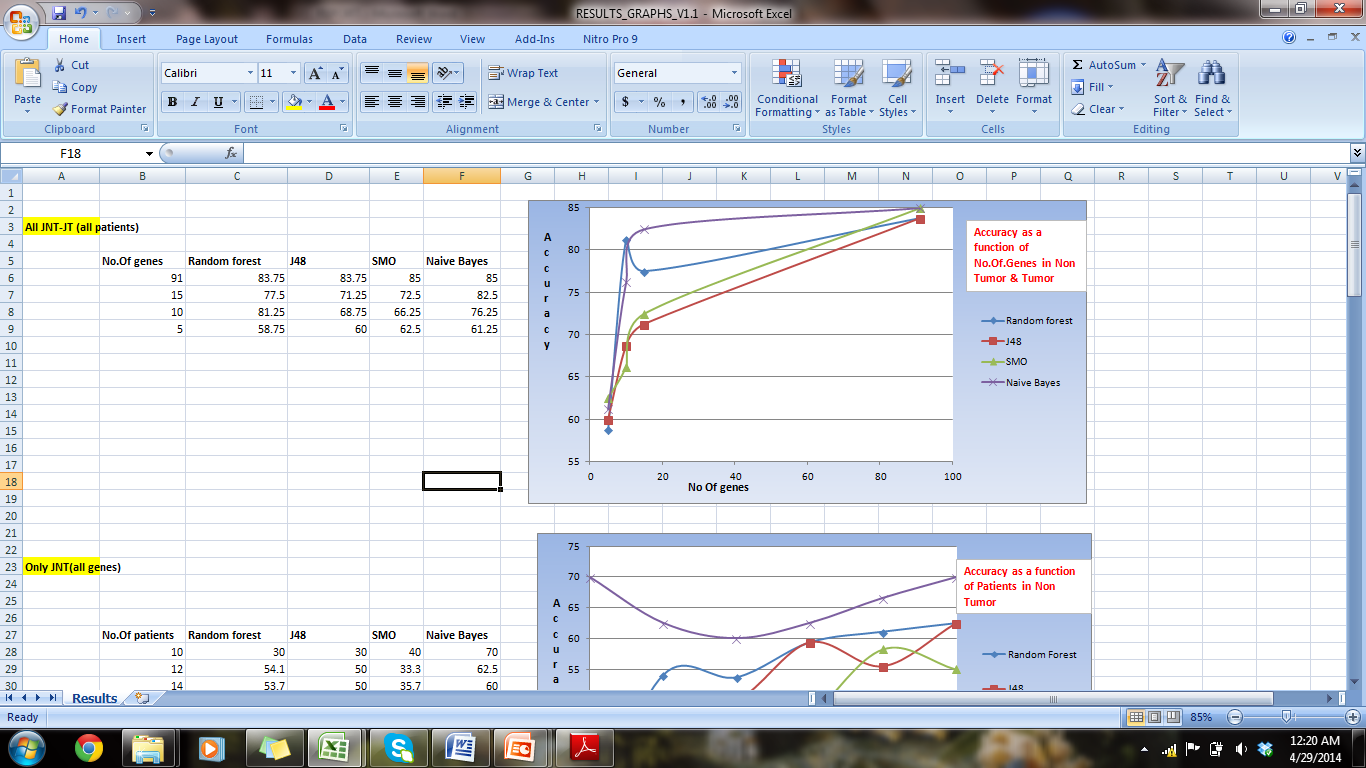
**Observations:**

The Tumor & Non-Tumor genes were taken & the classifier for this was Tumor Vs Non-Tumor. Weka was run with 4 different classifier algorithms & the %accuracy were recorded. we were unable to establish a tumor-derived predictor even with a much larger (hundreds of subjects) cohort. Although these results do not negate the value of tumor-derived expression profiling in predicting the outcomes of HCC, the data suggest that, at least in the cohort of HCV-positive HCC, tumor profiling alone may have limited value with regards to RFS

a) All Tumor & Non Tumor Patients :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No.Of genes | Random forest | J48 | SMO | Naive Bayes |
| 91 | 83.75 | 83.75 | 85 | 85 |
| 15 | 77.5 | 71.25 | 72.5 | 82.5 |
| 10 | 81.25 | 68.75 | 66.25 | 76.25 |
| 5 | 58.75 | 60 | 62.5 | 61.25 |

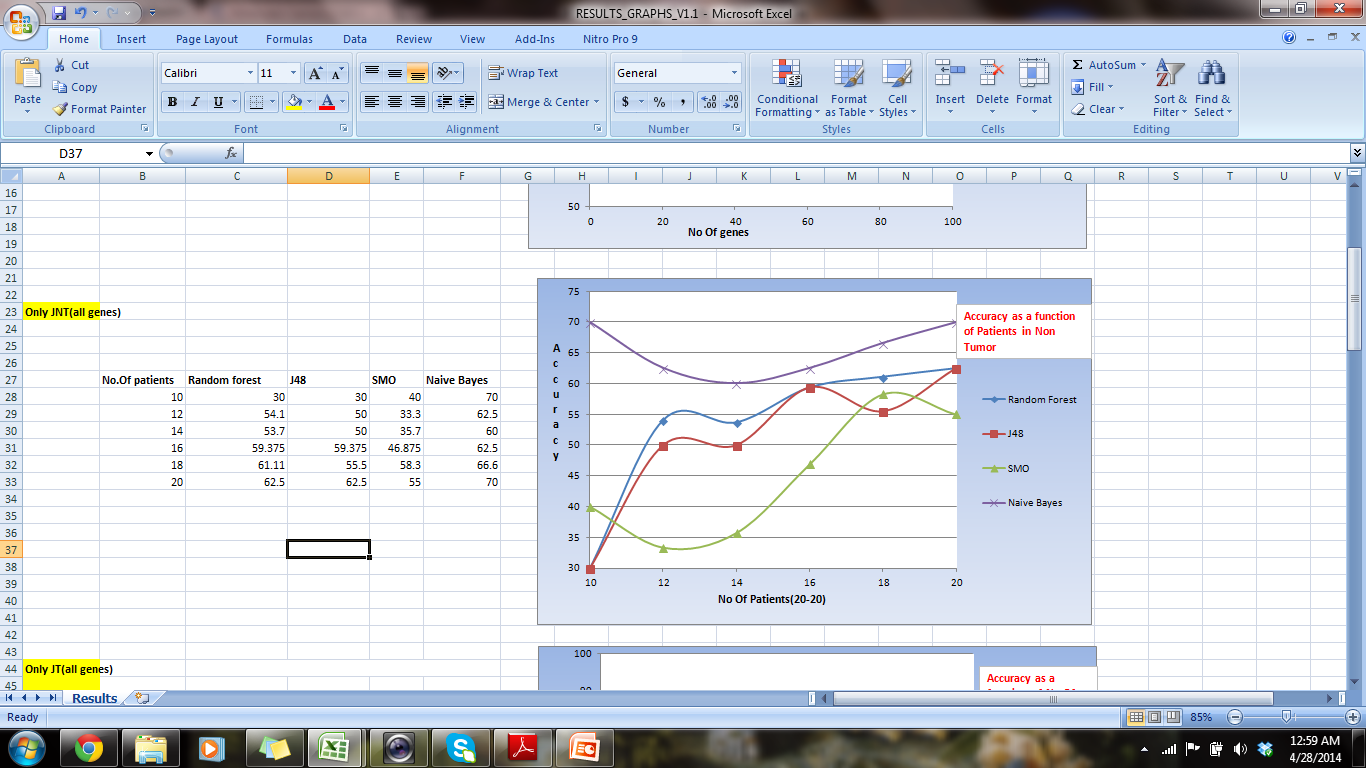
Graph:



1. Accuracy as a function of No. Of Patients (20-20) (Non-Tumor)

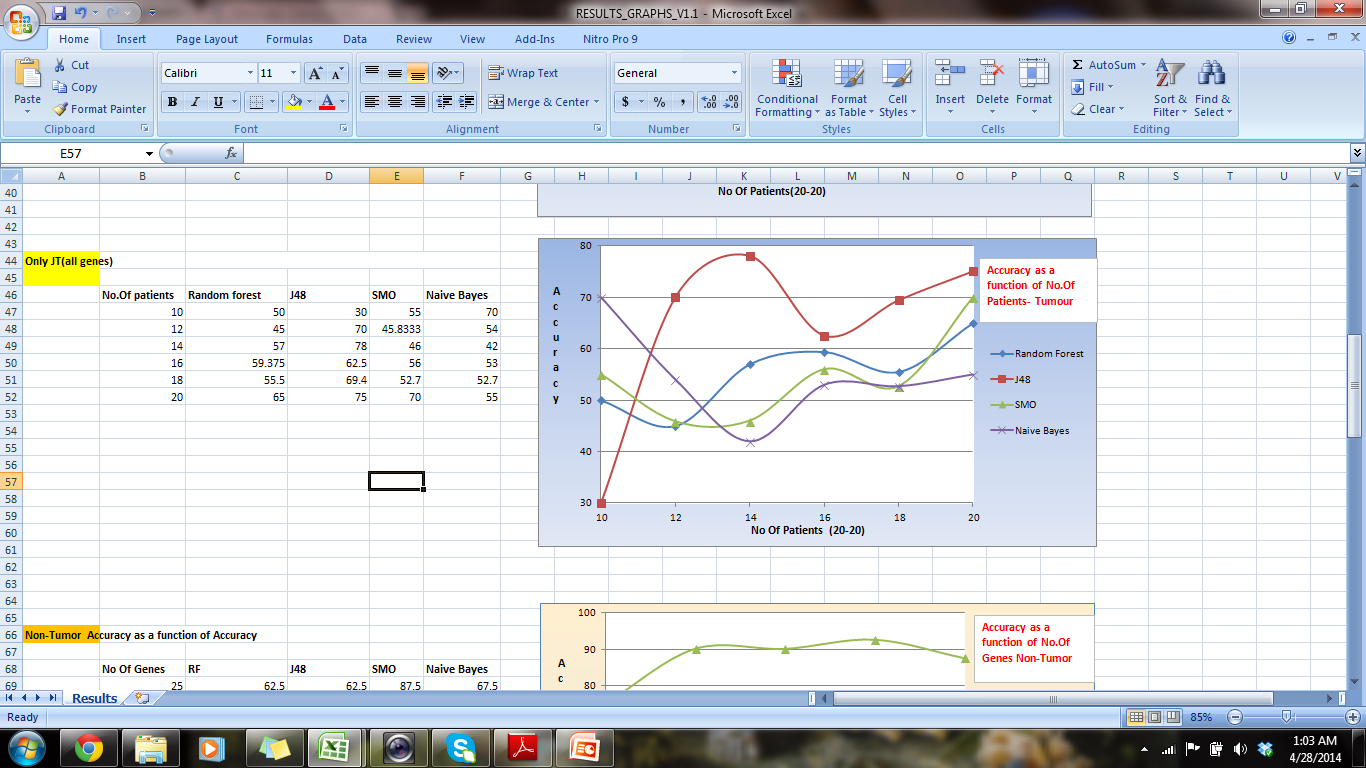
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No.Of patients | Random forest | J48 | SMO | Naive Bayes |
| 10 | 30 | 30 | 40 | 70 |
| 12 | 54.1 | 50 | 33.3 | 62.5 |
| 14 | 53.7 | 50 | 35.7 | 60 |
| 16 | 59.375 | 59.375 | 46.875 | 62.5 |
| 18 | 61.11 | 55.5 | 58.3 | 66.6 |
| 20 | 62.5 | 62.5 | 55 | 70 |

Graph :

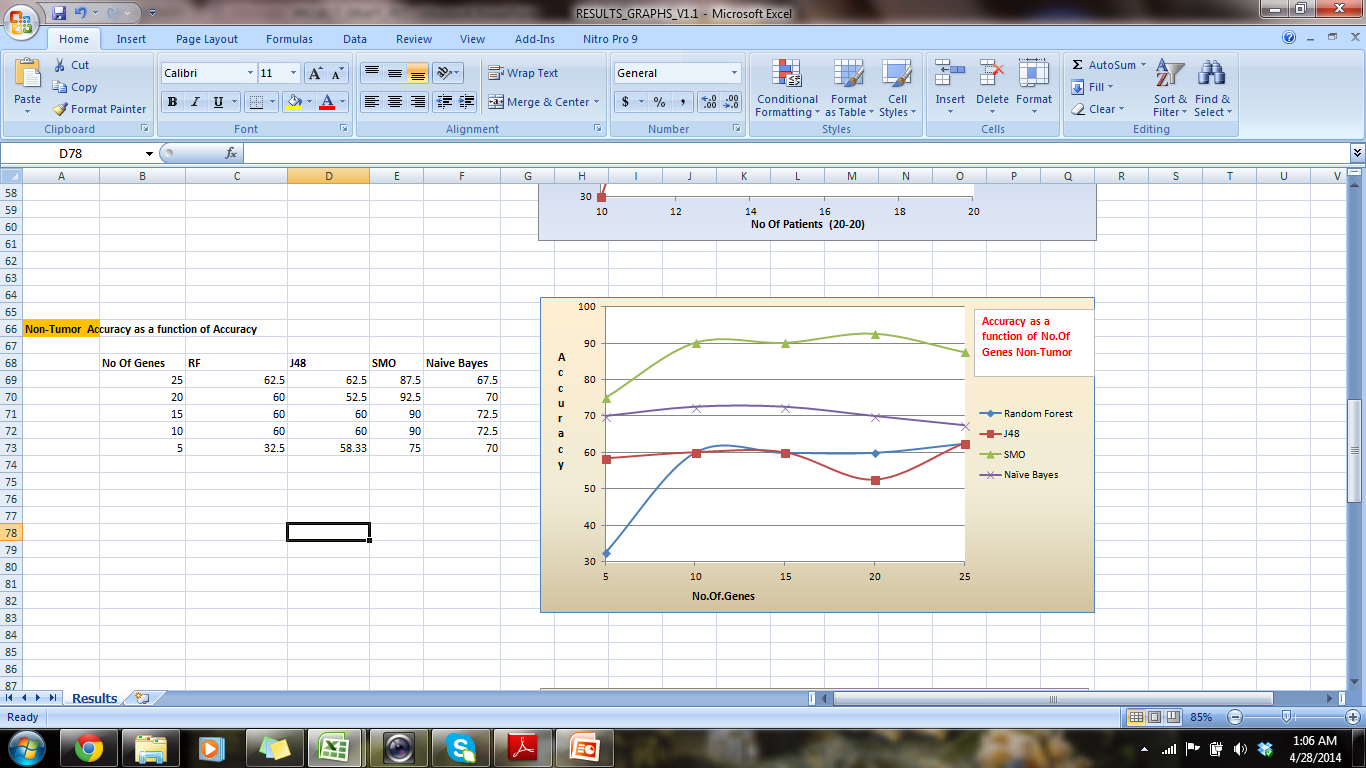


2. Accuracy as a function of No. Of Patients (20-20) (Tumor)

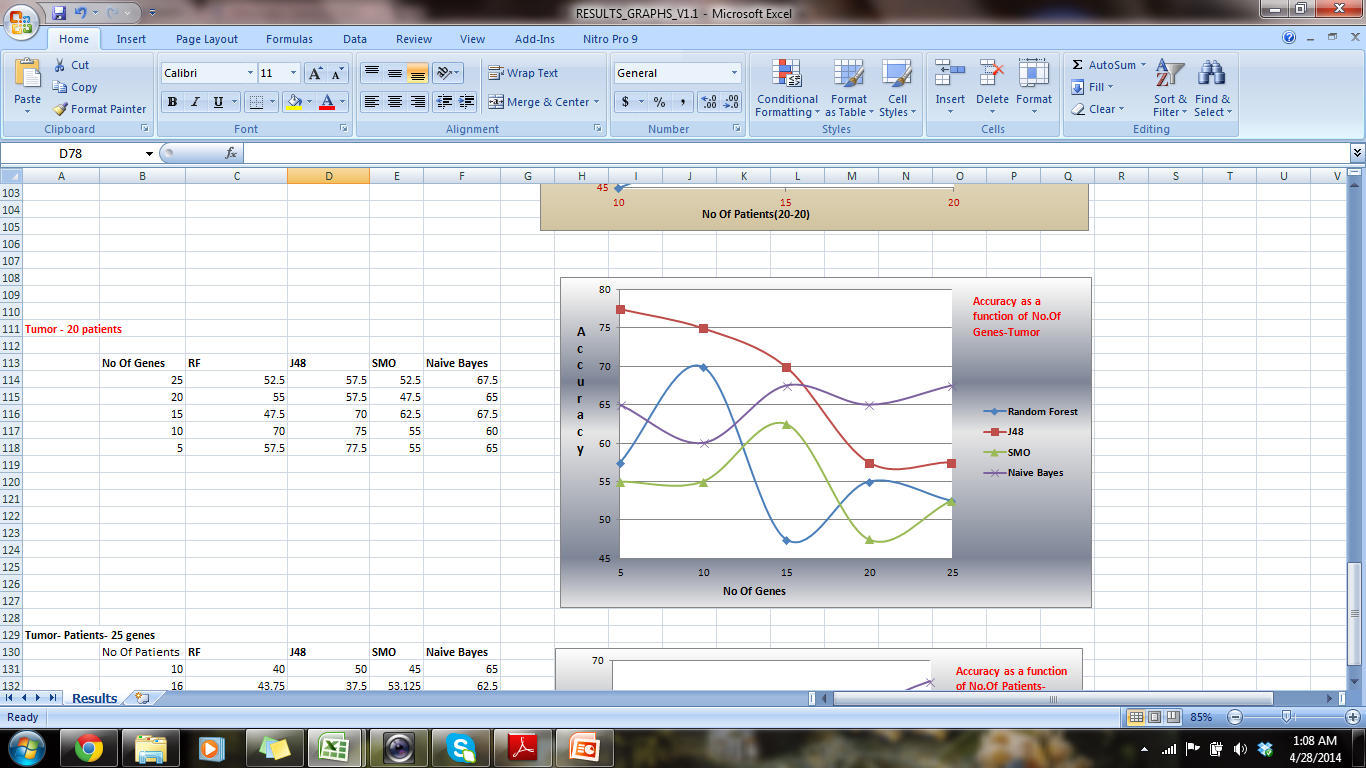
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No.Of patients | Random forest | J48 | SMO | Naive Bayes |
| 10 | 50 | 30 | 55 | 70 |
| 12 | 45 | 70 | 45.8333 | 54 |
| 14 | 57 | 78 | 46 | 42 |
| 16 | 59.375 | 62.5 | 56 | 53 |
| 18 | 55.5 | 69.4 | 52.7 | 52.7 |
| 20 | 65 | 75 | 70 | 55 |



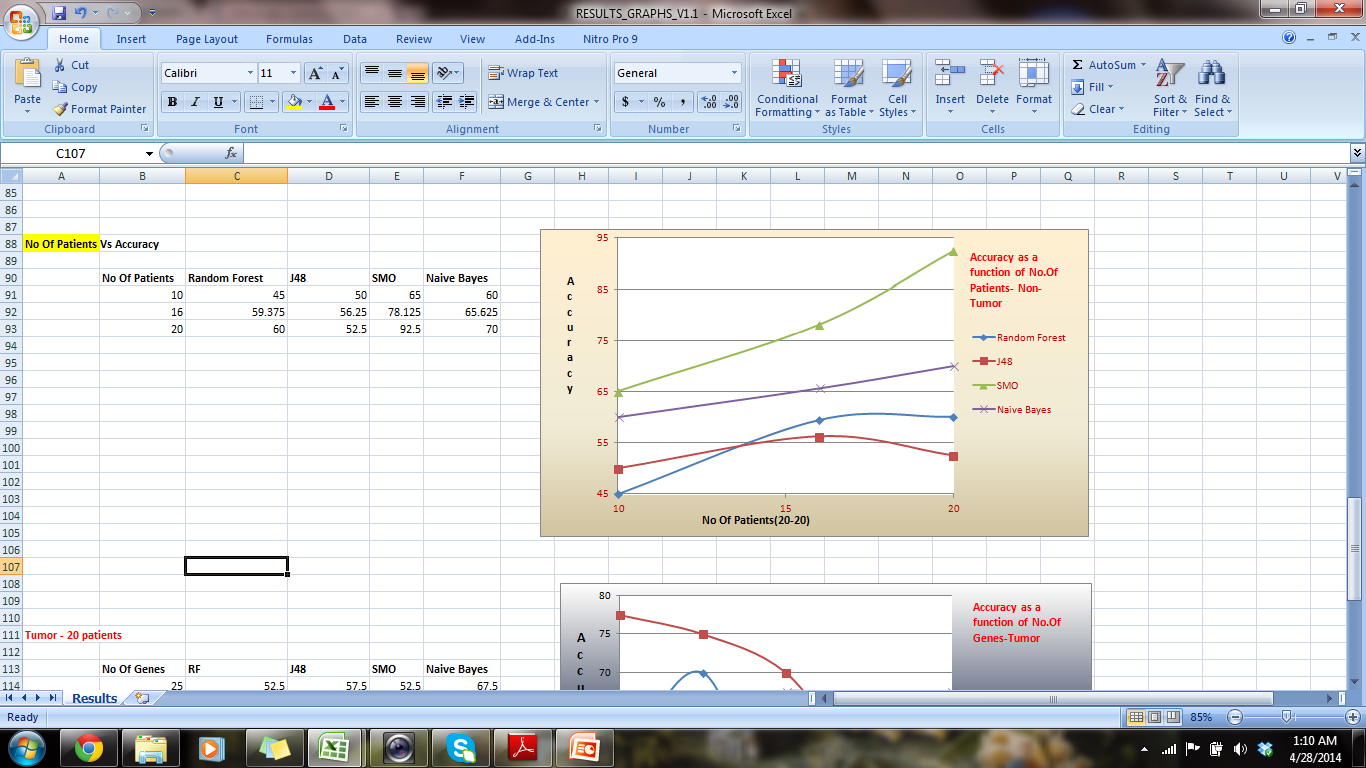
c) Accuracy as a function of No.Of Genes- Nontumor



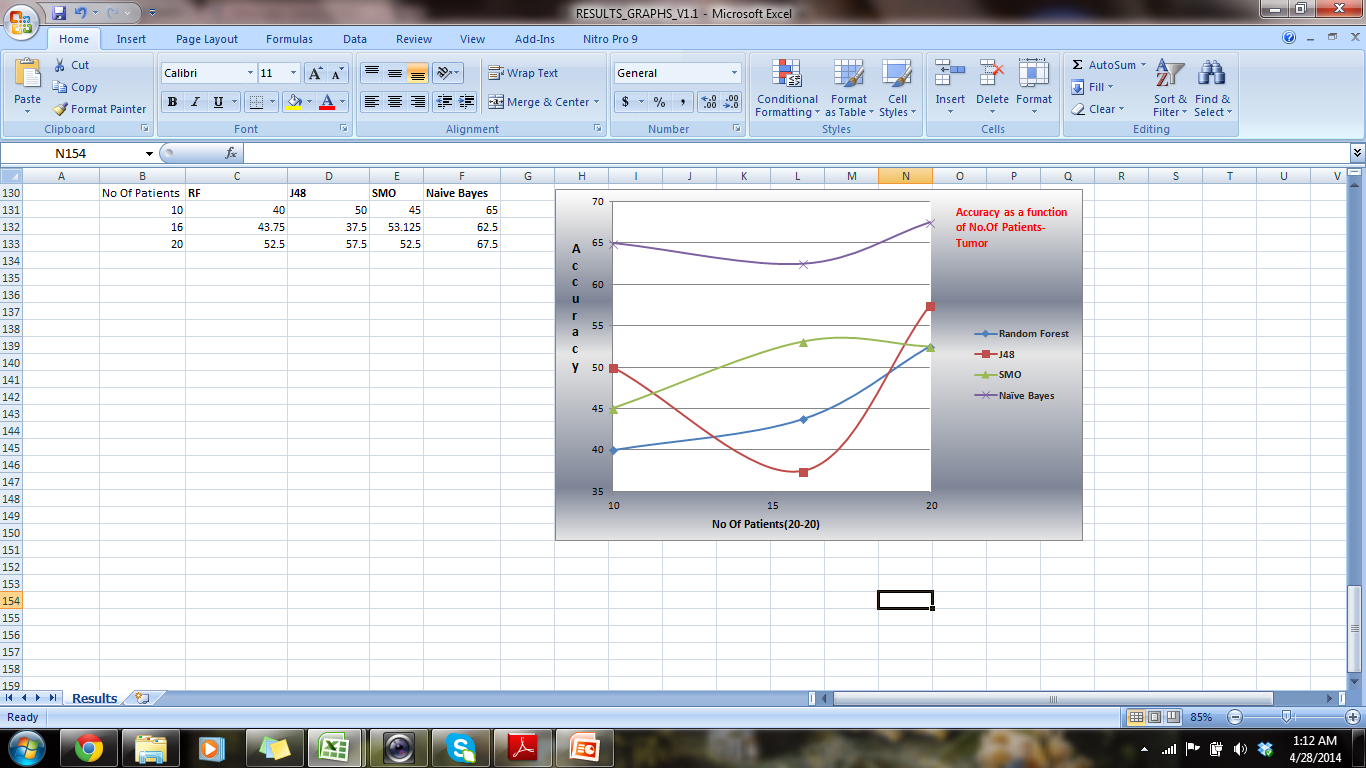
Accuracy as a function of No.Of Genes- Tumor



3) Accuracy as a function of No Of Patients - Non-Tumor



Accuracy as a function of No Of Patients - Tumor



**Note: All the graphs where the Y-axis is No. Of Patients actually mean a subset of the patients is depicted here. 20 patients subset means 20 short & 20 Long.**

**Conclusion:**

* **Tumor derived gene expression profiles did not produce a robust classifier of RFS in this project.**
* **With the accuracies, we can clearly discriminate Tumor & Non-Tumor data**
* **Based on Gene Expression profile , limited genes of 91 & 5,10, 15 taken : expression is not the same in Tumor & Non-tumor, which has no relation with survival model.**
* **The gene expression predictor may hold important insights into the patho-biology of HCC recurrence and de novo tumor formation in patients.**

Genes & Their data :

1) ABHD13 abhydrolase domain containing 13 [ Homo sapiens (human) ]

|  |  |
| --- | --- |
| Official Symbol | ABHD13provided by [HGNC](http://www.genenames.org/) |
| Official Full Name | abhydrolase domain containing 13provided by [HGNC](http://www.genenames.org/) |
| Primary source | [HGNC:20293](http://www.genenames.org/data/hgnc_data.php?hgnc_id=20293) |
| Locus tag | RP11-153I24.2 |
| See related | [Ensembl:ENSG00000139826;](http://www.ensembl.org/id/ENSG00000139826) [HPRD:12618;](http://www.hprd.org/protein/12618) [Vega:OTTHUMG00000017330](http://vega.sanger.ac.uk/id/OTTHUMG00000017330) |
| Gene type | protein coding |
| RefSeq status | VALIDATED |
| Organism | [Homo sapiens](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606) |

2) Persephin (PSPN)

|  |  |
| --- | --- |
| Summary | The protein encoded by this gene is a neurotrophic factor, belonging to the GDNF family. Neurotrophic factors are important for the proper development and maintenance of the nervous system. These factors promote neuronal survival and can prevent the neuronal degeneration associated with injury, toxin exposure, or neurodegenerative disease. The encoded protein has amino acid similarity to its other family members, glial cell line-derived neurotrophic factor and neurturin. This gene product promotes the survival of ventral midbrain dopaminergic neurons in culture and prevents their degeneration after 6-hydroxydopamine treatment in vivo. [provided by RefSeq, Jul 2008] |

3) ATAD3B ATPase family, AAA domain containing 3B [ Homo sapiens (human) ]

|  |  |
| --- | --- |
| Summary | ATAD3A (MIM 612316) and ATAD3B are mitochondrial membrane proteins that contribute to the stabilization of large mitochondrial DNA (mtDNA)-protein complexes called nucleoids (He et al., 2007 [PubMed 17210950]).[supplied by OMIM, Sep 2008] |

4) SNHG17 small nucleolar RNA host gene 17 (non-protein coding) [ Homo sapiens (human) ]

|  |  |
| --- | --- |
| Official Symbol | SNHG17provided by [HGNC](http://www.genenames.org/) |
| Official Full Name | small nucleolar RNA host gene 17 (non-protein coding)provided by [HGNC](http://www.genenames.org/) |
| Primary source | [HGNC:48600](http://www.genenames.org/data/hgnc_data.php?hgnc_id=48600) |
| See related | [Ensembl:ENSG00000196756](http://www.ensembl.org/id/ENSG00000196756) |
| Gene type | ncRNA |
| RefSeq status | PREDICTED |
| Organism | [Homo sapiens](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606) |
| Lineage | Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo |

5) PPM1A protein phosphatase, Mg2+/Mn2+ dependent, 1A [ Homo sapiens (human) ]

|  |  |
| --- | --- |
| Summary | The protein encoded by this gene is a member of the PP2C family of Ser/Thr protein phosphatases. PP2C family members are known to be negative regulators of cell stress response pathways. This phosphatase dephosphorylates, and negatively regulates the activities of, MAP kinases and MAP kinase kinases. It has been shown to inhibit the activation of p38 and JNK kinase cascades induced by environmental stresses. This phosphatase can also dephosphorylate cyclin-dependent kinases, and thus may be involved in cell cycle control. Overexpression of this phosphatase is reported to activate the expression of the tumor suppressor gene TP53/p53, which leads to G2/M cell cycle arrest and apoptosis. Three alternatively spliced transcript variants encoding distinct isoforms have been described. [provided by RefSeq, Jul 2008] |