Gene Prioritization for Inference of Robust Composite Diagnostic Signatures in the Case of Melanoma

Ioannis Valavanis, Kostantinos Moutselos, Ilias Maglogiannis, Aristotelis Chatziioannou

Gene Prioritization for Inference of robust composite diagnostic signatures in the case of Melanoma

Ioannis Valavanis[1], Kostantinos Moutselos[2], Ilias Maglogiannis[3], and Aristotelis Chatziioannou[1]*

[1]Institute of Biology, Medicinal Chemistry & Biotechnology, National Hellenic Research Foundation, Athens, Greece
[2]Department of Biomedical Informatics, University of Central Greece, Lamia, Greece
[3]University of Piraeus, Dept. of Digital Systems, Piraeus, Greece
*Corresponding author
{ivalavan, achatzi}@eie.gr
kmouts@ucg.gr, imaglo@unipi.gr

Abstract. An integrated dataset originating from multi-modal datasets can be used to target underlying causal biological actions that through a systems level process trigger the development of a disease. In this study, we use an integrated dataset related to cutaneous melanoma that comes from two separate sets (microarray and imaging) and the application of data imputation methods. Our goal is to associate low-level biological information, i.e. gene expression, to imaging features, that characterize disease at a macroscopic level. Using an average Spearman correlation measurement of a gene to a total of 31 imaging features, a set of 1701 genes were sorted based on their impact to imaging features. Top correlated genes, comprising a candidate set of gene biomarkers, were used to train an artificial feed forward neural network. Classification performance metrics reported here showed the proof of concept for our gene selection methodology which is to be further validated.

Keywords: multi-modal, microarray, gene, imaging feature, data imputation, correlation, artificial neural network

1 Introduction

The use of biomedical data from different sources, so called multi-modal datasets, is of known importance in the context of personalized medicine and future electronic health record management. Different data linked together can help towards a holistic approach. Especially, in cancer research data from clinical studies (age, sex, size or grade of tumor size) can be integrated with gene expression data from microarray experiments [1].

Integration can take place at different levels, e.g. across sub-systems (musculoskeletal, cardiovascular, etc.), or across temporal and dimensional scales (body, organ, tissue, cell) [2]. In the context of Virtual Physiological Human (VPH), an integrated framework should promote the interconnection of predictive models

pervading different scales, with different methods, characterized by different granularity. An integrated framework could produce system level information and enable formulation and testing of hypotheses, facilitating a holistic approach [3-4]. The framework should make it possible to interconnect predictive models defined at different scales, with different methods, and with different levels of detail, into systemic networks that provide a concretization of those systemic hypotheses [2,5].

An integrated framework studying multi-modal datasets can target underlying causal biological actions that through a systems level disease manifestation are translated to macroscopic disease related phenotypes. Motivated by this, in this study we aim to associate low-level biological information, i.e. gene expression, to imaging features using two different datasets related to cutaneous melanoma. The datasets used here come from two different sets of subjects that are described either by molecular features (gene expression) or imaging features. These sets have been previously used by authors in [3] to produce an integrated data set by applying data imputation methods to handle missing values in each of the sets. We actually re-use the produced dataset and our aim here is to find a robust gene signature that in whole influences the set of imaging features in the derived integrated dataset. We thus use spearman correlation measurements to derive the gene subsets that mostly affect the imaging features. The selected molecular features are then used to construct and evaluate artificial neural network classifiers that are trained to distinguish cutaneous melanoma cases from controls. Results show that the statistical selection of gene features using the multi-modal features' correlation can provide a robust signature that generalizes well when inputted to the classifiers.

2 Dataset

Two different datasets, one corresponding to microarray data and one to imaging data, were used. Since both sets are related to cutaneous melanoma, a brief introduction is firstly done in this section to the disease and then the two sets are described. Finally, the integrated dataset and how it has been produced is described.

2.1 Cutaneous Melanoma

Cutaneous Melanoma (CM) is considered a complex multigenic and multifactorial disease that involves both environmental and genetic factors. CM tumorigenesis is often explained as a progressive transformation of normal melanocytes to nevi that subsequently develop into primary cutaneous melanomas. The molecular pathways involved have been although little studied [6] and despite that genomic markers or gene signatures have been defined for various cancers (such as breast cancer), there has been no similar progress for malignant melanoma. Genomic studies that have been performed on CM exploit different microarray technological platforms applied in highly heterogeneous patient sets. These differences hurdle significantly comparisons, yielding cohorts of reduced total size and diversity.

Regarding the clinical diagnosis of melanoma, several approaches for analysis and diagnosis of lesions exist that use images for the analysis and diagnosis of lesions. The Menzies scale, the Seven-point scale, the Total Dermoscopy Score based on the

ABCD rule, and the ABCDE rule (Asymmetry, Border, Color, Diameter, Evolution) are some examples of these. As human interpretation of image content is fraught with contextual ambiguities, advanced computerized techniques can assist doctors in the diagnostic process [7].

2.2 Microarray data

The microarray dataset was taken from the Gene Expression Omnibus (GEO) [8], GDS1375. RNA isolated from 45 primary melanoma, 18 benign skin nevi, and 7 normal skin tissue specimens was used for gene expression analysis, using the Affymetrix Hu133A microarray chip containing 22,000 probe sets. Signal intensities were globally scaled so that the average intensity equals 600. The gene expression values across all categories were log transformed, and the mean values of all genes in the normal skin were calculated. Subsequently, the mean gene vector concerning the normal skin categories was subtracted from all replicate vectors of the other two categories (due to log-transformation the by normal skin category was replaced with a subtraction). The initial signal intensities provided thus ratios of differential expression, calculated by dividing the signal intensities of each category by the respective gene value of the normal category. The differentially expressed gene values of the melanoma versus skin, and nevi versus skin, were then analyzed. An FDR for multiple testing adjustment, p-value 0.001 and a 2-fold change thresholds were applied and thus 1701 genes were statistically preselected.

2.3 Imaging data

The dataset derived from skin lesion images contained 972 instances of nevus skin lesions and 69 melanoma cases. The following three types of features were analyzed: Border Features which cover the A and B parts of the ABCD-rule of dermatology, Color Features which correspond to the C rules, and Textural Features which are based on D rules [9]. A total of 31 features were produced (one feature was removed due to having zero variation across the samples). The relevant pre-processing for all features is described in [9].

2.4 Integrated Set

Microarray and imaging data sets were unified into one dataset using missing value imputation, as already described in [3]. The dataset prior to missing value imputation corresponded to a sparse matrix containing 1104 samples (benign or malignant samples, either from microarray data or imaging data) and a total of 1732 features (differential gene expression or imaging features). Prior to missing value imputation, examples originating from microarray dataset included missing values for imaging features, and examples originating from imaging dataset included missing values for gene expression measurements. A uniform missing value imputation methodology was used and a final integrated dataset (including no missing values) was produced. The procedure was followed twice, and two integrated datasets (Set 1 and Set 2) were created. One dataset (Set 1) was used for a triple scope: i) the selection of genes based on their correlation to imaging features ii) training an artificial neural network classifier to distinguish disease status (benigh vs. malignant) when inputted by selected genes or alternatively by all imaging features (see Section 4) and iii) testing

the classifier using 3-fold cross validation. Set 2 was used as an independent testing set for testing the classifier.

3 Methods

Using the pool of 1701 statistically pre-selected genes, we identified here the genes that are mostly correlated with the imaging features in whole. For correlation measurements, Spearman correlation was used (-1 implies negative correlation, 1 implies positive correlation). For a gene *i*, its correlation to an imaging feature *j* was calculated and marked as $Corr_{i,j}$ ($1 \leq i \leq 1701$, $1 \leq j \leq 31$, $-1 \leq Corr_{i,j} \leq 1$). The average values of absolute correlation measurements of gene *i* to all N=31 imaging features was used as a total correlation measurement ($Total\_Corr_i$) of gene *i* to imaging features (eq. 1). All genes were sorted in descending order according to total correlation and the most correlated genes were used as input to a feed-forward artificial neural network (ANN) that was trained and evaluated in distinguishing malignant from benign samples. Serially, the most correlated gene was used as input, than the two most correlated ones and the three most correlated ones. Top 5 and top 10 genes were also used as input, while the total set of imaging features was also used as input to the ANN for comparison reasons.

$$Total\_Corr_i = \frac{\sum_{j=1..N=31} abs(Corr_{i,j})}{N} \quad (1)$$

The ANN used here was trained using the back-propagation algorithm for 1000 epochs with a learning rate equal to 0.3 and momentum equal to 0.2. The hidden layer used sigmoid activation function and contained ((num of features+num of classes)/2 + 1) nodes. ANN was trained using Set 1 and classifier's performance in terms of total accuracy (number of samples correctly classified), and class sensitivity (number of true positives in a class that were correctly classified in this class) was measured using this set and 3-cross validation. The ANN was also trained using the whole Set 1 and was then used to classify samples in Set 2 and performance metrics were calculated as well. The classification and testing protocol was implemented within the stand-alone Rapidminer platform [10-11].

4 Results and Discussion

Table 1 presents the performance metrics (total accuracy, benign class sensitivity, malignant class sensitivity) measured when differential gene expression of various subsets of genes from the pool of 1701 were fed to ANN. Specifically, ANN was fed by the top 1, top 2, top 3, top 5 and top 10 genes according to total correlation

Table 1. Performance metrics obtained by ANN when fed by top gene(s) based on correlation to imaging features or the set of imaging features.

ANN input features (gene(s) based on correlation to imaging features or the set of imaging features) (vector dimension)	*Set 1 - Total Accuracy (3-cross validation)*	*Set 1 - Benign Class Sensitivity (3-cross validation)*	*Set 1 - Malignant Class Sensitivity (3-cross validation)*	*Set 2 – Total Accuracy*	*Set 2 - Benign Class Sensitivity*	*Set 2 - Malignant Class Sensitivity*
Top 1 gene (n=1)	96.47	97.88	83.33	95.74	99.6	62.82
Top 2 genes (n=2)	98.73	99.49	92.11	98.37	99.49	88.6
Top 3 genes (n=3)	99.46	99.9	95.61	98.82	99.6	92.11
Top 5 genes (n=5)	99.95	100	95.61	99.55	99.7	98.25
Top 10 gens (n=10)	100	100	100	99.64	99.7	99.12
Gene with the median Total_Corr value (n=1)	89.58	93.13	58.77	93.48	100	36.84
Gene Least correlated (n=1)	83.79	87.98	47.37	89.67	100	0
Imaging Features (n=31)	59.69	61.62	42.98	89.67	100	0

measurements to imaging features as described above (Table 1, Rows 2-6). Performance of ANN when inputted by the worst gene according to total correlation measurement, the gene featured the median total correlation measurement (sorted as top 50% in the sorted gene list) and the total of imaging features are reported as well for comparison reasons (Table 1, Rows 7-9) .

Results show that top genes can provide very good performance metrics and when serially adding top genes performance gets better. Eventually, almost all samples can be classified correctly when top 10 genes are used and this happens also for Set 2 that was not used in training process (see further discussion below). In general, little worst performance is obtained when ANN is evaluated in Set 2, while sensitivity measurements for malignant class are worse than the corresponding ones for benign class. This has to do with the much greater abundance of benign samples in the integrated dataset. The performance obtained when genes less correlated to imaging features are fed to ANN are much lower, showing the proof of concept for selecting gene features by taking into account their impact to imaging features. Results in Table 1 show also that the performance metrics obtained here by the top genes in terms of their correlation to imaging features are much higher than the ones obtained when imaging features are fed to the ANN. This shows that selected genes, actually being involved in the biological actions beneath melanoma phenotype, could comprise a molecular signature and a potential set of molecular biomarkers/predictors for the disease. This feature set describing low-level biomedical information seems to perform better than the set of macroscopic imaging features, but of course this is to be cross-validated by further tests.

It is to be noted that performances presented here may comprise over estimations of the ANN behavior and predicting ability. This has to do with the fact that similar patterns of features may exist within Set 1 and Set 2 or across these two sets, since missing data imputation has taken place to a great extent as regards the signal population of the integrated dataset (features and disease phenotype). This could not be avoided since the integrated dataset has originated from two separate datasets (microarray and imaging), while a multi-modal dataset based on a single set of subjects forming an epidemiological cohort yet remains elusive to the best of our knowledge. However, further cross-validation tests and the application of more missing value imputation methods represent tangible goals for future work.

References

1. Martin, C., Deters, H.G., Nattkemper, T.W.: Fusing Biomedical Multi-modal Data for Exploratory Data Analysis. ;In ICANN (2) 798-807 (2006)
2. Viceconti, M., Clapworthy, G., Testi, D., Taddei, F., and McFarlane, N. (2010). Multimodal fusion of biomedical data at different temporal and dimensional scales. Comp. Mtds and Progs Biomed, 102(3):227-237 (2010)
3. Moutselos, K., Chatziioannou, A., Maglogiannis, I.: Feature Selection Study on Separate Multi-modal Datasets: Application on Cutaneous Melanoma. ;In AIAI (2) 36-45, (2012)
4. Fenner, J.W., Brook, B., Clapworthy, G., Coveney, P.V., Feipel, V., Gregersen, H., Hose, D.R., Kohl, P., Lawford, P., McCormack, K.M., Pinney, D., Thomas, S.R., Van Sint Jan, S.,

Waters, S., Viceconti, M.: The EuroPhysiome, STEP and a roadmap for the virtual physiological human. Philos. Transact. A Math. Phys. Eng. Sci. 366, 2979–2999 (2008)
5. STEP Consortium. Seeding the EuroPhysiome: A Roadmap to the Virtual Physiological Human. (online) 5 July 2007, http://www.europhysiome.org/roadmap
6. Balázs, M., Ecsedi, S., Vízkeleti, L. et al., "Genomics of Human Malignant Melanoma " Breakthroughs in Melanoma Research, Breakthroughs in Melanoma Research Y. Tanaka, ed., InTech, (2011).
7. Ogorzałek M., Nowak L, Surowka G. et al., "Modern Techniques for Computer-Aided Melanoma Diagnosis," Melanoma in the Clinic - Diagnosis, Management and Complications of Malignancy, Melanoma in the Clinic - Diagnosis, Management and Complications of Malignancy M. Murph, ed., InTech, (2011).
8. Barrett T., Troup D.B, Wilhite S.E. et al., "NCBI GEO: archive for functional genomics data sets - 10 years on," Nucleic Acids Res, vol. 39, no. Database issue, pp. D1005-10, Jan, 2011.Chevalier, R.L.: Obstructive nephropathy: towards biomarker discovery and gene therapy. Nat. Clin. Pract. Nephrol. 2(3), 157--168 (2006).
9. Maragoudakis M., and Maglogiannis I., "Skin lesion diagnosis from images using novel ensemble classification techniques," in 10th IEEE EMBS International Conference on Information Technology Applications in Biomedicine, Corfu, Greece (2010)
10. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06) (2006)
11. http://rapid-i.com/