Inference of a robust diagnostic signature in the case of Melanoma: Gene Selection by Information Gain and Gene Ontology Tree Exploration

Ioannis Valavanis, *Member, IEEE*, Konstantinos Moutselos, Ilias Maglogiannis, *Member IEEE*, and Aristotelis A. Chatziioannou, *Member, IEEE*

Abstract— **Integrated datasets originating from multi-modal data can be used towards the identification of causal biological actions that through a systems level process trigger the development of a disease. We use, here, 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 select a subset of genes that comprise candidate biomarkers and compare these to imaging features, that characterize disease at a macroscopic level. Using information gain ratio measurements and exploration of Gene Ontology (GO) tree, we identified a set of 33 genes both highly correlated to the disease status and with a central role in regulatory mechanisms. Selected genes were used to train various classifiers that could generalize well when discriminating malignant from benign melanoma samples. Results showed that classifiers performed better when selected genes were used as input, rather than imaging features selected by information gain measurements. Thus, genes in the backstage of low-level biological processes showed to carry higher information content than the macroscopic imaging features.**

Keywords: multi-modal, microarray, gene, imaging feature, data imputation, information gain, classification, gene ontology tree

I. Introduction

Cutaneous Melanoma (CM) is a complex multifactorial disease involving both environmental and genetic factors [1]. Tumorigenesis of this cancer is explained as a progressive transformation of normal melanocytes to nevi that develop into primary cutaneous melanomas. Despite that genomic markers or gene signatures have been defined for various cancers (such as breast cancer), there has been no significant progress for malignant melanoma. One reason for this is that genomic studies that have been performed on CM exploit different microarray technological platforms applied in highly heterogeneous patient sets.

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 can be subjective, advanced computerized techniques can aid clinicians in the diagnostic process [2].

Regarding multifactorial diseases like CM, 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. The use of different data linked together can help towards a holistic approach of the disease, and for the evaluation and comparison of various subsets of markers (genetic/environmental factors, imaging features). Integration can take place at different levels, e.g. across sub-systems, or across temporal and dimensional scales (body, organ, tissue, cell) [3]. 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. The final goal would be the interconnections of predictive models defined at different scales into systemic networks that provide a holistic description of a disease [3-4]. In the case of cancer research, data from clinical studies (age, sex, size or grade of tumor size) can be integrated with gene expression data from microarray experiments [5].

An integrated framework studying multi-modal datasets can target underlying causal biological actions that through a systems level disease manifestation are translated to the disease phenotype. Motivated by this, we aim here to associate low-level biological information, i.e. gene expression, to CM disease status and compare the information content of genes to that of macroscopic CM disease descriptors, i.e. imaging features. Towards our goals, we use 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. Data has been previously used by authors in [6] to produce an integrated data set by applying data imputation methods to

Manuscript received July 30, 2013. This work was supported by the Entrepreneurial Program "Competitiveness and Entrepreneurship" Action COOPERATION entitled "PIK3CA Oncogenic Mutations in Breast and Colon Cancers: Development of Targeted Anticancer Drugs and Diagnostics" [POM].

I. Valavanis and A.Chatziioannou (corresponding author) are with the Institute of Biology, Medicinal Chemistry & Biotechnology, National Hellenic Research Foundation, Athens, Greece (phone: +30-210-7273768; fax: +30-210-7273751; e-mail: achatzi@eie.gr). I. Maglogiannis is with the Department of Digital Systems, Piraeus, Greece. K. Moutselos is with the Department of Biomedical Informatics, University of Thessaly, Lamia, Greece.

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

handle missing values in each of the sets. We actually use several instances of the produced dataset by the application of various data imputation methods. Using statistical entropy-based methods along with functional analysis, we object a gene signature for cutaneous melanoma, while imaging features selected again statistically are compared to the selected genes subset in terms of their information content. Comparison is done by feeding a series of classifiers with the selected features subsets. Results show the superiority of genetic features within the task of discriminating malignant from benign samples from the CM dataset used here.

II. Dataset

The integrated dataset used here comes from two different datasets (microarrays and imaging) that are described below. Then, the methodology for producing the dataset using data imputation methods is described.

A. Microarrays Data and Pre-Processing

The microarray dataset was found in the Gene Expression Omnibus (GEO) [6][7], 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. After global normalization and log transformation of gene expression values, the mean gene vector concerning the normal skin categories was subtracted from all replicate vectors of the other two categories. Thus, ratios of differential expression were provided, 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.

B. 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 calculated: 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 [8]. 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 [8].

C. Integrated Dataset

Microarray and imaging data sets were unified into single datasets using missing value imputation, as already described in [6]. 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. Two algorithms were used for missing value imputation: uniform data imputation and bootstrap data imputation [6]. Each algorithm was applied three times and a total of six (6) datasets were derived. All of the derived datasets are used as described in the methodology sections with a double scope: select statistically genes or imaging features and evaluate the selected subsets of features when inputted to classification algorithms.

III. Methods

A. Features Selection (Genes/Imaging) based on Functional Analysis and Entropy-Based Information Gain Estimation

Gene Ontology (GO) was used in order to explore the underlying functional content beneath the set of 1701 genes differentially expressed between the benign and malignant melanoma samples. To this end, the GORevenge algorithm [9], freely available in the web, was used. GORevenge exploits graph-theoretical algorithmic methodologies and systematically exploits the GO tree in order to output a series of functionally related genes/GO terms that may not be included in the list of genes/GO terms initially submitted to the algorithm. Thus, it can aid the elucidation of hidden functional regulatory effects among genes and can therefore promote a system's level interpretation. GORevenge was applied twice, once for the exploring GOs related to Molecular Function (MF) and once for GOs related to Biological Process (BP) aspect and was inputted by the set of 1701 genes. Resnik semantic similarity metric and the Bubble genes algorithm were utilized with a relaxation parameter equal to 0.30 [9]. Only the genes that were both outputted by GORevenge and were simultaneously contained in the set of diffentially expressed genes were selected for further processing.

To further increase our confidence in the context of the systemic analysis, the entropy-based information gain (IG), also known as Kullback–Leibler divergence [10], was also calculated for the set of diffentially expressed genes in respect to disease status (malignant vs. benign). Specifically, genes were ranked based on their IG ratio values [11], measured as the mean value of IG ratios in all six datasets (3 uniform and 3 bootstrap) derived by two data imputation methods as already described. Genes with an IG ratio value in the top 20% (340 genes) were selected as the most informative ones in terms of variability.

Genes with relatively high IG and genes with a central regulatory role in the underlying active molecular networks (after applying the GORevenge algorithm) represent genes with mutually independent (orthogonal) characteristics. Therefore, an intersection of the high IG valued gene set with the functionally related GORevenge gene sets may possibly reveal the highly informative critical molecular players involved in the ageing process in both genders. The two GORevenge output lists (MF, BP) were, thus, intersected with genes that presented high IG and genes found comprised the gene signature presented here.

The set of 31 imaging features was also evaluated in terms of IG ratio of the features in respect to disease status (malignant vs. benign). The selected imaging features were then compared to the selected genes subset in terms of their total informational content when discriminating malignant from benign samples.

B. Classification

The resulted genes subset and imaging features subset were tested for their classification potency (malignant vs. benign) using a series of popular classifiers. For each of the two subsets, four classifiers were constructed and evaluated in terms of generalization using separately all six datasets (3 uniform and 3 bootstrap) derived by the two data imputation methods used here. Specifically, two weighted *k*-nearest neighbor (*k*-NN) classifiers (*k*=1, 30) [12], a Decision Tree (DT) [13], and the Random Forest (RF) algorithm [14] were used. Their performance was measured using 3-cross validation resampling. Sensitivity measurements (True Positive/(True Positive + False Negative) for each of the two classes were calculated and used to compare the selected genes subset and imaging features subset. For our classification purposes, appropriate workflows were built using the Rapidminer platform [15-16].

IV. Results and Discussion

GORevenge analysis (MF and BP aspects) resulted in total of 179 genes found in the original list of 1701 differentially expressed genes. Out of this subset of genes,

TABLE I
GORevenge most Important Genes, corresponding to the 'Biological Process' Or Molecular Function Aspect, with a mean IG ratio in the top 20%

Gene	IG	Gene	IG
MAP3K4	0.943	STXBP1	0.838
CTSB	0.928	NOD2	0.829
PTPRF	0.924	ZBTB16	0.826
BCL2L1	0.921	AKR1C3	0.825
PARP1	0.912	AKT1	0.825
EGFR	0.870	PSEN2	0.825
GJA1	0.866	ADRB2	0.824
BBS4	0.862	CCND2	0.821
DMD	0.856	SCARB1	0.821
ITGAV	0.851	PBX1	0.819
CDK5	0.847	AKR1C1	0.816
CDK2	0.845	CSK	0.816
VAV3	0.843	PSMC4	0.814
AURKA	0.841	PCNA	0.814
STEAP3	0.839	FOXO1	0.812

TABLE II
Mean IG ratios of all 31 imaging features

FEATURE	IG	FEATURE	IG
I.mean	**0.340**	Asymetry	0.084
L.mean	**0.334**	Distance.std	0.080
B.std	**0.320**	std.B	0.073
mean.R	**0.310**	H.mean	0.066
Area	**0.302**	GMSM.mean	0.064
A.mean	**0.296**	GLSM.std	0.062
std.G	0.171	ASM	0.053
Complexity	0.161	A.std	0.051
Dissimilarity	0.159	H.std	0.040
mean.G	0.152	Perimeter	0.039
Grad.mean	0.136	I.std	0.037
S.mean	0.135	Grad.std	0.029
B.mean	0.126	Grad.max	0.027
std.R	0.119	Eccentricity	0.024
mean.B	0.118	L.std	0.023
S.std	0.114		

Top 6 are the imaging features selected and inputted to the classifiers

TABLE III
Sensitivity measurements (%) for malignant and benign classes for classifiers inputted by 33 selected genes (3-cross validation resampling)

Set	1-NN Benign Sensitivity	1-NN Malignant Sensitivity	30-NN Benign Sensitivity	30-NN Malignant Sensitivity	DT Benign Sensitivity	DT Malignant Sensitivity	RF Benign Sensitivity	RF Malignant Sensitivity
bootstrap 1	100.00	98.25	100.00	96.49	100.00	95.61	100.00	97.37
bootstrap 2	100.00	96.49	100.00	96.49	100.00	94.74	100.00	97.37
bootstrap 3	100.00	99.12	100.00	96.49	100.00	93.86	100.00	99.12
uniform 1	100.00	98.25	100.00	96.49	100.00	93.86	100.00	98.25
uniform 2	100.00	99.12	100.00	96.49	100.00	100.00	100.00	99.12
uniform 3	100.00	99.12	100.00	96.49	100.00	93.86	100.00	99.12

TABLE IV
Sensitivity measurements (%) for malignant and benign classes for classifiers inputted by 6 selected imaging features (3-cross validation resampling)

Set	1-NN Benign Sensitivity	1-NN Malignant Sensitivity	30-NN Benign Sensitivity	30-NN Malignant Sensitivity	DT Benign Sensitivity	DT Malignant Sensitivity	RF Benign Sensitivity	RF Malignant Sensitivity
bootstrap 1	94.34	50.88	100.00	0.00	95.56	37.72	99.90	1.75
bootstrap 2	94.14	36.84	100.00	0.00	93.94	50.88	100.00	0.00
bootstrap 3	94.14	50.88	100.00	0.00	94.55	45.61	99.80	0.00
uniform 1	99.80	0.00	100.00	0.00	95.25	31.58	100.00	0.00
uniform 2	94.14	36.84	100.00	0.00	93.94	50.88	100.00	0.00
uniform 3	94.14	50.88	100.00	0.00	94.55	45.61	99.80	0.00

33 genes were featured a mean IG ratio in the top 20% of all 1701 genes. These genes, presented in Table I, are featured a high information content measured with entropy-based statistics (mean IG ratio>0.81) and were also found to a have a central regulatory role in the GO tree as given by GORevenge analysis.

The imaging features were found to be less associated with disease status (malignant vs. benign) than genes according to IG ratios (Table II). The top 6 imaging features (mean IG ratio>0.29, for all other features a mean IG ratio<0.2 was measured) were selected as the ones to comprise an imaging features subset to be compared to the 33 genes selected by IG measurements and GORevenge analysis. Thus, both subsets were inputted to the classifiers used here. Sensitivity results for the two classes (malignant, benign) obtained for all six instances of the integrated dataset here are presented in Tables III and IV for genes and imaging features, respectively.

Results in Table III show that top genes can provide very good performance metrics. Almost all samples can be classified when selected genes are inputted to the classifiers, with 1-NN and RF classifier being the most well performing. Sensitivity measurements for malignant class are little worse than the corresponding ones for benign class and this is related to the much greater abundance of benign samples in the integrated dataset. When combining results in Table III and IV, it is obvious that performance metrics obtained here by the top genes are much higher than the ones obtained when imaging features are fed to the classifiers. Regarding the use of imaging features and classifiers' performance, the much greater abundance of benign samples in the integrated dataset appears to have a great effect here. Thus, and along with the lower information content of imaging features, it makes classifiers perform moderately when recognizing malignant samples (maximum malignant class sensitivity was found 51%).

The selected genes presented here, 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. They were obtained from a two step process, involving entropy based statistics and the exploration of regulatory mechanisms, and their potential role as biomarkers can be supported by the performance of classifiers when inputted by the corresponding expression values. It is worth noting that the genes subset presented here, describing low-level biomedical information, performs better than the set of macroscopic imaging features. This can of course be cross-validated by further tests. It has to be noted that performances presented here may comprise over estimations of the classifiers' predicting ability. This has to do with the fact that similar patterns of features may exist within the instances of the integrated dataset used here, since missing data imputation has taken place to a great extent as regards the signal population of the integrated dataset (genes, imaging features and disease phenotype).

Future work will include further tests to support the generalization ability of classifiers using the genes presented for discriminating malignant from benign samples. Gene enrichment analysis and the identification of enriched processes (based on GO and/or KEGG pathways), that will enable a system's level description of melanoma development, are also within our future plans.

References

[1] C.R. Rossi, M. Foletto, A. Vecchiato, S. Alessio, N. Menin, M. Lise, "Management of cutaneous melanoma M0: state of the art and trends", Journal of Cancer, Volume 33, Issue 14, 1997, pp. 2302-2312

[2] M. Ogorzałek, L. Nowak L, G. 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

[3] M. Viceconti, G. Clapworthy, D. Testi,, F. Taddei, N. McFarlane, "Multimodal fusion of biomedical data at different temporal and dimensional scales", Comp. Mtds and Progs Biomed, 2010,102(3):227-237

[4] STEP Consortium. Seeding the EuroPhysiome: A Roadmap to the Virtual Physiological Human. (online) 5 July 2007, http://www.europhysiome.org/roadmap

[5] C. Martin, H.G. Deters, T.W. Nattkemper, "Fusing Biomedical Multi-modal Data for Exploratory Data Analysis", In ICANN (2), 2006, 798-807

[6] K. Moutselos, A. Chatziioannou, I. Maglogiannis, "Feature Selection Study on Separate Multi-modal Datasets: Application on Cutaneous Melanoma", In AIAI, 2012, (2) 36-45

[7] T. Barrett, D.B. Troup, S.E. Wilhite et al., "NCBI GEO: archive for functional genomics data sets - 10 years on" Nucleic Acids Res, vol. 39, 2011, pp. D1005-10

[8] M. Maragoudakis, and I. Maglogiannis, "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, pp. 1-5

[9] K. Moutselos, I. Maglogiannis, I., A. Chatziioannou, "GOrevenge: A novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies", IEEE Trans Biomed Eng,. 2011;21846603

[10] S. Kullback, R. A. Leibler, "On Information and Sufficiency", Annals of Mathematical Statistics 22 (1), 1951, pp. 79–86.

[11] D. Jianhua, X. Qing, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification", Applied Soft Computing, Vol. 13 (1), 2013, pp. 211-221

[12] Y.Wu, K. Ianakiev and V. Govindaraju, "Improved k-nearest neighbor classification Pattern Recognition", 35, 2002, 2311-2318.

[13] L. Breiman, J.H. Friedman, R.A. Olson and C.J. Stone "Classification and Regression Trees", Wadsworth, Belmont, California, 1984.

[14] L. Breiman, "Random Forests", Machine Learning 45 (1), 2001, 5–32

[15] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, 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

[16] http://rapid-i.com/