sciendo

Int. J. Appl. Math. Comput. Sci., 2019, Vol. 29, No. 1, 169–178
DOI: 10.2478/amcs-2019-0013

amcs

MACHINE LEARNING TECHNIQUES COMBINED WITH DOSE PROFILES INDICATE RADIATION RESPONSE BIOMARKERS

ANNA PAPIEZ [a], CHRISTOPHE BADIE [b], JOANNA POLANSKA [a,*]

[a] Data Mining Group, Institute of Automatic Control
Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland
e-mail: joanna.polanska@polsl.pl

[b] Cancer Mechanisms and Biomarkers, Radiation Effects Department
Centre for Radiation, Chemical & Environmental Hazards, Public Health England
Chilton, Didcot, Oxfordshire OX11 ORQ, UK

The focus of this research is to combine statistical and machine learning tools in application to a high-throughput biological data set on ionizing radiation response. The analyzed data consist of two gene expression sets obtained in studies of radiosensitive and radioresistant breast cancer patients undergoing radiotherapy. The data sets were similar in principle; however, the treatment dose differed. It is shown that introducing mathematical adjustments in data preprocessing, differentiation and trend testing, and classification, coupled with current biological knowledge, allows efficient data analysis and obtaining accurate results. The tools used to customize the analysis workflow were batch effect filtration with empirical Bayes models, identifying gene trends through the Jonckheere–Terpstra test and linear interpolation adjustment according to specific gene profiles for multiple random validation. The application of non-standard techniques enabled successful sample classification at the rate of 93.5% and the identification of potential biomarkers of radiation response in breast cancer, which were confirmed with an independent Monte Carlo feature selection approach and by literature references. This study shows that using customized analysis workflows is a necessary step towards novel discoveries in complex fields such as personalized individual therapy.

Keywords: machine learning, gene profiling, radiation response, multiple random validation, transcription.

1. Introduction

Contemporary molecular biology implies the need for developing solutions for efficient data analysis due to the constantly growing amounts of data gathered in high throughput experiments. Simple and standard statistical procedures, though serving as a basis for processing workflows, are often no longer valid approaches in view of the complexity of experimental designs and high data dimensionality. Machine learning and statistical modeling are fields which are constantly being developed in service of biomarker detection. Machine learning techniques successfully enhance knowledge discovery in cancer research (Parmar *et al.*, 2015; Jagga and Gupta, 2015), radiotherapy adaptation (Guidi *et al.*, 2017; Fargeas *et al.*, 2015), and integrative studies aid exploration throughout transcriptomics (and other omics) data sets (Francescatto *et al.*, 2018).

One of the areas with mathematical modeling constantly in progress is radiation research, especially when it comes to health risks and opportunities. Ionizing radiation is an omnipresent factor, which has significant impact on many aspects of human life. Small doses are absorbed on a daily basis while using everyday equipment such as radios or microwave ovens, whereas higher doses occurring during accidents may have very detrimental effects (Abbott, 2015). However, high doses used under controllable conditions carry beneficial effects, i.e., they are used widely for therapeutic purposes. In fact, medical procedures such as X-ray imaging or radiotherapy constitute the main source of man-made radiation exposure (Ray *et al.*, 2012). For instance, 2 Gy of ionizing radiation, which is classified as a high dose, is a commonly used fraction of a total dose given to the patient during radiation therapy in various types of

*Corresponding author

cancer (Joiner, 2004). This is a standard, although, it is known that radiosensitivity is a trait specific for each individual and depending on its level the reaction to radiotherapy may be extremely different. Radiosensitive patients obtaining too high doses more frequently than required have a high chance of developing late adverse effects, while for radioresistant individuals the standard procedure may be insufficient for the healing effect to progress. Therefore, there is a pressing need for a thorough understanding of the processes underlying radiation response in the field of dosimetry, for enabling therapy personalization.

Many studies have been conducted in the domain of radiation research. These experiments are often costly and time consuming and that is why it is crucial to extract information entirely efficiently in single studies, but also to make the most of combining information from already available data and knowledge. In this work we examine two datasets from breast cancer patient samples. In one experiment these blood samples were treated with a therapeutic dose of 2 Gy and in the other with a high dose of 4 Gy. While this experimental setting may be useful to conduct differentiation analyses, at the same time it could pose a problem when attempting to merge data sets, for instance, in classification tasks, as the doses differ. Such factors may greatly affect results of analyses when not taken under consideration. However, current biological expertise allows us to assume a linear model of dependency between radiation dose and gene expression (Brenner *et al*., 2003). Although uncertainty persists when it comes to low radiation doses absorbed (Mullenders *et al*., 2009) (Fig. 1), 2 and 4 Gy doses are both classified as high according to UNSCEAR (UNSCEAR, 2000). These facts enable the use of appropriate mathematical tools, i.e., linear interpolation, to estimate and unify the expression values to correspond in both cases to a 2 Gy dose.

By means of dose response profiles and overrepresentation *in silico* analysis, the study provides guidance for insight into the processes activated and inhibited along with increased high doses of radiation used for medical purposes. This research addresses the task of biomarker identification by means of integrative transcriptomics data analysis. The non-standard statistical techniques include differentiation analysis between doses with respect to the response patterns and applying these dose profiles as filters in classification. The combination of statistical differentiation inference and profile modeling with machine learning methods for classification and interaction analysis allows identifying the most significant features, which may serve as a potential dose response signature.

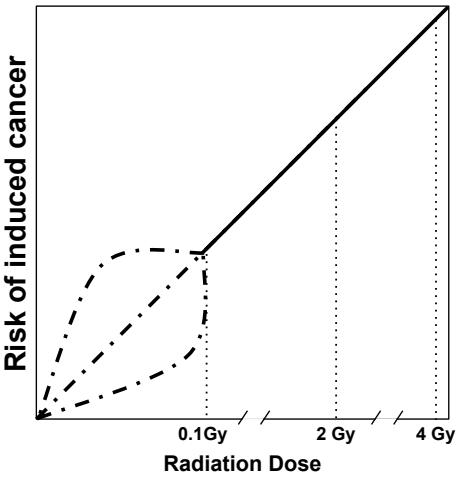

Fig. 1. Commonly accepted radiation dose model. The solid line indicates experimentally confirmed linear dependency between cancer risk and absorbed dose. This, however, holds for high doses from 0.1 Gy. The low-dose relation illustrated with possible models in intermittent lines is still under investigation.

2. Material and methods

2.1. Material. The data sets used in this study consist of two independently obtained expression sets from microarray experiments on the subject of radiosensitivity. The experiments were designed with the objective of identifying genes differentiating between radioresistant (RR) and radiosensitive (RS) women in a group of breast cancer patients undergoing radiotherapy. Blood samples were collected from the donors for the subsequent RNA extraction from lymphocytes for the microarray experiment. The first set was processed on the HuGene Affymetrix 1.0 ST oligonucleotide microarray platform, measuring 19,718 genes. There were 60 samples, of which 30 were labeled as radiosensitive and 30 as radioresistant. These samples were divided into two lots assigned to one of the two conditions: controls and irradiated with a therapeutic level dose of 2 Gy. The second experiment was performed using a custom Breakthrough 20K cDNA microarray chip (Finnon *et al*., 2012), measuring 19,959 genes. The study group consisted of 31 radiosensitive and 28 radioresistant patients and the treatment samples were subjected to a high dose of 4 Gy. The clinical description of the samples and radiosensitivity status assignment is available in the work of Yarnold *et al*. (2005).

2.2. Methods.

2.2.1. Preprocessing. In the first stage of analyses, the data sets were normalized separately according to their platform type. The Affymetrix oligonucleotide single

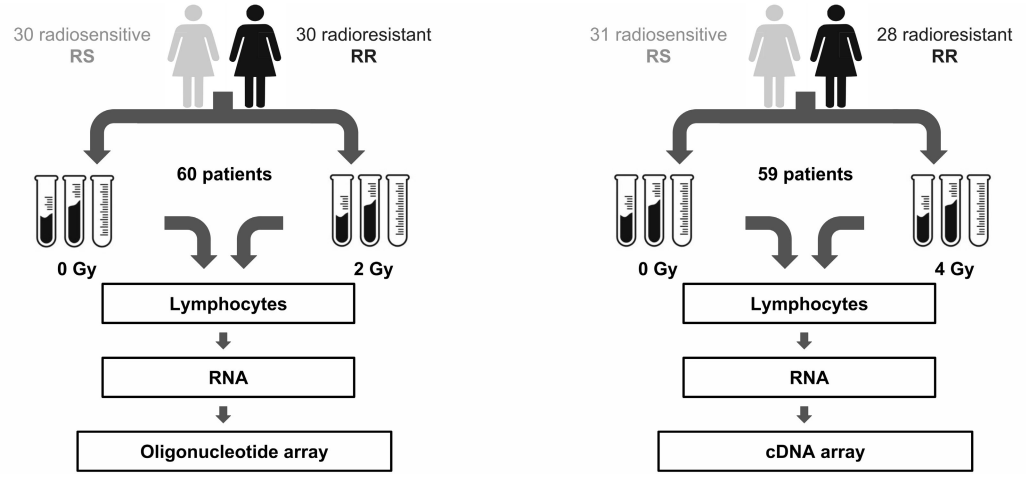

Fig. 2. Diagram presenting a comparison of experimental designs. The twin experiments were carried out using the same labeling of RR and RS patients with similar numbers of samples. They differ, nonetheless, with sample treatment doses and microarray experimental platforms. These issues had to be resolved during combined data processing.

channel data was normalized using the robust multichip average (RMA) method (Bolstad *et al*., 2003), specifically with background correction, quantile normalization and summarization using the median polish algorithm. Probes were reannotated with a Brainarray database custom chip description file (ENTREZG annotation, version 19.0.0) (Dai *et al*., 2005). The goal of this work was, however, to carry out a combined analysis of the data. Therefore, batch effect reduction through empirical Bayes methods was applied using the ComBat software (Johnson *et al*., 2007) for three batches (one for each of the two channels in cDNA data and one for oligonucleotide data) with no covariates (Papiez *et al*., 2014).

2.2.2. Differentiation analysis. In order to benefit from the joint analysis of two experiments, a common gene set for the two platforms was extracted for further processing. Statistical inference was performed with the use of differentiation tests between the dose groups: t-test, modified t-test and U Mann–Whitney test, according to the normality and variance homogeneity assumptions. Moreover, as an additional criterion for selection to the next steps, only genes which did not produce significant differences between controls in the two experiments were chosen. These genes were also investigated further for differentiation between 2 Gy and 4 Gy with division into radiosensitive and radioresistant samples. The differentially expressed genes between doses specific for radiosensitive and radioresistant patients were then examined in terms of the functional characteristics by studying overrepresented biological process Gene Ontology terms (Ashburner *et al*., 2000). Overrepresentation was assessed using Fisher's exact test implemented in the topGO R package (Alexa and Rahnenfuhrer, 2010) with Benjamini–Hochberg correction for multiple testing. The advantages of incorporating bioinformatics databases into biomarker discovery schemes have been previously shown in various studies (Meehan *et al*., 2013; Kong *et al*., 2014).

2.3. Trend testing. With two different doses, the genes were additionally assessed for the presence of trend with the Jonckheere–Terpstra test (Terpstra, 1952; Jonckheere, 1954). This test is an equivalent of the Kruskall–Wallis one for samples that may be sorted. Thus, the hypotheses are as follows:

$$H_0 : \Theta_1 = \Theta_2 = \cdots = \Theta_k, \tag{1}$$

$$H_A : \Theta_1 \leq \Theta_2 \leq \cdots \leq \Theta_k, \tag{2}$$

where Θ_i is the i-th sample median.

The genes marked significant at the level of 5% with strictly increasing and decreasing (i.e., not significantly monotonic) trends were analyzed for Gene Ontology term enrichment.

Furthermore, a more detailed insight into the nature of the trends was necessary. Hence, taking a step further, the analyzed genes, which did not present significant differences between controls in both experiments, were classified into one of the six types of response profiles (Fig. 3):

- irradiation related up-regulated,
- irradiation related down-regulated,
- dosimetry applicable up-regulated,
- dosimetry applicable down-regulated,
- high dose activation up-regulated,

- high dose activation down-regulated.

These response profiles were assigned based on the differentiation of expression levels between doses, e.g., in the irradiation related up-regulated group the expression levels were significantly differentiating between 0 Gy and 2 Gy, but no significant difference was observed between 2 Gy and 4 Gy.

2.4. Multiple random validation. In order to identify the potential biomarkers of radiation response, the samples were classified in a multiple random validation procedure. However, in this case simple separation between controls and irradiation samples was not possible, due to the inconsistent doses used in two experiments. Therefore, information gathered in the course of trend testing was used and the following procedure was executed on genes assigned into the irradiation related and dosimetry applicable categories. In the features that fell into the dosimetry applicable group, expression values in the 2 Gy dose point were substituted with a linear interpolation value between the control and 4 Gy values in the corresponding samples. In the irradiation related group, as there were no significant differences between values in 2 Gy and 4 Gy dose points, the values remained the same. In this way, a data set with two classes: controls and 2 Gy samples, was approximated. Results obtained using this novel method were compared to multiple random validation performed on unadjusted expression values.

Multiple random validation was carried out in 500 repetitions. In each repetition the data were randomly divided into training and test sets with a ratio of 7:3, and case/control proportions followed the actual proportion in the entire set. The classification was performed by means of logistic regression with forward stepwise feature selection using the Bayes factor (Berger and Pericchi, 1996) as a criterion for increasing the number of model features. In each repetition, genes forming the final model were recorded. The genes were ranked according to the frequency of their occurrence in a single signature. The resulting list served as the reference for further comparative analysis.

2.5. Monte Carlo feature selection validation. The entire data set was submitted to the Broadside tool (Krol, 2015), performing distributed Monte Carlo based feature selection (MCFS), to identify genes showing the most significant interaction networks in terms of radiation response. Broadside is a distributed feature selection and interaction mining algorithm and application designed for classification, regression and survival analysis problems. Interactions are captured by permuting pairs of variables, capturing the effect these permutations have on the model performance measure, and solving a linear equation system in order to perform a decomposition of feature total effects into the main and interaction ones. As a result, Broadside is not bound to a specific type of model and is more robust as it is free from the risk of misinterpreting unpruned decision tree structures as useful features and interactions. The most often occurring genes in the multiple random validation logistic models were compared to the results of the MCFS networks.

3. Results

3.1. Differentiation analysis. After extracting the common genes for both experimental platforms: Affymetrix oligonucleotide and custom cDNA, the joint analysis was carried out on a total of 9852 genes. Firstly, these genes were investigated with regard to the control samples in order to ensure the same base level for both experimental datasets. Among the control samples in the two studies, 7429 genes did not produce a significant difference between the normalized data sets. This set was then further investigated to determine gene sets indicating different patterns of response in radiosensitive and radioresistant patients. The overlap of the genes producing significant differences between expression levels in 2 Gy and 4 Gy is presented in Fig. 4. There were 1214 genes unique to the RS group and 730 genes for one RR.

On performing Gene Ontology enrichment analysis using Fisher's exact test with regard to all of the differentially expressed genes between 2 Gy and 4 Gy, after Benjamini–Hochberg correction, one significantly overrepresented term remained, i.e., the cellular amino acid metabolic process. Within the radiosensitive group, no significantly overrepresented terms were discovered; however, in the radioresistant patients 31 terms were statistically significant (see Table A2). GO BP terms linked to the radioresistant group included, among others, Stress response, Oxidative phosphorylation and Immune response regulation.

3.2. Trend testing. The numbers of features with increasing, decreasing and monotonic trends according to the results of the Jonckheere–Terpstra test are presented in Table 1. Additionally, genes were divided into strictly increasing and decreasing groups if they did not appear in the monotonic trend group. The genes with strictly increasing and decreasing dose response served as a basis for GO term enrichment analysis. The strictly up-trending genes yielded 99 significantly overrepresented terms and the down-trending—38 GO terms. Among the terms linked to decreasing features, there were processes related to hemopoiesis and homoeostasis, as well as GPI anchor metabolism and biosynthesis, whereas among the terms enriched with increasing trend genes cellular response to ionizing radiation could be found, along with Wnt

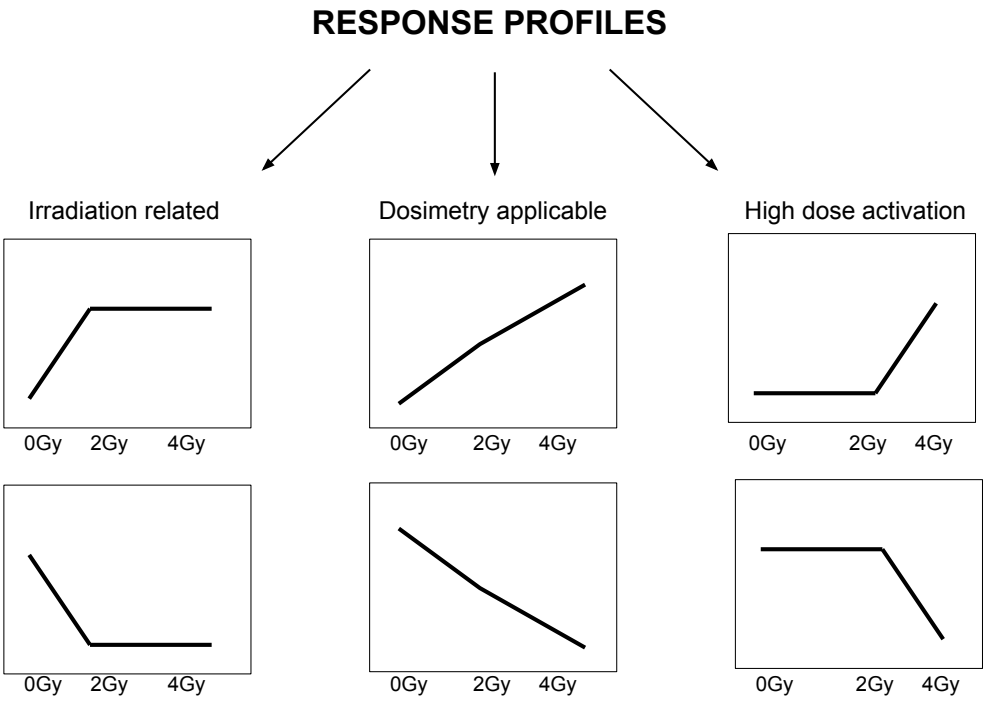

Fig. 3. Illustration of dose response profiles applied for gene grouping to enable accurate expression value interpolation.

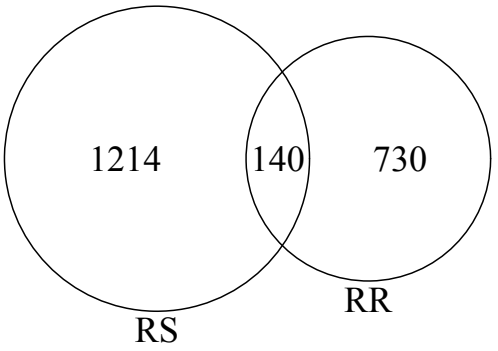

Fig. 4. Venn diagram presenting a comparison of the numbers of genes differentially expressed between 2 and 4 Gy doses in RR vs. RS samples.

signaling. The full lists of GO terms are available upon request from the authors.

The numbers of genes classified into the six types of response profiles are summarized in Table 2.

3.3. Multiple random validation. Average statistics were calculated over the course of 500 multiple random validation iterations: positive predictive value (PPV), negative predictive value (NPV) and overall accuracy. The classification was performed initially on original data, and afterwards on data adjusted using linear interpolation in the case of the dosimetry applicable type of gene profiles. There were 1,677 genes in the irradiation related category and 1,088 of the dosimetry applicable one. The original data results and the compared adjusted data results are presented in Table 3.

Table 1. Numbers of genes showing a significant dose trend. The strictly increasing and decreasing genes are those which do not appear in the monotonic trend group.

	Increasing	Monotonic	Decreasing
No. of genes	717	377	53
	Strictly increasing		Strictly decreasing
No. of genes	363		30

In the adjusted data, features selected for the logistic regression models were recorded. The top three most frequent genes included GADD45A, ZMAT3 and NAMPT. A complete list of the genes together with their occurrence frequencies is comprised in Table A1.

The entire initial set of features was processed independently using Monte Carlo feature selection, and the essential fragment of the ensuing network is presented in Fig. 5. It is clearly visible that genes with the highest numbers of interactions, and therefore the largest networks, are GADD45A, ZMAT3 and CCNG1.

4. Discussion

Basic analysis of differentially expressed genes under varying doses of radiation indicated potential changes in

Table 2. Numbers of genes grouped in to particular dose response profiles.

Number of genes in response profiles					
Irradiation related		Dosimetry applicable		High dose activation	
Up-No change	610	Up-Up	117	No change-Up	48
Down-No change	1067	Down-Down	969	No change-Down	319

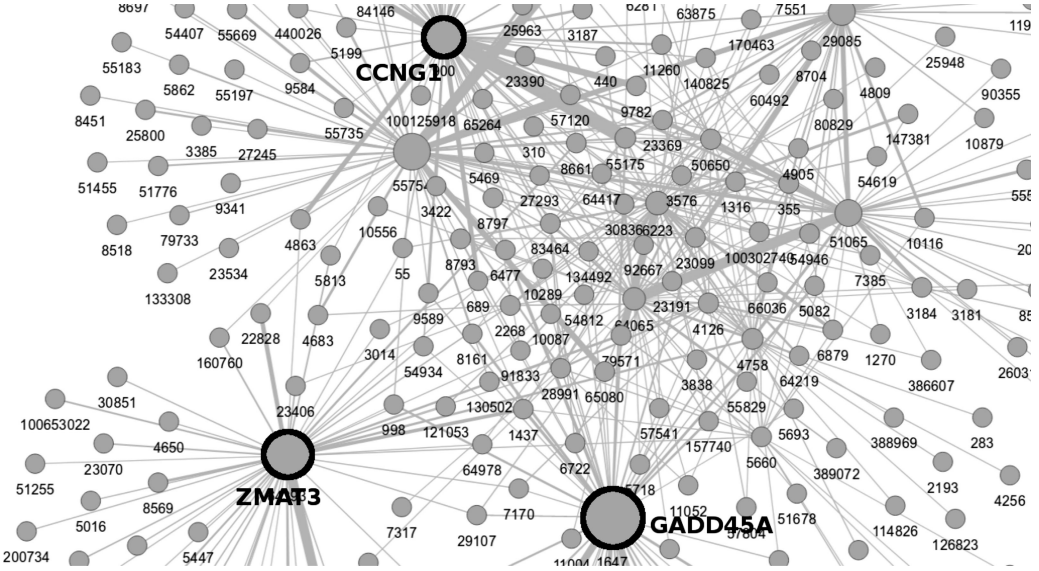

Fig. 5. Central fragment of a gene interaction network created as an illustration of Monte Carlo feature selection results on the entire data set. The genes in bold show the highest number and largest strength of interaction with other genes.

Table 3. Multiple random validation metric results for analysis conducted on original expression data values values adjusted using linear interpolation of the appropriate gene profiles.

Original expression data			
	Mean [%]	Lower CI [%]	Upper CI [%]
PPV	86.71	86.13	87.29
NPV	89.32	88.76	89.89
Accuracy	87.73	87.44	88.02
Interpolation adjusted data			
PPV	93.11	92.78	93.45
NPV	94.38	94.08	94.67
Accuracy	93.56	93.39	93.72

the expression profiles of radiosensitive and radioresistant patients. The genes specifically differentially expressed in the radiosensitive group participate in a wide range of biological processes, some of them being directly reported to play major roles in radiation response and tumor development (Weichselbaum *et al*., 1994; Park *et al*., 2014; Reinhardt *et al*., 1997). Meanwhile, the radiosensitive group gene set shows a lack of activity in processes known to be of biological importance. This is in coherence with the issue of key processes being silenced in radiosensitive patients and points to an area of further experimental investigation.

In the trend analysis, genes with strictly increasing trend were overrepresented in biological processes such as explicitly cellular response to ionizing radiation, but also Wnt signaling, which has been reported to be linked to breast cancer mechanisms in a study comprising a large dataset analyzed in a non-customary manner going beyond differential expression (Schmid *et al*., 2012). The down-trending genes tend to be more engaged in hemopoiesis and homoeostasis, which have been previously shown to play a role in stem cell injury from ionizing radiation (Shao *et al*., 2014). Also, GPI anchors, being important apoptosis regulators when deregulated by ionizing radiation, may have a considerable impact on cellular resistance (Brodsky *et al*., 1997).

The use of a non-standard approach to data classifying was highly beneficial. While simple logistic regression classification in a multiple random validation scheme on unadjusted data yielded satisfactory results in the context of separating control and dose-treated samples, there was room for improvement. Considering the twofold nature of the doses applied to samples in both experiments, data in the 4 Gy timepoint were linearly interpolated to 2 Gy. However, instead of executing this

in a non-selective approach, the genes were processed according to their dose response profile. The adjusted data gave significantly supreme results in comparison with the situation where dose values were not taken into account. Adjusted data classifying surpassed the simple approach in terms of positive and negative predictive values, as well as accuracy. This proves that, when possible, not only increasing sample size enhances classifier potential, but also using tailored solutions based on the knowledge of the underlying models to adjust data may be necessary.

The genes most often occurring in the logistic regression models were examined in terms of their biological function, to justify their contribution to the underlying processes. They include the following:

- GADD45A is a member of a group of genes whose transcript levels are increased following stressful growth arrest conditions and treatment with DNA-damaging agents. The DNA damage-induced transcription of this gene is mediated by both p53-dependent and independent mechanisms (Zhan, 2005). It has been previously proven to be a biomarker of radiation response (Kabacik *et al*., 2015).

- ZMAT3 mRNA and the protein are up-regulated by wildtype p53 and overexpression of this gene inhibits tumor cell growth, suggesting that this gene may have a role in the p53-dependent growth regulatory pathway (Bersani *et al*., 2014).

- NAMPT is thought to be involved in many important biological processes, including metabolism, stress response and aging. It has been shown to play a key role in radiotherapy treatment (Elf *et al*., 2017).

Moreover, an independent feature selection method was applied to the entire data set. The MCFS method as a rule based algorithm points out genes with regard to their number and strength of interactions. In the illustration (Fig. 5) it is clearly visible that three genes hold most interactions (number of lines starting in a node) and also the strongest ones (width of interaction line). These key features in the case of the two merged experiments were mainly GADD45A, ZMAT3 and CCNG1. Cyclin G1 (CCNG1) is a gene associated with G2/M phase arrest in response to DNA damage. It acts as an intermediate by which p53 mediates its role as an inhibitor of cellular proliferation and has been found to be linked with radiation response (Kabacik *et al.*, 2015; 2011; Manning *et al.*, 2013; Cruz-Garcia *et al.*, 2018).

The independent identification of key features important for modeling radiation response further justifies the use of a tailored non-standard data processing technique for classification purposes. The two most significant features, GADD45A and ZMAT3, were supported by means of another feature selection algorithm, as well as literature research. Furthermore, this may provide an incentive towards scientific research of the less frequent logistic regression model genes as to their possible importance not as single biomarkers of radiation response, but rather in terms of the impact they have when functioning in a network.

5. Conclusions

In this work a customized approach to high-throughput transcriptomic data analysis was proposed, based on statistical tools and current knowledge of the biological mechanisms. The data were obtained in the course of twin experiments with varying details, i.e., radiation dose and microarray platform. The analysis stages consisted of

1. identifying differentiating genes between radioresistant and radiosensitive patients, and the related biological processes;

2. assigning genes to their response patterns using the Jonckheere–Terpstra test and a custom profiling classification scheme;

3. applying the gene profiles as a filter to adjust data by means of linear interpolation to enable efficient classification in a multiple random validation setting.

These steps led to successful determination of potential biomarkers of radiation response, which were confirmed with an independent computational approach (MCFS) and literature study. Moreover, the differentiation and trend analyses confirmed the participation of genes deemed significant in biological processes linked with radiation response and cancer.

In silico machine learning analysis combined with classic statistical techniques with functional validation and profile modeling is a comprehensive solution for elucidating potential dose response mechanisms and revealing the most significant features to form a signature. Application of this kind of tailored procedures is a step towards enabling personalized individual therapy. It narrows down the search area for experts, potentially saving time and effort and allowing improvement in planing the design of future biological experiments established in order to study the impact of specific doses on breast cancer radiotherapy.

Acknowledgment

The authors would like to thank Lukasz Krol for providing the Broadside software package for Monte Carlo feature selection and ongoing support related with its usage.

The work was funded by National Science Center grants OPUS 10: UMO-2015/19/B/ST6/01736 (Anna Papiez) and BK-200/RAU1/2018/8 (Joanna Polanska).

Calculations were carried out using the GeCONiI infrastructure (POIG 02.03.01-24-099).

References

Abbott, A. (2015). Researchers pin down risks of low-dose radiation, *Nature* **523**(7558): 17–8.

Alexa, A. and Rahnenfuhrer, J. (2010). topGO: Enrichment analysis for gene ontology, *R Package Version 2.30*.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology, *Nature Genetics* **25**(1): 25.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association* **91**(433): 109–122.

Bersani, C., Xu, L., Vilborg, A., Lui, W. and Wiman, K. (2014). Wig-1 regulates cell cycle arrest and cell death through the p53 targets FAS and 14-3-3σ, *Oncogene* **33**(35): 4407.

Bolstad, B.M., Irizarry, R.A., Åstrand, M. and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics* **19**(2): 185–193.

Brenner, D.J., Doll, R., Goodhead, D.T., Hall, E.J., Land, C.E., Little, J.B., Lubin, J.H., Preston, D.L., Preston, R.J., Puskin, J.S., Ron, E., Sachs, R.K., Samet, J.M., Setlow, R.B. and Zaider, M. (2003). Cancer risks attributable to low doses of ionizing radiation: Assessing what we really know, *Proceedings of the National Academy of Sciences* **100**(24): 13761–13766.

Brodsky, R.A., Vala, M.S., Barber, J.P., Medof, M.E. and Jones, R.J. (1997). Resistance to apoptosis caused by PIG-A gene mutations in paroxysmal nocturnal hemoglobinuria, *Proceedings of the National Academy of Sciences* **94**(16): 8756–8760.

Cruz-Garcia, L., O'Brien, G., Donovan, E., Gothard, L., Boyle, S., Laval, A., Testard, I., Ponge, L., Woźniak, G., Miszczyk, L., Candéias, S.M., Ainsbury E., Widlak, P., Somaiah, N. and Badie, C. (2018). Influence of confounding factors on radiation dose estimation in in vivo validated transcriptional biomarkers, *Health Physics* **115**(1): 90–101.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J. and Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, *Nucleic Acids Research* **33**(20): e175–e175.

Elf, A.-K., Bernhardt, P., Hofving, T., Arvidsson, Y., Forssell-Aronsson, E., Wängberg, B., Nilsson, O. and Johanson, V. (2017). NAMPT inhibitor GMX1778 enhances the efficacy of 177Lu-DOTATATE treatment of neuroendocrine tumors, *Journal of Nuclear Medicine* **58**(2): 288–292.

Fargeas, A., Albera, L., Kachenoura, A., Dréan, G., Ospina, J.-D., Cologner, J., Lafond, C., Delobel, J.-B., De Crevoisier, R. and Acosta, O. (2015). On feature extraction and classification in prostate cancer radiotherapy using tensor decompositions, *Medical Engineering and Physics* **37**(1): 126–131.

Finnon, P., Kabacik, S., MacKay, A., Raffy, C., AHern, R., Owen, R., Badie, C., Yarnold, J. and Bouffler, S. (2012). Correlation of in vitro lymphocyte radiosensitivity and gene expression with late normal tissue reactions following curative radiotherapy for breast cancer, *Radiotherapy and Oncology* **105**(3): 329–336.

Francescatto, M., Chierici, M., Dezfooli, S.R., Zandonà, A., Jurman, G. and Furlanello, C. (2018). Multi-omics integration for neuroblastoma clinical endpoint prediction, *Biology Direct* **13**(1): 5.

Guidi, G., Maffei, N., Vecchi, C., Gottardi, G., Ciarmatori, A., Mistretta, G. M., Mazzeo, E., Giacobazzi, P., Lohr, F. and Costi, T. (2017). Expert system classifier for adaptive radiation therapy in prostate cancer, *Australasian Physical & Engineering Sciences in Medicine* **40**(2): 337–348.

Jagga, Z. and Gupta, D. (2015). Machine learning for biomarker identification in cancer research—developments toward its clinical application, *Personalized Medicine* **12**(4): 371–387.

Johnson, W.E., Li, C. and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostatistics* **8**(1): 118–127.

Joiner, M.C. (2004). A simple α/β-independent method to derive fully isoeffective schedules following changes in dose per fraction, *International Journal of Radiation Oncology Biology Physics* **58**(3): 871–875.

Jonckheere, A.R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika* **41**(1/2): 133–145.

Kabacik, S., Mackay, A., Tamber, N., Manning, G., Finnon, P., Paillier, F., Ashworth, A., Bouffler, S. and Badie, C. (2011). Gene expression following ionising radiation: Identification of biomarkers for dose estimation and prediction of individual response, *International Journal of Radiation Biology* **87**(2): 115–129.

Kabacik, S., Manning, G., Raffy, C., Bouffler, S. and Badie, C. (2015). Time, dose and ataxia telangiectasia mutated (ATM) status dependency of coding and noncoding RNA expression after ionizing radiation exposure, *Radiation Research* **183**(3): 325–337.

Kong, X., Liu, N. and Xu, X. (2014). Bioinformatics analysis of biomarkers and transcriptional factor motifs in down syndrome, *Brazilian Journal of Medical and Biological Research* **47**(10): 834–841.

Krol, L. (2015). Distributed Monte Carlo feature selection: Extracting informative features out of multidimensional problems with linear speedup, *in* S. Kozielski *et al.* (Eds.), *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*, Springer, Cham, pp. 463–474.

Manning, G., Kabacik, S., Finnon, P., Bouffler, S. and Badie, C. (2013). High and low dose responses of transcriptional biomarkers in ex vivo X-irradiated human blood, *International Journal of Radiation Biology* **89**(7): 512–522.

Meehan, T.F., Vasilevsky, N.A., Mungall, C.J., Dougall, D.S., Haendel, M.A., Blake, J.A. and Diehl, A.D. (2013). Ontology based molecular signatures for immune cell types via gene expression analysis, *BMC Bioinformatics* **14**(1): 263.

Mullenders, L., Atkinson, M., Paretzke, H., Sabatier, L. and Bouffler, S. (2009). Assessing cancer risks of low-dose radiation, *Nature Reviews Cancer* **9**(8): 596.

Papiez, A., Finnon, P., Badie, C., Bouffler, S. and Polanska, J. (2014). Integrating expression data from different microarray platforms in search of biomarkers of radiosensitivit, *International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, Spain*, Vol. 1, pp. 484–493.

Park, B., Yee, C. and Lee, K.-M. (2014). The effect of radiation on the immune response to cancers, *International Journal of Molecular Sciences* **15**(1): 927–943.

Parmar, C., Grossmann, P., Bussink, J., Lambin, P. and Aerts, H.J. (2015). Machine learning methods for quantitative radiomic biomarkers, *Scientific Reports* **5**(13087): 13087.

Ray, M., Yunis, R., Chen, X. and Rocke, D.M. (2012). Comparison of low and high dose ionising radiation using topological analysis of gene coexpression networks, *BMC Genomics* **13**(1): 190.

Reinhardt, M.J., Kubota, K., Yamada, S., Iwata, R. and Yaegashi, H. (1997). Assessment of cancer recurrence in residual tumors after fractionated radiotherapy: A comparison of fluorodeoxyglucose, L-methionine and thymidine, *The Journal of Nuclear Medicine* **38**(2): 280.

Schmid, P.R., Palmer, N.P., Kohane, I.S. and Berger, B. (2012). Making sense out of massive data by going beyond differential expression, *Proceedings of the National Academy of Sciences* **109**(15): 5594–5599.

Shao, L., Luo, Y. and Zhou, D. (2014). Hematopoietic stem cell injury induced by ionizing radiation, *Antioxidants & Redox Signaling* **20**(9): 1447–1462.

Terpstra, T.J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking, *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen* **55**(1): 327–333.

UNSCEAR (2000). *Sources and Effects of Ionizing Radiation*, Vol. 1, United Nations Publications, New York, NY.

Weichselbaum, R.R., Hallahan, D., Fuks, Z. and Kufe, D. (1994). Radiation induction of immediate early genes: Effectors of the radiation-stress response, *International Journal of Radiation Oncology, Biology, Physics* **30**(1): 229–234.

Yarnold, J., Ashton, A., Bliss, J., Homewood, J., Harper, C., Hanson, J., Haviland, J., Bentzen, S. and Owen, R. (2005). Fractionation sensitivity and dose response of late adverse effects in the breast after radiotherapy for early breast cancer: Long-term results of a randomised trial, *Radiotherapy and Oncology* **75**(1): 9–17.

Zhan, Q. (2005). GADD45A, a p53-and BRCA1-regulated stress protein, in cellular response to DNA damage, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **569**(1): 133–143.

Anna Papiez, a member of the Data Mining Group at the Silesian University of Technology, received her MSc in bioinformatics there. She is currently enrolled in a PhD program in biocybernetics and biomedical engineering. Her projects involve developing high-throughput data analysis and integration techniques in cooperation with Public Health England and Helmholtz Center Munich.

Christophe Badie is the head of the Cancer Genetics and Cytogenetics Group at the Centre for Radiation, Chemical and Environmental Hazards, Public Health England. His interest areas include research into the chromosomal and molecular mechanisms that underlie leukaemia initiation and development and investigation of specific genetic factors that influence haematopoietic cell radiosensitivity and cancer susceptibility. Also, his attention is directed towards new biomarkers of exposure and long term effects focusing on transcriptional modification. He is an author of 90+ articles, cited over 1,200 times.

Joanna Polanska is the head of the Data Mining Group at the Silesian University of Technology and the director of the Upper Silesian Center for Computational Science and Engineering. Her research interests lie in developing bioinformatics tools for omics data analysis, applications of Gaussian mixture models, mathematical modeling for diabetes type I, gene network analyses, and signal processing in molecular imaging. She has been continuously involved in numerous international multidisciplinary projects, including CARDIORISK Euroatom, GENEPI low-RT, or EUBIROD. She is an author of 300+ articles, cited over 2,800 times.

Appendix

Table A1. Gene occurrence frequency in multiple random validation iterations for genes incorporated in the model at least 10 times.

Gene symbol	Frequency
GADD45A	413
ZMAT3	87
NAMPT	64
COL4A3BP	51
TRAP1	49
SYNCRIP	45
G3BP1	44
SRSF8	42
DDX39A	38
NAALAD2	36
KAT5	25
RNPS1	22
XPOT	22
SPON1	15
DENND4A	13
ARHGAP19	10

Table A2. GO terms enriched with genes differentiating 2 Gy and 4 Gy dose response in radioresistant patients.

Term	Annotated	Significant	Expected	Fisher p-value
cellular metabolic process	4063	559	489.21	2.90E-09
cellular response to stress	764	123	91.99	5.20E-08
immune response-regulating cell surface receptor	156	34	18.78	3.10E-07
metabolic process	4400	589	529.79	3.20E-07
macromolecular complex subunit organization	820	143	98.73	6.70E-06
protein modification by small protein conjugation	404	76	48.64	7.20E-06
nucleobase-containing compound metabolic process	2233	350	268.87	9.00E-06
oxidative phosphorylation	55	19	6.62	1.30E-05
antigen processing and presentation	106	28	12.76	1.30E-05
response to stress	1496	195	180.13	1.80E-05
cell cycle	723	114	87.05	3.50E-05
cellular nitrogen compound metabolic process	2491	385	299.93	3.50E-05
electron transport chain	58	18	6.98	4.20E-05
RNA processing	304	69	36.60	4.40E-05
mitochondrial transport	131	30	15.77	4.90E-05
mRNA metabolic process	271	63	32.63	5.40E-05
translational elongation	50	18	6.02	5.70E-05
mitochondrion organization	231	55	27.81	7.30E-05
translational termination	39	16	4.70	8.80E-05
DNA-templated transcription, elongation	59	19	7.10	1.20E-04
mitochondrial translation	45	18	5.42	1.20E-04
cellular macromolecule catabolic process	364	74	43.83	1.50E-04
regulation of establishment of protein localization	53	16	6.38	1.60E-04
proteolysis involved in cellular protein catabolic process	259	53	31.19	1.90E-04
immune response-regulating signaling	221	38	26.61	2.00E-04
cellular protein catabolic process	270	57	32.51	2.40E-04
regulation of axon extension	33	9	3.97	2.40E-04
nitrogen compound metabolic process	2686	398	323.41	2.50E-04
T cell receptor signaling pathway	68	25	8.19	3.20E-04
positive regulation of ubiquitin-protein transferase activity	38	13	4.58	3.20E-04
immune response-activating cell surface receptor	145	34	17.46	3.40E-04

Received: 28 February 2018
Revised: 28 June 2018
Accepted: 18 October 2018