Exploration of ovarian cancer microarray data focusing on gene expression patterns relevant to survival using artificial neural networks

Clare Coveney*, Dong L. Tong, David J. Boocock, Robert C. Rees and Graham R. Ball

The John van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, United Kingdom.
{clare.coveney, dong.tong, david.boocock, robert.rees, graham.ball}@ntu.ac.uk

Abstract. A stage III ovarian cancer diagnosis yields a 22% 5-year survival rate, this applies to over half of the 7000 new cases diagnosed each year in the UK. Stratification of patients with this heterogeneous disease, based on active molecular pathways in their cancer, would aid a targeted treatment and improve prognosis. Hundreds of genes have been significantly associated with ovarian cancer, few have yet been verified. Exploration of published microarray data sets using Artificial Neural Networks confirmed the robustness of PRELP as a biomarker for survival time from stage III ovarian cancer, and generated a new panel of 44 genes that significantly predicted survival length of a blind validation set (p=0.00073).

Keywords: artificial neural networks·ovarian cancer·survival time·gene microarray·PRELP

1 Introduction

With approximately 7000 new diagnoses each year, ovarian cancer is the 5th most common cancer in the UK. A 92% 5 year survival can be expected from a stage I diagnosis, this drops to 22% at stage III. Unfortunately, due to the asymptomatic nature of the early stages of the disease and the lack of a sensitive screening tool, over half ovarian cancer cases are diagnosed at stage III or above **[1]**.

Despite a wealth of data and information being produced from ovarian cancer patient material, little has changed in the diagnostic, prognostic or treatment care for patients with the disease **[2]**. Thus, stratification of patients suffering this heterogeneous disease, based on active molecular pathways in their cancer, would aid a targeted treatment and improve prognosis.

Numerous genes have been significantly associated with ovarian cancer, yet few have been fully validated as biomarkers. In a comprehensive and systematic assessment of the online data available, Braem *et al* **[3]** reported that of 1065 genetic variants investigated, 200 were statistically significant; of these, 105 were included in

replication studies of which only 19 have been exclusively positively replicated. However, no attempt has been made to validate the remaining statistically significant genetic variants. This suggests an opportunity for confirming or refuting potential biomarkers published to date as an alternative to generating new ones. In addition, exploration of published data using different analytical approaches and meta-analysis of comparable cohorts will eliminate genes that were incidentally flagged as significant due to regional variation between sample cohorts, and focus on the genes and molecular pathways that show a consistent significant association between ovarian cancer and survival times.

This study presents the use of different analytical approaches to interpret ovarian microarray data. In this paper artificial neural networks (ANNs) were used to identify biomarkers for ovarian cancer. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins), a freely available online databank of genomic and proteomic relationships, was used to filter out biological components based on reported relevance.

2 Materials and Methods

2.1 Ovarian dataset

The stage III ovarian microarray data with accession number E-GEOD-13876 available in the ArrayExpress website **[4]** was selected and downloaded for analysis. This microarray data reported by Crijins *et al*. **[5]** contains 157 samples collected from ovarian cancer patients who have been treated with the same care pathway, was categorized into 3 groups; short-term survival (below 14 months of survival, 48 samples), long-term survival (over 25 months of survival, 37 samples) and in-between (14 – 25 months of survival, 72 samples). The samples were hybridized to high-density oligonucleotide microarrays using Operon v3.0 technology and replicated into 2 sets, i.e. Set 1 and Set 2. Each of these sets contains 34592 gene probe expression levels. Detailed information on the sample preparation on this dataset can be found in the original study **[5]** and from the ArrayExpress website. To maximize the chance of deriving a panel of markers that can distinguish between survival times, it was decided to compare short- and long-term survival groups. In this paper, the Set 1 was used for data modeling and the Set 2 served as blind validation set.

2.2 Artificial neural network analysis

An in-house designed artificial neural network (ANN) algorithm **[6,7]** was used to identify a set of gene probes which can correctly predict survival times (i.e. short- and long-term survival) for ovarian cancer patients. To screen candidate markers, a 3-layered backpropagation ANN model with the structure of 1-2-1 was applied. The Set 1 dataset was randomly partitioned into training, test and validation sets in which the training set was used to train the network, the test set was used to stop the network

when the optimum classification performance of the network was achieved and the validation set was used to test the predictive performance of the network. The ratio for the training, test and validation sets was 0.6:0.2:0.2, respectively. The samples were randomly allocated into each of these groups 50 times each time a new network model was created to avoid any bias on the reported results.

The following exhaustive search strategy was embedded into the algorithm: a new probe set id is selected as the input node in the network input layer each time a new network model was created; the sigmoid activation function was applied in the models; three hundred (300) epochs were used for the training process and 100 epochs for testing window stopping if the mean square error (MSE) failed to improve less than 0.01 over the window; a single input node was deliberately applied in this analysis to ensure that all probe sets in the data were thoroughly examined by the ANN.

2.3 Protein interaction database

The online protein interaction database Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) was used to visualize the association between identified markers. STRING **[8]** is an online, comprehensive, biological database of known and predicted protein-protein interactions. It is freely available and contains listings of proteins linked by; localization, homology, text-mining, databases, experiments, co-expression, co-occurrence and gene fusion. Lists of gene or protein identifiers can be mapped in STRING and interactions between those listed are generated and displayed diagrammatically.

3 Results and Discussion

Two hundred (200) gene probes with significant expression values were highly ranked by ANN. The list of genes that were found to be significantly different between long- and short-term survivals was compared to those found by Crijins *et al*. **[5]**. One of the 86 gene probes mentioned in Crijins ranked in the significant top proportion (PREPLP ranked 12th out of 34592 gene probes; 0.03%), however the remaining 85 genes appeared spread across this rank order (34592 gene probes).

A rudimentary literature search was conducted; counting the number of publications in which the gene code or the gene name (of which there can be multiple) of the 200 genes of interest occurred with terms such as "*ovarian* OR *ovary*", "*cancer*", "*ovarian neoplasm*", "*ovarian cancer*". Many were documented to be related to cancer and other diseases, few already linked to ovarian cancer and some were not linked via literature to cancer or ovarian disease.

The identified list of 200 gene probes was further condensed to those that were already annotated with a gene code. These were entered into STRING to identi-

fy any relationships amongst the proteins they code for. Forty four (44) were found to be linked already at least by co-mention in literature, as shown in Fig. 1 and Table 1.

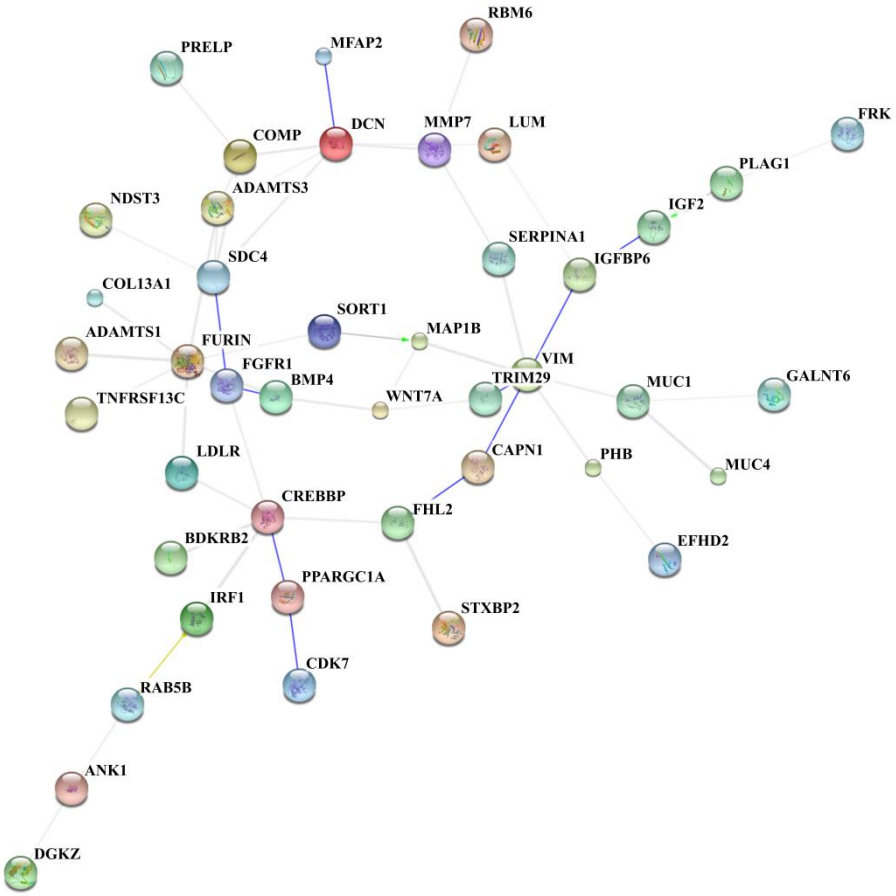

Fig. 1. Gene-gene interaction of the 44 associated genes in STRING database

Table 1. Summary of the mapped 44 genes in STRING database.

Gene Symbol	Database Entry	Description
ADAMTS1	NM_006988;AL162080	ADAM Metallopeptidase With Thrombospondin Type 1 Motif, 1
ADAMTS3	NM_014243;AB002364	ADAM metallopeptidase with thrombospondin type 1 motif, 3
ANK1	X16609;NM_020477	Ankyrin 1, Erythrocytic
BDKRB2	AF378542;NM_000623	Bradykinin Receptor B2
BMP4	NM_001202;D30751	Bone Morphogenetic Protein 4
CAPN1	BC015091	Calpain 1, (mu/I) large subunit
CDK7	BC005298;NM_001799	Cyclin-dependent kinase 7
COL13A1	AJ293624;NM_080812	Collagen, type XIII, alpha 1
COMP	S79500;NM_000095	Cartilage oligomeric matrix protein
CREBBP	U89355;NM_004380	CREB binding protein
DCN	NM_133506;BC005322	Decorin
DGKZ	U94905;NM_003646	Diacylglycerol kinase, zeta
EFHD2	BC014923;NM_024329	EF-hand domain family, member D2
FGFR1	M34188;NM_023111	Fibroblast growth factor receptor 1
FHL2	NM_001450;U60117	Four and a half LIM domains 2
FRK	BC012916;NM_002031	Fyn-related kinase
FURIN	NM_002569;A06939	Furin (paired basic amino acid cleaving enzyme)
GALNT6	Y08565;NM_007210	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6 (GalNAc-T6)
IGF2	M22373;NM_000612_X07868	Insulin-like growth factor 2 (somatomedin A)
IGFBP6	M69054;NM_002178	Insulin-like growth factor binding protein 6
IRF1	BC009483;NM_002198	Interferon regulatory factor 1
LDLR	BC014514;NM_000527	Low Density Lipoprotein Receptor
LUM	BC007038;NM_002345	Lumican
MAP1B	L06237;NM_032010	Microtubule-associated protein 1B
MFAP2	BC015039;NM_017459	Microfibrillar-associated protein 2
MMP7	BC003635;NM_002423	Matrix metallopeptidase 7 (matrilysin, uterine)
MUC1	L38597	Mucin 1, cell surface associated
MUC4	AJ242544;NM_004532	Mucin 4, cell surface associated
NDST3	NM_004784;AF074924	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 3
PHB	no link in adrf file	Prohibitin
PLAG1	U65002;NM_002655	Pleiomorphic adenoma gene 1
PPARGC1A	AF106698;NM_013261	Peroxisome proliferator-activated receptor gamma, coactivator 1 alpha
PRELP	NM_002725;BC032498	Proline/arginine-rich end leucine-rich repeat protein
RAB5B	AF267863;NM_002868	RAB5B, member RAS oncogene family
RBM6	U50839;NM_005777	RNA binding motif protein 6
SDC4	BC030805;NM_002999	Syndecan 4
SERPINA1	M26123;NM_000295	Serpin Peptidase Inhibitor, Clade A (alpha-1 antiproteinase, antitrypsin), member 1
SORT1	X98248;NM_002959	Sortilin 1
STXBP2	NM_006949;U63533	Syntaxin binding protein 2
TNFRSF13C	NM_052945;AF373846	Tumor necrosis factor receptor superfamily, member 13C
TRIM29	NM_058193;AF230389	Tripartite motif containing 29
VIM	NM_003380;M25246	Vimentin
WNT7A	NM_004625;BC008811	Wingless-type MMTV integration site family, member 7A

The predictive power of these 44 genes discriminating short- and long-term survivors were further validated using a separate set of technical replicates (Set 2). Amongst the 85 blind samples 56 (65.9%) were correctly classified as short- or long-term survivors with the 14 month and 25 month boundaries. The predictive performance of the model was assessed, Fig. 2 shows the predicted survival against actual survival.

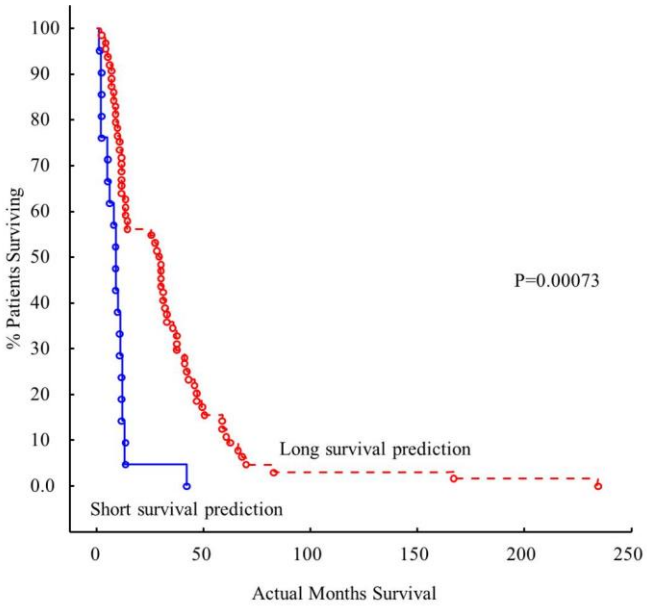

Fig. 2. Kaplan and Meier plot comparing the predicted survival group against the actual survival time of patients based on the 44 gene panel

Crijins *et al* **[5]**, who published the microarray data reanalyzed in the current study, used a continuous prediction algorithm to generate a list of 86 genes that can determine a favorable or unfavorable survival prognosis. In this paper, an ANN combined with STRING was used to generate a list of 44 genes that could successfully predict short- or long-term survival of a validation set (p=0.00073). Proline/arginine-rich end leucine-rich repeat protein (PRELP) appears in both lists of genes, further implicating its involvement in molecular pathways activated in ovarian cancer.

Furthermore, by applying different filtering tools to determine significantly differentially expressed genes; a new list of 44 genes has been generated. The finding of genes already associated with ovarian cancer in the 44 gene panel, namely IGF2 and BMP4 **[9,10,11,12,13]**, provides confidence in the methods used and warrants deeper investigation into those which are not. Both lists add to an ever building body of evidence of gene expression in ovarian cancer.

Despite the high number of publications producing and analyzing data from microarray experiments, there is no consensus on how the data should be pre-processed, processed and mined **[14, 15, 16]**. These results are an example of how published data can be complemented by reanalysis. Findings that are found to be of significant interest through a separate evaluation become less likely to be an artefact of the subsampling of cases or computational method used.

4 Conclusions and Future Work

To conclude, exploration of published data using a different non-linear analytical strategy offers robustness and highlights PRELP as a putative biomarker for stage III ovarian cancer, and, generated a panel of 44 genes that significantly predicted survival length of a blind validation set (p=0.00073). These results warrant further research, primarily a meta-analysis with similar datasets. This adds to the growing body of genomic information relating to ovarian cancer and contributes to the ability to predict a patient's likelihood to respond to a care pathway and would help clinicians navigate patients through therapy. In this instance, revealing a patient's prognosis in the standard care pathway may warrant more radical treatment options.

Further investigation and testing of the 44 gene model on a different patient cohort with the same diagnosis would provide an opportunity for validation and meta-analysis. Further exploration on different curated databases would confirm the biological links between these genes.

Acknowledgements

The authors would like to thank the John and Lucille van Geest Foundation who fund this research.

5 References

1. Cancer Research UK, http://info.cancerresearchuk.org/cancerinfo/cancerstats/types/ovary/.

2. Hays, J.L. et al.: Proteomics and ovarian cancer: integrating proteomics information into clinical care. J. Proteomics. 73, 10, 1864–72 (2010).

3. Braem, M.G.M. et al.: Genetic susceptibility to sporadic ovarian cancer: A systematic review. Biochim. Biophys. Acta. 1816, 2, 132–46 (2011).

4. ArrayExpress, http://www.ebi.ac.uk/arrayexpress/.

5. Crijns, A.P.G. et al.: Survival-related profile, pathways, and transcription factors in ovarian cancer. PLoS Med. 6, 2, e24 (2009).

6. Lancashire, L.J. et al.: A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. Breast Cancer Res. Treat. 120, 1, 83–93 (2010).

7. Lancashire, L.J. et al.: Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. Artif. Intell. Med. 43, 2, 99–111 (2008).

8. Szklarczyk, D. et al.: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39, Database issue, D561–8 (2011).

9. Kanety, H. et al.: Increased insulin-like growth factor binding protein-2 (IGFBP-2) gene expression and protein production lead to high IGFBP-2 content in malignant ovarian cyst fluid. Br. J. Cancer. 73, 9, 1069–73 (1996).

10. Sayer, R. a et al.: High insulin-like growth factor-2 (IGF-2) gene expression is an independent predictor of poor survival for patients with advanced stage serous epithelial ovarian cancer. Gynecol. Oncol. 96, 2, 355–61 (2005).

11. Lee, E.-J. et al.: Insulin-like growth factor binding protein 2 promotes ovarian cancer cell invasion. Mol. Cancer. 4, 1, 7 (2005).

12. Thériault, B.L. et al.: BMP4 induces EMT and Rho GTPase activation in human ovarian cancer cells. Carcinogenesis. 28, 6, 1153–62 (2007).

13. Shepherd, T.G. et al.: Autocrine BMP4 signalling regulates ID3 proto-oncogene expression in human ovarian cancer cells. Gene. 414, 1-2, 95–105 (2008).

14. Allison, D.B. et al.: Microarray data analysis: from disarray to consolidation and consensus. Nat. Rev. Genet. 7, 1, 55–65 (2006).

15. Leung, Y.F., Cavalieri, D.: Fundamentals of cDNA microarray data analysis. Trends Genet. 19, 11, 649–59 (2003).

16. Bolstad, B.M. et al.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19, 2, 185–93 (2003).