BIOINFORMATICS

Vol. 18 no. 12 2002
Pages 1617–1624

A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer

Ann-Marie Martoglio[1,,†], James W. Miskin[2,†], Stephen K. Smith[1] and David J.C. MacKay[2,*]*

[1]Reproductive Molecular Research Group, Department of Pathology and Department of Obstetrics and Gynaecology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QP, UK and [2]Cavendish Astrophysics Group, Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge, CB3 0HE, UK

Received on February 8, 2002; revised on May 8, 2002; accepted on June 6, 2002

ABSTRACT

Motivation: A number of algorithms and analytical models have been employed to reduce the multidimensional complexity of DNA array data and attempt to extract some meaningful interpretation of the results. These include clustering, principal components analysis, self-organizing maps, and support vector machine analysis. Each method assumes an implicit model for the data, many of which separate genes into distinct clusters defined by similar expression profiles in the samples tested. A point of concern is that many genes may be involved in a number of distinct behaviours, and should therefore be modelled to fit into as many separate clusters as detected in the multidimensional gene expression space. The analysis of gene expression data using a decomposition model that is independent of the observer involved would be highly beneficial to improve standard and reproducible classification of clinical and research samples.

Results: We present a variational independent component analysis (ICA) method for reducing high dimensional DNA array data to a smaller set of latent variables, each associated with a gene signature. We present the results of applying the method to data from an ovarian cancer study, revealing a number of tissue type-specific and tissue type-independent gene signatures present in varying amounts among the samples surveyed. The observer independent results of such molecular analysis of biological samples could help identify patients who would benefit from different treatment strategies. We further explore the application of the model to similar high-throughput studies.

Availability: Supporting details of the decomposition model can be found at http://www.inference.phy.cam.ac.uk/mackay/abstracts/icagenes.html and the ovarian cancer study data can be found at http://www.path.cam.ac.uk/~angio/publications/martoglioetal2002/ovcaica.html.

Contact: amm53@cam.ac.uk; mackay@mrao.cam.ac.uk

INTRODUCTION

Complementary DNA arrays generate a large volume of data representing relative gene transcript abundance in the samples surveyed. To date, a variety of algorithms and mathematical models have been used for the management, analysis, and interpretation of high-density array data. However, the experimental basis and range of assumptions needed place a limit in the field. Interpretation of the data is aimed at defining genes with similar expression patterns, with the underlying assumption that co-expressed genes are functionally related (Eisen *et al.*, 1998; Spellman *et al.*, 1998), or may share similar regulatory systems (Heyer *et al.*, 1999). Thus, depending on the *a priori* information available on the genes surveyed and the nature of the experimental design, computational approaches can be designed to study the causative response in an experimental system, or to predict the functional nature of novel genes. Reliance on *a priori* or intuitive knowledge is a heavy limitation to the application of computational methods in areas such as molecular diagnostics. No two (or more) genes are likely to exhibit precisely the same co-expression pattern in different pathophysiological samples. Also, many genes may be involved in a number of distinct behaviours defined in separate clusters in the multidimensional gene expression space. Many underlying conditions in a given sample (e.g. angiogenesis, tumorigenesis, apoptosis) may have yet-to-be-defined hallmark gene expression profiles that would be masked or unnoticed unless a hybrid of supervised and unsupervised clustering methods was correctly applied (Tamayo *et al.*, 1999). Although there are many

*To whom correspondence should be addressed.
† These authors contributed equally to this work.

mathematical models that can be applied and tested to analyze gene expression data, there is a current lack in the quantitative definition of probabilities for gene expression patterns. These should emerge not only on a gene-to-gene, or gene–network basis, but also in a system-specific environment. That is, the analysis and interpretation of gene expression data should be greatly facilitated once the quantitative probability of gene co-expression profiles is defined in tissue-, disease-, and pathophysiology-specific contexts. These questions should start to be answered once reproducible data sets from large sample groups, such as biopsy tissue samples, cell cultures modelling different pathophysiological systems, and so on, become available.

In the attempt to avoid the use of *a priori* knowledge or probabilistic predictions to pre-define clustering parameters, we tested an unsupervised model based on Independent Component Analysis (ICA; Bell and Sejnowski, 1995; Martoglio, 2000; Miskin, 2001). Whereas a standard clustering method assigns each gene to *one* cluster of genes which have correlated expression patterns, we believe it is more reasonable to expect each gene to be influenced by *several* transcription factors, each of which influences several genes. Thus we used a model in which each gene can participate, to varying degrees, in many independent patterns of covariation. We called these co-expression patterns 'gene signatures'. The overall gene expression profile from each hybridized array is considered to be a linear combination of independent latent gene expression profiles, which are discovered by bilinear decomposition. By employing this method (see description in **Systems and methods**), a number of problems inherent in multivariate analysis are bypassed: (1) genes are flexibly identified in as many co-expression patterns (or, in analogy, clusters) as necessary, rather than reducing their allocation to any one particular set of observations; (2) detailed information is maintained by identifying a number of latent variables underlying the whole data set, rather than reducing interpretation to a rigid set of cluster possibilities; and (3) the number of potentially valid underlying features, or gene co-expression patterns, can be assessed *after* the learning process rather than having to pre-set the number of possible clusters and having to limit the analytical model around these. It would therefore be possible to classify samples by 'blind', or observer independent, separation with minimal imposition of functional structure or limiting clauses to the analytical learning process. This allows the disclosure of hidden gene co-expression patterns, or gene signatures, in the data sets that would have otherwise not been identified with other methods.

Each gene signature within the data defines a combination of genes that behave in a similar fashion, and is represented in a quantitative manner per tissue sample. In other words, each gene signature represents a distinct gene co-expression pattern that distinguishes variations amongst the samples. The ability of the model to uncover hidden gene signatures within the data results in multiple tagging, or classification, of the samples. This, in turn, allows conditional sub-categorization by the observer, such that distinct types of pathophysiological behaviours can be captured and assessed within a given sample or sample sets. The method is exemplified using data from an ovarian cancer study employing tailored cDNA arrays with 175 gene targets (Martoglio *et al.*, 2000; http://www.path.cam.ac.uk/∼angio/publications/martoglioetal2002/ovcaica.html).

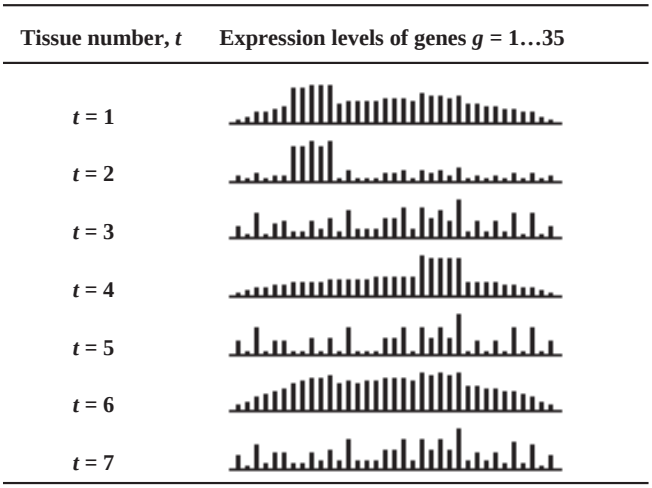
SYSTEMS AND METHODS

The working data set consisted of a total of 17 hybridization profiles, corresponding to normal ovary ($n = 5$), serous papillary adenocarcinoma ($n = 5$), poorly differentiated serous papillary adenocarcinoma ($n = 4$), benign serous cystadenoma ($n = 1$), and benign mucinous cystadenoma ($n = 2$). These were numbered tissue 1–17, respectively. Details on sample preparation, labelling, hybridization to cDNA arrays, and reproducibility controls were as described in Martoglio *et al.* (2000). Results emerging from the same data set using non-ICA methods are also presented in Martoglio *et al.* (2000).

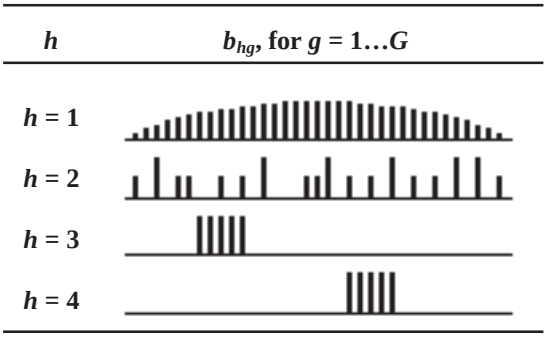
Independent component analysis (ICA) modelling

We assume that the data to be modelled is a matrix D of gene expression levels in different tissues (samples). The entry d_{tg} gives the expression level of gene g in tissue t. A cartoon of such a data matrix D is shown in Figure 1a. We assume that gene expression levels vary under the influence of H latent variables, $a_1, a_2, \ldots, a_H$. The number of latent variables (H) there are is something we will attempt to infer from the data. Each gene expression level may be influenced by one or more of these latent variables. Which genes are influenced by the hth latent variable, and how strongly, is described by a vector b_{hg}, with $g = 1 \ldots G$. This vector describes a set of genes that are expected to co-vary—they will tend to go up and down together. A list of $H = 4$ such vectors is shown in Figure 1b. We call these vectors 'gene signatures'. The first latent variable in this example influences all genes. If the latent variable a_1 is large, all genes are expressed at higher levels, though some more than others. We might call such a latent variable a housekeeping latent variable (see **Discussion and conclusion**). Latent variable number $h = 3$ influences just five of the genes, equally; latent variable $h = 4$ influences another five; and latent variable $h = 2$ influences about half of the genes.

We assume that the gene expression levels in each tissue t are generated as follows: the latent variables $a_1, a_2, \ldots, a_H$ are set to particular levels $a_{t1}, a_{t2}, \ldots, a_{tH}$, where a_{th} specifies the amount of pattern h present in tissue t; the gene expression levels d_{tg}



(a)



(b)

Fig. 1. (a) Representation of data matrix D, modelling mock data from a DNA array containing 35 genes ($G = 35$), hybridized with 7 tissue samples ($T = 7$). Each tissue shows a different overall pattern for all 35 gene hybridization intensities. (b) Gene signatures revealed by independent component analysis (ICA) of the mock data presented in (a). Each gene signature (latent variable h) represents a set of genes that are expected to co-vary in the samples tested. Gene signature $h = 1$ in this example influences all genes, and may be considered a 'housekeeping', or ubiquitous, gene signature, expected to be present in a relatively high amount in all samples. Gene signature $h = 2$ influences approximately half of the genes on the array, while gene signature $h = 3$ influences just five of the genes, equally, and gene signature $h = 4$ influences another 5 genes from the arrayed set.

are then found by adding up the patterns b_{hg}, weighted by the amounts a_{th}. We assume that the data are noisy, with noise n_{tg} in the measurement of d_{tg}. So,

$$d_{tg} = \sum_h a_{th} b_{hg} + n_{tg}.$$

Or, using matrix notation,

$$D = AB + N.$$

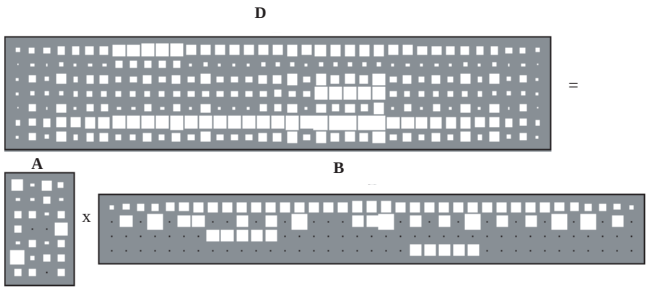


Fig. 2. Hinton diagram representations of matrix D (microarray data set), decomposed to B and A (gene signatures and corresponding weights, or quantification) in independent component analysis (ICA).

This is exactly how the mock data in Figure 1a were generated. The patterns B are those shown in Figure 1b; the amounts of the patterns are:

$$A = \begin{bmatrix} 3 & 0.1 & 2 & 0.62 \\ 0.1 & 0.1 & 1 & 0.1 \\ 1 & 1 & 0.1 & 1 \\ 1 & 0 & 0 & 4 \\ 0.1 & 1 & 0.1 & 1 \\ 5.1 & 0.2 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

and there was no noise. Another way of representing the matrices D, A and B is by Hinton diagrams in which the magnitude of a number is shown by the size of a square (Figure 2). Here, all entries in the matrices A and B are positive, and all the squares are white; negative entries would be represented by black squares. If a model with a number of latent variables H significantly smaller than G fits the data well, the model can evidently compress the data substantially, since each extra tissue's G expression level are captured by just H latent variables.

For microarray studies in which each spot on the array yields two measurements, an experimental measurement m_{tg} and a control measurement l_{tg}, d_{tg} is defined to be the ratio m_{tg}/l_{tg}. We would prefer to replace all such 'normalizations' of data by inferences of the implicit variables. One reason not to work with the experimental ratio m_{tg}/l_{tg} is because the noise in the ratio comes from noise in m and from noise in l. Those noise variables might be Gaussian, but the noise in the ratio certainly won't be. These issues and others, such as inference, probability, and approximation in the model presented can be found at http://www.inference.phy.cam.ac.uk/mackay/abstracts/icagenes.html.

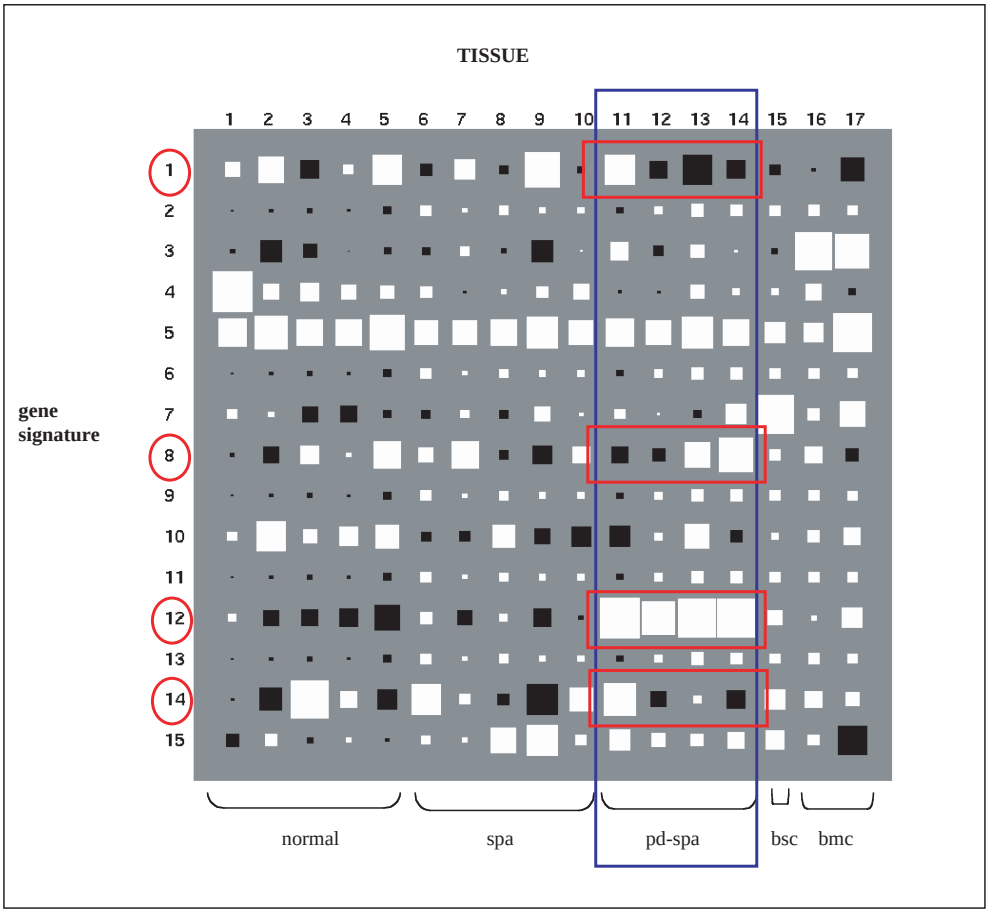


Fig. 3. Gene signatures revealed by independent component analysis (ICA) of ovarian cDNA array data. Independent component analysis (ICA) of ovarian cDNA array results revealed tissue type-specific gene signatures as well as tissue type-independent gene signatures present in varying amounts among the ovarian samples surveyed (red circles). For example, gene signature12 can be observed consistently predominant in the pd-spa sample group (delineated in blue), while inherent variations within the histologically similar samples (i.e. pd-spa) can be observed at gene signatures 1, 8, and 14 (red rectangles). Molecular analysis of tissue samples in such multi-parameter schemes may help to classify tissues based on a number of underlying disease processes. This could help identify patients who would benefit from different treatment strategies.

RESULTS

Tissue-specific gene signatures: testing the matrix

The working data set consisted of a total of 17 hybridization profiles, as described above (see **Systems and methods**). Fifteen of the samples were selected in the first instance and submitted to the independent component analysis, without any tissue detail entry into the modelling process. An overall assessment of the results indicated the possible definition of tissue-specific gene signatures such as gene signature 3: benign mucinous cystadenoma (bmc), gene signature 7: benign serous cystadenoma (bsc), and gene signature 12: poorly differentiated serous papillary adenocarcinoma (pd-spa) (see 'Learning run', http://www.path.cam.ac.uk/~angio/publications/martoglioetal2002/ovcaica.html). If the gene signatures were true definers of tissue type, they could be fixed and used to test further independent samples. That is, observer-independent analysis could be used to classify ovarian tissue samples based on their gene signature-specific profile. In order to assess the ICA-based results, the defined gene signatures were fixed, such that the A matrix remained constant (see **Systems and methods**), and the two ovarian samples that had not been submitted to the initial analysis (i.e. tissue 2 (normal ovary) and tissue 13 (poorly differentiated serous papillary adenocarcinoma)) were tested. Tissue 13 showed prominent mapping to the pd-spa gene signature (gene signature 12) defined in the learning run, while tissue 2 expressed similar amounts of that profile as

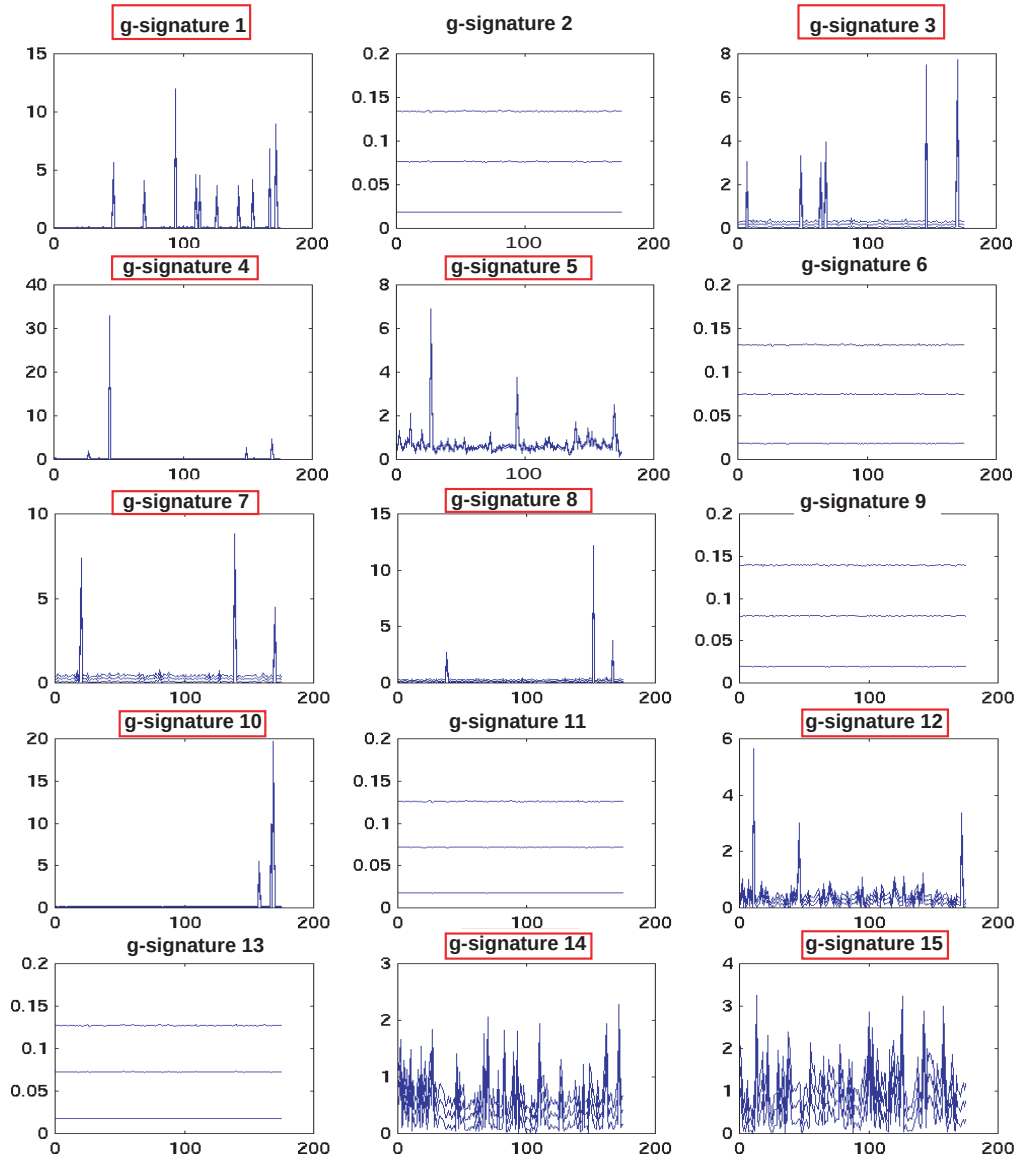


Fig. 4. Graphical representation of gene (*g*) signatures unveiled by independent component analysis (ICA) of ovarian tailored cDNA array data. Each graph shows the gene expression profile (genes along the *x*-axis; relative signal intensity along the *y*-axis), defining each gene signature. Error bars are shown for visual assessment of the signal-to-noise ratio. Gene signatures 1, 3, 4, 5, 7, 8, 10, 12, 14, and 15 (red boxes) show consistent representation (minimal variance) in the ovarian samples tested. Analysis of the leading genes for these gene signatures may indicate inherent patho-physiological processes distinctively measurable among the ovarian cancer specimens.

other normal ovarian tissues (Figure 3; see 'Fixed matrix tests', http://www.path.cam.ac.uk/~angio/publications/martoglioetal2002/ovcaica.html). These results supported the definition of a tissue-specific gene signature able to distinguish poorly differentiated serous papillary adenocarcioma specimens amongst mixed ovarian samples in the experimental setting described.

The definition of gene signatures is dependent on the representation of a given gene co-expression profile amongst all tissue samples tested. Assessment of the signal-to-noise ratio, ($\log_{10}$(mean(average^2/variance))), for each potential gene signature (row) is therefore necessary prior to descriptive comparisons across the samples (columns) tested. A graphical representation of each potential gene signature is presented in Figure 4, and corresponding signal-to-noise ratios can be seen in Table 1 at http://www.path.cam.ac.uk/~angio/publications/martoglioetal2002/ovcaica.html. Gene signatures 2, 6,

9, 11, and 13 revealed a high level of variance, and were therefore not considered futher. On the other hand, gene signatures 1, 3, 4, 5, 7, 8, 10, 12, 14, and 15 all showed clear definition of a hidden gene expression profile. Assessment of these specific profiles may hold important leads to conditions inherent in all the ovarian tissue samples, which vary depending on the behaviour of the tissue with regard to the pathophysiological state or condition that the gene signatures represent. The leading genes defining each gene signature are listed in Table 2 at http://www.path.cam.ac.uk/~angio/publications/martoglioetal2002/ovcaica.html, although we emphasize that signatures are based on the relative co-expression of all 175 genes on the array.

In the process of defining gene signatures for future applications, a larger sample data set is suggested. That is, the larger the sample set used for the initial 'learning run', the stronger the robustness of the emerging gene signatures. Nevertheless, it is interesting to note that the ICA-based method was able to extract meaningful hidden gene co-expression profiles from a limited 15-sample set, pointing to its potential for diagnostic, prognostic, or therapeutic applications.

Biological interpretation of cDNA array-derived gene signatures

We note again that larger studies incorporating larger sample and gene sets will lead to more conclusive definitions of emerging gene signatures. In the meantime, a number of important observations from the present study are noteworthy and may be considered in future experimental designs, as follows.

Poorly differentiated serous papillary adenocarcinoma (pd-spa) gene signature. Of all ovarian tissue samples surveyed, tissues 11, 12, 13, and 14, all pd-spa samples, were consistently seen to have the most prominent abundance for gene signature 12. This specific gene co-expression pattern included highly expressed Tie1, placental growth factor, HLA I, HLA-DR, cadherin-3, cadherin-11, gp130 and cofilin (see Table 2 at http://www.path.cam.ac.uk/~angio/publications/martoglioetal2002/ovcaica.html). Although histopathological classification is supported by the tissue type-specific gene signature, heterogeneities amongst the pd-spa samples were captured by other gene signatures, such as gene signature 1, 8, and 14 (Figure 3). The pathophysiological traits represented by these gene signatures is not yet defined. Nevertheless, they indicate that phenotypic tissue type- classification is not enough to capture important differences inherent amongst similar tissue samples. The use of a combination of tissue type- and pathophysiology-specific gene signatures holds promising application for the sub-classification of patients beyond current binary disease diagnostic methods. This could lead to improved diagnostic and prognostic interpretations, and the development of a targeted selection of treatment strategies.

Benign mucinous cystadenoma (bmc) gene signature. Gene signature 3, defined by leading genes metalloproteinase inhibitor 1 (TIMP-1), 5-hydroxytryptamine 2B receptor (5-HT2B), AMP deaminase 3, calgranulin B, and mucin-1, was seen to be most prominent in tissues 16 and 17, both of benign mucinous cystadenoma classification, amongst all the ovarian tissue samples surveyed. Normal samples (tissues 1–5) demonstrated a consistently subtracted profile for this gene signature; interestingly, tissue 4, the 'normal' contra-lateral ovary from the same patient as bmc-ridden tissue 16, demonstrated the least subtracted profile amongst all normal samples.

Pre-menopause gene signature. Gene signature 4, defined by leading genes endothelin-1 receptor (EDNRA), cadherin-6, and Cu/Zn-superoxide dismutase, was distinctively prominent in tissue 1, a member of the normal ovaries group, but originating from a pre-menopausal patient. This observation showed the sensitivity of the ICA-based method in discriminating functionally distinct samples not detected by conventional intra-class comparative methods (Martoglio *et al.*, 2000).

Chemoresistance/matrix integrity gene signature. The leading genes defining gene signature 15 were characteristic of a chemoresistant profile (e.g. cofilin, glutathione *S*-transferase P, previously presented Martoglio *et al.*, 2000) and had a predominance of matrix metalloproteinases (MMP-3, MMP-7, MMP-10), known to degrade the cellular matrix in a number of pathophysiological conditions, including vessel invasion and metastasis (Sang, 1998; Yamashita *et al.*, 1998; Ishii and Asuwa, 2000; von Lampe *et al.*, 2000). Thus, inspection of the molecular profile of a specimen based on this gene signature could indicate the invasive state of the tissue. This type of check-point could aid in treatment decisions that would, for example, interfere with TIMPs or other related molecules, as in recent anti-angiogenic treatment strategies (Blavier *et al.*, 1999; Indraccolo *et al.*, 1999).

The results presented show promising application to future molecular diagnostic, prognostic, and therapeutic applications. Further studies incorporating more tissue samples would be needed to confirm and assign the combined relevance of the observed gene signatures.

DISCUSSION AND CONCLUSION

Assessment of the ICA-based results revealed a number of gene signatures representing potential inherent pathophysiological processes in the various ovarian tissue samples, as described above. Some gene signatures were

highly prominent in tissue-specific samples, such as the pd-spa gene signature and bmc gene signature, indicating potential application in observer-independent tissue classification. Other gene signatures, such as gene signatures 1, 8, 14, and 15, were present at varying levels among the ovarian tissue samples, with no consistent patterns within a given tissue type.

Gene signatures defining a given pathophysiological or disease state are expected to be present at similar levels among the representative samples of a given tissue type. This would be true if histopathological classification identified tissues undergoing a finite regime of molecular changes leading to the disease phenotype. The mixed representation of some gene signatures among the ovarian samples may be due to a number of reasons: (1) gene signatures may be identifying inherent pathophysiological processes in the ovary that vary independently of the normal or tumorigenic state of the tissue, or (2) some tissues may have been incorrectly classified by traditional histopathological methods. Should (1) hold true, larger studies and further comparative parameters are necessary to define the molecular trait defined by the specific gene signatures present. The second explanation proposed is in line with the reported incidence of observer-dependent variance in tissue classification (Martoglio *et al.*, 2000), and further instigated by the lack of emergence of a gene signature able to define the spa tissue sample group. To test this hypothesis, an overall tissue type probability test was performed on the ICA results, using all gene signatures defined to model clustering of the tissue samples into normal or diseased groups. The results showed that tissues 6, 7, and 10, expected to be of spa type, had a partial overall molecular profile similar to that of the normal ovary group. Interestingly, tissues 4 and 5, which were samples of 'normal' ovary of patients with benign mucinous cystadenoma and benign serous cystadenoma in the contra-lateral ovary, respectively, showed a partial overall molecular profile similar to that of the diseased ovary group (results not shown). This supports the possibility that mis-classification of some of the spa samples may be hindering disclosure of a spa-specific gene signature. The application of gene signatures may contribute to the development of improved diagnostic, prognostic, and therapeutic screening regimes.

Exploring data normalization issues

Amongst the differentially expressed gene signatures discussed above, ICA revealed a gene signature ubiquitously expressed in all tissues queried (Figure 3, gene signature 5). Elongation factor 1-alpha and actin were amongst the leading genes in the ubiquitous gene signature, in agreement with previous independent reports (van de Corput *et al.*, 1998; leading genes listed in Table 2 at http://www.path.cam.ac.uk/~angio/publications/martoglioetal2002/ovcaica.html). It is proposed that normalization of cDNA array data against single genes endogenously expressed in the samples could hinder proper comparative analysis of the results. Quantitative comparison of a fixed co-expression profile or signature common to all samples may provide a better means of assessing inherent sample variablity.

Future ICA-based developments

ICA is used in a variety of data mining schemes, as it has many profitable properties, such as analytical robustness and scalability to large data sets. A similar model to that described above has been applied by Liebermeister to analyze cell cycle-related gene expression in the yeast and human lymphocyte gene expression data (Liebermeister, 2002). With the analysis of further data sets from larger high throughput studies, the functional significance of each gene signature (or 'mode', as referred to by Liebermeister) can be described and tested. The analytical method can be similarly applied to protein high throughput studies. In the future, treatment or prognostic interpretations could be based on a combination of inherent disease indicators, tested from a single sample specimen, resulting in multi-targeted solutions appropriate to the individual patient.

ACKNOWLEDGEMENTS

We thank Dr Anthony Corps, Dr Cristin Print, and Dr David Ward for critical comments and discussion. AMMS was supported in part by the Cambridge Commonwealth Trust, JWM and DJCM were supported in part by the Gatsby Foundation.

REFERENCES

Bell,A.J. and Sejnowski,T.J. (1995) An information maximization approach to blind separation and blind deconvolution. *Neural Comput.*, **7**, 1129–1159.

Blavier,L., Henriet,P., Imren,S. and Declerck,Y.A. (1999) Tissue inhibitors of matrix metalloproteinases in cancer. *Ann. N. Y. Acad. Sci.*, **878**, 108–119.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, **9**, 1106–1115.

Indraccolo,S. *et al.* (1999) Generation of expression plasmids for angiostatin, endostatin and TIMP-2 for cancer gene therapy. *Int. J. Biol. Markers.*, **14**, 251–256.

Ishii,T. and Asuwa,N. (2000) Collagen and elastin degradation by matrix metalloproteinases and tissue inhibitors of matrix metalloproteinase in aortic dissection. *Hum. Pathol.*, **31**, 640–646.

Liebermeister,W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, **18**, 51–60.

Martoglio,A.M., Tom,B.D., Starkey,M., Corps,A.N., Charnock-Jones,D.S. and Smith,S.K. (2000) Changes in tumorigenesis- and angiogenesis-related gene transcript abundance profiles in ovarian cancer detected by tailored high density cDNA arrays. *Mol. Med.*, **6**, 750–765.

Martoglio,A.M. (2000) *Tailored high-throughput cDNA arrays for investigations in ovarian cancer*, PhD thesis, Department of Obstetrics and Gynaecology and Department of Pathology, University of Cambridge.

Miskin,J.W. (2001) *Ensemble learning for independent component analysis*, PhD thesis, Department of Physics, University of Cambridge.

Sang,Q.X. (1998) Complex role of matrix metalloproteinases in angiogenesis. *Cell Res.*, **8**, 171–177.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridisation. *Mol. Biol. Cell*, **9**, 3273–3297.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

van de Corput,M.P. *et al.* (1998) Sensitive mRNA detection by fluorescence in situ hybridisation using horseradish peroxidase-labeled oligodeoxynucleotides and tyramide signal amplification. *J. Histochem Cytochem.*, **46**, 1249–1259.

von Lampe,B., Barthel,B., Coupland,S.E., Riecken,E.O. and Rosewicz,S. (2000) Differential expression of matrix metalloproteinases and their tissue inhibitors in colon mucosa of patients with inflammatory bowel disease. *Gut.*, **47**, 63–73.

Yamashita,K., Azumano,I., Mai,M. and Okada,Y. (1998) Expression and tissue localization of matrix metalloproteinase 7 (matrilysin) in human gastric carcinomas. Implications for vessel invasion and metastasis. *Int. J. Cancer*, **79**, 187–194.