BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 3 2007, pages 298–305
doi:10.1093/bioinformatics/btl597

Gene expression

Integrating transcription factor binding site information with gene expression datasets

Ian B. Jeffery[1,*], Stephen F. Madden[1], Paul A. McGettigan[1], Guy Perrière[2], Aedín C. Culhane[3] and Desmond G. Higgins[1]

[1]UCD Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland, [2]Laboratoire de Biométrie et de Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard Lyon 1, 43 boulevard du 11 Novembre, 1918, 69622 Villeurbanne Cedex, France and [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Mayer 232, 44 Binney Street, Boston, MA 02115, USA

Received on June 15, 2006; revised on September 12, 2006; accepted on November 21, 2006
Advance Access publication November 24, 2006
Associate Editor: David Rocke

ABSTRACT

Motivation: Microarrays are widely used to measure gene expression differences between sets of biological samples. Many of these differences will be due to differences in the activities of transcription factors. In principle, these differences can be detected by associating motifs in promoters with differences in gene expression levels between the groups. In practice, this is hard to do.

Results: We combine correspondence analysis, between group analysis and co-inertia analysis to determine which motifs, from a database of promoter motifs, are strongly associated with differences in gene expression levels. Given a database of motifs and gene expression levels from a set of arrays, the method produces a ranked list of motifs associated with any specified split in the arrays. We give an example using the Gene Atlas compendium of gene expression levels for human tissues where we search for motifs that are associated with expression in central nervous system (CNS) or muscle tissues. Most of the motifs that we find are known from previous work to be strongly associated with expression in CNS or muscle. We give a second example using a published prostate cancer dataset where we can simply and clearly find which transcriptional pathways are associated with differences between benign and metastatic samples.

Availability: The source code is freely available upon request from the authors.

Contact: Ian.Jeffery@ucd.ie

1 INTRODUCTION

The identification and understanding of transcriptional regulatory networks and their interactions are major challenges in biology. A major point of control in the transcription of a mRNA is the binding of one or more transcription factors (regulatory proteins) to sites in their promoters called transcription factor binding sites (TFBSs). The identification and characterization of these TFBS has been a major activity over the past 10 years. Some of this TFBS information has been stored in databases such as TRANSFAC (Wingender *et al*., 1996) where examples of experimentally verified sites are given. Most TFBS however are less well characterized. Many have yet to be discovered i.e. there are known transcription factors with no clear TFBS. In other cases, TFBS are computationally predicted using comparative genomics.

Several methods are available to look for overrepresented TFBS in the promoter sequences of a set of genes. These methods include searching for over represented short sequence motifs that are present in the promoter regions of a set of genes without a priori knowledge, such as MEME, and Gibbs sampler (Lawrence *et al*., 1993; Bailey and Elkan, 1994). While other techniques such as CONFAC, oPOSSUM and Toucan (Aerts *et al*., 2003; Karanam and Moreno, 2004; Ho Sui *et al*., 2005) search for known TFBS motifs from databases such as TRANSFAC and Jaspar (Sandelin *et al*., 2004). These approaches use statistical ranking functions, (e.g. Fisher exact test) to look for overrepresentation in each subset or grouping of genes.

Combinatorial interactions of multiple transcription factors play an important role in gene regulation. These take the form of *cis*-regulatory modules (CRMs) where combinations of TFBS can be linked to sets of co-expressed genes. For example, a large number of skeletal muscle associated TFBS have been experimentally characterized. In this tissue it was possible to link CRMs generated from five muscle associated transcription factor position specific scoring matrix (PSSMs) with skeletal muscle gene expression (Wasserman and Fickett, 1998). Software tools for statistically assessing the significance of the combinations of TFBSs are being developed in increasing numbers. These tools tend to rely either on reference collections of well-characterized CRMs or search for significant combinations of sites where CRM information is not available or both. The use of CRMs does not lend itself to the approach used in this paper, however, as a precompiled matrix of all possible interactions of all 1236 motifs with varying distances between each motif would be too cumbersome.

All these methods require a gene list. To generate a gene list from microarray data a gene expression threshold level must be assigned. All genes above this threshold are considered to be turned on, while all the genes below this threshold are said to be switched off. Therefore a gene may only have two values, on or off. This represents an enormous loss of detailed information compared with the quantitation in the original microarray data.

In this paper we take a simple yet highly effective multivariate analysis approach that uses as input, entire microarray datasets

*To whom correspondence should be addressed.

which are cross-referenced with known and predicted TFBS target information. We use correspondence analysis (CA) as a general-purpose dimension reduction and data exploration method (Benzécri, 1976). It can be used to find patterns of array grouping in gene expression data sets (Fellenberg *et al*., 2001) or it can be used to find patterns of motif occurrence in groups of genes.

Co-Inertia analysis (CIA, Dolédec and Chessel, 1994) is used to find trends in common between two different datasets that describe different characteristics of objects under investigation. In this case, you carry out CA so as to find axes from the two datasets that have maximum co-variance. It is a useful general-purpose data integration technique where you have two different sets of measurements carried out on the same objects.

In this paper we use CIA to integrate between two different sets of measurements on large collections of human genes: (1) counts of motif occurrences in gene promoters; (2) gene expression levels from combined microarray datasets (Gene Atlas) or a single dataset (arrays from different types of prostate cancer). This allows us to perform an unsupervised analysis that identifies motifs that are most associated with the main patterns of gene expression differences. Finally, we can use a technique called Between Groups Analysis (BGA, Dolédec and Chessel, 1987) to carry out a supervised analysis where groupings of arrays or tissues of a priori interest are contrasted with the rest.

2 SYSTEM AND METHODS

2.1 TFBSs

A combined total of 1236 TFBS from three different sources were used: TRANSFAC 6.2 Professional (Wingender *et al*., 1996) (414 vertebrate motifs), Jaspar (Sandelin *et al*., 2004) (81 motifs) and (Xie *et al*., 2005) (741 motifs). The promoter sequences were taken from the four-way conserved genome sequence alignments (human, mouse, rat and dog) as generated by (Xie *et al*., 2005). These data cover 2 kb upstream and 2 kb downstream of the annotated transcription start site for 14 590 human genes. The tffind program from the Piptools package (Elnitski *et al*., 2002) was used to generate a frequency matrix of the hits of the 1236 motifs in each of the 14 590 genes. Tffind requires the motifs to be present in the aligned promoter of all four species, rather than from the human promoter alone. This 'phylogenetic foot printing' greatly reduces the false positive rate. Tffind was run at four different PSSM thresholds, 0.7, 0.75, 0.8 and 0.85. The search parameters for tffind were: motifs were searched individually; no gaps were allowed; all sequences within the alignment were required to match the pattern.

2.2 Tissue expression data

The human tissue expression data were obtained from Gene Atlas v2.0 (Su *et al*., 2002). Gene Atlas gives combined expression levels for genes in each of 79 tissues. We excluded fetal and cancer tissues and only used normal adult tissue, leaving us with 67 tissues. There were 5 muscle tissue types (heart, tongue, skeletal muscle, smooth muscle and cardiomyocytes) and 16 CNS tissues (temporal lobe, globuspallidus, cerebellum peduncles, cerebellum, caudate nucleus, whole CNS, parietal lobe, medulla oblongata, amygdala, prefrontal cortex, occipital lobe, hypothalamus, thalamus, subthalamic nucleus, cingulate cortex and pons). The intersection between the aligned promoter sequences and the Gene Atlas gene set was 11 241 genes.

2.3 Cancer expression data

The prostate cancer dataset was obtained from Varambally *et al*. (2005). The data were downloaded from http://www.ncbi.nlm.nih.gov/geo/ (Gene Expression Omnibus, accession number: GSE3325) as raw data files (.cel files). Gene expression values were called using the robust multi-chip average method (Irizarry *et al*., 2003) and data were quantile normalized using the Bioconductor package, affy. The original data contained 54 675 affy probes and 19 samples. This dataset covered four groups of samples from prostate tumours: benign (6 samples), primary (7 samples), metastatic androgen receptor positive (2 samples) and metastatic androgen receptor negative (4 samples). In this paper we compare the two metastatic groups to the benign. The RefSeq ids that corresponded to the affy probes were obtained using the hgu133plus2 annotation library. Probes that hit multiple genes were filtered out. The intersection between the aligned promoter sequences and the hgu133plus2 gene set was 11 061 genes, If there were multiple probes for the same gene, the probes were averaged for that gene.

2.4 Rank products

The Rank Products method was developed for identifying differentially expressed genes in cDNA expression data (Breitling *et al*., 2004; Breitling and Herzyk, 2005). It is based on the argument that a gene in an experiment examining n genes in k replicates, has a probability of being ranked first (Rank 1) of $1/n^k$ if the lists were entirely random. Therefore, it is unlikely for a single gene to be in the top position in all replicates if this gene was not differentially expressed. More generally, for each gene g in k replicates i, each examining n_i genes, one can calculate the corresponding combined probability as a rank product

$$\mathrm{RP}_g{}^{\mathrm{up}} = \prod_{i=1}^{\mathrm{k}} (r_{i,g}{}^{\mathrm{up}}/n_i)$$

Where $r_{i,g}{}^{\mathrm{up}}$ is the position of gene g in the list of genes in the i-th replicate sorted by decreasing fold change, i.e. $r^{\mathrm{up}} = 1$ for the most strongly upregulated gene, etc. The genes can then be sorted according to their RP value.

In this work, we used a combination of multivariate analysis methods to make ranked lists of TFBS that appear to be associated with gene expression differences between sets of tissue samples. We used 4 different PSSM thresholds with tffind giving four different ranked lists of TFBS. The Rank Products method was then used to assign a value to each TFBS according to how it ranked in the four lists. This allows TFBS information at different stringencies to be combined in a practical manner and allows us to incorporate TFBS from a range of information contents. The significance of the RP scores is assessed by comparison to 100 randomized datasets. These give P-values for each motif. Using these P-values as probabilities assumes that the 4 lists from the 4 PSSM thresholds are independent. This is clearly not so, and we just use the P-values to rank the motifs. Any motifs with a P-value of ≤ 0.0001 are reported in this paper.

2.5 CIA

To study simultaneously two linked data tables, we used CIA, a coupling approach that was first introduced to study ecological data (Dolédec and Chessel, 1994). We already used this method to perform a cross-platform comparison of gene expression data measurements (Culhane *et al*., 2003). It is similar to the partial least square (PLS) regression method (Höskuldsson, 1988) and, like PLS, it can be used when the number of variables is greater than the number of observations, which is usually the case with gene expression data.

CIA must be used in combination with ordination methods like CA or principal component analysis. Those methods summarize a table by searching for orthogonal axes on which the projection of the sampling points (rows) have the highest possible variance. Using the axes produced by the analyses computed on the original data tables, CIA searches for successive pairs of axes that maximize their covariance.

The mathematical basis of CIA is summarized below. Let X and Y be the original data tables, with n rows, and respectively p and q columns. The two statistical triplets produced by the ordination methods performed on the data sets are noted (X, D_n, D_p) and (Y, D_n, D_q), with D_n and D_p the diagonal matrices containing rows and columns weights for X, and D_n and D_q the

diagonal matrices containing the rows and columns weights for Y. After diagonalization let u and v be a pair of eigenvectors for (X, D_n, D_p) and (Y, D_n, D_q) respectively. The projection of the multidimensional space associated with X on to vector u generates n coordinates in a column matrix:

$$\xi = XD_p u \quad (1)$$

The projection of the multidimensional space associated with table Y on to vector v generates n coordinates in a column matrix:

$$\psi = YD_q v \quad (2)$$

Co-inertia associated with the pair of vectors u and v is equal to:

$$H(u, v) = \xi^t D\psi \quad (3)$$

If the initial data tables are centered, then the co-inertia is the covariance between the two new scores:

$$\mathrm{Cov}(\xi, \psi) = \mathrm{Corr}(\xi, \psi)\sqrt{\mathrm{Iner}_1(u)\,\mathrm{Iner}_2(v)} \quad (4)$$

with $\mathrm{Iner}_1(u)$ the projected inertia on to vector u (i.e. the variance of the new scores on u), $\mathrm{Iner}_2(v)$ the projected inertia on to vector v (i.e. the variance of the new scores on v), and $\mathrm{Corr}(\xi, \psi)$ the correlation between the two coordinate systems. A CIA axis associated with a pair of eigenvectors u and v will maximize $\mathrm{Cov}(\xi, \psi)$.

BGA is a supervised classification method used in combination with ordination methods and similar to discriminant analysis (Culhane *et al*., 2002; Dolédec and Chessel, 1987). Once an ordination has been computed on a pre-ordered data set in which the different groups are defined, BGA finds the linear combination of the axes that maximizes between-group variance and minimizes within-group variance. Like CIA it can be used when the number of variables is greater than the number of observations.

We apply a combination of BGA and CIA to matched tables of motif occurrences and tissue or array expression levels. This forces the analysis to rank arrays or tissues along a single axis that best discriminates the two groups of samples (e.g. metastatic versus benign prostate cancer samples) and a second axis with ranked motifs. The two axes are found as the ones that maximize their covariance. The motifs that are ranked at the ends of the motif axis will be the ones that are most associated with the genes that are highly expressed in the samples at the same end of the sample axis. Thus, for each split in the data that we specifiy, using BGA, we get a ranked list of motifs. We get a separate ranked list of motifs for each of the four tffind PSSM thresholds that we use. The results from the four are combined using the Rank Products statistic. This produces a single ranked list of motifs. All calculations were carried out using the MADE4 library (Culhane *et al*., 2005; Thioulouse *et al*., 1997) of the open source *R* package. MADE4 can be downloaded freely from the Bioconductor web site (www.bioconductor.org).

2.6 Comparing co-inertia to oPOSSUM

Opossum is a method for the identification of over-represented TFBSs in sets of co-expressed genes (Ho Sui *et al*., 2005). Unlike CIA, this technique requires a list of related genes for analysis. To compare our method to oPOSSUM we reran the prostate dataset analysis using only the TFBS from JASPAR. This contains 81 vertebebrate motifs. The feature selection method SAM (Tusher *et al*., 2001) was used to generate a gene list of 300 refseq ids from the prostate cancer dataset. oPOSSUM was run with default settings. This selects the top 10 motifs associated with prostate cancer, calculated by the *Z*-score statistic and the top 10 motifs calculated with the Fisher *P*-value.

3 IMPLEMENTATION

3.1 Co-inertia applied to gene atlas data

We first applied CIA to find TFBS motifs associated with the different tissue types present in Gene Atlas (Su *et al*., 2002).

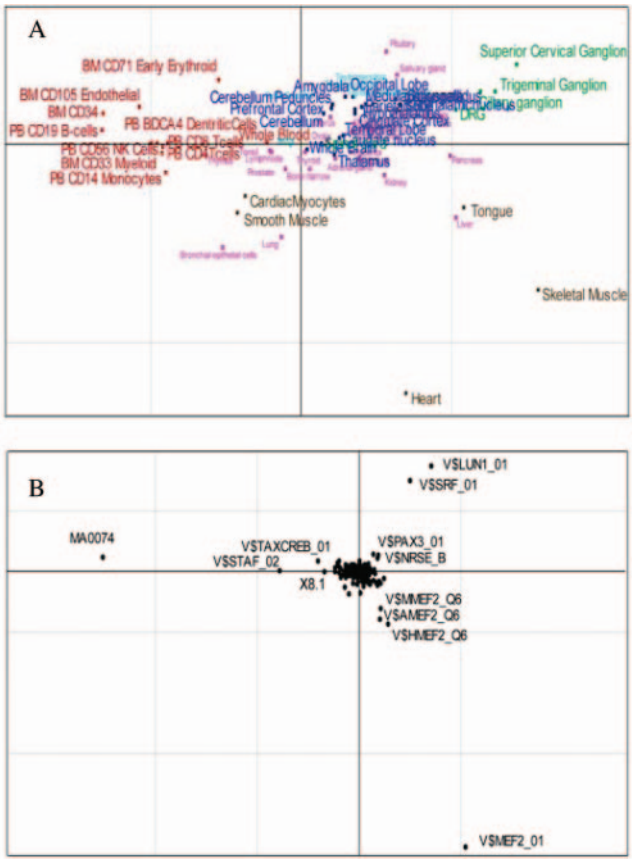

Fig. 1. Axes 1 (horizontal) and 3 (vertical) of the unsupervised CIA for the gene atlas dataset. A transcription factor motif matrix produced with a PSSM threshold of 0.85 was used. (A) shows the projection of the tissue samples. The blood (red), CNS (blue) and the peripheral nervous system (green) samples are split along the first axis, while the muscle (brown) samples are separated from these by the third axis. (B) Shows the projection of the TFBS motifs. Motifs that are in the same orientation (direction from the origin) as a group of samples are associated with those samples.

Unsupervised CIA was applied to gene expression data for 67 normal adult tissues and the associated TFBS motif/gene matrix. The axes of the resulting CIA were visually inspected to check for biologically interesting associations. Axes one and three are used here for illustration and are plotted in Figure 1.

In Figure 1a, the blood (red), CNS (blue) and the peripheral nervous system (green) samples are split along the first axis, while the muscle (brown) samples are separated from these by the third axis. Figure 1b shows the motifs associated with these trends. The most extreme motifs along each axis are labelled and named. For muscle, the first four labelled motifs from the bottom right hand corner, moving towards the origin, are binding sites for MEF2, which is a characteristic muscle specific transcription factor (Pollock and Treisman, 1991). LUN1 has significant expression levels in the CNS, although little work has been done on the role of this transcription factor in neural tissue. MA0074 refers to the vitamin D nuclear receptor which is expressed in a number of different blood tissues, specifically those associated with the immune system (Bhalla *et al*., 1983). Also, the muscle characteristic transcription factors SRF (V$SRF_01, serum response factor) and

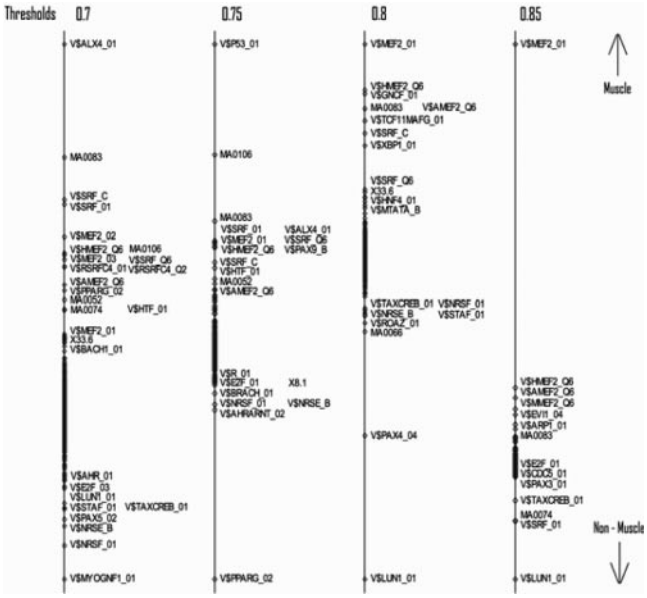

Fig. 2. Supervised co-inertia results for the muscle dataset. The co-inertia analyses were preformed using the tissue expression data and a transcription factor motif matrix. Each axis shows the projection of the motifs that are associated with the muscle tissues produced with four PSSM thresholds of 0.7, 0.75, 0.8 and 0.85 respectively.

PAX3 (V$PAX3_01, paired box 3) are associated with the CNS in Figure 1. This is consistent with the biological information as PAX3 is a key factor in CNS development (Chalepakis *et al*., 1993) and SRF is required for neuronal synapse plasticity (Ramanan *et al*., 2005).

3.2 Motif/gene thresholds

The CIA in Figure 1 was performed using a motif/gene matrix that was produced using a PSSM threshold of 0.85. This PSSM threshold produces a relatively clean signal but has a high false negative rate with high information content motifs.

To overcome this problem we perform four supervised co-inertia analyses, at PSSM thresholds of 0.85, 0.8, 0.75 and 0.7. This is done in order to get a range of signal to noise ratios. The highest PSSM thresholds will have little signal for high information content motifs, while the signal for the low information content motifs in the lower threshold matrices will be indiscernible from the background noise. These four thresholds are combined with supervised co-inertia analyses and the Rank Products statistic (Breitling *et al*., 2004).

3.3 Supervised CIA

We used a combination of CIA and BGA to discriminate groups of samples instead of the individual samples (e.g. muscle tissues versus all other tissues). BGA is applied to the samples of the gene expression data matrix. CIA is applied to this BGA at each of the four motif/gene matrices (one for each PSSM threshold). This returns four lists of promoter motifs that are ranked based on the motif's association with the group of interest (Fig. 2).

We grouped the muscle (5 tissues) and the CNS (16 tissues) related tissues and used this technique to identify what motifs are most characteristic of each of these tissue groups relative to all other tissues. Table 1 shows motifs returned by the comparisons.

Table 1. Motifs associated with muscle and CNS tissues in Gene Atlas

TF	Motif ID	Description
Muscle		
MEF2	V$MEF2_01 V$HMEF2_Q6 V$AMEF2_Q6 V$RSRFC4_01 V$MEF2_02 MA0052 V$MMEF2_Q6 V$MEF2_03 V$RSRFC4_Q2	MADS box transcription enhancer factor 2
SRF	MA0083 V$SRF_C V$SRF_Q6 X121.1	Serum response factor
Unknowns	X33.6 X121.8 X33.7	
MTATA	V$MTATA_B	The muscle TATA box
NR2F2	V$ARP1_01	Nuclear receptor subfamily 2, group f, member 2
HNF-1	MA0046	Hepatocyte nuclear factor 1
EVI-1	V$EVI1_04	Ecotropic viral integration site 1
NR2F1	MA0017	Nuclear receptor subfamily 2, group f, member 1
TATA	V$TATA_01	The TATA box
TEF-1	MA0090	Transcriptional enhancer factor 1
CNS		
NRSF	V$NRSF_01 V$NRSE_B	Neuron-restrictive silencer factor
LUN1	V$LUN1_01	Lung 1
MIBP1/RFX1	V$MIF1_01	The myc intron binding protein/regulatory factor X1 heterodimer
EBF	V$ROAZ_01	Early B-cell factor
Unknowns	X67.5 X43.2 X43.1	
AP-4	V$AP4_01	Activator protein 4
PBX1	V$PBX1_02	Pre-B-cell leukemia transcription factor 1

The motif ID codes are from the three parent databases for all significant motifs [X*n.n* are from Xie *et al*. (2005); MA*n* are from Sandelin *et al*. (2004); V$*n* are from Wingender *et al*. (1996)].

Previous studies identified five classes of transcription factors that are characteristic of skeletal muscle-specific expression (Wasserman and Fickett, 1998), three of which are well represented in our results: (1) MEF-2, which is strongly expressed in both cardiac and skeletal muscle (Pollock and Treisman, 1991); (2) SRF, a mitogen-responsive factor required for the activation of key muscle proteins such as cardiac alpha-actin (Vandromme *et al*., 1992; Chang *et al*., 2003); (3) TEF-1, a co-factor essential for muscle specific gene regulation via the MACT element (Gunther *et al*., 2004). It must be pointed out that this combined set of motifs from three sources is redundant. Many motifs are effectively duplicated between the datasets. These can be combined and clustered in various ways but this is complicated as the duplicated motifs usually differ slightly from each other.

The absence of the two other classes from our list (MYOD and Sp1) can be explained by the fact that our list is a composite of five

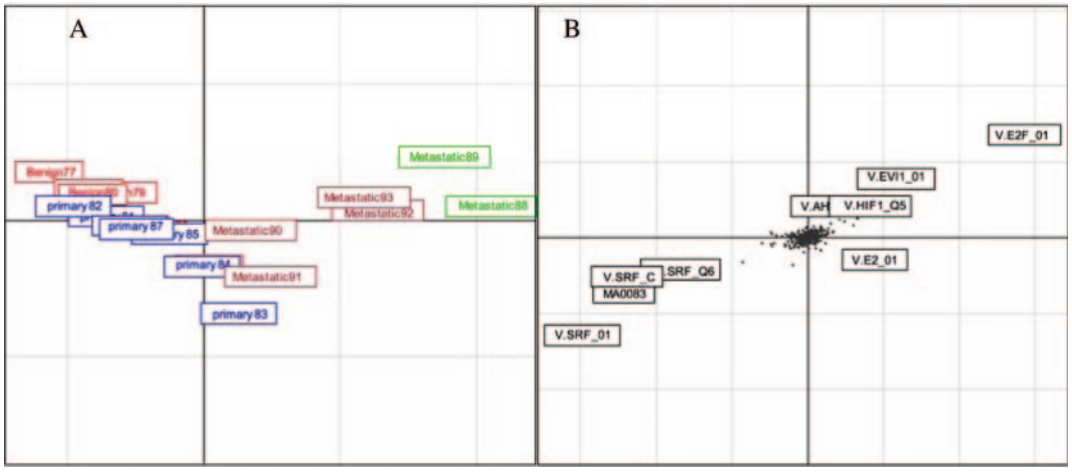

Fig. 3. Unsupervised CIA for the prostate dataset. A transcription factor motif matrix produced with a PSSM threshold of 0.85 was used. (**a**) Shows the projection of the tissue samples. The benign (red), primary (blue) and the AR+ metastatic (green) and AR− metastatic (brown) samples are split along the *x*-axis. The benign (red) samples are separated from the the primary samples (blue) by the *y*-axis. (**b**) Shows the projection of the TFBS motifs. Motifs that are in the same orientation (direction from the origin) as a group of samples are associated with those samples.

muscle tissue types rather than just skeletal muscle. Also, previous studies have found it difficult to discriminate Sp1 because of the ubiquitous nature of its action and the relatively low information content of its motif (Ho Sui *et al*., 2005). MyoD may be missed because it binds to the E-box sequence which is found in many non-muscle genes (Goswami *et al*., 1993).

Table 1 also lists the motifs characteristic of the CNS, the most striking of which is NRSE (neuron restrictive silencer element). This is the DNA-binding motif of the NRSF/REST protein which restricts the expression of neural-related genes to the CNS (Schoenherr and Anderson, 1995). All of the other transcription factors relating to the CNS are known to regulate genes in tissues of the CNS [LUN1 (Bredel *et al*., 2005), MIBP1/RFX1 (Fukuda *et al*., 2002; Ma *et al*., 2006), EBF (Tsai and Reed, 1997), PBX1 (Roberts *et al*., 1995), AP-4 (Dilaver *et al*., 2003)].

3.4 Prostate cancer dataset

In the previous section we described results for a large repository of gene expression data for a range of tissues. A second scenario is where two sets of arrays are from two disease states (e.g. normal versus disease or disease type 1 versus disease type 2). In the example we use here, we wish to find transcriptional pathways associated with metastasis in prostate cancer.

Figure 3 shows the unsupervised CIA. The *x*-axis separates the benign and the primary tumours, and the *y*-axis separates these two groups from the metastatic groups. Using the supervised co-inertia technique, we compared both metastatic groups (six samples) together against benign (six samples). The resulting axes are shown in Figure 4. Table 2 shows the motifs that were returned associated with the metastatic samples. 10 of the 13 TFs identified by these TFBS motifs relate to transcription factors with known associations to prostate cancer E2F (Foster *et al*., 2004), HIF1 (Du *et al*., 2003), EBF (Regnauld *et al*., 2002), LXR (Fukuchi *et al*., 2004), c-Myc/Max (Buttyan *et al*., 1987), AR (Benson *et al*., 1985), TP53 (Thomas *et al*., 1993), RXR/VDR (Taylor *et al*., 1996), ARNT and AHR/ARNT (Kashani *et al*., 1998).

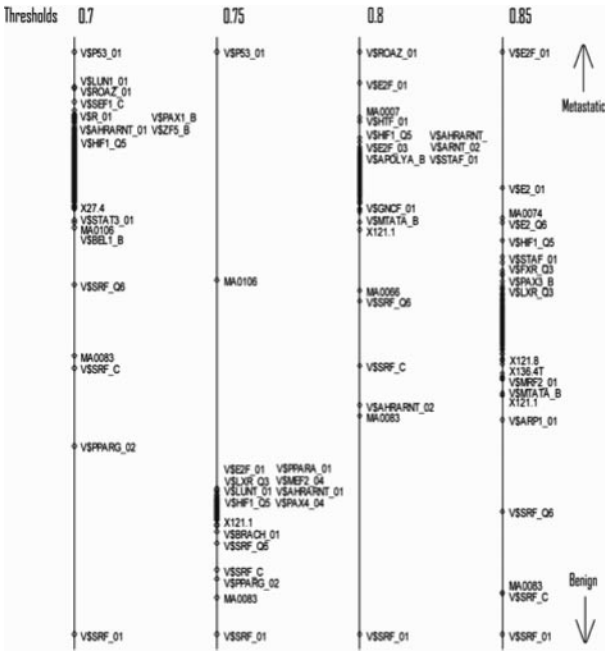

Fig. 4. Supervised co-inertia results for the prostate dataset. The co-inertia analyses were preformed using the prostate expression data and a transcription factor motif matrix. Each axis shows the projection of the motifs that are associated with metastasis. These were produced with four PSSM thresholds of 0.7, 0.75, 0.8 and 0.85 respectively.

V$NMYC_01, V$STAF_01, V$ZF5_B are the motifs not previously associated with prostate cancer. N-myc is a member of the MYC family of oncogenes that encode nuclear proteins serving as transcription factors. Although has not been shown to be involved with prostate cancer, its TFBS is similar to the c-myc binding site and so there is a high overlap in their target genes. V$STAF_01 is the SPH motif which is bound by ZFP143 and Zinc finger protein 76. ZFP143 is the human ortholog of the Xenopus Staf protein.

Table 2. Motifs associated with metastatic prostate tumours when compared with benign tumours

TF	Motif ID	Description
E2F	V$E2F_01 V$E2F_03	E2F transcription factor 1
HIF1	V$HIF1_Q5	Hypoxia induced factor 1
AHR/ARNT	V$AHRARNT_01	Aryl hydrocarbon receptor nuclear transporter
EBF	V$ROAZ_01	Early B-cell factor
ARNT	V$ARNT_02	Aryl hydrocarbon receptor nuclear transporter
LXR	V$LXR_Q3	Liver X receptor
c-Myc/Max	V$MYCMAX_03	Avian myelocytomatosis viral oncogene homolog/ MYC-accociated factor X heterodimer
n-Myc	V$NMYC_01	N-MYC avina myelocytomatosis viral related oncogene homolog
AR	MA0007	Androgen receptor
TP53	V$P53_01	Tumor protein p53
ZNF143 ZNF76	V$STAF_01	SPH motif is bound by Zinc finger protein 143 and Zinc finger protein 76
ZFP161	V$ZF5_B	Zinc finger protein 161
RXR/VDR	MA0074	Retinoid X receptor/vitamin D receptor heterodimer

The motif ID codes are the same as in Table 1.

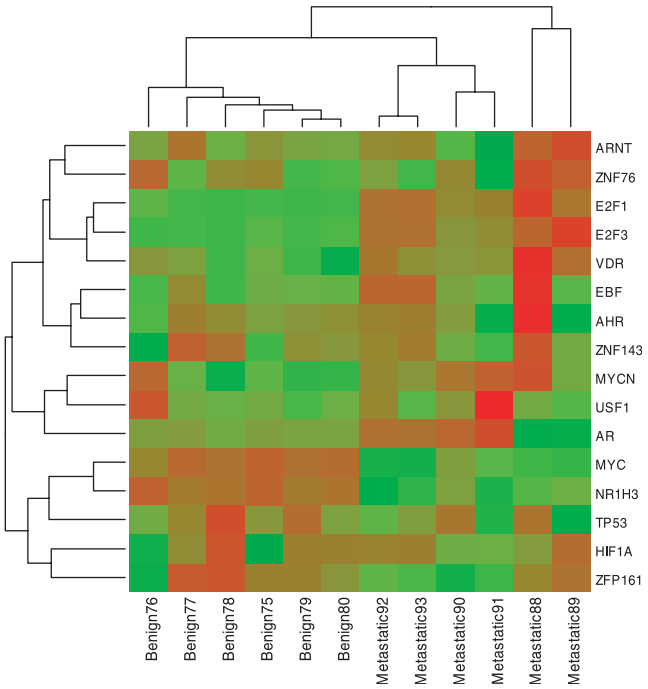

Fig. 5. Heatmap of the gene expression data of oncogenic transcription factors that were found to be associated with metastatic prostate tumours.

ZFP76 is a novel protein related to staf (Myslinski *et al*., 1998). They are thought to be involved in normal and abnormal cellular proliferation and differentiation. V$ZF5_B is bound by ZFP143 which is a transcriptional repressor of c-myc whose expression is known to be associated to prostate cancer (Kaplan and Calame, 1997).

Figure 5 is a heatmap of the expression data for most of the transcription factors, whose TFBS were identified above. We find that the gene expression data of the oncogenic transcription factors identified, do appear to be differentially expressed between the disease states. The AR− and the AR+ metastatic groups are clearly identifiable, both from the topology of the dendrogram and the expression of the AR transcription factor. Also, a clear divide between the benign and metastasis groups can be seen.

3.5 Comparison with oPOSSUM

Submitting the 300 genes to oPOSSUM returned 13 motifs that were in either list (there was overlap between the *Z*-score list and the Fisher *P*-value list). Six of these motifs are consistent with those obtained with the CIA, when applied to the JASPAR dataset. Five of these have been shown to be involved in Prostate cancer [USF (Chen *et al*., 2006), Max (Buttyan *et al*., 1987), RXR/VDR (Taylor *et al*., 1996), Ahr-ARNT and ARNT (Kashani *et al*., 1998)]. N-Myc was found by both methods but is not known to be involved in prostate cancer. Motifs that CIA finds that are not present in the oPOSSUM results but are involved with prostate cancer are AR (Benson *et al*., 1985), Myc-Max (Buttyan *et al*., 1987), p50 (Palayoor *et al*., 1999), E2F (Foster *et al*., 2004), AP2alpha (Zi *et al*., 2000). Pax6 and staf were identified by us but are not known to be involved in this disease. Staf was also found in the original analysis and is believed to be involved in cellular proliferation and differentiation. oPOSSUM identifies TP53, Elk-1, NRF-2, HLF, deltaEF1, cEBP and CREB as associated with the metastatic group. TP53 is ranked at number 15 by CIA and is involved in the disease (Thomas *et al*., 1993). Elk-1 (Xiao *et al*., 2002) and CREB (Xiao *et al*., 2002) are also involved in the disease but were poorly ranked by CIA.

4 DISCUSSION

One complication with combining sets of TFBS from different sources is the redundancy between different datasets. Usually, you have to employ some kind of clustering method to combine motifs. This is not trivial and each clustering will result in a different set of motifs. CIA, as used here, has the great advantage of not requiring any clustering to work effectively. The dimension reduction that comes from the underlying CA, sorts out any redundancy. This allows you to combine data sets as you find them.

We applied our method to the gene atlas tissue microarray dataset. It was observed that the TFBS motifs returned were of known importance to the tissues being investigated. Of the nine transcription factors associated with muscle, three were previously described muscle transcription factors (Wasserman and Fickett, 1998). The other six transcription factors had significant muscle expression profiles.

Our findings with CNS tissues were similar, with 11 TFBS motifs returned, of which 7 were of known function. The list included two neural specific motifs, and all are expressed in the CNS.

Investigation of the prostate cancer dataset identified motifs that are bound by transcription factors that are known to be linked to the

pathology of the disease, as well as a small number of novel putative therapeutic targets.

Although this technique is clearly useful in the identification of TFBS associated with both tissues and disease states, it is limited by the availability of transcription factor motifs. At present, it is roughly estimated that there are between 1500 and 3000 transcription factors in the human genome, only a small proportion of which have well characterized TFBS. Even though we used 1236 motifs from three different collections the majority of these have not been assigned to a transcription factor. Nonetheless, this situation will rapidly become clearer over the next few years as more data become available. For example, in vertebrates there are on-going genome-sequencing efforts for over a dozen mammals.

Many methods provide the option to change the PSSM threshold. This is because of the degenerate nature of the TFBS. High information content TFBS will be rare at higher PSSM thresholds, while low information TFBS will be ubiquitous at lower PSSM thresholds. We use four thresholds of 0.7, 0.75, 0.8 and 0.85 and by looking for complete or partial consistency in the results we believe we get a better sensitivity and specificity, then using any single PSSM threshold.

Our method also provides uniquely powerful visualization tools due the ordination techniques employed. This enables the researcher to investigate in an unsupervised manner the strongest trends between the gene expression data and the TFBS.

ACKNOWLEDGEMENTS

We thank Ailís Fagan for helpful advice. We also gratefully acknowledge funding from the European Union Fifth Framework Programme and Science Foundation Ireland.

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al*. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*., **31**, 1753–1764.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **2**, 28–36.

Benson,R.C.,Jr *et al*. (1985) Androgen receptor binding activity in human prostate cancer. *Cancer*, **55**, 382–388.

Benzécri,J. (1976) *L'Analyse de Données. II. L'Analyse des Correspondances*. Dunod, Paris.

Bhalla,A.K. *et al*. (1983) Specific high-affinity receptors for 1,25-dihydroxyvitamin D3 in human peripheral blood mononuclear cells: presence in monocytes and induction in T lymphocytes following activation. *J. Clin. Endocrinol. Metab*., **57**, 1308–1310.

Bredel,M. *et al*. (2005) High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res*., **65**, 4088–4096.

Breitling,R. *et al*. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*., **573**, 83–92.

Breitling,R. and Herzyk,P. (2005) Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J. Bioinform. Comput. Biol*., **3**, 1171–1189.

Buttyan,R. *et al*. (1987) Enhanced expression of the c-myc protooncogene in high-grade human prostate cancers. *Prostate*, **11**, 327–337.

Chalepakis,G. *et al*. (1993) Pax: gene regulators in the developing nervous system. *J. Neurobiol*., **24**, 1367–1384.

Chang,J. *et al*. (2003) Inhibitory cardiac transcription factor, SRF-N, is generated by caspase 3 cleavage in human heart failure and attenuated by ventricular unloading. *Circulation*, **108**, 407–413.

Chen,N. *et al*. (2006) Tumor-suppression function of transcription factor USF2 in prostate carcinogenesis. *Oncogene*, **25**, 579–587.

Culhane,A.C. *et al*. (2002) Between-group analysis of microarray data. *Bioinformatics*, **18**, 1600–1608.

Culhane,A.C. *et al*. (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**, 59.

Culhane,A.C. *et al*. (2005) MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, **21**, 2789–2790.

Dilaver,G. *et al*. (2003) Colocalisation of the protein tyrosine phosphatases PTP-SL and PTPBR7 with beta4-adaptin in neuronal cells. *Histochem. Cell Biol*., **119**, 1–13.

Dolédec,S. and Chessel,D. (1987) Rhythmes saisonniers et composantes stationelles en milieu aquatique I—Description d'un plan d'observations complet par projection de variables. *Acta Oecologica Oecologica Generalis*, **8**, 403–426.

Dolédec,S. and Chessel,D. (1994) Co-inertia analysis: an alternative method for studying species-environment re-lationships. *Freshw. Biol*., **31**, 277–294.

Du,Z. *et al*. (2003) Expression of hypoxia-inducible factor 1alpha in human normal, benign, and malignant prostate tissue. *Chin. Med. J. (Engl.)*, **116**, 1936–1939.

Elnitski,L. *et al*. (2002) PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. *Genomics*, **80**, 681–690.

Fellenberg,K. *et al*. (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 10781–10786.

Foster,C.S. *et al*. (2004) Transcription factor E2F3 overexpressed in prostate cancer independently predicts clinical outcome. *Oncogene*, **23**, 5871–5879.

Fukuchi,J. *et al*. (2004) Antiproliferative effect of liver X receptor agonists on LNCaP human prostate cancer cells. *Cancer Res*., **64**, 7686–7689.

Fukuda,S. *et al*. (2002) Characterization of the biological functions of a transcription factor, c-myc intron binding protein 1 (MIBP1). *J. Biochem. (Tokyo)*, **131**, 349–357.

Goswami,S.K. *et al*. (1993) MyoD transactivates angiotensinogen promoter in fibroblast C3H10T1/2 cells. *Cell. Mol. Biol. Res*., **39**, 125–130.

Gunther,S. *et al*. (2004) VITO-1 is an essential cofactor of TEF1-dependent muscle-specific gene regulation. *Nucleic Acids Res*., **32**, 791–802.

Ho Sui,S.J. *et al*. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*., **33**, 3154–3164.

Höskuldsson,A. (1988) PLS regression methods. *J. Chemomet*., **2**, 211–228.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kaplan,J. and Calame,K. (1997) The ZiN/POZ domain of ZF5 is required for both transcriptional activation and repression. *Nucleic Acids Res*., **25**, 1108–1116.

Karanam,S. and Moreno,C.S. (2004) CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. *Nucleic Acids Res*., **32**, W475–W484.

Kashani,M. *et al*. (1998) Expression of the aryl hydrocarbon receptor (AhR) and the aryl hydrocarbon receptor nuclear translocator (ARNT) in fetal, benign hyperplastic, and malignant prostate. *Prostate*, **37**, 98–108.

Lawrence,C.E. *et al*. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Ma,K. *et al*. (2006) The transcription factor regulatory factor X1 increases the expression of neuronal glutamate transporter type 3. *J. Biol. Chem*., **281**, 21250–21255.

Myslinski,E. *et al*. (1998) ZNF76 and ZNF143 are two human homologs of the transcriptional activator Staf. *J. Biol. Chem*., **273**, 21998–22006.

Palayoor,S.T. *et al*. (1999) Constitutive activation of IkappaB kinase alpha and NF-kappaB in prostate cancer cells is inhibited by ibuprofen. *Oncogene*, **18**, 7389–7394.

Pollock,R. and Treisman,R. (1991) Human SRF-related proteins: DNA-binding properties and potential regulatory targets. *Genes Dev*., **5**, 2327–2341.

Ramanan,N. *et al*. (2005) SRF mediates activity-induced gene expression and synaptic plasticity but not neuronal viability. *Nat. Neurosci*., **8**, 759–767.

Regnauld,K. *et al*. (2002) G-protein alpha(olf) subunit promotes cellular invasion, survival, and neuroendocrine differentiation in digestive and urogenital epithelial cells. *Oncogene*, **21**, 4020–4031.

Roberts,V.J., van Dijk,M.A. and Murre,C. (1995) Localization of Pbx1 transcripts in developing rat embryos. *Mech. Dev*., **51**, 193–198.

Sandelin,A. *et al*. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.

Schoenherr,C.J. and Anderson,D.J. (1995) The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science*, **267**, 1360–1363.

Su,A.I. *et al*. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.

Taylor,J.A. *et al*. (1996) Association of prostate cancer with vitamin D receptor gene polymorphism. *Cancer Res*., **56**, 4108–4110.

Thioulouse,J. *et al*. (1997) ADE-4: a multivariate analysis and graphical display software. *Stat. Comput*., **7**, 75–83.

Thomas,D.J. *et al.* (1993) p53 expression and clinical outcome in prostate cancer. *Br. J. Urol.*, **72**, 778–781.

Tsai,R.Y. and Reed,R.R. (1997) Cloning and functional characterization of Roaz, a zinc finger protein that interacts with O/E-1 to regulate gene expression: implications for olfactory neuronal development. *J. Neurosci.*, **17**, 4159–4169.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Vandromme,M. *et al.* (1992) Serum response factor p67SRF is expressed and required during myogenic differentiation of both mouse C2 and rat L6 muscle cell lines. *J. Cell. Biol.*, **118**, 1489–1500.

Varambally,S. *et al.* (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, **8**, 393–406.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Wingender,E. *et al.* (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Xiao,D. *et al.* (2002) GRP receptor-mediated immediate early gene expression and transcription factor Elk-1 activation in prostate cancer cells. *Regul. Pept.*, **109**, 141–148.

Xie,X. *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Zi,X. *et al.* (2000) Impairment of erbB1 receptor and fluid-phase endocytosis and associated mitogenic signaling by inositol hexaphosphate in human prostate carcinoma DU145 cells. *Carcinogenesis*, **21**, 2225–2235.