Functional modules integrating essential cellular functions are predictive of the response of leukaemia cells to DNA damage

Katrin Sameith1,†, Philipp Antczak1, Elliot Marston2, Nil Turan1, Dieter Maier3, Tanja Stankovic2 and Francesco Falciani1,*

1School of Biosciences and Institute of Biomedical Research (IBR), 2CRUK Institute for Cancer Studies, University of Birmingham, Birmingham, B152TT, UK and 3Biomax Informatics AG, Lochhamer Str. 9, 82152 Martinsried, Germany

Received on March 12, 2008; revised on September 5, 2008; accepted on September 12, 2008
Advance Access publication September 17, 2008

ABSTRACT
Motivation: Childhood B-precursor lymphoblastic leukaemia (ALL) is the most common paediatric malignancy. Despite the fact that 80% of ALL patients respond to anti-cancer drugs, the patho-physiology of this disease is still not fully understood. mRNA expression-profiling studies that have been performed have not yet provided novel insights into the mechanisms behind cellular response to DNA damage. More powerful data analysis techniques may be required for identifying novel functional pathways involved in the cellular responses to DNA damage.

Results: In order to explore the possibility that unforeseen biological processes may be involved in the response to DNA damage, we have developed and applied a novel procedure for the identification of functional modules in ALL cells. We have discovered that the overall activity of functional modules integrating protein degradation and mRNA processing is predictive of response to DNA damage.

Availability: Supplementary material including R code, additional results, experimental datasets, as well as a detailed description of the methodology are available at http://www.bip.bham.ac.uk/vivo/fumo.html.

Contact: f.falciani@bham.ac.uk

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Childhood B-precursor lymphoblastic leukaemia (ALL) is the most common paediatric malignancy and it is curable in 80% of the patients with chemotherapy. However, the patho-physiology of ALL, particularly in relation to the mechanism behind response to treatment is still not fully understood. We have recently demonstrated that transcriptional response of leukemic cells to ionizing irradiation (IR)-induced DNA damage is complex and includes a range of pro-apoptotic as well as pro-survival genes (Stankovic et al., 2004). Furthermore, microarray studies have been performed with the objective of identifying genes associated to response to DNA damaging chemotherapeutic agents (Hollemann et al., 2004). Despite the efficacy in identifying molecular markers of response to chemotherapy agents these approaches have not provided novel insights into the mechanisms behind cellular response to DNA damage. This may be at least in part a consequence of the low sensitivity of the statistical methods being employed and of the intrinsic difficulties in the interpretation of the results of large-scale datasets. In order to address these issues a number of research groups have proposed methodologies, which do not rely on the analysis of individual mRNA expression profiles. Instead, they are based on a measure of overall activity of gene modules (Park et al., 2007). Hartwell et al. (1999) first defined a functional module as a ‘discrete unit whose function is separable from those of other modules’. However, the exact meaning of modularity depends on the network under consideration. A protein complex or a metabolic pathway may be defined as a highly interconnected region of a larger network (Bader and Hogue, 2003) characterized by a functional similarity between interacting proteins (Lubovac et al., 2006). On the other hand, a transcriptional module may be considered as a set of genes that is regulated in concert by a common regulation program as observed in mRNA expression profiles (Segal et al., 2003). In the last few years methods to identify functional modules integrating physical interaction networks and mRNA expression data have been developed with the purpose of identifying sub-networks, which are enriched of genes sharing a given property. Ideker et al. (2002) proposed a method designed to identify modules of connected proteins enriched of corresponding genes, which are differentially regulated over a range of conditions. More recently, Maraziotis et al. (2007) and Ulitsky and Shamir (2007) developed methods to identify gene interaction networks whose members are also transcriptionally coupled. These methods, in their initial implementation, are based on Pearson correlation coefficient to estimate the degree of relationship between different genes. They start from seed proteins to grow sub-networks of corresponding highly correlated genes via a local optimization procedure. In this study, we report the development and the application of a new methodology for module identification, which applies iterated simulated annealing to discover potentially overlapping sub-networks in a large network of physical interactions whose entities show a significant high co-expression on the transcriptional level. The methodology we have developed has two major differences with respect to the previously mentioned methods.
The first is the fact that it is based on mutual information (MI). The use of MI offers the advantage that non-linear relationships can also be detected. Moreover, recent validation studies performed both on experimental and simulated data (Basso et al., 2005) have demonstrated that MI is superior to other correlation measures to infer biologically relevant relationships from gene expression data. The second difference is that the optimization strategy is based on simulated annealing, a global optimization method that may allow a more efficient exploration of the search space. In order to apply this approach to identify new associations between molecular functions and response to DNA damage in ALL, we have also performed a new study representative of the gene expression profiles of ALL cells before and shortly after exposure to IR as a DNA damaging agent. This experimental model is clinically relevant, since there is a high degree of relationship between responses to IR-induced DNA damage and clearance of leukaemia blasts after chemotherapy treatment in vivo (Stankovic et al., 2004). Our computational analysis identifies 37 modules representing a range of cellular functions, which are active in ALL cancer cells. Here, we show that the transcriptional activity of a subset of these modules (10) is predictive of ALL cells response to IR. In addition to biological functions already known to be modulated in this process (i.e. cell cycle and apoptosis), we have identified a highly predictive network representing the interaction between the mRNA processing machinery, targeted protein degradation and immune system receptor signalling.

2 DATASETS AND METHODS

2.1 Experimental system: response to DNA damage in B-precursor ALL tumour samples

In this study, we have used response to IR in B-cell precursor ALL tumour samples as a relevant experimental model (Stankovic et al., 2004). Diagnosis of ALL was based on standard morphologic and immune-phenotypic evaluation. Cells representing a population of 22 paediatric B-precursor ALL tumour samples immediately before and 8 h after exposure to IR have been profiled using the AffymetrixTM microarray platform (Array U133A). Data have been normalized and pre-processed as described in the Supplementary Material. The final dataset contained measurements of 10,795 genes in 44 biological samples.

2.2 Integration of molecular interaction and pathway databases

The module discovery procedure we have developed has been applied to a network of biomolecular interactions representing genes included in the microarray dataset. A unique network has been constructed by the incorporation of interactions reported in protein–protein and protein–DNA databases (BIND, DIP, BioGRID and HPRD) as well as in a metabolic and signalling pathway database (KEGG) (http://bind.ca, http://dip.doe-mbi.ucla.edu, http://www.biogrid.org, http://www.hprd.org, http://www.genome.jp/kegg; Avila-Campillo et al., 2007). In the Supplementary Material, we supply R code for functional module discovery that allows using the outputs of two strategies of integrating information from these databases: the BioXM knowledge management tool (http://www.biomax.com) and the Cytoscape plug in BioNetBuilder (Avila-Campillo et al., 2007). The latter one was used for the analysis described here. Genes and their corresponding proteins were mapped together on the same node in the interaction network, which contained 6793 nodes connected by 29,209 edges.

2.3 FuMO: a procedure for the identification of functional modules

The methodology we have developed, namely functional module optimizer (FuMO), aims to identify ‘active’ sub-networks in a large graph of experimentally verified biologic interactions. An ‘active’ sub-network includes genes, whose mutual interactions are strongly supported both at the transcriptional and molecular interaction or pathway levels in a given experimental system. The procedure is structured into three separate steps: (i) the definition of a scoring function, (ii) a module search procedure to navigate through the interaction network and (iii) a final refinement stage, which defines precise module borders. A comprehensive and detailed description of the methodology introduced here can be found in the Supplementary Material.

2.3.1 Scoring sub-networks and standardization for random effects

In order to rate the strength of a particular sub-network, we first compute the degree of intra-module gene–gene co-expression. This is achieved using ARACNe, a well-validated network inference algorithm (Basso et al., 2005). For a given sub-network A in the gene interaction graph, we apply ARACNe on an mRNA expression-profiling dataset to compute the average MI I between the mRNA expression profiles of each pair of genes

$$score_A = \frac{1}{|O_1\cup O_2\cup O_3|} \sum_{O_1 \cup O_2 \cup O_3} I(G_i, G_j)$$

where the genes $G_i$ and $G_j$ correspond to the genes or their products $O_1$ and $O_2$ contained in the sub-network A and n is the number of gene pairs considered. This basic module score is further corrected for the likelihood of being observed by chance. On this account, 1000 gene subsets of size K (drawn from the same mRNA expression data but independently from the network structure) were iteratively generated, and their mean score $\mu_k$ and SD $\sigma_k$ derived for each possible module size k. For a given sub-network A of size K = k, the z-score is then defined as:

$$z_{score_A} = \frac{score_A - \mu_k}{\sigma_k}$$

Given the basic score, we expect the z-score to be monotonically increasing with increasing module size k. However, depending on the data at hand, several peaks may be observed. We therefore applied an additional smoothing step (using the pspline package in R) on $\mu_k$ and $\sigma_k$ until no peaks were observed for modules with a basic score of 1 (as it is the most extreme) and size 1 ≤ k ≤ 50.

2.3.2 Searching for high-scoring sub-networks

In order to identify a representative collection of network modules, it is not sufficient to be able to score highly significant networks. Since evaluating every possible module is computationally intractable, an efficient procedure to search for high-scoring networks is required. Simulated annealing (van Laarhoven and Aarts, 1987) is an excellent method for both discrete and continuous optimization problems. The search procedure is started from a seed module consisting of a randomly chosen gene. Modules are constructed in a stepwise fashion. During each step, a node adjacent to at least one working module is added or a node contained in a working module is removed. Hence, there are four possible scenarios: a module may be enlarged, several modules may be merged, a module may be reduced or split into smaller modules. The resulting new module(s) is (are) accepted with the probability

$$p(\Delta score) = \frac{1}{e^{\Delta score / s}}$$

where $\Delta score$ is the difference between the (maximum) z-score of the new and the working solution(s), and $s$ is a control parameter that decreases with time, i.e. the number of steps $s$ performed in the search procedure.

$$T(s) = e^{-0.1s}$$

If no further improvement is obtained for a given number of steps, the procedure stops. The working modules are saved, the seed module is...
reinitialized, and the search procedure is restarted. This process is repeated until a defined number of steps is reached.

2.3.3 Module refinement The process of 'simulated annealing' may not explore every possible local modification in the module borders. For this reason, a deterministic local optimization procedure ('quenching') has been subsequently applied to refine the selected modules. Iteratively, all possible local changes (removal/addition of single nodes from/to the modules) are tested. The best modification is accepted if and only if the appropriate z-score is improved. If, for any module, no local modification can increase its z-score any further, the algorithm stops.

2.3.4 Final module selection Despite the fact that the procedure uses a z-score as an objective function it is still possible that a number of low scoring modules are identified by random chance. It may be desirable to eliminate these from the list of selected modules to maximize the chance that the modules have biological relevance. For this reason the search procedure has been applied to a scrambled dataset. The procedure, originally proposed by Ideker et al. (2002), works by randomly assigning MI values (as calculated from the original mRNA expression values), to pairs of proteins in the physical interaction network. The scores of the modules identified by our procedure on the scrambled dataset have been used to define a threshold to select highly significant modules identified from the original dataset. In this case the threshold has been defined as the highest scoring module obtained from the scrambled search (>7). As a result, one of the modules identified by the search procedure in the ALL dataset has been removed (Fig. S8 in the Supplementary Material).

2.3.5 Method validation on a simulated search space To validate our method, as well as to test diverse parameter settings, a simulated search space was created. Based on a second mRNA expression dataset containing measurements of 2435 probesets in 233 tumours [Yeoh et al., 2002; data pre-processed as described in Tavazoie and Falciani (2006)], an integrated network comprising 1127 nodes and 2875 edges was constructed. Within this network, six modules (i.e. sub-networks, three at five and three at 20 nodes) have been randomly selected. The degree of co-expression between different genes is simulated such that genes are only associated by high MI values if they (or their products) are part of the same a priori defined module. Modules of size 5 and 20 nodes have been simulated to represent gene–gene interactions each of them with a relatively low, high and very high pairwise MI values. Subsequently, the search algorithm was tested using different parameter settings, which allowed to accurately identify the six a priori defined modules. As a consequence of the validation study, search parameters found to provide good results were chosen for the functional module search procedure on real data as described in the previous paragraphs. A detailed description of the validation study can be found in the Supplementary Material.

2.4 Module representation and functional annotation

To represent the relationship between different modules, we have used multidimensional scaling (MDS) on a matrix representing the degree of overlap between modules (expressed as the relative number of genes included in the smaller module but excluded in the largest one). Modules have been functionally annotated using ‘biological process’ Gene Ontology (GO) terms using the functional clustering tool available in the software application DAVID (Huang et al., 2007). This methodology tests for enrichment of functional terms (i.e. GO terms) in a given gene list and uses an agglomeration algorithm to condense a list of associated terms into organized classes (functional clusters) with related biology. The more significant GO term (if associated to an FDR <1%) of the highest scoring cluster of coherent GO terms has been selected to describe a given module.

2.5 Defining module activity

The overall activity of a module has been defined by its first two principal components (PCs) computed using the mRNA expression profiles (in the 37 modules identified by our procedure these explain on average 76% of the variance). To assess the predictive power of individual modules, we have used these two PCs as inputs of a linear discriminant analysis (LDA) classifier. The classifier accuracy has been computed using leave one out cross-validation (LOOCV), a procedure, which is suitable for estimating classification accuracy with small datasets. In order to establish the probability of a given module to predict IR response with a given accuracy by random chance, we have computed an empirical P-value based on the distribution of accuracies of 1000 datasets whose sample labels were permuted. In order to develop a statistical model representative of the interaction between the different modules in response to DNA damage, we have used the first two components of all the modules (a total of 74 variables) as inputs of a multivariate variable selection procedure coupled with a LDA classification algorithm as described in Trevino and Falciani (2006).

2.6 Testing for differential gene expression

In order to help interpret the modules M_18 and M_20 genes have been tested for differential expression using a paired t-test. Genes with a P < 0.05 have been annotated in Figure 3.

3 RESULTS

3.1 Overview of the analysis strategy

Previous studies using an mRNA expression-profiling approach to characterize the response of ALL cells to DNA damage have relied on the identification of genes, which are differentially expressed between control and cells exposed to DNA damaging agents (Holleman et al., 2004). In this study, we applied a modularization procedure, which identified multigene complexes whose components are transcriptionally coupled in ALL cells. Subsequently, we tested whether the overall activity of these modules is predictive of cellular response to DNA damage. The modularization strategy we have developed is described in a schematic form in Figure 1. The first step in the analysis is the construction of two networks representing different types of gene interactions. A fully connected network is constructed from the ALL cells expression data using ARACNe, a reverse engineering algorithm that measures the link between genes using MI of every pair of mRNA expression profiles (Fig. 1A). The second network is constructed by integrating molecular interaction and pathway databases (Fig. 1B). It represents the whole spectrum of

Fig. 1. Schematic representation of the analysis strategy.
known possible relationships between genes or their corresponding products, respectively. A sub-network in this molecular interaction network (Fig. 1D) is defined as a functional module if its components have a high degree of co-expression as represented in the full network constructed from the mRNA expression data. It is computationally impractical to explore every possible combination of modules, therefore a search strategy based on simulated annealing is used to navigate through the large interaction network (the need of searching for high scoring networks is represented in Fig. 1B with a curved trajectory across the network structure). The procedure identifies functional modules representative of functions, which are active in ALL cells. Ultimately, we would like to identify modules whose activity is associated to IR exposure. In order to achieve this, we summarize the activity of each module (Fig. 1E) by its first two PCs (Fig. 1F) and use these as input variables of a supervised classification algorithm. In our analysis, we have also tested the predictive power of the combination of modules using a multivariate variable selection strategy based on genetic algorithms (Trevino and Falciani, 2006). The use of dimensionality reduction techniques is an established procedure to summarize the transcriptional activity of a functional module (Ortega et al. (2008)). In the development of the multi-module classifier this procedure effectively allows the selection of entire pathway components in the model and reduces the search space in which the variable selection procedure operates. Compared to a gene-based model this procedure maximizes the chance of finding an optimal solution and improves biological interpretability.

3.2 Construction of a ‘functional Module Map’ of the interaction network
Our modularization procedure identifies 37 network modules of variable size (49 genes on average). For the sole purpose of visualizing the high-level relationship of these modules within the larger interaction network, we have used MDS, a projection technique designed to summarize the degree of relationship between highly multidimensional objects in a two-dimensional space. Modules with a large overlap in gene content would be located in close proximity, whereas modules sharing a very small percentage of genes would be at the opposite end of the two-dimensional map. Modules have been functionally annotated using the functional clustering tool available in the software application DAVID (Huang et al., 2007). Among the represented modules, there are different cell-cycle associated functions (i.e. DNA replication initiation, mitosis, mitotic cell-cycle and regulation of cell cycle), immune response (i.e. response to external stimulus and immune system process), protein metabolism and mRNA processing.

3.3 The activity of individual modules representing essential cellular functions is predictive of exposure to IR
The modular map we have described in the previous paragraph represents the activity of a broad range of cellular processes, which is indicative of the physiology of a tumour cell. The experimental data we have used contain samples of cells exposed to IR. Consequently, it may be reasonable to expect the activity of a subset of modules to be predictive of response to IR exposure. In order to address this question, we have computed a measure of overall module activity (based on the first two PCs of the gene expression matrix associated with the module). Modules have been functionally annotated using the functional clustering tool available in the software application DAVID (Huang et al., 2007). Among the represented modules, there are different cell-cycle associated functions (i.e. DNA replication initiation, mitosis, mitotic cell-cycle and regulation of cell cycle), immune response (i.e. response to external stimulus and immune system process), protein metabolism and mRNA processing.

The table lists modules, which have a significant ability to discriminate between non-irradiated and irradiated cells. The first column represents the module identifier (same as in Fig. 2). The second column describes the number of genes in the module and the third the percentage of variance explained by the first two components of the module and the fourth the representative GO term. The columns %GO and %Acc represent the percentage of genes representing the most significant GO term and the accuracy of LDA as a percentage.

![Functional modules in ALL DNA damage response](image)

**Table 1. Functional modules predictive of IR exposure**

<table>
<thead>
<tr>
<th>Module</th>
<th>Size</th>
<th>Var (%)</th>
<th>Significant GO term</th>
<th>%GO (%)</th>
<th>%Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M18</td>
<td>31</td>
<td>73</td>
<td>mRNA metabolic process</td>
<td>26</td>
<td>82</td>
</tr>
<tr>
<td>M22</td>
<td>47</td>
<td>70</td>
<td>DNA replication Initiation</td>
<td>13</td>
<td>80</td>
</tr>
<tr>
<td>M7</td>
<td>22</td>
<td>85</td>
<td>Mitotic cell cycle</td>
<td>45</td>
<td>73</td>
</tr>
<tr>
<td>M20</td>
<td>22</td>
<td>78</td>
<td>mRNA processing</td>
<td>27</td>
<td>73</td>
</tr>
<tr>
<td>M24</td>
<td>31</td>
<td>77</td>
<td>Mitosis</td>
<td>26</td>
<td>73</td>
</tr>
<tr>
<td>M30</td>
<td>31</td>
<td>80</td>
<td>Immune system process</td>
<td>32</td>
<td>70</td>
</tr>
<tr>
<td>M45</td>
<td>50</td>
<td>65</td>
<td>Response to external stimulus</td>
<td>18</td>
<td>70</td>
</tr>
<tr>
<td>M6</td>
<td>50</td>
<td>75</td>
<td>Regulation of cellular process</td>
<td>54</td>
<td>68</td>
</tr>
<tr>
<td>M9</td>
<td>50</td>
<td>78</td>
<td>Regulation of cellular process</td>
<td>54</td>
<td>68</td>
</tr>
<tr>
<td>M19</td>
<td>71</td>
<td>77</td>
<td>Phosphorus metabolic process</td>
<td>24</td>
<td>66</td>
</tr>
</tbody>
</table>

The table lists modules, which have a significant ability to discriminate between non-irradiated and irradiated cells. The first column represents the module identifier (same as in Fig. 2). The second column describes the number of genes in the module and the third the percentage of variance explained by the first two components of the module and the fourth the representative GO term. The columns %GO and %Acc represent the percentage of genes representing the most significant GO term and the accuracy of LDA as a percentage.

![A functional module map representative of ALL cells.](image)

**Fig. 2.** A functional module map representative of ALL cells. The figure shows the relationship between functional modules (expressed as a MDS plot projection of a matrix representing the degree of gene overlap) identified by our procedure in the ALL dataset. The diameter of the circles is proportional to the number of genes represented in a module. Closed circles represent modules which are also highly predictive of IR exposure. Three functional modules do not overlap with any other module and hence, are listed in a separate box.

to a given module) and tested a module’s ability to predict response to IR using an LDA classifier. In order to select modules, which have a significant predictive power, we have compared the classification accuracy of a given module with a distribution of accuracies obtained by permuting the sample classes 1000 times. Using this method, we have computed an empirical $P$-value associated to a given Module. Table 1 lists 10 modules with significant predictive power ($P < 0.05$) ranked by cross-validated classification accuracy. These modules cluster in one distinct area of the module map we have previously described (Fig. 2). As expected, a number of
modules representing cell-cycle-related functions ($M_{22}, M_{24}, M_7$) are present in the list of significant modules. However, it is of interest that two closely associated predictive modules ($M_{20}$ and $M_{18}$) represent the combination of important cellular functions, which were not previously associated to IR response. These functions represent two important receptors in the immune system [B-cell receptor (BCR), CSF3R] and six associated signalling molecules, the mRNA processing machinery and components of the proteasome, a multiprotein complex involved in targeted protein degradation.

In order to better understand the implications of the coordinate regulation of these functional complexes, we have performed a t-test to identify the genes in modules $M_{20}$ and $M_{18}$, which are differentially regulated in response to IR. Figure 3 shows the structure of the overlapping between modules $M_{18}$ and $M_{20}$ and the direction of change of individual gene responses to IR ($P < 0.05$). The coordinate up-regulation of components of the proteasome and down-regulation of mRNA processing and immune receptor signalling appears to be the major feature of this composite network.

### 3.4 A multivariate model predictive of response to IR integrates the activity of individual modules

In the previous paragraph, we have shown that transcriptional activity of 10 of the 37 modules we have identified is predictive of exposure to IR in ALL cells. In order to test whether the combination of the activity of individual modules may provide a stronger predictor of IR exposure, we have applied a multivariate variable selection strategy to develop multimodule predictors. Our approach identifies a highly predictive (91% of cross-validated accuracy) representative model based on the combination of the second PCs of modules $M_{18}$, $M_{20}$ and $M_{13}$. This result is consistent with the fact that the second components of modules $M_{18}$, $M_{20}$ and $M_{13}$ appear to discriminate better between pre- and post-irradiation samples (Figs S10–S12, Supplementary Material). Figure S13 shows the degree of separation of the three samples in the second PCs space. Note that the accuracy of a classifier based on individual genes is comparable to the overall module-based classifier we have developed (data not shown).

### 4 DISCUSSION AND CONCLUSIONS

Despite considerable progress in cure rates of paediatric B-precursor ALL this disease remains a common cause of death in childhood. Conventional therapy is based on DNA damaging agents and it is effective in a large percentage of the patients. However, still a number of patients do not respond to therapy and understanding the transcriptional network of DNA damage response could potentially facilitate the discovery of novel therapeutic targets, which are capable of increasing sensitivity of leukemic cells to DNA damage. In this project, we have used exposure to IR as a clinically relevant experimental model to investigate the response to DNA damage. Depending on the level of DNA damage and the integrity of appropriate response pathways, tumour cells stop cell-cycle progression and undergo apoptosis. mRNA expression-profiling studies have been performed in ALL to characterize the response of a variety of cancer cells to chemicals inducing DNA damage (Holleman et al., 2004). These studies have confirmed that cells respond by modulating genes involved in cell division, DNA repair and apoptosis. Using a new network modularization approach we have identified active modules of a gene interaction network, which are predictive of cellular response to exposure to IR. Our analysis has shown that the transcriptional activity of modules representing immune system receptors (CD79A, CD19, CD22 and CSF3R),
mRNA processing and targeted protein degradation is predictive of IR response. Interestingly this process involves the up-regulation of the proteasome components and the coordinate down-regulation in the expression of the other two functions. In this context, it is reasonable to hypothesize the existence of a process which would be part of a ubiquitination-based program directed to silence the activity of an irradiated cell. This hypothesis is consistent with a recent report describing a mechanism for p53 regulation by targeted protein degradation (Marchenko and Moll, 2007). Interestingly, module 30 contains a multiprotein complex formed by a set of five components of the proteasome and p53 itself (Figs S17 and S18). Our method performs well with respect to the analogous modularization procedure developed by Ulitsky and Shamir (2007). When applied to our dataset, the two methods identified a largely different set of modules (Fig. S15). Modules identified by our procedure however, displayed a higher degree of inter-module co-expression (Fig. S16). The procedure we have developed has therefore the potential to be applied to other biological systems as it provides an efficient method for the definition of interaction modules with a higher degree of inter-module correlation. The recent development and application of several techniques designed to identify ‘interesting’ modules within a large network of gene interactions supports the general applicability of these methodologies to a variety of biological systems. With the development of integrated databases combining several public repositories it is therefore likely that module-based analysis will become a standard tool in Bioinformatics for mining large-scale datasets.

ACKNOWLEDGEMENTS

We would like to thank Prof. Mike Salmon (Birmingham), Dr Victor Trevino and Dr Chris Bunce (Birmingham) for useful comments on the article.

Funding: European Union Grant BIOBRIDGE; Leukaemia research Fund UK; BBSRC (grant S20214). The analysis was carried out by the Birmingham Biosciences high performance computer cluster funded by the BBSRC REI grant BB/D524624/1.

Conflict of Interest: none declared.

REFERENCES