BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 189–196
doi:10.1093/bioinformatics/btt625

Gene expression

Advance Access publication November 28, 2013

Finding consistent disease subnetworks using PFSNet

Kevin Lim* and Limsoon Wong

School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: Microarray data analysis is often applied to characterize disease populations by identifying individual genes linked to the disease. In recent years, efforts have shifted to focus on sets of genes known to perform related biological functions (i.e. in the same pathways). Evaluating gene sets reduces the need to correct for false positives in multiple hypothesis testing. However, pathways are often large, and genes in the same pathway that do not contribute to the disease can cause a method to miss the pathway. In addition, large pathways may not give much insight to the cause of the disease. Moreover, when such a method is applied independently to two datasets of the same disease phenotypes, the two resulting lists of significant pathways often have low agreement.

Results: We present a powerful method, PFSNet, that identifies smaller parts of pathways (which we call subnetworks), and show that significant subnetworks (and the genes therein) discovered by PFSNet are up to 51% (64%) more consistent across independent datasets of the same disease phenotypes, even for datasets based on different platforms, than previously published methods. We further show that those methods which initially declared some large pathways to be insignificant would declare subnetworks detected by PFSNet in those large pathways to be significant, if they were given those subnetworks as input instead of the entire large pathways.

Availability: http://compbio.ddns.comp.nus.edu.sg:8080/pfsnet/

Contact: kevinl@comp.nus.edu.sg

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on July 21, 2013; revised on September 24, 2013; accepted on October 28, 2013

1 INTRODUCTION

Discovering the mechanisms regulating a disease requires researchers to understand the cascade of events leading to a particular phenotype. These mechanisms are often depicted in biological pathways in the literature. While microarray technology provides researchers a glimpse of the underlying characteristics that differentiate phenotypes—e.g. normal samples versus diseased samples—it has been shown that methods for microarray analysis often produce results that are inconsistent and unreproducible when applied independently on independent datasets (Zhang *et al.*, 2009) and, in some cases, are even no better than randomly produced gene signatures (Venet *et al.*, 2011).

Clearly, a method that produces a higher agreement between two datasets of the same disease lends more confidence to the real cause of the disease. SNet (Soh *et al.*, 2011) is one of the rare exceptions in gene-expression-analysis methods that produce pathways and genes that are consistently significant when independently applied to multiple independent datasets of the same disease phenotypes, even when these datasets are not restricted to the same platform. The percentage of agreement is higher than other methods compared.

In this article, we identify some shortcomings of SNet, and present two refinements which we incorporate in the new method PFSNet. We have tested PFSNet on six datasets from three disease types. Our method produces an improvement of 25–64% and 31–51% in gene agreement and pathway agreement, respectively, when compared to SNet, and even more when compared to earlier methods like GSEA, GGEA, SAM and *t*-test.

2 BACKGROUND

In the past decade and a half, many microarray analysis approaches have been proposed to identify genes whose expression profile is useful for the diagnosis of a disease, the prognosis of a treatment or deciphering the cause of a disease. These genes are usually differentially expressed between the disease phenotypes and normal tissues. Using statistics to identify differentially expressed genes is one of the earliest and still popular methods. For example, the *t*-test identifies individual genes whose mean expression value is higher in one class of samples than the other. The problem with using any statistical test is that performing the same hypothesis test multiple times results in a large number of false positives. For example, if the cut-off threshold for statistical significance is 0.05, then performing the test on 30 000 genes results in 1500 expected false positives. This leads to a number of works that correct the *P*-values (Tusher *et al.*, 2001).

For several years now, there has been a paradigm shift from looking at individual genes to gene sets. Such methods avoid large multiple-hypothesis testing by preselecting gene sets using biological knowledge. These gene sets are often termed 'pathways' in the literature, and are groups of genes that perform a specific function. These methods can be classified into four categories, described in separate sections below.

2.1 Overlap analysis

Overlap analysis methods (Khatri and Draghici, 2005) test whether the proportion of genes over-expressed in one phenotype in a pathway is different from the proportion in the whole array. This is usually done using the hyper-geometric test with the null hypothesis that there is no difference between the proportions of differentially expressed genes in the pathway compared to a random gene set.

*To whom correspondence should be addressed.

The drawback of overlap-based methods is that differentially expressed genes may be caused by genes that are not themselves differentially expressed. For example, a gene may have a mutation that does not affect its expression level but affects the shape of its protein product and, thus, the ability of the protein product to interact with its down-stream partners. Such a causal gene has the ability to effect a change in many differentially expressed genes. The biological explanation usually implicates the causal gene, which by virtue of not being itself differentially expressed, is not detected by single-gene-based methods and often also overlap analysis methods. Moreover, the threshold applied to decide whether a gene is differentially expressed is arbitrarily chosen. This can have a large effect on the set of differentially expressed genes, hence causing a large impact on the hyper-geometric test (Fig. 1b).

2.2 Direct-group analysis

Direct-group methods like GSEA (Subramanian *et al.*, 2005) and FCS (Goeman *et al.*, 2004) use a suitable statistical test on the whole pathway instead of preselecting differentially expressed genes. These methods circumvent the problems of overlap methods mentioned previously and are usually less sensitive to arbitrary changes (due to different threshold values) in the number of genes which are differentially expressed.

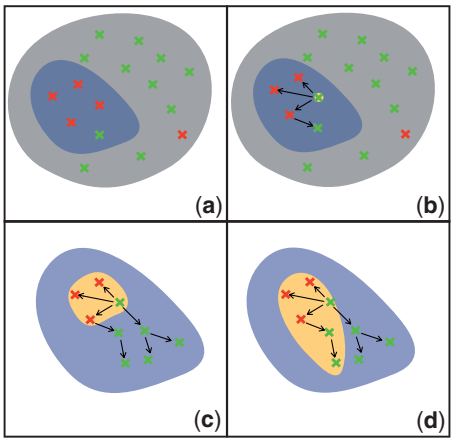

Fig. 1. Methods that rely on gene sets are categorized into three groups. (a) The differentially expressed genes (darker shade; red, see online), and genes with no expression difference (lighter shade; green, see online). Overlap analysis methods compute the statistical significance of whether differentially expressed genes are more likely to exist in a pathway or in a random set of genes. (b) A different threshold is used to select differentially expressed genes, this difference is large enough for overlap analysis methods to miss the pathway, whereas direct-group methods are less sensitive to such changes. (c) A large number of non-causal genes contaminates the score in direct-group methods, whereas network-based methods restrict themselves to smaller parts of the affected pathway (lighter shade; yellow, see online) and are still able to detect them. Some network-based methods, which look around a small neighborhood of regulatory elements, produce subnetworks that do not sufficiently explain the biological cause of the disease. (d) In contrast, the explanation usually involves a cascade of genes whose upstream genes exert an effect on downstream genes

GSEA (Subramanian *et al.*, 2005) computes a Kolmogorov–Smirnov-like statistic and FCS (Goeman *et al.*, 2004) computes a Mean-Log*P* statistic for each pathway to determine its statistical significance. The significant pathways hence have genes overrepresented in one phenotype but not the other.

The problem with approaches in this category is that when the pathway contains too many non-causal genes, the statistical score can be largely affected. It is likely that these methods identify pathways that contain sufficiently large proportion of disease-related genes but pathways that contain only a few disease-related genes can be missed out (Fig. 1c).

2.3 Network-based analysis

In order to address the problems that arise from direct-group analysis, network-based methods identify a subset of genes that might be most relevant to a phenotype in each pathway. This breaks up the pathway into smaller parts, which we call 'subnetworks' in this article. Methods in this category include NEA (Sivachenko *et al.*, 2007) and SNet (Soh *et al.*, 2011).

In NEA (Sivachenko *et al.*, 2007), the subnetworks are constructed based on the immediate neighborhood around each regulatory element. Each subnetwork is then tested for significance using, e.g. statistical methods mentioned in the direct-group analysis approaches. The problem is that the results usually only provides a partial explanation to the cause of disease. In contrast, diseases are usually explained by a cascade of genes whose upstream genes exert an effect on downstream ones (Fig. 1d).

In a recent publication, another network-based method, SNet (Soh *et al.*, 2011), has the capability to extract meaningful subnetworks, which are not necessarily restricted to a small neighborhood around some genes. The subnetworks in SNet are generated by setting a threshold on the gene expression levels. The authors first selected the top α% genes in one phenotype. From this set of genes, they used majority voting to select genes that occur in at least β% of the given phenotype (by default $\alpha = 10\%$ and $\beta = 5\%$). In each pathway, genes not in the selected list are removed, causing the pathway to fragment into smaller pieces. The pieces exceeding a certain size are considered candidate subnetworks for subsequent significance analysis (Fig. 2).

One practical consideration when using SNet is the choice of the thresholds. In their article, α and β thresholds are set at 10%

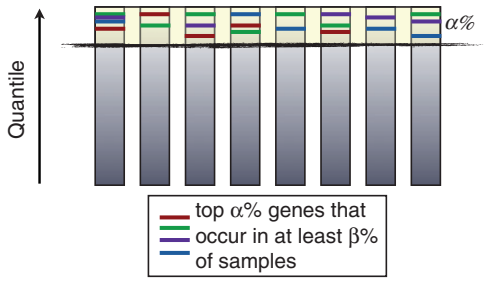

Fig. 2. In SNet, the top % of genes of each sample in phenotype D is highlighted (lighter shade; yellow, see online). A subset of these genes that are represented in color bands are in at least % of the samples are then taken to induce subnetworks

and 50%, respectively. The shortcoming of SNet is that genes that lie only slightly below the top α% can get missed out because they occur in less than β% of the samples. One might attempt to increase α% threshold but many false positives would be included. In order to address this shortcoming, we propose fuzzification on the gene-expression values, adapted from Geistlinger *et al*. (2011), where a weight between 0 and 1 is assigned so that genes below the thresholds can be also considered but given lesser importance.

2.4 Model-based analysis

Model-based methods are a category of gene-set-based methods that attempt to learn parameters for a dynamic model of any given pathway using one phenotype, say, normal tissues. For example, pathways are modeled using linear regression in SRI (Zampieri *et al*., 2011) and Petri nets in GGEA (Geistlinger *et al*., 2011). The constructed model is simulated on both phenotypes (normal and diseased tissues). Differentially expressed pathways are identified if the model is consistent in one phenotype but not the other.

These methods provide a fine-grain model of the pathways because the parameters learnt during model construction ensure that the expression values obey certain functions specified by the model. This enables researchers to do more sophisticated analysis like predicting the expression of certain genes under some predefined conditions. However, the drawback of these methods is that the parameters are hard to estimate when creating such a fine-grain model of pathways. Thus, while large databases of pathways exist, there are far fewer fine-grain pathway models available for use.

3 METHODS

3.1 Subnetwork generation

We conjecture that one of the reasons SNet is consistent across datasets is due to the fact that absolute gene-expression values were not used. Instead, by considering the top 10% of the highly expressed genes, the gene ranking is preserved across two independent datasets.

In our approach, we first assign a fuzzy value, $fs(e_{g_i,p_j})$, to each gene, g_i, based on the ranking of its expression value e_{g_i,p_j} within the sample p_j. We use two thresholds θ_1 and θ_2 to separate the upper and lower quantiles of the genes within a sample so that any expression value that falls between these quantiles is given a weight between 0 and 1 (Fig. 3).

The fuzzy value is assigned to each gene for every sample. This allows us to perform majority voting to select genes to include in the subnetworks. Each sample p_j gives a partial vote $fs(e_{g_i,p_j})$ (of value between 0 and 1) for each gene g_i. In contrast, in SNet, every sample p_j either gives the gene g_i a total vote (of value 1) if g_i's expression is ranked in the top 10% in p_j, or gives g_i no vote if its expression is not in the top 10%.

Our goal in this step is to compute a gene list, which segregates the pathways into smaller components. The voting criteria that determine whether the gene g_i is accepted into this gene list are given below:

$$\sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|} > \beta \tag{1}$$

where D is the phenotype for which the subnetwork is generated, p_j ranges over the patients of phenotype D and fs is the fuzzy function which converts the gene expression value e_{g_i,p_j} to a value between 0 and 1.

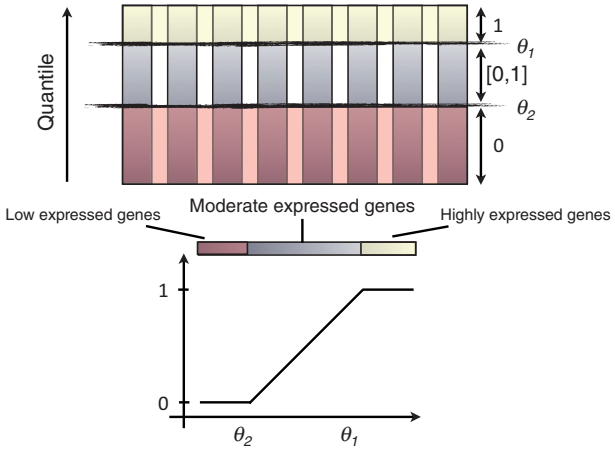

Fig. 3. Our fuzzification function maps expression values to numbers between 0 and 1 according to their ranks within a sample

Once the gene list is computed, subnetworks in each reference pathway are generated by taking connected components induced by the genes in this list. We ignore subnetworks that are less than size 5.

3.2 Subnetwork scoring

The goal in this step is to generate a subnetwork score for each sample so that the population of scores for phenotype D and $\neg D$ can be differentiated by using a suitable statistical test, e.g. t-test.

The first method introduced in this paper is FSNet (Fuzzy SNet). In FSNet, the acceptance criteria in Equation (1) are in fact an average of partial votes for a particular gene across the phenotype D. We let each gene in the subnetwork be scored by this average partial vote.

Let $\beta^*(g_i)$ denote the average partial vote described in Equation (1), i.e.

$$\beta^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|}. \tag{2}$$

Then, the score computed for sample k, for a particular subnetwork S is:

$$\text{Score}^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta^*(g_i). \tag{3}$$

In each sample p_k, the score of a subnetwork S is basically the sum of weighted average partial votes of genes in S that are highly expressed in p_k. The average partial votes can be thought of as a measure of agreement between the samples of phenotype D. In SNet, when total votes are used, the average total vote is the percentage of samples having gene g_i in the top 10%.

The second method introduced in this paper is PFSNet (Paired Fuzzy SNet). In PFSNet, instead of computing the gene scores with respect to phenotype D, we also compute the gene scores with respect to phenotype $\neg D$. Hence, each node is given scores which we denote as $\beta_1^*(g_i)$ and $\beta_2^*(g_i)$, computed as follows:

$$\beta_1^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|}, \beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{fs(e_{g_i,p_j})}{|\neg D|}. \tag{4}$$

Accordingly, for every subnetwork S, each patient of phenotype D can be scored under β_1^* and β_2^*, as follows:

$$\text{Score}_1^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_1^*(g_i), \tag{5}$$

$$\text{Score}_2^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_2^*(g_i). \tag{6}$$

3.3 Statistical test

In FSNet, we can simply compute subnetwork scores in Equation (3) for the two phenotype populations. This gives us two distributions which we want to discriminate from each other. A t-test is done under the null hypothesis that the mean score for the phenotype D and $\neg D$ populations are not different. The alternative hypothesis is that the mean score for phenotype D is greater than the mean score for phenotype $\neg D$. The significant subnetworks identified at this step are enriched in phenotype D. To discover subnetworks enriched in phenotype $\neg D$, we compute the node scores in Equation (2) using the $\neg D$ samples, and repeat the same process of statistical testing.

In PFSNet, for a subnetwork S that behaves differently between the two phenotypes, we expect the two scores $\text{Score}_1^{p_k}(S)$ and $\text{Score}_2^{p_k}(S)$ in Equations (5) and (6) to differ from each other for any sample p_k. Since these scores arise from the same sample p_k and subnetwork S, we can do a paired t-test, under the null hypothesis that the difference in scores gives us a distribution with mean = 0. To find subnetworks that are enriched in the other phenotype, we compute the node scores in Equation (4) using the $\neg D$ samples.

3.4 Permutation test

As the null distribution may not really be represented by the theoretical distribution (Gatti *et al*., 2010; Goeman and Buehlmann, 2007; Venet *et al*., 2011), we generate the null distribution by a permutation procedure. We randomly swap the class labels (1000 times) for each dataset—i.e. randomly assigning a sample to belong to either phenotype D or $\neg D$ while maintaining the original proportion of D and $\neg D$ samples—and obtain a distribution of subnetwork scores. From this null distribution, we estimate at 5% significance level on one-tail of the distribution, whether a subnetwork which we compute for our original dataset is statistically significant.

4 RESULTS

We use pathways from PathwayAPI (Soh *et al*., 2010), a database that unifies popular pathway databases like KEGG (Kanehisa *et al*., 2012; Kanehisa and Goto, 2000), Wikipathways (Kelder *et al*., 2012) and Ingenuity (Ingenuity, 1998).

For each of the three disease types studied here—Leukemia (Armstrong *et al*., 2002; Golub *et al*., 1999), Childhood Acute Lymphoblastic Leukemia (ALL subtype) (Ross *et al*., 2004; Yeoh *et al*., 2002) and Duchenne Muscular Dystrophy (DMD) (Haslett *et al*., 2002; Pescatori *et al*., 2007)—we obtain two independent datasets which are produced using different microarray platforms. For each disease type, we run PFSNet, FSNet, SNet, GSEA, GGEA, SAM and t-test on the two datasets independently and obtain two corresponding outputs.

We compare the results from the two datasets using two measures of Jaccard-like agreement, defined below.

We use the subnetwork generation procedure mentioned in Section 3.1 to generate the subnetworks in dataset 1. We then test these subnetworks for statistical significance using the procedure mentioned in Sections 3.2–3.4 on datasets 1 and 2 independently. Let the significant subnetworks identified by datasets 1 and 2 be SN_1 and SN_2, respectively. Then the subnetwork-level agreement is defined as

$$\frac{\text{SN}_1 \cap \text{SN}_2}{\text{SN}_1 \cup \text{SN}_2}. \tag{7}$$

When testing GSEA, which identifies pathways instead of subnetworks, we measure the pathway-level agreement which is defined analogously.

We also measure the agreement between the genes in the output generated by the two independent datasets. Let the genes in SN_1 and SN_2 be G_{SN_1} and G_{SN_2}, respectively, then the gene-level agreement is defined as

$$\frac{G_{\text{SN}_1} \cap G_{\text{SN}_2}}{G_{\text{SN}_1} \cup G_{\text{SN}_2}}. \tag{8}$$

4.1 Comparing PFSNet, FSNet and SNet

FSNet is flexible enough to be able to emulate SNet by setting $\theta_1 = \theta_2 = 10\%$. In this way, genes above the 90th percentile are given a total vote and genes below the 90th percentile are given no vote at all. This is equivalent to setting SNet with $\alpha = 10\%$.

When comparing PFSNet, FSNet and SNet, we fix $\theta_1 = 5\%$ and vary θ_2 between 5% and 50% for PFSNet and FSNet. We also vary α between 5% and 50% for SNet. This allows more genes to be considered in the subnetworks in all the methods. To emulate majority voting, β is set at 50% (Fig. 4).

Our experiments show that when θ_1 (in FSNet and PFSNet) or α (in SNet) is low, the subnetworks may not be a true representation of the disease simply because too few genes are chosen to induce the subnetworks. But when too many genes are considered, there will be false positives showing up in the subnetworks. For example, when the value for θ_2 is set at the extreme ends (5% and 50%), the subnetworks have very low agreement between datasets in all three methods. In the Leukemia dataset, FSNet achieves the maximum subnetwork agreement of 100% ($\theta_2 = 20\%$) whereas SNet achieves the maximum subnetwork agreement of 77% ($\alpha = 15\%$). In the DMD dataset, FSNet achieves maximum subnetwork agreement of 90% ($\theta_2 = 10\%$) whereas SNet achieves maximum subnetwork agreement of 73% ($\alpha = 10\%$). In the ALL subtype dataset, FSNet achieves maximum subnetwork agreement of 38% whereas SNet only achieves 26%.

The results also show that giving genes that are not in the top 5% a partial vote is better than giving them a total vote. As we allow more and more genes to be considered, FSNet generally gives better subnetwork agreement than SNet. FSNet is thus more robust towards noise when incorporating more genes. For example, when $\theta_2 = \alpha = 50\%$, FSNet is able to get 69% subnetwork agreement but SNet only manages 40% in the Leukemia dataset. Similarly in DMD, FSNet achieves 59% whereas SNet achieves 29%.

In PFSNet, we get even higher subnetwork-level agreement than both FSNet and SNet. This shows the node scores obtained from the opposite phenotype play an important role in contributing towards consistent subnetworks. In particular, while both FSNet and SNet do not have very good subnetwork-level agreement in the ALL subtype dataset (38% and 25%, respectively), PFSNet is able to achieve 74%.

We also measure the gene-level agreement to check whether the significant subnetworks contain similar genes in the two datasets. We see a similar trend that PFSNet performs better than FSNet which in turn performs better than SNet. In particular, for the ALL subtype dataset which has the worst pathway-level

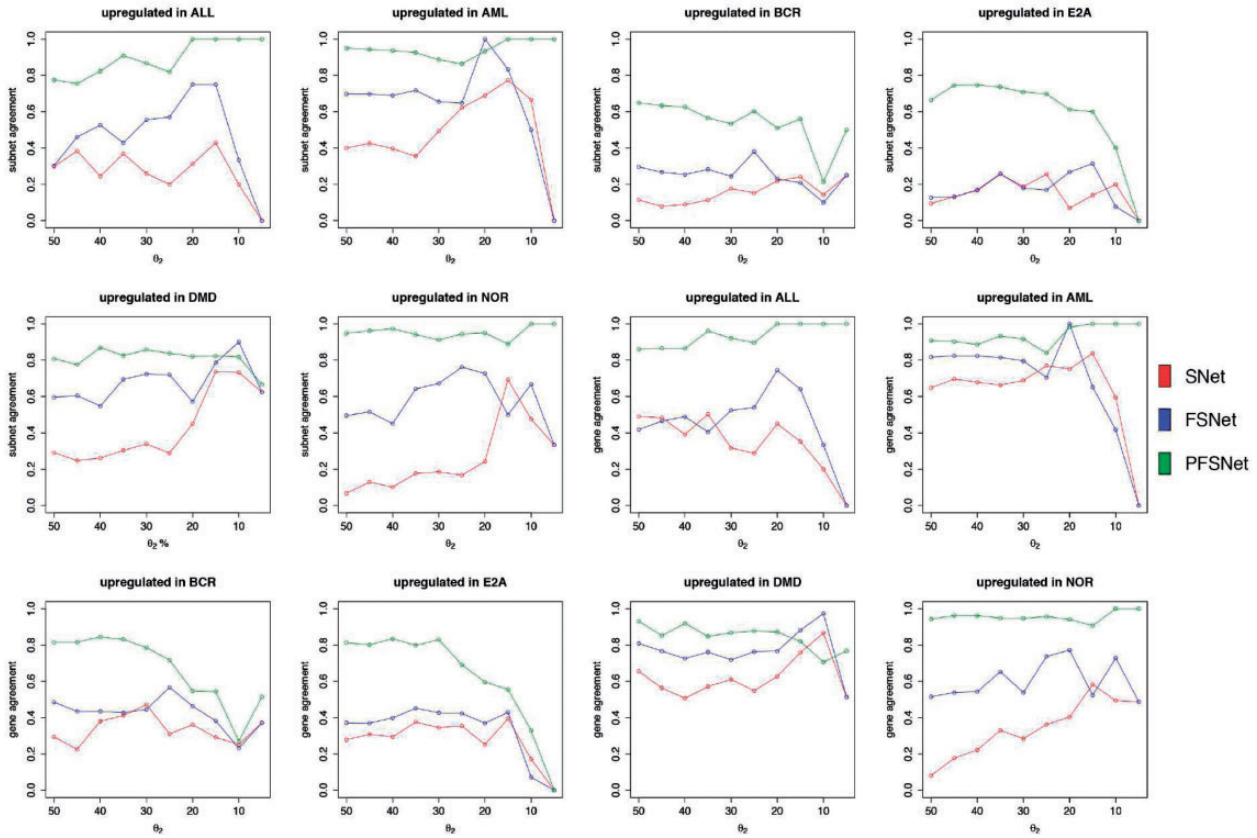

Fig. 4. Consistency of subnetworks and their genes in Leukemia (ALL/AML), ALL Subtype (*BCR*/*E2A*) and DMD dataset (DMD/NOR)

agreement reported above, the maximum gene-level agreement for PFSNet, FSNet and SNet are 84%, 57% and 47%, respectively.

4.2 Comparing with GSEA, GGEA, SAM and *t*-test

We compare our methods with GSEA, GGEA, SAM and *t*-test. We run GSEA and GGEA on both datasets and measure the level of pathway agreement between the two datasets. In general, we achieve higher pathway-level agreement than GSEA and GGEA. PFSNet has a pathway-level agreement between 56–100%, FSNet has a pathway-level agreement between 38–75%, GSEA has a pathway-level agreement between 12–57% and GGEA has a pathway-level agreement between 18–51% (Table 1).

We also measure the gene-level agreement from significant subnetworks between each pair of datasets. In order to compare this with GSEA, we computed the gene-level agreement from the 'leading edge' gene sets in each pair of datasets. The 'leading edge' genes are those genes that appear in GSEA's ranked list at the point at which the Kolmogorov–Smirnov running sum reaches its maximum deviation from zero (Subramanian *et al.*, 2005). We also compare gene-level agreement with SAM and *t*-test which identifies individual differentially expressed genes. PFSNet has a gene-level agreement between 54–100%, FSNet has a gene-level agreement between 38–88%, SNet has a gene-level agreement between 29–76%. In contrast, GSEA, SAM and *t*-test have much worse agreement at the 5% significance level. GSEA has a gene-level agreement between 4–44%, SAM has a gene-level agreement between 8–50% and *t*-test has a gene-level agreement between 8–41% (Table 2).

Table 1. Comparing pathway-level agreement of PFSNet, FSNet, GGEA and GSEA

Dataset	PFSNet	FSNet	GSEA	GGEA
Leukemia	1.00	0.75	0.12	0.18
ALL (subtype)	0.56	0.38	0.34	0.37
DMD	0.82	0.79	0.57	0.51

For PFSNet and FSNet, threshold values of $\theta_1 = 5\%$, $\theta_2 = 15\%$ are used.

4.3 Comparing pathways and subnetworks

As pathways are often large, many analyses involving the whole pathways do not give consistent results. For example, when we tested GSEA/GGEA in the previous subsection using pathways, the level of agreement was generally low.

One of the contributions in SNet, FSNet and PFSNet is the ability to break large pathways into smaller subnetworks. We select significant subnetworks from SNet, FSNet and PFSNet,

Table 2. Comparing gene-level agreement of PFSNet, FSNet, SNet, GSEA, SAM and *t*-test

Dataset	PFSNet		FSNet		SNet		GSEA		SAM(5% sig)		SAM(top 100)		*t*-test(5% sig)		*t*-test(top 100)	
	D	¬*D*	*D*	¬*D*	*D*	¬*D*	*D*	¬*D*	*D*	¬*D*	*D*	¬*D*	*D*	¬*D*	*D*	¬*D*
Leukemia	1.00	0.81	0.64	0.42	0.35	0.58	0.12	0.20	0.50	0.47	0.01	0.01	0.35	0.29	0.19	0.07
ALL (subtype)	0.54	0.70	0.38	0.41	0.29	0.57	0.04	0.04	0.19	0.27	0.12	0.21	0.08	0.10	0.01	0.00
DMD	0.82	0.72	0.88	0.75	0.76	0.54	0.44	0.20	0.34	0.08	0.27	0.19	0.41	0.19	0.11	0.25

Note: For PFSNet and FSNet, threshold values of $\theta_1 = 5\%$, $\theta_2 = 15\%$ are used. For GSEA, the 'leading edge genes' were used. For SAM and *t*-test, we took genes at 5% significance level and also the top *n* genes indicated in brackets. *D*, subnetworks enriched in phenotype D; ¬*D*, subnetworks enriched in phenotype ¬D.

and test them using GSEA. We discover that many of these subnetworks are also considered significant by GSEA/GGEA, even though GSEA/GGEA had earlier considered the original whole pathways from which these subnetworks were derived to be insignificant.

We next test whether these subnetworks are consistently declared significant in two independent datasets by GSEA/GGEA (Table 3). Subnetworks taken from PFSNet give the highest agreement of ∼100%, subnetworks taken from FSNet give the highest agreement of ∼71% and the subnetworks taken from SNet give the highest agreement of ∼50%. In contrast, using large pathways, GSEA and GGEA have an agreement of ∼57% and 51%, respectively.

Table 3. Testing subnetworks from PFSNet, FSNet and SNet using GSEA and GGEA

	PFSNet	FSNet	SNet
Leukemia (GSEA)	0.50	0.00	0.00
Leukemia (GGEA)	0.67	0.50	0.50
ALL subtype (GSEA)	1.00	0.15	0.11
ALL subtype (GGEA)	1.00	0.47	0.35
DMD (GSEA)	0.90	0.57	0.50
DMD (GGEA)	0.54	0.71	0.45

4.4 Biologically significant subnetworks

We also check the subnetworks consistently detected by PFSNet for biological relevance. We discover that many subnetworks and their genes are involved in relevant disease-related processes known in the literature. Some of these subnetworks predicted as significant in PFSNet are not discovered by SNet. We report these subnetworks ranked according to the *P*-value computed by PFSNet in Table 4 (also see Supplementary Material). We will describe three example subnetworks for the respective diseases to demonstrate their relevance to the diseases.

For DMD, the subnetwork responsible for striated muscle contraction is shown in Supplementary Figure S5a. The cause of Duchenne muscular dystrophy is well known to stem from the gene Dystrophin, which codes for a protein attached to the cell membrane (sacrolemma) of striated muscle cells (Goldstein and McNally, 2010). When its expression is perturbed, the cell membrane becomes fragile and permits an amplification in calcium signals into the muscle cell causing a cascade of signals to induce cell death. Our subnetwork is generated around the Dystrophin gene and implicates other genes belonging to the Myosin (*MYBPC1*, *MYBPC2*) and Troponin (*TNNI1*, TNNI2) family. The Myosin and Troponin genes are responsible for controlling muscle contractions. The down-regulation of Troponin in DMD patients might help explain muscle contracture, a condition in which the muscle shortens. This is because with lower abundance of Troponin, Myosin is able to bind to Actin. This mechanism together with the amplification of calcium causes the muscle to constantly contract, shortening over time (Goldstein and McNally, 2010; Krans, 2010).

For the Leukemia dataset (in which patients are either classified to have acute lymphoblastic leukemia or acute myeloid leukemia), one of the significant subnetworks that is biologically relevant is part of the Interleukin-4 signaling pathway (Supplementary Fig. S6b). The binding of Interleukin-4 to its receptor (Cardoso *et al*., 2008) causes a cascade of protein activation involving JAK1 and STAT6 phosphorylation. STAT6 dimerizes upon activation and is transported to the nucleus and interacts with the RELA/NFKB1 transcription factors, known to promote the proliferation of T-cells (Rayet and Gelinas, 1999). In contrast, acute myeloid leukemia does not have genes in this subnetwork up-regulated and are known to be unrelated to lymphocytes.

For the ALL subtype dataset, the patients are categorized to either having the *BCR-ABL* oncogene or *E2A-PBX1* oncogene. Antigen-processing pathway is one of the significant subnetworks. This suggests that lymphocytes elicit different response in the two ALL subtypes. The immunophenotypic characteristics of acute leukemias have been described in the literature (Hruak and Porwit-MacDonald, 2002; Giunta and Pucillo, 2012). ALL belonging to the *BCR-ABL* subtype express the cluster of differentiation (CD) markers CD9 and CD10, whereas those belonging to the *E2A-PBX1* subtype express the CD19 and CD45 markers.

5 CONCLUSION

Recent methods for analyzing microarray data that focus on identifying biological processes and pathways are superior to

Table 4. Top five subnetworks that have biological significance

Leukemia	ALL subtype	DMD
Proteasome degradation	Wnt signaling[a,b]	Striated muscle contraction[a,b]
IL-4 signaling[a,b]	Antigen processing	Integrin signaling
Antigen processing[a]	Jak-STAT signaling[a,b]	VEGF signaling[a]
B-cell receptor signaling[b]	T-cell receptor signaling	Tight junction
Wnt signaling[a,b]	Adherens junction[a,b]	Actin cytoskeleton signaling

[a]Subnetworks that were not found in SNet.
[b]Pathways that were missed by GSEA.

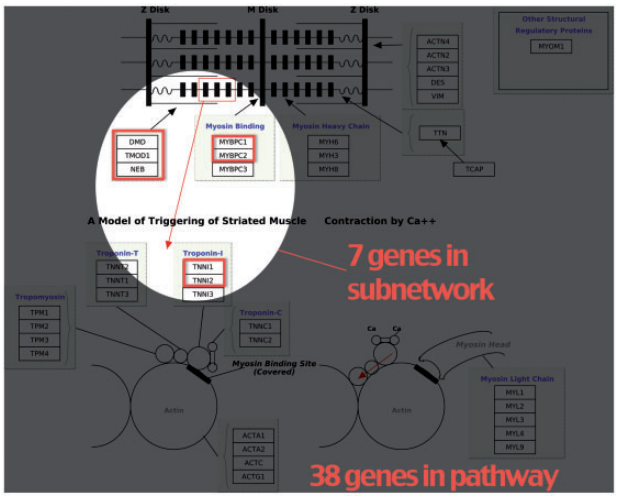

Fig. 5. An example of a significant subnetwork in the stirated muscle contraction pathway. Large pathways containing many genes can be easily missed by pathway-based methods. In contrast, PFSNet is able to identify subnetworks within large pathways. (Pathway image from Wikipathways)

the traditional method of testing individual genes for two main reasons. First, gene sets represented by pathways make discovery more interpretable. Second, the ability to identify gene sets whose members might have only slight changes in individual gene expression values.

However, pathways become too generalized when they are large. Such a pathway may also be missed even when a subnetwork within it is indeed important for the disease (Fig. 5). Breaking down pathways to subnetworks provides even better interpretability. Moreover, we discover that previous methods used in microarray data analysis produce 'significant' gene sets or pathways that are not very consistent across datasets of the same disease.

Methods that analyze subnetworks like SNet also have their shortcomings that it is hard to decide what thresholds to use because a too-relaxed threshold will include some non-relevant genes and a too-stringent threshold will exclude some relevant genes.

In this article, we have introduced two improvements to SNet: by incorporating the fuzzification technique (FSNet), and by computing paired *t*-statistic based on the fuzzy score of two phenotypes (PFSNet). We have found that subnetworks identified by FSNet and PFSNet shows higher consistency across independently obtained datasets than other methods.

Funding: This work is supported by National University of Singapore (research scholarship to Lim, in part); and Singapore Ministry of Education (tier-2 grant MOE2012-T2-1-061).

Conflict of Interest: none declared.

REFERENCES

Armstrong,S.A. *et al.* (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41–47.
Cardoso,B.A. *et al.* (2008) Interleukin-4 stimulates proliferation and growth of t-cell acute lymphoblastic leukemia cells by activating mtor signaling. *Leukemia*, **23**, 206–208.
Gatti,D. *et al.* (2010) Heading down the wrong pathway: On the influence of correlation within gene sets. *BMC Genomics*, **11**, 574.
Geistlinger,L. *et al.* (2011) From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, **27**, i366–i373.
Giunta,M. and Pucillo,C. (2012) *BCR-ABL* rearrangement and hla antigens: A possible link to leukemia pathogenesis and immunotherapy. *Revista Brasileira de Hematologia e Hemoterapia*, **34**, 323–324.
Goeman,J.J. and Buehlmann,P. (2007) Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, **23**, 980–987.
Goeman,J.J. *et al.* (2004) A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.
Goldstein,J.A. and McNally,E.M. (2010) Mechanisms of muscle weakness in muscular dystrophy. *J. Gen. Physiol.*, **136**, 29–34.
Golub,T.R. *et al.* (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
Haslett,J.N. *et al.* (2002) Gene expression comparison of biopsies from duchenne muscular dystrophy (dmd) and normal skeletal muscle. *Proc. Natl Acad. Sci. USA*, **99**, 15000–15005.
Hruak,O. and Porwit-MacDonald,A. (2002) Antigen expression patterns reflecting genotype of acute leukemias. *Leukemia*, **16**, 1233–1258.
Ingenuity (1998). www.ingenuity.com (29 November 2013, date last accessed).
Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
Kelder,T. *et al.* (2012) Wikipathways: Building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.
Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
Krans,J.L. (2010) The sliding filament theory of muscle contraction. *Nat. Edu.*, **3**, 66.
Pescatori,M. *et al.* (2007) Gene expression profiling in the early phases of DMD: A constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression. *FASEB J.*, **21**, 1210–1226.

Rayet,B. and Gelinas,C. (1999) Aberrant rel/nfkb genes and activity in human cancer. *Oncogene*, **18**, 6938–6947.

Ross,M.E. *et al.* (2004) Gene expression profiling of pediatric acute myelogenous leukemia. *Blood*, **104**, 3679–3687.

Sivachenko,A.Y. *et al.* (2007) Molecular networks in microarray analysis. *J. Bioinform. Comp. Biol.*, **5**, 429–456.

Soh,D. *et al.* (2010) Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinform.*, **11**, 449.

Soh,D. *et al.* (2011) Finding consistent disease subnetworks across microarray datasets. *BMC Bioinform.*, **12(Suppl 13)**, S15.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

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

Venet,D. *et al.* (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.*, **7**, e1002240.

Yeoh,E.J. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancel cell*, **1**, 133–143.

Zampieri,M. *et al.* (2011) A system-level approach for deciphering the transcriptional response to prion infection. *Bioinformatics*, **27**, 3407–3414.

Zhang,M. *et al.* (2009) Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, **25**, 1662–1668.