# 1 Introduction

It is still popular to find differentially expressed genes by microarray analysis. Independent of the platform and the analysis methods used, the result of a microarray experiment is, in most cases, a list of differentially expressed genes [2]. There are many techniques used to identify differentially expressed genes. These techniques can be divided into three categories: individual genes, gene pathways and gene classes approaches [1]. In the past, people’s focus is in the function of one gene. However, the expression level of this gene in the cells (including tumor cells) is not so high. What really works in cell activity is not a single gene, but a group of genes that participate in a function. So it is necessary to introduce the systematic biology perspective to slightly “macroscopically” observe and evaluate the function of the cells. Recently, more and more research is devoted to gene sets instead of individual genes. Gene sets are group of differentially expressed genes, these genes are usually relating to diseases. Here, we call the differentially expressed genes between patients and normal for the ‘disease genes’. The gene sets blur the “powerful” function of the individual gene, and is more concerned with the role of the whole functional group, thus closer to the normal condition of the cell, which is also its advantage.

Studying gene sets is more conducive to understanding the cause of disease and find out which part of biological functions is affected in the disease. However, the increasing complexity of gene expression data presents several challenges for researchers [2]. A big challenge faced by the researchers is how to choose disease genes according to patients’ gene expression value and how to translate disease genes into a better understanding of the underlying biological process. There have been many microarray analysis methods proposed to solve the problem. Those methods are useful for diagnosis of a disease. Such as GSEA, SNet, PFSNet and so on. But there are also some shortcomings for traditional methods: GSEA is a method based on pathways. Its enrichment analysis was carried out to find out relevant pathways. The essence of this method is the whole pathways are considered and each pathway is scored by a variation of Kolmogorov-Smirnov statistic which determines its importance. But pathways are often large and the statistical score is easily affected by the genes in each pathway. When a pathway contains too few useful genes, the score of it will be low and finally this pathway will be missed, hence causing a big impact on the results. As for SNet, it is a network-based method. In this method, differentially expressed genes is selected by setting a threshold on the gene expression levels. SNet use gene rank value rather than the absolute gene expression values to select genes. This approach makes the results more consistent in two independent datasets. The shortcoming of this method is the threshold is hard to choose and genes around threshold is easily missed leading the results are not accurate. Based on SNet, there is another method called PFSNet. In PFSNet, the genes around threshold are considered by setting two threshold to choose significant genes. Another meaningful improvement of PFSNet is PFSNet do not use the absolute gene expression values but use a “fuzzy value” to substitute gene expression level according to gene ranks. The “fuzzy value” is a number between 0 and 1. This method scores each subnet of pathways using a paired t-statistic based on the fuzzy score of two phenotypes. Finally the p-value of every single subnet is estimated within the subnets list and keep those which are significant. This principle of “network-based” or “pathway-based” association (Califano et al., 2012) is now being applied to effectively map the genetics underlying complex phenotypes, including cancer and other common diseases. However, the scoring process of these methods do not consider the structure of each subnet. The significance of each biological process is measured by gene sets, which only contain differentially expressed genes. So the question is: Is it possible that the causative gene is not highly expressed but affects the expression of genes that are closely linked surrounding it?

This paper looks at a problem of finding biological process according to gene sets based on GO. Different from the previous method, the gene sets we study here are composed of differential expressed genes and some new genes. These new genes have potential to play the same function as differential genes. Directly, the gene sets are first mapped onto knowledge of GO terms; affected subnets are then statistically associated with the disease phenotype. We identify some shortcomings of the previous methods in finding consistent disease subnets. Thinking they did not consider the biological process’ structural information and just focus on the differentially expressed genes. In this article we present our technique, DNet, to identify significant biological process within a phenotype of microarray experiments. The method drives its power by focusing on gene sets as well as GO term network structure information. This method can greatly increase our power to identify relevant associations between phenotype and biological process[4].

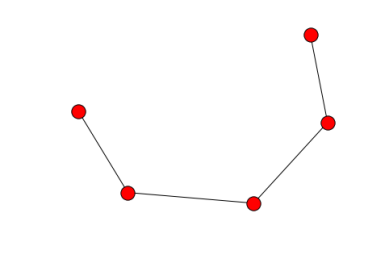
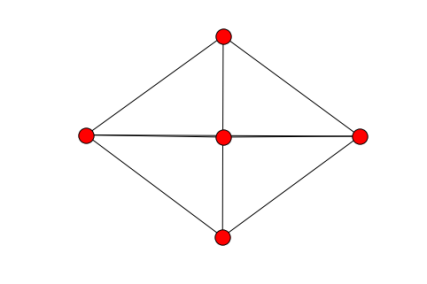
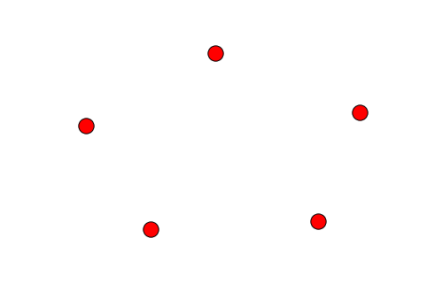
# 2 Method

We blend information on GO and genes from GO annotations. The GO network we study here is a directed graph (DAG) of terms and hierarchical relations. It consists of three branches: biological processes, cellular components and molecular functions. Here, we focus on researching biological process. The graph is made up of vertices and nodes. It is defined as G= (V, E). V is the vertex set whose elements are the nodes of the graph. This set is often denoted V(G) or just V. E is the edge set whose elements are edges of the graph. This set is often denoted E(G) or just E [5]. Each vertices represents a GO term and each edges represent the relationship between two terms.

The subnets we found here is an undirected graph, defined as S = (V’, E’). Each vertices represents a gene and each edges represent the relationship between the two genes. Every subnet belong to a term and represent a small biological process. Each subnets with more than five nodes.

For undirected simple graphs, the graph density is defined as:

where E is the number of edges and V is the number of vertices in the graph. The maximum number of edges for an undirected graph is ½ |V| (|V|−1), so the maximal density is 1 (for complete graphs) and the minimal density is 0(Fig. 1) [6].

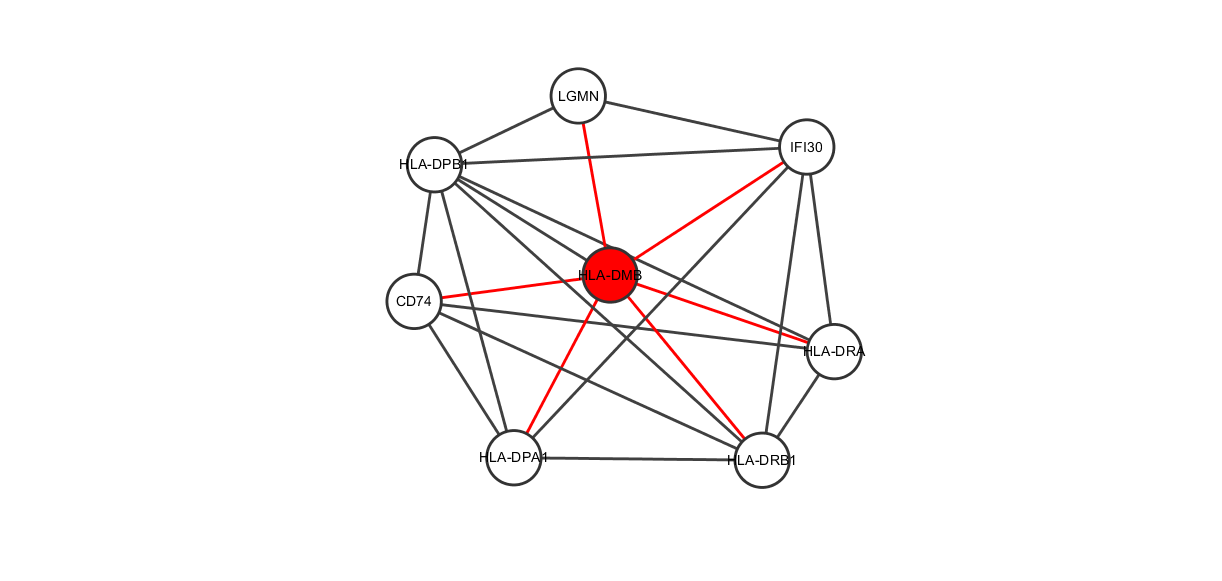
 

**Fig. 1.** (a) The maximum number of edges for an undirected graph is ½ |V| (|V|−1), so the maximal density is 1. (b) 最小连通子图的有N-1个边 (c) The minimal density of a graph is 0.

## 2.1 Subnets generation

For each disease we study here, we consider test patients and normal patients gene expression level together. The phenotype of test patients is defined as D and the phenotype of normal patients is defined as ¬D. We rank each gene for each patients based on gene expression data. After ranking, we give each gene a weight value w( gi, pj ). Genes whose rank is up 95% get a weight value 1 and genes whose rank is below 85% get a weight value 0. Genes between 85% and 95% are given a weight value between 0 and 1 based on their rank. Then we calculate each gene’s average weight value and chose genes whose average weight value greater than 0.5 into the gene list L. These genes are considered to be differentially expression genes in the disease because their weight value are high in most patients.

When we complete the construction of the gene list, the next step is to generate subnets. In this step, we first generate the subnets SS according to L. Then we add genes which are tightly connected to subnets into S. Because the differentially expressed genes may be caused by genes that are not themselves differentially expressed. For example, a gene has a mutation which impact on the shape of its protein product but its expression level are not affected. In this case, it can affect the surrounding genes which closely linked to it. Finally this gene has a great impact on many differentially genes but its expression level has no change. Thus it is easily to be ignored since the traditional method only pick genes that have great changes in expression. So it is more likely that the genes which have more connection to gene sets L plays a similar function and we should pay attention to these “extra” genes as well. E.g. for subnet S with n nodes, when a node that don’t belong to S connect to more than 2\*n/3 nodes in the subnet S, we add this node into S. (Fig. 2)



**Fig. 2**. We add nodes that closed to subnets into subnets S

## 2.2 Subnets scoring

For every subnet S, each patients of phenotype D can be scored as follows:

 （2-1）  （2-2）

Where is the phenotype for which the subnet is generate and is the phenotype for normal patients. ranges over the patients of phenotype and  is the weight value for each patient of each gen.

In this step, we get two scores for each subnet. When we get two scores that both describe the network S, we expect the scores calculated by D and ¬D is really diffident since they come from different phenotype. So a paired t-test was done under the null hypothesis that the difference in scores gives us a distribution with mean=0. While doing a t-test, we considered the structural information of the subnet as well. It is easy to think that sparse networks and dense networks are really different. A dense network has more edges and is more likely to have informations. The nodes in dense network are closely connected and have the tendency to play a same function. Thus, we introduce the density of the graph to measure the score of each network. The density of the graph can reflect the closeness of the nodes in the fraph. So a variant t-test is used to score each subnet. The final score of subnets not only use the gene average weight value but also the density of the subnet. For every subnet S, the score is defined as follows:

(2-3)

(2-4)

Where is the density of subnet S.

## 2.3 Subnets output

# 3 Results

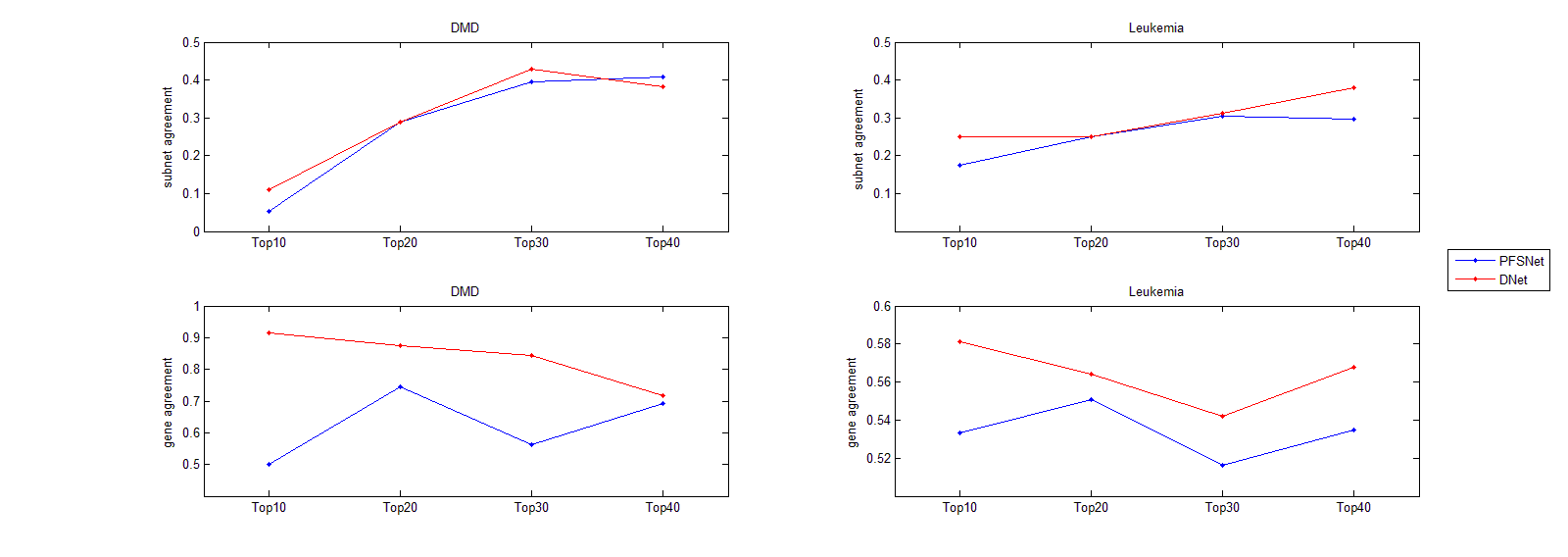
The GO project has develop three structured ontologies that describe gene products in term of their associated biological processes, cellular components and molecular functions in a species-independent manner. A biological process is a recognized series of events or molecular functions. Here, we study the biological process across GO database. We use GO from Gene Ontology Consortium, a project that address the need for consistent description of gene products across databases.

We tested DNet on independent datasets of two diseases. For each of the two disease types studied here—Leukemia (Armstrong et al., 2002; Golub et al., 1999) 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 DNet and PFSNet on the two datasets independently and obtain a corresponding outputs about disease affected biological process. We compared the results form DNet and PFSNet. The results indicate that our technique gives a higher consistently for each disease.

## 3.1 comparing DNet and PFSnet

When comparing DNet and PFSNet, we set = 5% and = 15% for PFSNet. In this way, genes above 95th percentile are given a total vote and genes below 85th are given no vote at all. This allows same genes to be considered in DNet and PFSNet. In the results, we compare top 10, top 20, top 30 and top 40 significant subnets from two datasets using Jaccard similarity coefficient method. It is defined as follows:

We run DNet and PFSNet on two datasets and analysis the results of top 10, top 20, top 30 and top 40 subnets. In DNet, we get even higher subnet-level agreement than PFSNet in the two dataset (Fig.3). In the top 10 subnets of the results, DNet achieves the maximum subnet agreement of 25% in Leukemia datasets and 11.1% in DMD datasets whereas PFSNet achieves the maximum subnet agreement of 17.6% in Leukemia datasets and 5.3% in DMD datasets. This shows the subnet structure plays an important role in consistent disease subnet. Besides, as we allow more genes which closely related to the subnets to be considered, we also measure the gene-level agreement from significant subnets between two datasets to see whether these genes are similar. The result shows that adding these genes into subnets makes the gene-agreement of DNet even higher. In the top 10 subnets of results, In the Leukemia dataset, DNet achieves the maximum gene agreement of 58% whereas PFSNet achieves maximum gene agreement of 53.3%. In the DMD datasets, DNet achieves the maximum gene agreement of 91.5% whereas PFSNet achieves maximum gene agreement of 50%. The results also tell us that the extra genes we find is similar to the differentially expressed genes.

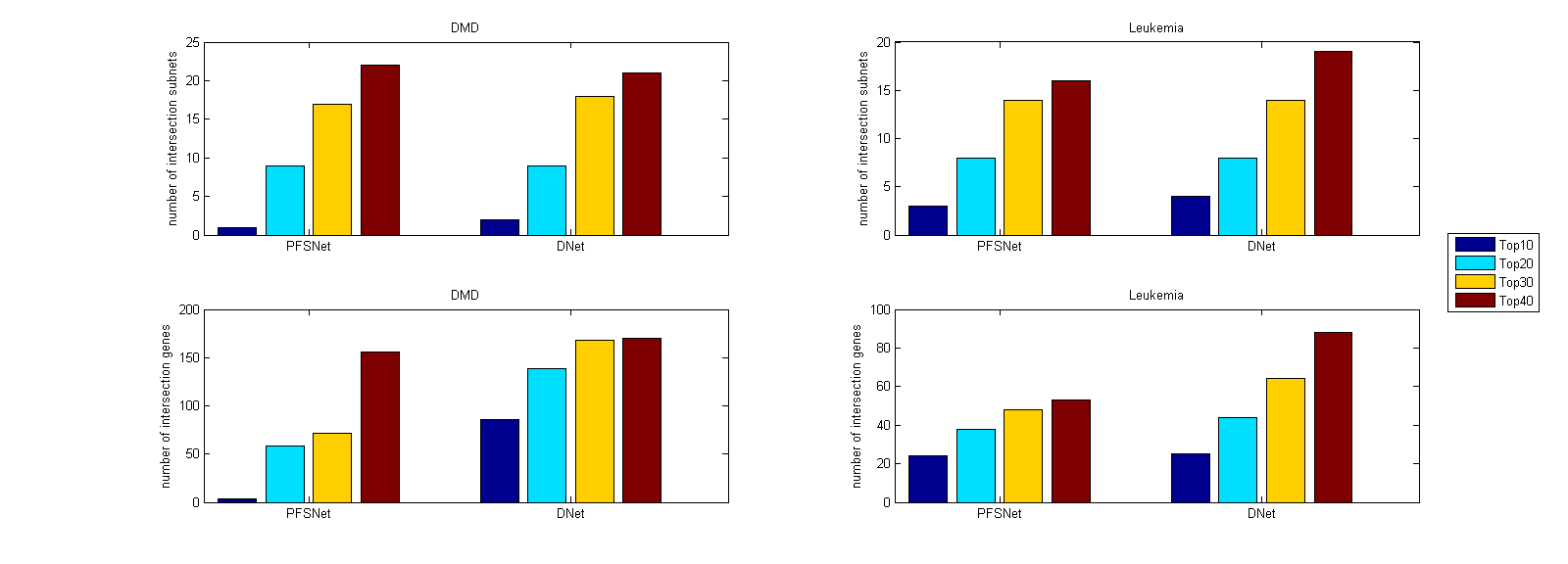


**Fig. 3.** Consistency of subnets and their genes in Leukemia and DMD dataseet.

## 3.2 comparing GO and subnets

GO is a big directed graph. It contains more than 40000 terms and 80000 relationships. The GO data provides a very effective way of linking biological knowledge with the analysis of the large datasets of post-genomics research [7]. As for GO terms, each GO term can be seen as a network made up by genes. We chose terms whose annotation genes between 50 and 100 in our methods. These terms total contain more than 7000 genes and almost 100000 relationships. After that, we break each term’s network into small subnets to see whether it is associated with diseases in two independent datasets.

As there are two independent datasets for each disease, we can get two results for each disease. For the subnets both appear in two datasets, we also compare our methods with PFSNet about the intersection genes in these subnets for each disease. (Fig.4). Comparison results shows that whether it is DMD or Leukemia, the number of intersection subnets and genes in DNet is more than PFSNet. We obtain a low result overlap in the PFSNet. For example, in the DMD datasets, the number of intersection subnets in DNet is 2 whereas the number in FSNet is 1. We compare the number of intersection genes in the intersection subnets as well. The result is the number of intersection genes in DNet is 86 while FSNet can only find 4 genes. It is obvious that the structure of each subnet affects its ranking. When we consider the density of each subnet, we can get a higher agreement among independent datasets. Also, the genes that not differentially expressed in biological process but closely connect to gene sets is similar to differentially expressed genes, consider these genes leading the gene agreement become higher.



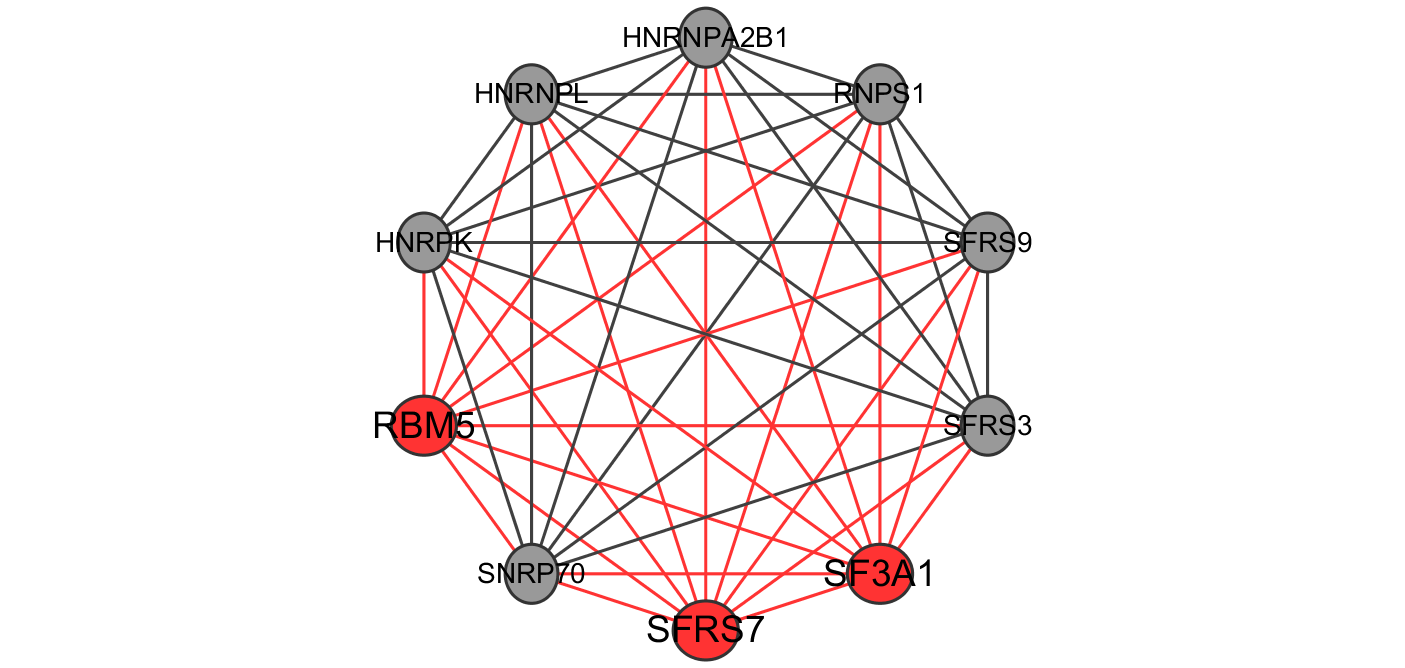
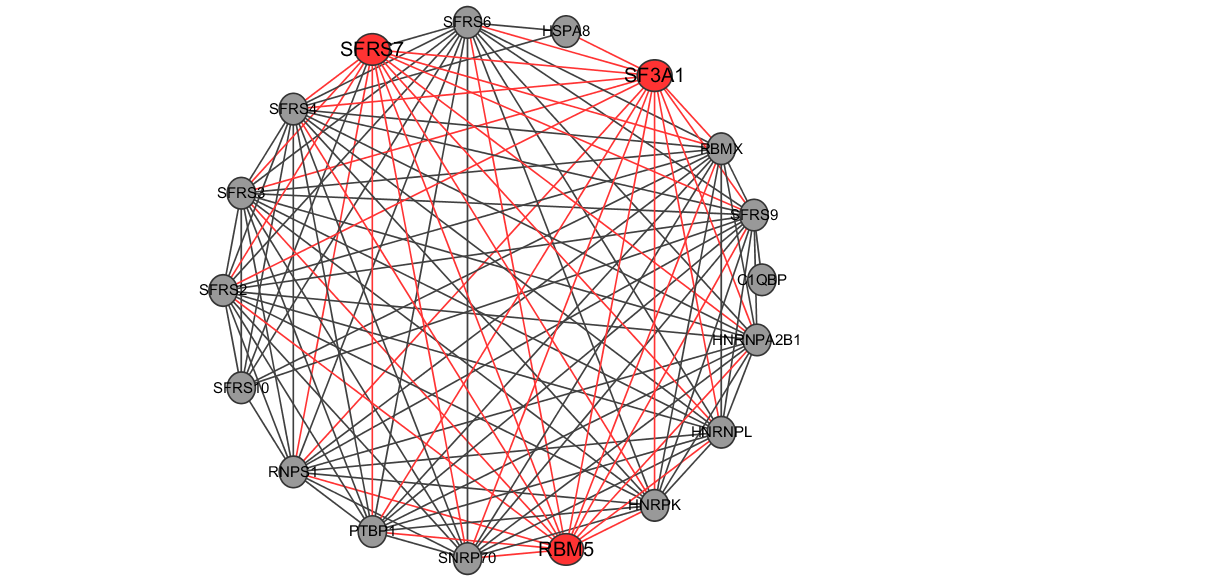
**Fig. 4.** Consistency of subnets and their genes in Leukemia and DMD dataseet.

## 3.3 biologically significant subnets

For the subnets and extra genes we find here, we check them for biologically significance. We discover that the genes we add to subnets participate in many process of the disease.

For Leukemia dataset, one of the significant subnet we find is term “GO: 0048024” . This subnet is associate with This subnet is also in the result of PFSNet. We compare the output of this subnet’s nodes between DNet and PFSNet and get two gene list. Comparing the gene lists we find four genes that appear in DNet while PFSNet do not contain. They are RBM5, SFRS7 and SF3A1 (Fig. 5).

Serine/arginine-rich splicing factor 7 (SFRS7) is related to Leukemia. The protein encoded by it is a member of the serine/arginine (SR)-rich family of pre-mRNA-splicing factors, which constitute part of the spliceosome (Table 1) [11][12]. The SR family plays an important role in the alternative splicing and it is closely related to the occurrence and development of tumor. In the past years, the study of DNA and RNA sequencing (RNAseq) has been very mature. Comparative DNA and RNA sequencing studies have revealed that humanspecific distal regulatory elements, RNA editing, and alternative splicing play key roles in human embryonic stem cell (hESC) self-renewal and cell fate determination. Several of the phosphoproteins regulated during differentiation are components of the posttranscriptional RNA modification machinery, including double-stranded RNA-specific adenosine deaminase (ADAR) and serine/arginine-rich splicing factor 7 (SFRS7) [10]. In many studies of Leukemia, various spliceosome gene mutations were detected and SFRS7 is one of the members[17]. Recent studies have shown that some spliceosome genes involved in the early steps of U2-dependent splice site recognition are commonly mutated in hematologic malignancies and solid cancers. For example, exome-sequencing studies found that SRSF7 was mutated in the patients with chronic lymphocytic leukemia.



**Fig. 5.** RBM5, SFRS7 and SF3A1 in two datasets.

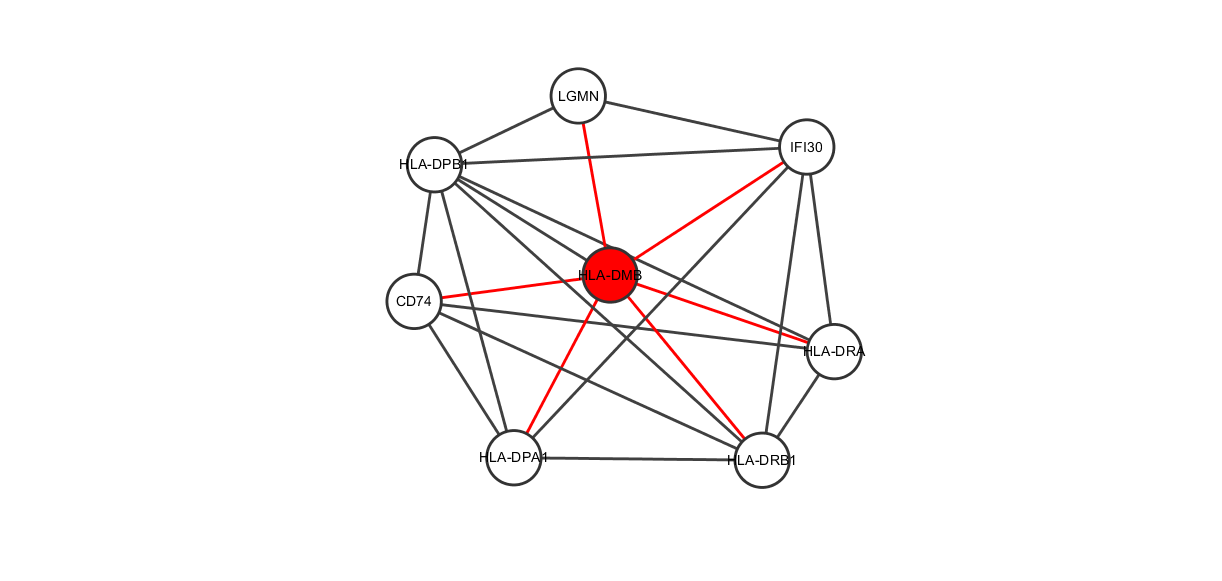
**Table 1.** SR-rich family

|  |  |  |  |
| --- | --- | --- | --- |
| **Human genes encoding SR proteins** | | | |
| Gene name | SR protein | Chromosomal location | UniProt |
| SFRS1 | SF2/ASF/SRp30a | 17q21.3-q22 | Q07955 |
| SFRS2 | SC35/SRp30b | 17q25.1 | Q01130 |
| SFRS3 | SRp20 | 6p21.31 | P84103 |
| FRS4 | SRp75 | 1p35.3 | Q08170 |
| SFRS5 | SRp40 | 14q24.2 | Q13243 |
| SFRS6 | SRp55 | 20q13.11 | Q13247 |
| SFRS7 | 9G8 | 2p22.1 | Q16629 |
| SFRS9 | SRp30c | 12q24.23 | Q13242 |
| SFRS11 | SRp54 | 1p31.1 | Q05519 |

The RNA maturation is an important and complex biological process. It requires several small nuclear ribonucleoproteins (snRNPs) that comprise the two forms of spliceosomes. The major form of spliceosome (U2-type) is composed of U1, U2, U4/6 and U5 snRNPs, and catalyzes most splicing events in metazoans.Mutations of genes, such as SF3B1, SRSF2, U2AF1, ZRSR2, and to a lesser extent SF1, SF3A1, U2AF2 or PRPF40B, encoding spliceosome compounds have been found to occur at high frequencies in myelodysplastic syndromes (MDS) and chronic lymphocytic leukemia (CLL). Subsequently, SF3B1 mutations were also found in solid tumors such as endometrial, lung, bladder, pancreatic and breast carcinomas and cutaneous melanomas [14]. In myelodysplastic syndrome, spliceosome genes were reported to be mutated in 45–85% of patients; mutations were found in SF3A1, PRPF40B and so on. These findings illustrate that RNA splicing-related genes appear to be associated with cancer[15].

The study about RBM5 suggest that RBM6-RBM5 transcription-induced chimerism might be a process that is linked to the tumour-associated increased transcriptional activity of the RBM6 gene. It appears that none of the transcription-induced chimeras generates a protein product; however, the novel alternative splicing, which affects putative functional domains within exons 3, 6 and 11 of RBM6, does suggest that the generation of these chimeric transcripts has functional relevance. Finally, the association of chimeric expression with diseases suggests that RBM6-RBM5 chimeric expression may be a potential tumour differentiation marker [16]. Thus, the disruption of these genes results in splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells.

For DMD, one of the significant subnet is term “GO: 0019886”. In this subnet, we find gene HLA-DMB is associate with subnet. We also check its biological significance.



# 5 conclusion

Analyzing gene sets rather than individual genes has many advantages. The most fundamental thing is that the disease is not caused by a single gene but a group of interacting genes. A variety of methods have been developed for discover biological process about disease based on the differential expression genes. Pathways and GO is always used to analysis. They are all represent the biological process in a form of network. Here, we are committed to finding GO terms related to differentially expression genes. But GO is very large and it describe the whole process of human life. Each term is related to a process. The total network is not conductive to analysis because they will cause false positives. Thus, through the choose of gene sets, we map the gene sets to the subnets. Methods that analysis subnets such as SNet, PFSNet and so on have their shortcomings. First, the gene sets in these methods are composed of differentially expression genes only. Genes that do not change expression value is filtered. Second, when scoring the subnets, the process of scoring is only related to the expression value of each gene. As a graph, the structure is as important as the nodes’ value.

In this article, we improved the above two shortcomings by consider the closely connected genes and the density of each subnets. We have found that the results turns to be more consistent between two independent datasets and prove that our approach is right.

# 6 reference

[1] Finding consistent disease subnetworks across microarray datasets

[2] Ontological analysis of gene expression data: current tools,limitations, and open problems

[3] Visualization and analysis of microarray and gene ontology data with treemaps

[4] Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems

[5] Coleman, Thomas F.; Moré, Jorge J. (1983), "Estimation of sparse Jacobian matrices and graph coloring Problems", SIAM Journal on Numerical Analysis, 20 (1): 187–209, doi:10.1137/0720013

[6] Biggs, N.; Lloyd, E.; Wilson, R. (1986), Graph Theory, 1736–1936, Oxford University Press.

[7]Improvements to cardiovascular gene ontology.

[10] Reversion to an embryonic alternative splicing program enhances leukemia stem cell self-renewal

[11] Gene signatures of drug resistance predict patient survival in colorectal cancer

[12] Lessons from the Cancer Genome

[13] SRSF7 serine and arginine rich splicing factor 7

[14] A common alternative splicing signature is associated with SF3B1 mutations in malignancies from different cell lineages

[15] SF3A1 and pancreatic cancer: new evidence for the association of the spliceosome and cancer

[16] RBM6-RBM5 transcription-induced chimeras are differentially expressed in tumours

[]17] The significance of spliceosome mutations in chronic lymphocytic leukemia