# 1 Introduction

In recent years, more and more research is devoted to gene sets instead of individual gene. Gene sets are group of differentially expressed genes, these genes are usually relating to diseases. 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. We call the differentially expressed genes between patients and normal for the ‘disease genes’. It is still popular to find disease 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 disease genes. These techniques can be divided into three categories: individual genes, gene pathways and gene classes approaches [1]. 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 gene into a better understanding of the underlying biological process.

In the past years, there have been many microarray analysis methods proposed to identify differentially expressed genes. Those methods are useful for diagnosis of a disease. Such as GSEA, SNet, FCS, PFSNet and so on.

The such lists of differentially regulated genes Thus, the use of Gene Ontology enables researchers to summarize results of analyses in this framework [3].

This paper looks at a problem of finding biological process according to gene sets based on GO. Directly, the gene sets are first mapped onto knowledge of GO networks; 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. In this article we present our technique, DNet, to identify significant biological process within a phenotype of a microarray experiment. The method drives its power by focusing on gene sets as well as GO network structure information. This method can greatly increase our power to identify relevant associations between differentially expressed genes and phenotype. 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 [4].

# 2 background

# 3 Method

## 3.1 Subnets generation and scoring

DNet considers genes expression level as well as the genes networks structure information.

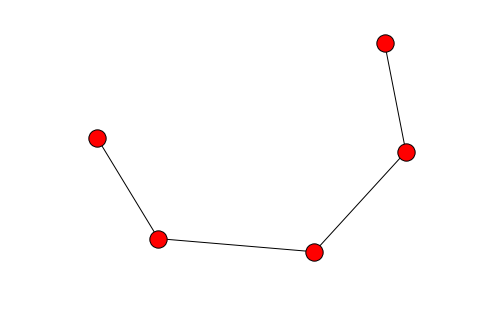
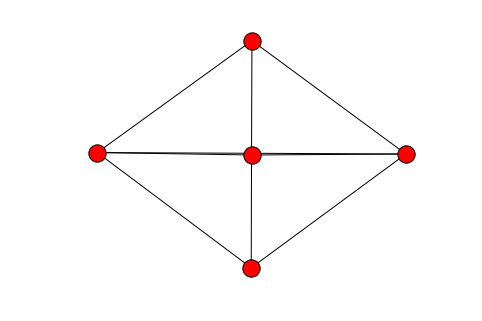
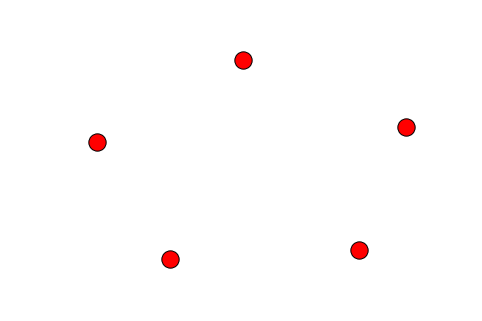
Here, we use the gene ontology network to find disease subnets and biological process by DNet.

The GO network we study here is a directed graph (DAG) of terms and hierarchical relations. 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 means a small biological process.

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

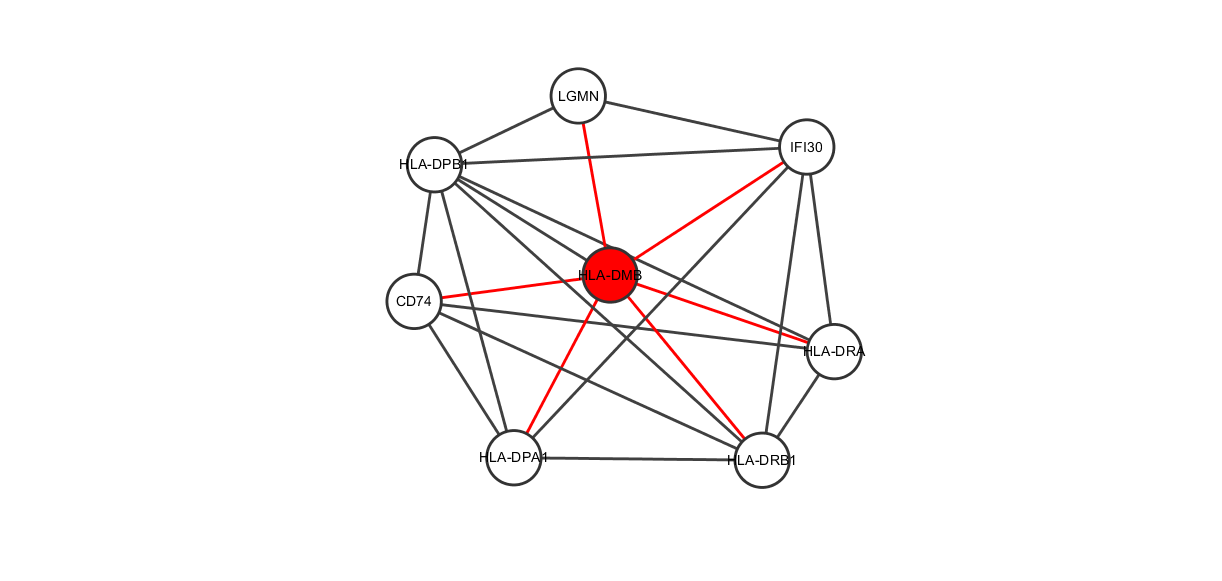
where E is the number of edges and V is the number of vertices in the graph. The maximum number of edges for an undirected graph is ½ |V| (|V|−1), so the maximal density is 1 (for complete graphs) and the minimal density is 0(Fig. 1) [6].

**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

For disease we studied here, we consider test patients and normal patients gene expression data together. They are defined as D and ¬D. We rank each gene for each patients based on 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. We chose genes whose average weight value greater than 0.5 into a gene list. They are considered to be high expression genes in the disease because they are high expressed in most patients.

The next step is to generate subnets using this gene list. In this step, we also considered genes which tightly connected to subnets. Because the differentially expressed genes may be caused by genes that are not themselves differentially expressed. It is more likely that the genes which have more connection to gene sets plays a similar function. E.g. for subnet S with n nodes, when a node that don’t belong to subnet 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

For every subnet S, each patients of 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 patient. ranges over the patients of phenotype and  is the weight value for each patient of each gen.

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 very different. A dense network has more edges and is more likely to have informations. 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.

# 4 Results

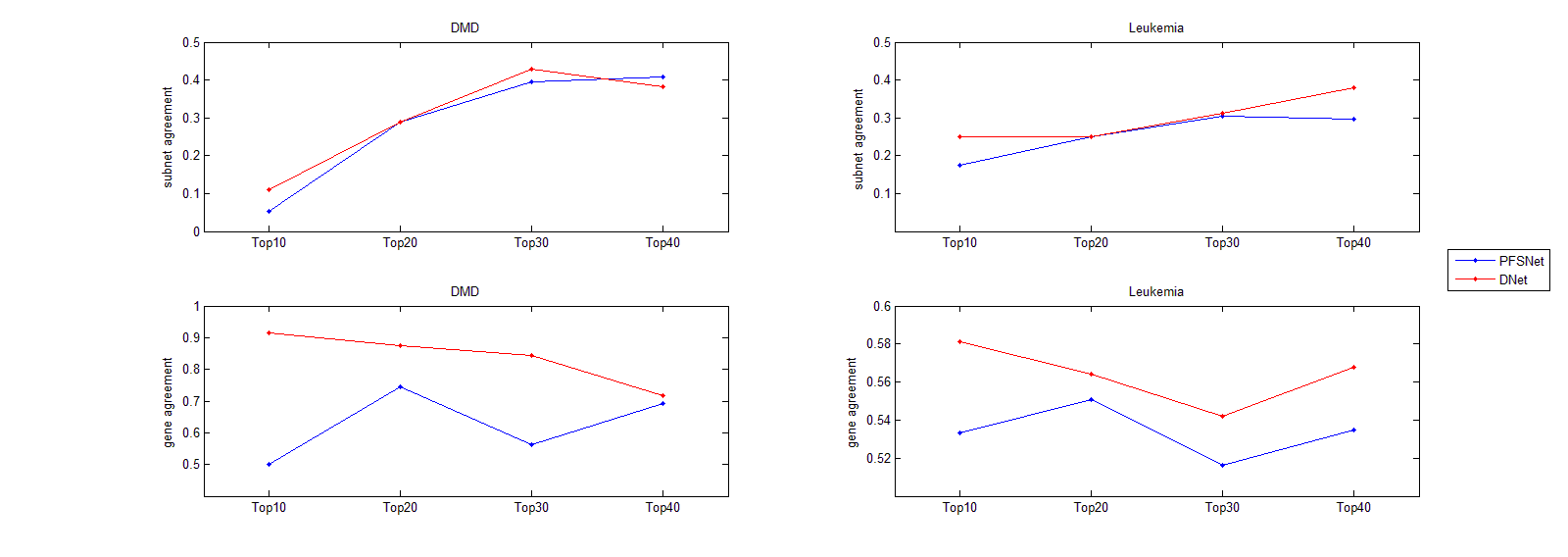
We use GO from Gene Ontology Consortium, a project that address the need for consistent description of gene products across databases. The GO project has develop three structured ontologies that describe gene products in term of their associated biological processes, cellular components and molecular 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.

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 PFSNet and DNet on the two datasets independently and obtain a corresponding outputs about disease affected biological process.We compare the results from PFSNet and DNet.

## 4.1 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:

In DNet, we get even higher subnet-level agreement than PFSNet in the two dataset (Fig.3). 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. The results also tell us that the extra genes we find is similar to the differentially expressed genes. For example, DNet is able to get 25% gene agreement in Leukemia dataset and 11.1% in DMD datasets while PFSNet get only 17.6% and 5.3%.

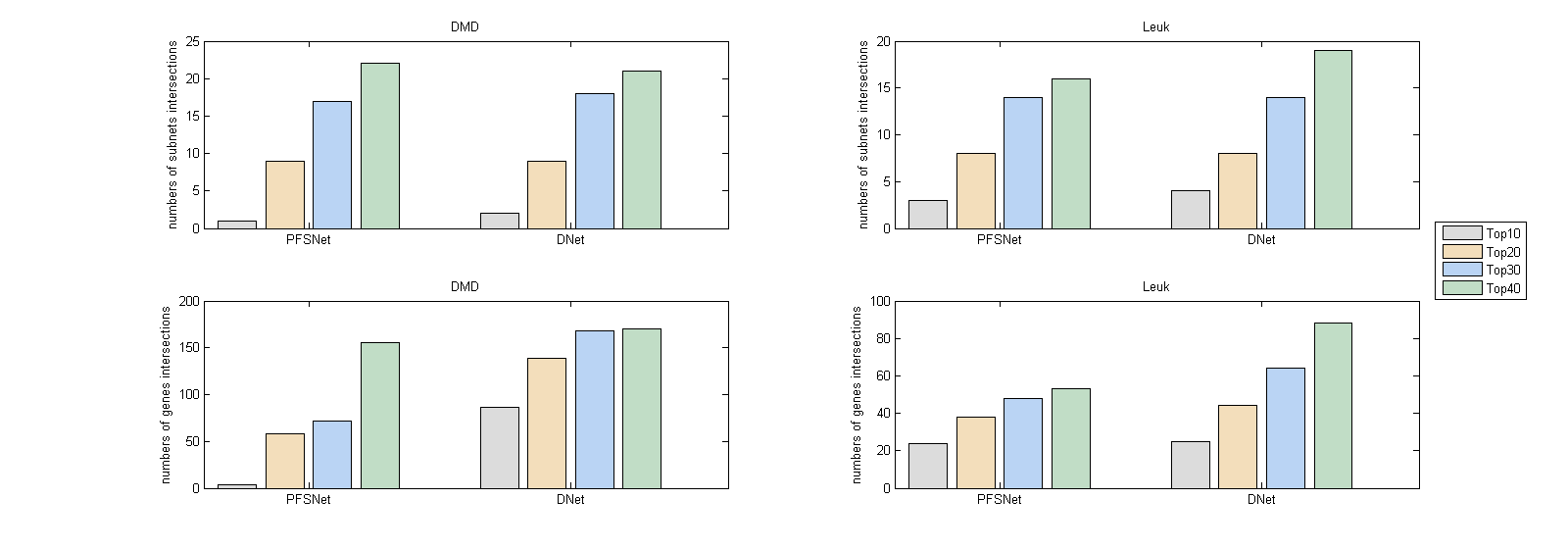


**Fig. 3.** Consistency of subnets and their genes in Leukemia and DMD dataseet.

## 4.2 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 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. We compare our methods with PFSNet. We run PFSNet on both datasets and analysis the results of top 10, top 20, top 30 and top 40 subnets between two datasets (Fig.4). Comparison results shows that whether it is DMD or Leukemia, the number of intersections of subnets and genes in DNet is more than PFSNet.



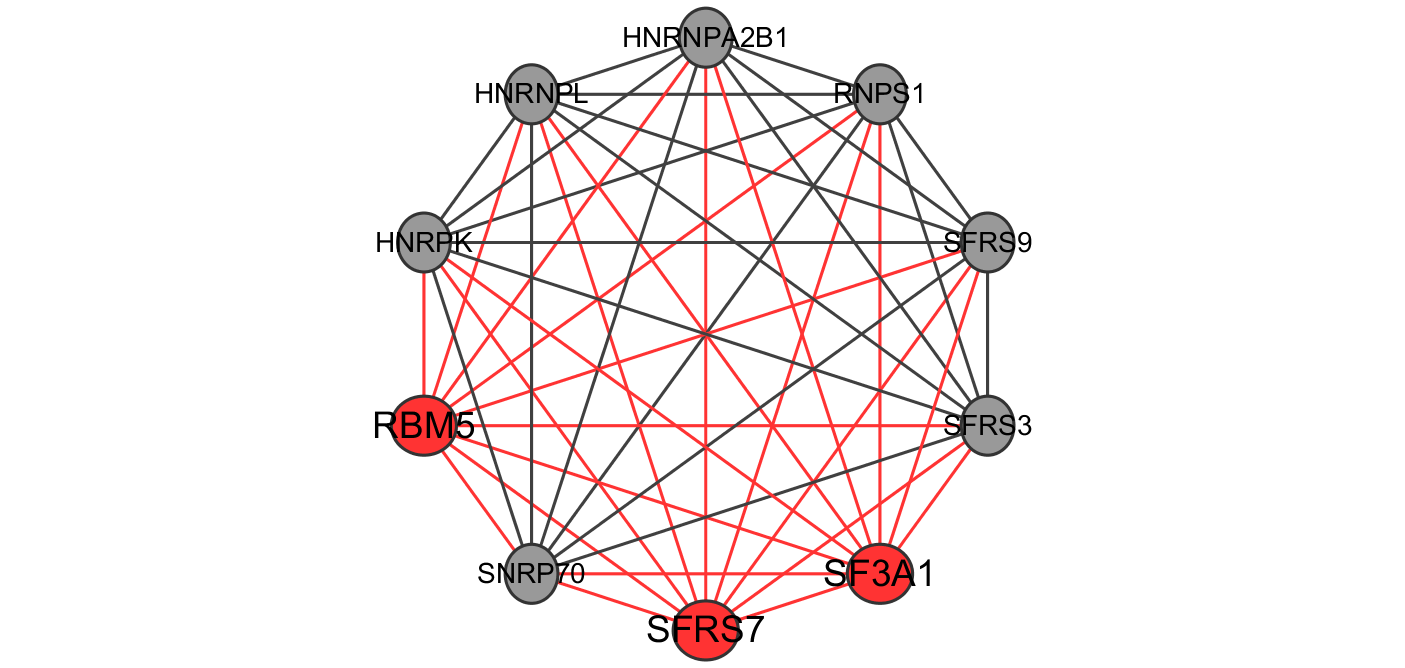
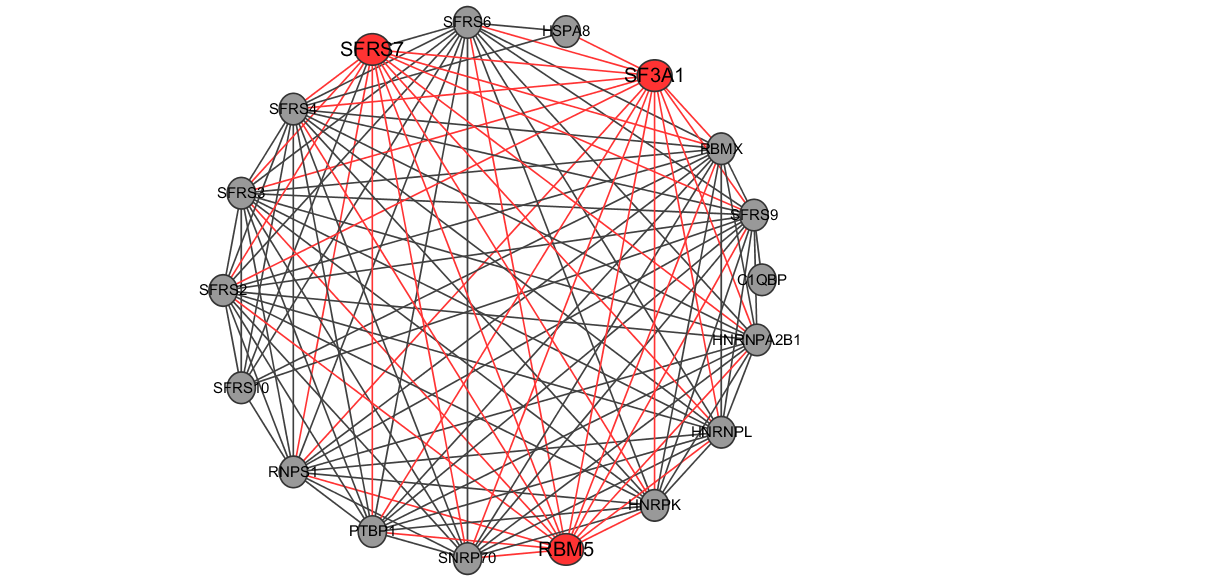
**Fig. 4.** Consistency of subnets and their genes in Leukemia and DMD dataseet.

## 4.3 biologically significant subnets

We check the subnets and genes outputed by DNet for biologically significance.

For Leukemia dataset, one of the significant subnet we find is term GO:0048024. We compare the outputs of this subnet’s nodes between DNet and PFSNet. In the output gene list, we can find four genes that is associated with this subnets while PFSNet can’t find. They are RBM5, SFRS7 and SF3A1 (Fig. 5).

SFRS7 is a member of the serine/arginine-rich family of pre-mRNA-splicing factors, which constitute part of the spliceosome[11].The protein encoded by this gene is a member of the serine/arginine (SR)-rich family of pre-mRNA splicing factors, which constitute part of the spliceosome (Table 1).[12]. Comparative DNA and RNA sequencing (RNAseq) 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].



**Fig. 5.** RBM5, SFRS7 and SF3A1

**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].

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 SF3B1 was mutated in 10–15% of patients with chronic lymphocytic leukemia, while other spliceosome genes (e.g. SRSF1, SRSF7 and U2AF65) were mutated at lower (but still detectable) frequencies in chronic lymphocytic leukemia patients. In myelodysplastic syndrome, spliceosome genes were reported to be mutated in 45–85% of patients; mutations were commonly found in SF3B1, SRSF2 and U2AF35, and also found (albeit at lower frequencies) in SF3A1, PRPF40B, U2AF65 and SF1. These findings illustrate that RNA splicing-related genes appear to be associated with cancer[15].

In the past years, 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].

# 5 conclusion

introduction

[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