Chapter 10

Pathway analysis

In many genomic data analysis, the output is a set of genes associated with disease (e.g. DE gene analysis from microarray data) or a set of co-expressed genes (e.g. from microarray cluster analysis). Although such candidate marker detectioin is useful to narrow down targets for further investigations, the long list of hundreds of genes may contain little unifying biological theme. This leads to difficulty in interpretation and further hypothesis generation. The gene set analysis (a.k.a. pathway analysis) has been pursued for functional annotation of a candidate gene list or an ordered gene result (e.g. ordered by p-values or q-values).

10.1 Pathway database

Many pathway databases are publicly available (Gene Ontology, KEGG, Biocarta, Reactome, MSigDB, Pathway Interaction Database etc). Most of them are in the form of gene sets (i.e. each pathway is represented as a set of genes). Some of them have gene-gene interaction network structure from currated literature information. For example, KEGG (http://www.genome.jp/kegg/) and PID (http://pid.nci.nih.gov/) contain hundreds of carefully constructed pathway networks that reflect accumulated biological knowledge in the field (see Figure 10.1). Embedding such complicated network structure in the analysis is often difficult. In the pathway analysis we describe in this chapter, we only consider pathways as gene sets.

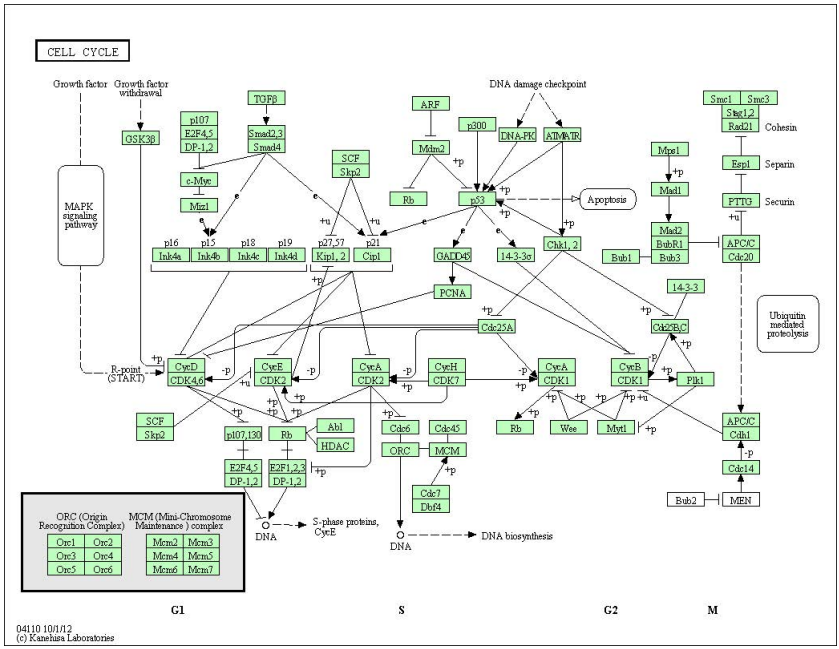

Figure 10.1: Gene-gene interaction network structure of "Cell Cycle" from KEGG.

10.2 Discrete approach: Fisher's exact test

The earliest approach used for pathway analysis is by testing a 2×2 contingency table using either Chi-square test or Fisher's exact test. Consider an expression data set with G genes and S samples, and a pathway of P genes. Suppose analysis of the data set generates a candidate gene list of N genes. Of the N genes, m belongs to the pathways and $N - m$ does not belong to the pathway. A 2×2 table is generated below.

	in pathway	not in pathway	total
in candidate gene list	m	N-m	N
not in candidate gene list	P-m	G-N-P+m	G-N
total	P	G-P	G

Under the null hypothesis, the event of a gene belonging to the pathway and the event it belonging to the candidate gene list are independent (i.e. the candidate gene list is not associated with the pathway). One may perform chi-squared test or Fisher's exact test for such a hypothesis test-

ing. We skip the introduction of these two tests here but describe their pros and cons. (?? add details of the two tests later??) The chi-squared test is easy to calculate but the null distribution is derived approximately. The test is accurate only if the sizes of the pahtwy and candidate gene list are large enough. On the other hand, Fisher's exact test is an exact test under any scenario. Its inference and p-values calculation are, however, slow for large gene sets.

Although the discrete approache described above is useful, it has a few assumptioins and drawbacks. Firstly, it assumes that a candidate gene list is given. Such a gene list is often derived from differentially expressed (DE) gene analysis and a false discovery rate threshold is imposed to generate a candidate gene list. As a result, the selection of threshold is arbitrary and can impact the pathway analysis result. An extreme situation can happen when all genes in the given pathway have moderate p-values (e.g. p=0.05). In this situation, no gene in the pathway can be selected to the candidate gene list after multiple comparison but the pathway is apparently biologically meaningful. Such an arbitrary threshold is relaxed by the continuous approaches introduced in the next paragraph.

10.3 Continuous approach: Kolmogorov−Smirnov test

Continuous approaches differ from discrete approaches in that we do not need arbitrary threshold to produce a candidate gene list for pathway analysis. Instead, the gene order and maybe the magnitude of DE evidence of the genes in the entire genome are considered. We use the famous KolmogorovSmirnov test (KS-test) as an example in this section.

Consider an expression data set with G genes and S samples, and a pathway $\mathbf{P}$ with P genes. Assume an ordered gene list $L = \{g_1, \cdots, g_G\}$ according to DE evidence is available (e.g. ordered by p-values) and the ordered association scores are $R = \{r_1, \cdots, r_G\}$ (e.g. p-values). Denote by the gene sets inside the pathway and outside the pathways as $L_{hit} = \{g_i, g_i \in \mathbf{P}\}$ and $L_{miss} = \{g_i, g_i \notin \mathbf{P}\}$, and assume the corresponding association scores are $R_{hit} = \{r_i, r_i \in \mathbf{P}\}$ and $R_{miss} = \{r_i, r_i \notin \mathbf{P}\}$. Suppose the empirical distributions of R_{hit} and R_{miss} are denoted as $\hat{F}_{hit}(x)$ and $\hat{F}_{miss}(x)$. The KS-test is defined as

$$D = sup_x |\hat{F}_{hit}(x) - \hat{F}_{miss}(x)|.$$

Under the null hypothesis, the DE evidence R has no association with

the pathway. Thus, the two empirical distributions $\hat{F}_{hit}(x)$ and $\hat{F}_{miss}(x)$ should be very similar and the KS-statistics D should be close to 0. Asymptotic theorem can show that the null distribution of D follows a distribution of brownian bridge when G and P is large enough. In practice, the exact null distribution and p-value assessment can be calculated (available in R).

The KS-test can be treated from another angle. Consider the ordered gene list from 1 up to J. Denote by $B_{hit}(\mathbf{P}, J) = \frac{\sum_{j \le J} 1_{\{g_j \in \mathbf{P}\}}}{P}$ and $B_{miss}(\mathbf{P}, J) = \frac{\sum_{j \le J} 1_{\{g_j \notin \mathbf{P}\}}}{G-P}$. We can easily show that

$$D = max_{1 \le J \le G} B(J) = max_{1 \le J \le G} |B_{hit}(\mathbf{P}, J) - B_{miss}(\mathbf{P}, J)| \quad (10.1)$$

Note that the new formulation in (10.1) shows that KS-test is invariant under any monotone transformation of $R = \{r_1, \cdots, r_G\}$. In other words, the test result is identical no matter p-values or t-statistics are used and only the rank by DE evidence matters. We also note that $B(0) = B(G) = 0$ and under null hypothesis, D again should be close to 0.

Example: Consider DE analysis result of 10 genes. In the ordered DE gene list, four genes $L_{hit} = (1, 2, 3, 5)$ are inside a specific pathway and six genes $L_{miss} = (4, 6, 7, 8, 9, 10)$ are outside the pathway. (?? Draw $\hat{F}_{hit}(x)$, $\hat{F}_{miss}(x)$ and $B(J)$??).

10.4 Gene set enrichment analysis

The KS-test described above has two major weaknesses. Firstly, the test is performed for each gene independently. To alleviate this assumption, we may adopt only the KS-statistic and perform permutation analysis to generate null distribution and assess the statistical significance. Secondly, only gene order is accounted for in the KS-test and the strength of DE evidence (i.e. association scores R) is ignored. Gene set enrichment analysis (GSEA) was proposed (Subramanian et al., 2005) to alleviate these two weaknesses and have been a popular tool for pathway analysis. Below we describe detailed procedures of GSEA.

Input data for GSEA:

1. Expression data with G genes and S samples, and a phenotype of interest.

2. Designate a ranking procedure (e.g. from any DE gene analysis such as SAM or LIMMA) to produce an ordered gene list $L = \{g_1, \cdots, g_G\}$ and the corresponding association score of each gene $R = \{r_1, \cdots, r_G\}$. The association score of each gene with the phenotyp of interest can be obtained from Pearson correlation or p-values of two-sample test (e.g. t-test) or linear regression. In GSEA, correlation is the default.

3. Independently obtained or derived gene sets $\mathbf{P_1}, \mathbf{P_2}, \cdots, \mathbf{P_M}$ with $p_1, \cdots, p_M$ genes (e.g. from Gene Ontology or KEGG).

Enrichment score $ES(\mathbf{P_i})$

1. Evaluate the fraction of genes in $\mathbf{P_i}$ ("hits") weighted by their association scores and the fraction of genes not in $\mathbf{P_i}$ ("misses") present up to a given position J in L:

$$T_{hit}(\mathbf{P_i}, J) = \sum_{g_j \in \mathbf{P_i}, j \leq J} \frac{|r_j|}{N(P_i)}, where N(P_i) = \sum_{g_j \in \mathbf{P_i}} |r_j|$$

$$T_{miss}(\mathbf{P_i}, J) = \sum_{g_j \notin \mathbf{P_i}, j \leq J} \frac{1}{G - p_i}$$

Finally, the ES score is defined as $ES(\mathbf{P_i}) = \max_J B(\mathbf{P_i}, J) = \max_J T_{hit}(\mathbf{P_i}, J) - T_{miss}(\mathbf{P_i}, J)$. We note that similar to KS-test, $T_{hit}(\mathbf{P_i}, 0) = T_{miss}(\mathbf{P_i}, 0) = 0$, $T_{hit}(\mathbf{P_i}, G) = T_{miss}(\mathbf{P_i}, G) = 1$ and $B(\mathbf{P_i}, 0) = B(\mathbf{P_i}, G) = 0$ (a property similar to Brownian bridge). In fact, when the weights r_g are all assigned to one, this enrichment score equals KS-test in (10.1).

Finally, the statistical significance and multiple hypothesis testing are assessed via permutation analysis. In the below section, we will discuss issues of permutation in pathway analysis.

10.5 hypothesis setting and permutation analysis

According to Tian et al. (2005), two hypotheses Q_1 and Q_2 are considered in the literature for pathway analysis (cited from the orginial paper).

1. Hypothesis Q_1: The genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes.

2. Hypothesis Q_2: The gene set does not contain any genes whose expression levels are associated with the phenotype of interest.

In general, permuting genes in the analysis is aimed to pursue Q_1 and permuting samples is for Q_2. In the former case, the association scores are deterministic and the gene set structure is random and vice versa for the latter case. (?? go through the appendix in Tian et al., 2005??)

10.6 Conclusioin

Pathway analysis is a powerful tool to link new findings from the analysis with existing biological knowledge. It provides better interpretation of the data and is useful to generate new biological hypothesis. Many methods have been developed (e.g. GSA, random set method etc). Nam and Kim (2008) provides a comprehensive review of methods (Table 1), software packages (Table 2) and pathway databases (Table 3). Other user-freindly packages also exist, such as Ingenuity Pathway Analysis (IPA), (DAVID) from NIH and MetaCore.

Related reading:

- Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad SciUSA 2005;102:1554550.

- Tian L, Greenberg SA, Kong SW, et al. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 2005;102:135449.

- Dougu Namand Seon-Young Kim. Gene-set approach for expression pattern analysis. BRIEFINGS IN BIOINFORMATICS. 2008; 9:189-197.