Poster L-14

Combining DNA copy number and gene expression data to reveal recurrent genetic abnormalities in pancreatic cancer

Authors:
Liviu Badea (*AI department, National Institute for Research in Informatics*)
Doina Tilivea (*AI department, National Institute for Research in Informatics*)

Short Abstract: We use the Pollack lab pancreatic cancer dataset for determining a set of genes with consistent expression changes in pancreatic cancer and relate these genes with the DNA copy number changes observed in the individual samples. A rank-covering algorithm provides an aggregated view of the candidate genes with DCN alterations that potentially explain the observed phenotype.

Long Abstract:

1. Motivation

DNA copy number and gene expression data are complementary for deciphering the molecular-level mechanisms of cancer, but combining them is highly non-trivial. Most biologically-oriented papers use human expert evaluations of the data without any automated means of combining the two. On the other hand, the few existing computationally-oriented studies (e.g. [2]) have mainly focused on determining the correlation between DNA copy numbers (DCN) and gene expression, without a detailed analysis of a cancer dataset and without linking isolated DCN changes with the phenotype.
Here we use the pancreatic cancer dataset of Bashyam et al [1] for determining a set of genes with consistent expression changes in pancreatic cancer and relate these genes with the DNA copy number changes observed in the individual samples. A rank-covering algorithm is used to obtain an aggregated view of the candidate genes with DCN alterations that potentially explain the observed phenotype. The algorithm recovered a set of genes with well-known roles in pancreatic cancer, but also suggested an additional set of genes, which seem to be involved in cancer, but whose exact role in pancreatic cancer remains to be determined experimentally.

2. The data set

Bashyam et al. [1] performed simultaneous array Comparative Genomic Hybridization and microarray expression measurements on a set of 23 human pancreatic cell lines using cDNA microarrays.
We retrieved the normalized intensity ratios from the Stanford Microarray Database and used the CGH-Miner software as in [1] to identify DNA copy number gains and losses.

3. A simple model of “common genes” and recurrent DCN alterations

As the different samples seem to have underwent distinct genetic alterations in each sample, we assume that there exists a set of “common genes” responsible for the disease phenotype,

which are either over- or under-expressed in all samples and are directly or indirectly affected by all these different sample-specific DCN changes. Our goal consists in identifying sets of genes with DCN alterations that are potentially responsible for the up-or down-regulation of these “common genes”. The very limited number of samples (23) will most probably not allow us to determine all DCN changes responsible for tumorigenesis in all samples (also since no data about point mutations, e.g. SNPs, or promoter hypermethylation is available).
We are mainly interested in genes affected by recurrent DCN abnormalities, i.e. genes that not only have DCN abnormalities in certain samples, but also have expression levels that are well correlated with the phenotype over the majority of samples (thus representing potential therapeutic targets).

3.1. Common genes

We constructed two lists of “common” up- and respectively down-regulated genes: Common+ (65 clones corresponding to 43 named genes), Common- (466 clones, 319 named genes). Many of the genes with a known role in pancreatic cancer are among these, for example the Bcl2-family genes BCL2, BCL2A1, BCL11B.
However, the genes with a known role in pancreatic cancer make up only a minority of the “common” genes. Understanding the precise relationships between the rest, as well as the distinct mechanisms whereby the “common genes” are affected in each sample is a daunting task. We have performed an extensive search of the literature for links between “common genes”. BCL2 and FOS are hubs in this network.

3.2. Genes with DCN alterations

Explaining the sample-specific changes leading to the observed expression changes in the common genes requires first the determination of the genes with DCN alterations that are matched by corresponding expression changes. More precisely, we construct two lists of genes with DCN amplifications (DCN+: 499 genes) and deletions respectively (DCN-: 465 genes). The overlap of Common+ and DCN+ contains 14 genes, while that of Common- and DCN- has 30 genes. A few such genes such as SMAD4, ERBB2, BCL2L1 are well known, but the vast majority is still very poorly characterized. The main contribution of this work consists in proposing a method for relating “common” genes to DCN alterations in the individual samples.

4. Relating common genes to DCN alterations in the individual samples

We start by constructing a matrix of potential causes for the observed expression changes of the common genes in the individual samples. Such potential causes for a common gene G and a sample s are genes g with DCN alterations in s whose expression levels are well-correlated with those of G over the entire set of samples. More precisely, we recorded – for each common gene G and each sample s – the set of genes g with DCN alterations in s having the best N=10 absolute uncentred correlations with G.
The matrix of potential causes is quite large and an aggregated view for each sample may be useful. More precisely, we may want to determine the main DCN alterations in each sample (responsible for the observed expression changes in the common genes). For this purpose, we use a so-called rank-covering algorithm. In each sample s, the algorithm starts with the

best correlated “DCN genes” g for each “common gene” G (g are called genes of rank 1). Then, the common genes G covered by each distinct g are determined and the genes g are sorted according to the number of common genes covered. Such a covering of “common genes” with “DCN genes” may serve as a potential link between DCN alterations and the common phenotype. However, due to the inherent noisy nature of microarray data, the correlations of the genes gi covering G with G may be quite close to one another (typical absolute values are around 0.9), thus rendering small rank differences meaningless.
We take this into account by considering the covers of “DCN genes” with progressively higher ranks k. Of course, such higher rank covers are larger and more robust to small differences in correlations.

4.1. Results

A number of genes with known involvement in pancreatic cancer stand out: BCL2, TCF4, SMAD4, etc.
The results also suggest potential roles in pancreatic cancer for some other genes: HOXB5, CUGBP2 (a critical regulator of the apoptotic response), TUSC3, etc. (more details in the poster).

References

[1] Bashyam et al. Neoplasia 2005 Jun 7(6):556-62.
[2] Lipson et al. Proc. WABI 2004, LNCS 3240, p.135, Springer 2004.