Making sense of microarray gene lists using text mining and over-representation analysis

Hui Sun Leong, Peter J. Giles, Suraj Menon and David Kipling
Department of Pathology, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK
Contact email: leongh@cardiff.ac.uk

Background

A major challenge in microarray data analysis is to place the list of differentially expressed genes in biological context and make functional interpretation. A common approach to address this is over-representation analysis (ORA), which uses hypergeometric distribution (or its binomial variants) to evaluate whether a particular functionally defined group of genes is represented more than expected by chance within the gene list. However, existing applications of ORA are largely limited to pre-defined terminologies (e.g. GO, KEGG). Therefore, the aim of this study is to expand the current ORA protocol to a wider mining of free-text, initially in the form of PubMed abstracts.

Results

Our initial explorations revealed an unappreciated annotation bias, which causes experimentally-derived gene lists to have a greater degree of associated abstracts than an equivalently-sized gene list created by random sampling from the chip (**Figure 1**). This bias skewed the hypergeometric *p*-values and leads to an apparent over-representation of many common and uninformative terms, alongside terms that convey biological insight.

To address the annotation bias problem, an ORA approach based on the detection of outliers was developed. The fundamental idea underlying this approach is: abstract terms pertinent to the biology under study often have lower background frequency than expected and assumes a distribution that is different from the remaining observations, thus appear as outliers. These 'interesting' terms can be identified using the standard outlier detection method based on variance stabilisation and local Z-score calculation. However, this simple Z-score based outlier identification method is susceptible to the so-called 'masking' effect (i.e. the less extreme outlier becomes undetected because of the most extreme outlier), which occurs when applied to extremely long and highly-annotated gene lists. This limitation was subsequently addressed by incorporating the Fisher non-central hypergeometric distribution model into our initial outlier detection protocol to give a better estimation of the location and scale parameters that define the outlier criteria.

When the improved outlier detection method was tested on a list of experimentally-defined interferon-stimulated genes (ISG) published by Sanda *et al.* (2006), biologically meaningful terms such as '*interferon*' and '*MHC*' are identified as significantly enriched in the gene list (**Figure 2**). This result is also comparable to that produced by mining GO terms using the

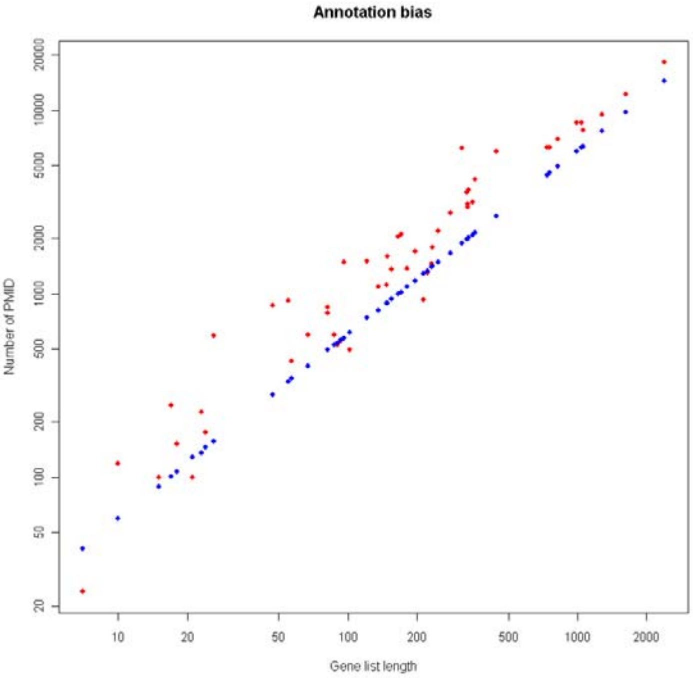

Figure 1 Comparison of the amount of annotation associated with experimentally-derived and random gene lists. 52 gene lists performed on the Affymetrix HG-U133A platform were collected from published literatures and these are referred to as the experimentally-derived gene lists (red points). Random gene lists that match the experimentally-derived gene lists in gene list length were created by random sampling from the whole chip (blue points). The numbers of PubMed articles associated with them were then calculated.

popular online tool EASEonline, but interestingly, none of the significant GO terms indicates the involvement of interferon. This suggests that mining PubMed abstracts could potentially reveal additional biological insight that is not possible by mining pre-defined ontology.

It was found that the proposed text mining approach is able to recapitulate manually determined themes from microarray studies. As an example, Nishimura *et al.* (2003) found that the *pmr4* mutant plant is more resistant to pathogens rather than becoming more susceptible and, based on evidence from gene expression profiling, they concluded that the basis for such resistance was due to enhanced activation of the salicylic-acid (SA) signal transduction pathway. The same gene list was re-analysed with the improved outlier detection method and terms such as '*salicylic*' and '*SA*' are among the most significant hits (**Table 1**), providing powerful proof-of-principle.

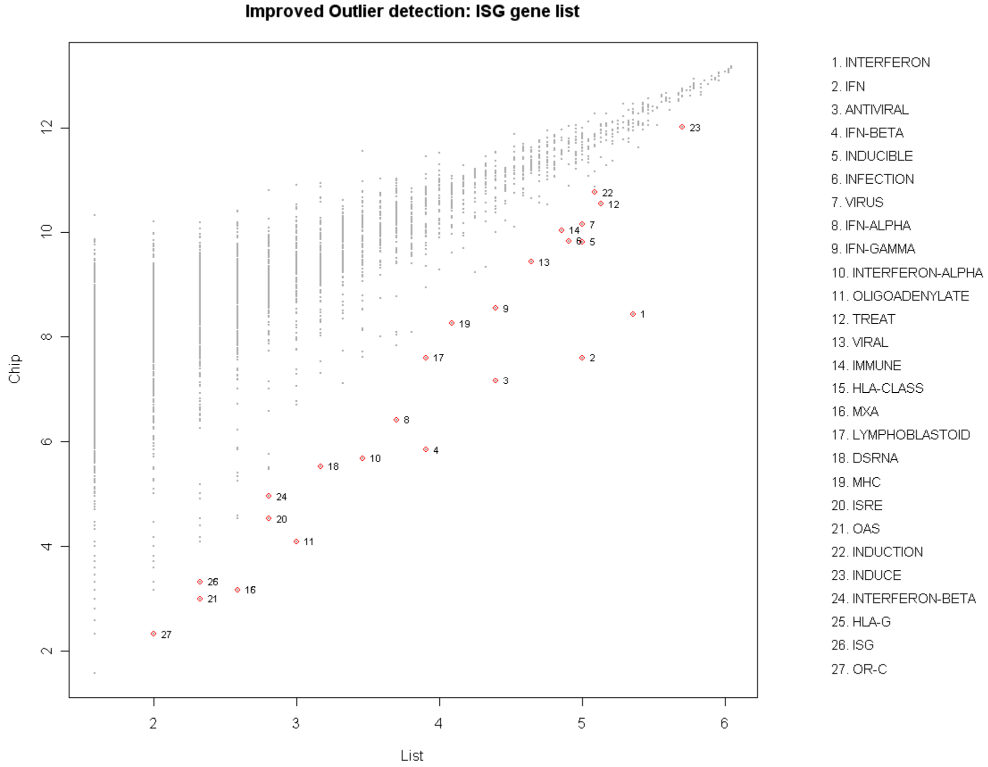

Figure 2 Over-represented abstract terms in the ISG gene list. The results are ordered by increasing value of *p*-value and the number in front of the terms represents the ranking. The significant terms can be detected as outliers in the *Chip*-versus-*List* plot (red circles), where each data point represents an abstract term, *Chip* (y-axis) represents the number of genes associated with each term on the whole chip, and *List* (x-axis) represents the number of genes associated with each term in the ISG gene list. Both axes are on a base-2 logarithmic scale.

Table 1 Over-represented abstract terms in Nishimura study

Token	Chip	List	Z score	*p*-value	Bonferroni *p*-value	Ranking
SALICYLIC	41	11	-5.33163	4.87E-08	3.27E-05	1
PSEUDOMONAS	36	9	-4.29282	8.82E-06	0.005928	2
SA	21	7	-4.09307	2.13E-05	0.014303	3
RESISTANCE	113	15	-3.83798	6.20E-05	0.041681	4
PHYTOALEXIN	5	4	-3.8	7.23E-05	0.048618	5
CAMALEXIN	5	4	-3.8	7.23E-05	0.048618	6

References

Sanda *et al.* (2006). Differential gene induction by type I and type II interferons and their combination. *J Interferon Cytokine Res.*, 26(7), 462-72.

Nishimura *et al.* (2003). Loss of a callose synthase results in salicylic acid-dependent disease resistance. *Science.*, 301(5635):969-72.