Poster L-9

Discovering co-occurrences of gene annotations in large lists of genes

Authors:
Pedro Carmona-Saez (*BioComputing Unit. National Center of Biotechnology*)
Monica Chagoyen (*BioComputing Unit. National Center of Biotechnology*)
Jose M. Carazo (*BioComputing Unit. National Center of Biotechnology*)
Francisco Tirado (*Computer Architecture Department. Universidad Complutense de Madrid*)
Alberto Pascual-Montano (*Computer Architecture Department. Universidad Complutense de Madrid*)

Short Abstract: We have developed a methodology to integrate biological annotations from several sources and extract sets of categories that frequently appear together in a list of genes. This method can be used to analyze large lists of genes to get insights into the associated biological mechanisms.

Long Abstract:
High-throughput techniques such as DNA microarrays, genome wide location analysis or proteomics methods have opened new ways to study biological processes from a global perspective. In many cases the net result of these experimental procedures consists of large lists of genes, e.g. genes that are differentially expressed among pathological and normal tissues.
Several methods have been proposed to analyze such lists of genes in order to get insights into the biological processes that are associated to the experimental conditions or analyzed phenotypes. Most of these applications find biological annotations that are significantly enriched in a list of genes with respect to a reference set, usually the whole genome or, in the context of gene expression, those genes used in a microarray. Using a specific source of biological annotations such as Gene Ontology those methods first find all the terms associated with the set of analyzed genes. The number of appearances of each term is then determined in the input and reference lists and a statistical test is used to compute p-values which can be subsequently adjusted for multiple testing. The result of this analysis consists of a list of single terms from a given ontology with their corresponding p-values. A review of such methods has been recently published by Khatri and Draghici (Khatri and Draghici, 2005).
However, identification of co-occurrences of diverse annotations in a set of differentially expressed genes can provide additional information to help in interpreting the biological mechanisms that are relevant to the experimental system. For example functional information and cellular localization can be integrated in order to define sets of genes that are involved in the same biological mechanisms and located in the same compartments, such as peroxisomal genes involved in fatty acid degradation (Koerkamp et al., 2002; Smith et al., 2002).
We have developed a methodology to integrate annotations from several sources and extract sets of annotations that frequently appear together and are enriched in a list of genes with respect to a reference set. This methodology is based on a variant of the classical a priori algorithm (Carmona-Saez et al., 2006) able to perform an exhaustive search and obtaining

all sets of annotations that co-occur in at least x genes in the analyzed list. Briefly, in a first step the procedure starts by determining the set of all single annotations that appear in at least x genes from the input set of genes. Using these annotations, two-element sets are generated and the database is scanned again to explore each gene and counting the frequency of each pair of annotations. The procedure is continued until no more combinations of annotations are possible. When all sets of annotations have been generated, a p-value is calculated based on the hypergeometric distribution (Boyle et al., 2004; Castillo-Davis and Hartl, 2003). Combinations of annotations with low p-values are significantly associated to the list of genes under study and can provide meaningful insights into the underlying biological mechanisms.

This method provides a novel functionality for functional profiling of gene lists based on the integrated analysis of annotations from several sources. This approach complements the functionality of similar data mining tools. We hope that this novel methodology can help researchers in the analysis of large lists of genes derived from high-throughput experimental techniques such as DNA-microarrays, SAGE, Genome-Wide location analysis or proteomics techniques.

Acknowledgements

This work has been partially funded by the Spanish grants CICYT BFU2004-00217/BMC, GEN2003-20235-c05-05, TIN2005-5619, PR27/05-13964-BSCH and a collaborative grant between the Spanish CSIC and the Canadian NRC (CSIC-050402040003). PCS is recipient of a grant from CAM. APM acknowledges the support of the Spanish Ramón y Cajal program.

References

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. and Sherlock, G. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics, 20, 3710-3715.

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M. and Pascual-Montano, A. (2006) Integrated analysis of gene expression by association rules discovery. BMC Bioinformatics, 7, 54.

Castillo-Davis, C.I. and Hartl, D.L. (2003) GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics, 19, 891-892.

Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21, 3587-3595.

Koerkamp, M.G., Rep, M., Bussemaker, H.J., Hardy, G.P., Mul, A., Piekarska, K., Szigyarto, C.A., De Mattos, J.M. and Tabak, H.F. (2002) Dissection of transient oxidative stress response in Saccharomyces cerevisiae by using DNA microarrays. Mol Biol Cell, 13, 2783-2794.

Smith, J.J., Marelli, M., Christmas, R.H., Vizeacoumar, F.J., Dilworth, D.J., Ideker, T., Galitski, T., Dimitrov, K., Rachubinski, R.A. and Aitchison, J.D. (2002) Transcriptome profiling to identify genes involved in peroxisome assembly and function. J Cell Biol, 158, 259-271.