BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 7 2003, pages 891–892
DOI: 10.1093/bioinformatics/btg114

GeneMerge—post-genomic analysis, data mining, and hypothesis testing

Cristian I. Castillo-Davis and Daniel L. Hartl*

Department of Organismic and Evolutionary Biology, Harvard University, Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received on August 27, 2002; revised on October 28, 2002; accepted on December 9, 2002

ABSTRACT

Summary: GeneMerge is a web-based and standalone program written in PERL that returns a range of functional and genomic data for a given set of study genes and provides statistical rank scores for over-representation of particular functions or categories in the data set. Functional or categorical data of all kinds can be analyzed with GeneMerge, facilitating regulatory and metabolic pathway analysis, tests of population genetic hypotheses, cross-experiment comparisons, and tests of chromosomal clustering, among others. GeneMerge can perform analyses on a wide variety of genomic data quickly and easily and facilitates both data mining and hypothesis testing.

Availability: GeneMerge is available free of charge for academic use over the web and for download from: http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html.

Contact: ccastillo-davis@oeb.harvard.edu

Supplementary information: http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html

INTRODUCTION

In the face of ever-growing genomic and proteomic data, researchers are shifting their attention to post-genomic analysis—the interpretation and synthesis of thousands of data points from a chemical, biological, or evolutionary perspective. Simple and flexible software that can take advantage of diverse genomic and proteomic data for both data mining and hypothesis testing is required.

A great deal of genomic and proteomic information is available. For many genes, something is known about their molecular and biological function (for example, The Gene Ontology Consortium, 2000), pathway membership (Kanehisa *et al.*, 2002), physical chromosomal location, level of polymorphism, RNAi phenotypes, disease phenotypes, and rate of molecular evolution. For non–coding regions, data are often available concerning the presence of known or putative transcription binding sites, levels of DNA methylation or acetylation, and GC content.

Unfortunately, this information is often difficult to integrate into a given genomic analysis since no common platform exists for analysis of such data. Ideally, genomic information should be available for analysis within a unified framework where sets of genes from any experiment can be interrogated easily. Currently, such a framework does not exist. We have developed software in order to meet this need called GeneMerge.

Are functional pathways up-regulated during bacterial infection also up-regulated during fungal infection? What are they? Are fast evolving genes preferentially located in areas of high recombination? Do co-regulated genes show non-random enrichment for a certain family of transcription factor binding sites? These and other questions can be answered quickly and easily with GeneMerge.

FRAMEWORK

GeneMerge returns descriptive information regarding genes under investigation and statistically-based rank scores regarding over-representation of descriptors in a given set of genes. Functional or categorical descriptive data is associated with genes in *gene-association* files. These text files link each gene in a genome with a particular datum of information. For example, the name of a gene and its chromosomal location, molecular function, or its identity as over-expressed in a particular type of cancer. Some currently available gene-association files are listed in the Supplementary information.

Gene-association data are many and varied and will undoubtedly grow as genomic and proteomic investigations accelerate. To deal with this explosion of data requires both a clear analytical framework and the flexibility to incorporate new data as soon as they become available. GeneMerge addresses the first requirement by performing a simple statistical test to answer a straightforward question, are particular functions or categories over-represented in the study data set? GeneMerge addresses the second requirement, easy incorporation of newly available data, with its simple gene-association file format. GeneMerge gene-association files are easy to create such that almost any worker can generate an

*To whom correspondence should be adressed.

association file for use in their study (see Supplementary information). This means that as new information about genes and proteins is generated (publicly or privately) it can be quickly and easily incorporated into an analysis.

GeneMerge takes four input files: (1) Study set gene file; (2) Population set gene file; (3) Gene-association file; (4) Description file. The study set is comprised of genes that are currently under investigation. The population set is comprised of those genes from which the study set was drawn, often a genome. The gene-association file links gene names with a particular datum of information using a shorthand identifier (ID). Finally, the description file contains human-readable descriptions of gene-association IDs.

Output is a tab-delimited text file that can be opened in most spreadsheet programs. It contains functional or categorical data associated with each gene in the study set and rank scores for over-represented functions/categories, as well as other pertinent data. Two sample GeneMerge analyses and further documentation is available in the Supplementary information.

STATISTICS

Rank scores for functional or categorical over-representation within the study set of genes is obtained using the hypergeometric distribution (1). The hypergeometric distribution describes the discrete probability of selecting r items of one kind in a sample of size k from a population of size n, where p is equal to the proportion of r-type items in the population, and sampling is without replacement (Sokal and Rohlf, 1995).

$$\Pr(r|n,\, p, k) = \frac{\binom{pn}{r}\binom{(1-p)n}{k-r}}{\binom{n}{k}} \qquad (1)$$

The hypergeometric distribution thus gives a quantification of the level of one's 'surprise' at finding over-representation for a particular item in a given sample of size k drawn from a larger population, size n. In GeneMerge, k is always the study set of genes and n is the population set, the set from which k is drawn, usually a genome or all genes on a particular DNA array. The study set k may be genes found to be significantly up- or down-regulated in a microarray experiment or a list of genes deemed interesting for another reason. Genes in the sample k are associated with particular identifiers, for example functions, processes, or states. The number of genes with a particular identifier is r. The fraction p is the proportion of genes in the population n associated with the particular identifier under investigation. By summing over all less likely cases, the hypergeometric gives the exact probability of drawing r genes with a particular identifier from a sample of size k from a population of size n given that the identifier exists in fraction p in the population set of genes.

Because GeneMerge assesses over-representation for all categories within a given study set of genes, a correction is necessary to account for over-representation that will invariably occur by chance when multiple tests are carried out. A strict and very conservative correction for multiple tests is the Bonferroni correction (Sokal and Rohlf, 1995). Here we use a modified Bonferroni correction based on the number of terms examined in each analysis.

Not all terms associated with genes are scored since they may represent 'singletons' in either the study set or the population set. For instance the term 'saccharopine dehydrogenase', is associated with only one gene in the *Saccharomyces* genome (population set). Likewise, a particular molecular function may be associated with only one gene among those up-regulated in a particular experiment (study set). In such cases, over-representation of the particular function or category is not possible and over-representation scores are not calculated.

In the case of population set singletons, over-representation is a logical impossibility and scoring is ruled out before the analysis begins; thus these terms are not applied to the Bonferroni correction. However, even though scores for terms that appear only once in the *study set* of genes are not calculated, we apply them conservatively to the Bonferroni correction term since their ex post facto exclusion is not blind.

Uncorrected and corrected scores are called raw e-scores (e_r) and e-scores (e_s) respectively, as a reminder that, in some cases, these values will not reflect true P-values. In particular, a gene may belong to multiple categories simultaneously, for example, in the case of genetic pathway data, a gene may often function in several biochemical cascades. In such cases, e-scores will not correspond to P-values. However, for one-to-one gene association data, which make up a majority of association data (chromosomal location data, deletion viability data, etc.), e-scores correspond explicitly to P-values and may treated as such.

ACKNOWLEDGEMENTS

Due to space constraints, our many thanks are found in the Supplementary information.

REFERENCES

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.

Sokal,R.R. and Rohlf,F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, Third edition, Freeman, New York.