The use of microarray chips in gene expression analysis requires
There are many FS methods which can be used, but how do

AGRA: analysis of gene ranking algorithms

Simon Kocbek1,*, Rune Sætre2,3, Gregor Stiglic1, Jin-Dong Kim2, Igor Pernek1, Yoshimasa Tsuruoka4, Peter Kokol1, Sophia Ananiadou5,6 and Jun’ichi Tsujii2,6,4

1Faculty of Health Sciences, University of Maribor, Maribor, Slovenia, 2Department of Computer Science, University of Tokyo, Tokyo, Japan, 3School of Computer Science, University of Manchester, Manchester, Manchester, UK, 4School of Information Science, JAIST, Nomi, Japan, 5National Centre for Text Mining (NaCTeM) and 6Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway

ABSTRACT

Summary: Often, the most informative genes have to be selected
from different gene sets and several computer gene ranking
algorithms have been developed to cope with the problem. To
help researchers decide which algorithm to use, we developed the
analysis of gene ranking algorithms (AGRA) system that offers a
novel technique for comparing ranked lists of genes. The most
important feature of AGRA is that no previous knowledge of gene
ranking algorithms is needed for their comparison. Using the text
mining system finding-associated concepts with text analysis, AGRA
defines what we call biomedical concept space (BCS) for each gene
list and offers a comparison of the gene lists in six different BCS
categories. The uploaded gene lists can be compared using two
different methods. In the first method, the overlap between each pair
of two gene lists of BCSs is calculated. The second method offers a
text field where a specific biomedical concept can be entered. AGRA
searches for this concept in each gene lists’ BCS, highlights the rank
of the concept and offers a visual representation of concepts ranked
above and below it.

Availability and Implementation: Available at http://agra.fzv.uni-
mb.si/; implemented in Java and running on the Glassfish server.

Contact: simon.kocbek@uni-mb.si

Received on December 14, 2010; revised on January 26, 2011;
accepted on February 14, 2011

1 INTRODUCTION

DNA microarray is a technology that can simultaneously measure
the expression levels of thousands of genes in a single experiment.
The use of microarray chips in gene expression analysis requires
an enormous amount of data to be analysed and often, while at the
same time, selecting the most informative genes from different gene
sets.

One of the possible ways to rank the genes is to use a feature
selection (FS) method. FS is a machine learning-based technique
used to select the most important features for building a robust
learning model. The same FS techniques are now widely used in
bioinformatics for identification of biomarkers or lists of relevant
genes from DNA microarray-based gene expression measurements.
There are many FS methods which can be used, but how do
researches know which one is the best? Several different methods
were proposed to estimate the ‘goodness’ of the ranked gene lists
(Ma, 2006; Qu et al., 2006). However, these methods usually
need computer experts who know how FS methods and learning
algorithms work. We describe a novel system called analysis of gene
ranking algorithms (AGRA), which allows biologists and other
experts with low or no previous computer knowledge to compare
different FS methods with the help of evidence mined from PubMed.
AGRA uses finding-associated concepts with text analysis (FACTA),
an online text search engine for MEDLINE abstracts that can quickly
compute the association strengths between a query and different
types of biomedical concepts based on their textual co-occurrence
statistics (Tsuruoka et al., 2008). While other similar systems exist,
such as XplorerMed (Perez-Iratxeta et al., 2002), MedlineR (Lin et al.,
2004), LitMiner (Maier et al., 2005) and Anis (Jelier et al., 2008),
FACTA was chosen because of its ability to pre-index words and
concepts, which result in fast, real-time responses of the system.
AGRA needs to process high amount of data, and fast response of
the underlying service is crucial for the efficient delivery of the
results.

AGRA extracts biomedical concepts using FACTA and thus
defines a biomedical concept space (BCS) for each gene list. BCS
is defined as six categories (gene/protein, disease, symptom, drug,
enzyme and chemical compound) of ranked biomedical concepts.
To compare the quality of different FS methods, AGRA calculates
the overlap for each pair of two gene list of BCSs. This way, gene
lists which are the product of different gene ranking algorithms
can be compared with a gold standard list. Finally, experts can use
their domain knowledge to search for a specific biomedical concept in
the ranked gene lists and decide which FS method outputs the most
relevant genes.

2 METHODS

Figure 1 shows AGRA’s main interface with uploaded gene lists. The
application offers a novel way to compare ranked lists of genes with the
help of BCS. BCS is a set of ranked biomedical concepts gathered through
FACTA where they are grouped into six different categories. FACTA can be
queried by inputting a word (e.g. P53), a concept ID (e.g. UNIPROT: P04637)
or a combination of these [UNIPROT: P04637 AND (lung OR gastric)].
AGRA calculates BCS for a single gene list in three steps: (i) calculation
of protein BCS; (ii) calculation of gene symbol BCS; and (iii) calculation of
gene list BCS.

To achieve this, each gene symbol from the gene lists is associated
with its protein(s) and their Uniprot IDs are extracted with help of the
Affymetrix annotation file (HG-U133 Plus 2 Annotations, Release 31).
AGRA then queries FACTA with these Uniprot identifiers and maximum number of biomedical concepts. Overlap is computed by summarizing values from all gene symbol BCS categories of each associated protein. The overlap for single concept xi is calculated as:

\[
w_x(i) = \frac{(n-i)}{n}
\]

where \(n\) is number of all its concepts and \(i\) represents the rank of the gene that concept belongs to in the gene list (starting from 1). When BCSs for all gene lists are extracted, AGRA calculates the overlap values for every combination of two BCSs to evaluate the effectiveness of FS methods. Overlap is a simple method to measure similarity between two BCSs where biomedical concepts that appear in both BCSs are counted and divided by the number of concepts in the shorter BCS. Another way to compare FS methods is to search for the position of relevant biomedical concepts in the final gene list BCS. This is done by summarizing values from all gene symbol BCS categories of each associated protein. The usage of AGRA is simple and only basic computer skills are required. The application consists of three different tabs: main, overlap and position. The main tab is used for uploading the gene lists and starting the analysis. The user should upload the lists in a CSV file where the first row represents gene list names and other rows represent ranked genes with the most important gene on the top and the least important gene on the bottom of the list. When the file is uploaded, the ranked genes for each list are displayed in a table next to each other so they can be visually compared. Then the user can enter a specific concept (e.g. ‘breast cancer’) and select in which BCS category AGRA should look for the concept. The system can be started with the start button which is disabled during the analysis. When finished, the results can be accessed through the overlap and position tabs. The overlap tab offers a visual analysis of overlap values for each pair of uploaded gene lists. Six tables represent six different categories. The first column and the first row of each table contain gene list names and each cell contains an overlap value between two corresponding lists. The value is coloured according to the overlap success rate where dark red colour indicates the lowest and light green indicates the highest overlap. The position tab offers an analysis of the position of the searched concept in each gene list’s BCS. With the help of a chart and a table, the user can inspect which concepts were found by AGRA for each gene list and how they were ranked. The position of the searched concept is marked.

**ACKNOWLEDGEMENTS**

We would like to thank Paul Thompson of NaCTeM for comments on the article.

**Funding**: Slovenian Research Agency (BI-JP09-11-002); Japan Society for the Promotion of Science Bilateral Joint Project; JISC-funded national centre for text mining.

Conflict of Interest: none declared.

**REFERENCES**


