BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 8 2011, pages 1185–1186
doi:10.1093/bioinformatics/btr097

Data and text mining
Advance Access publication February 23, 2011

AGRA: analysis of gene ranking algorithms

Simon Kocbek[1,*], Rune Sætre[2,3], Gregor Stiglic[1], Jin-Dong Kim[2], Igor Pernek[1], Yoshimasa Tsuruoka[4], Peter Kokol[1], Sophia Ananiadou[5,6] and Jun'ichi Tsujii[2,6,*]

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

Associate Editor: Jonathan Wren

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; Qiu *et al.*, 2006). However, these methods usually need computer experts who know how FS methods and learning algorithms work. We describe a novel system, 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 XplorMed (Perez-Iratxeta *et al.*, 2002), MedlineR (Lin *et al.*, 2004), LitMiner (Maier *et al.*, 2005) and Anii (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

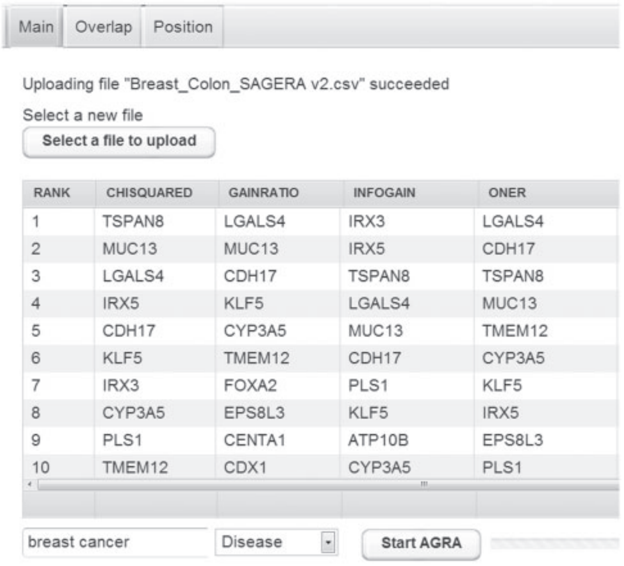
Figure 1 shows AGRA's main interface with uploaded gene lists. The application offers a novel way to compare ranked lists of genes with the help of BCS. BCS is a set of ranked biomedical concepts gathered through FACTA where they are grouped into six different categories. FACTA can be queried by inputting 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).

[*]To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.



Uploading file "Breast_Colon_SAGERA v2.csv" succeeded

Select a new file

RANK	CHISQUARED	GAINRATIO	INFOGAIN	ONER
1	TSPAN8	LGALS4	IRX3	LGALS4
2	MUC13	MUC13	IRX5	CDH17
3	LGALS4	CDH17	TSPAN8	TSPAN8
4	IRX5	KLF5	LGALS4	MUC13
5	CDH17	CYP3A5	MUC13	TMEM12
6	KLF5	TMEM12	CDH17	CYP3A5
7	IRX3	FOXA2	PLS1	KLF5
8	CYP3A5	EPS8L3	KLF5	IRX5
9	PLS1	CENTA1	ATP10B	EPS8L3
10	TMEM12	CDX1	CYP3A5	PLS1

breast cancer

Disease

Fig. 1. AGRA's main interface.

AGRA then queries FACTA with these Uniprot identifiers and maximum 50 most important biomedical concepts (ranked by their frequencies of appearing in the MEDLINE abstracts) from each category are extracted. Concepts that are gathered in this step represent six BCS categories of each associated protein.

Next, BCS categories for the gene symbol are calculated. If the gene symbol is associated with only one protein, its BCS is identical to the protein's one. When the symbol is associated with more than one protein, the average values of the frequencies in each category are calculated.

In the final step, the six categories for each gene list BCS are calculated. This is done by summarizing values from all gene symbol BCS categories from the list. Because the order of the gene symbols in the list is crucial, AGRA weights each gene symbol BCS according to the gene symbol position in the list. The weight w for single symbol x_i is defined as $w(x_i)=(n-(i-1))/n,$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).

Finally, to avoid sending queries to the FACTA system too often, AGRA saves BCSs in a local database. Whenever a gene symbol, for which BCS has not been defined yet, appears in one of the gene lists, the system queries FACTA, calculates its BCS and saves it locally.

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 BCS 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. Position of a single biomedical concept is defined as it is ranked number among all the concepts in one of the categories. This way, researchers can decide which FS method selects the most important concepts and ranks them higher compared with other methods.

3 USAGE OF THE APPLICATION

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. Due to the calculation complexity and limitation of the FACTA+ system, the input file should contain maximum 7 different gene lists with maximum 100 genes in each 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.

4 LIMITATIONS

In future work, we will address a number of AGRA's current limitations. Currently, FACTA uses its internal dictionary for associating proteins with their UniProt IDs, thus not every gene is associated with all of its proteins. Newer versions of FACTA will address this issue. Furthermore, some of the biomedical concepts found by the system indicate the same term (e.g. 'cancer' and 'neoplasm') but they can be ranked in different ways which can affect the quality of the final results.

ACKNOWLEDGEMENTS

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

Funding: Slovenian Research Agency (BI-JP/09-11-002); Japan Society for the Promotion of Science Bilateral Joint Project; JISC-funded national centre for text mining.

Conflict of Interest: none declared.

REFERENCES

Jelier,R. *et al.* (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.*, **9**, R96.

Lin,S.M. *et al.* (2004) MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics*, **20**, 3659–3661.

Ma,S. (2006) Empirical study of supervised gene screening. *BMC Bioinformatics*, **7**, 537–550.

Maier,H. *et al.* (2005) LitMiner and WikiGene: identifying problem-related key playersof gene regulation using publication abstracts. *Nucleic Acids Res.*, **33**, W779–W782.

Perez-Iratxeta,C. *et al.* (2002) Exploring MEDLINE abstracts with XplorMed. *Drugs Today*, **38**, 381–389.

Qiu,X. *et al.* (2006) Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, **7**, 29–34.

Tsuruoka,Y. *et al.* (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, **24**, 2559–2560.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016