Artificial Intelligence in Medicine (2005) **35**, 49—60

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

SiMCAL 1 algorithm for analysis of gene expression data related to the phosphatidylserine receptor

Daniel Dvorkin [a,b,*], Valerie Fadok [c], Krzysztof Cios [a,d,e]

[a] *University of Colorado at Denver and Health Sciences Center, Denver, CO 80217, USA*
[b] *University of Minnesota School of Public Health, Minneapolis, MN 55455, USA*
[c] *National Jewish Medical and Research Center, Denver, CO 80206, USA*
[d] *University of Colarado at Boulder, CO 80309, USA*
[e] *4C Data LLC, Golden, CO 80401, USA*

Received 11 October 2004; received in revised form 9 January 2005; accepted 12 January 2005

KEYWORDS
Clustering;
Jarris-Patrick;
Inflammation;
Microarray;
Nearest neighbor;
Nonparametric;
Phosphatidylserine;
PSR;
SiMCAL;
TGF-β

Summary

Objective: SiMCAL 1 (simple multilevel clustering and linking, version 1) is a novel clustering algorithm for time-series microarray data, presented here with an application to a specific data set. The purpose of the algorithm is to present a complete feature set not found in either Jarvis-Patrick clustering, from which it is derived, or in other popular clustering methods such as hierarchical and *k*-means. The data concern the activity of the phosphatidylserine receptor (PSR) which is believed to be a crucial molecular switch in the mediation of inflammatory response in apoptosis and lysis. By analyzing the behavior of PSR-related genes in mouse macrophages, we hope to elucidate the mechanisms involved in this important biological process.
Methods and materials: SiMCAL 1 is implemented in the Python programming language using the Numerical Python extensions, and the data are stored using the MySQL database management system. The data are derived from exposures of multiple Affymetrix mouse gene microarray chips to elevated levels of PSR antibody and control conditions. Code and data are available at http://www.dvorkin.com/daniel/Simcal1.zip (accessed: 17 January 2005).
Results: The algorithm meets its objectives: it is *simple*, in that it is computationally inexpensive; it is *multilevel*, in that it provides a small number of clearly defined hierarchical levels of clusters; and it offers *linking* between clusters at the same level in each hierarchy. Clustering and linking results indicate previously unknown co-regulation for genes expressing PGH synthase (COX2) and PGE2, appear to confirm increased production of proteins for clearance of apoptotic cells in the presence of PSR antibody, and correspond to other findings regarding the temporal relationship between PGE2 production and B cell proliferation and differentiation. These results are promising but should be taken as highly preliminary.

* Corresponding author at: 1138 E. 14th Ave. Unit 8, Denver, CO 80218, USA. Tel.: +1 303 548 4568; fax: +1 303 607 9430.
E-mail address: daniel.dvorkin@gmail.com (D. Dvorkin).

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2005.01.010

Conclusion: Both the algorithm and its application to this problem show great potential for future development. We plan to improve and extend the SiMCAL family of algorithms, and to obtain new data so that the algorithm(s) may be further applied to this and other problems of interest.

1. Introduction

The introduction is divided into biological and computational background sections. Although the computational aspect depends on the biological aspect, each may be considered in some degree of isolation: the biology is an area of ongoing ''wet-lab'' research as well as computational analysis, while the computation is based on data mining techniques with a wide variety of applications, and which are not specific to the problem at hand.

1.1. Biological background

Normal cell death, or *apoptosis*, requires regular housekeeping activity within all multicellular organisms. In vertebrates, this activity is performed by macrophages, which must be able to distinguish between three classes of cells: healthy cells, which the macrophages must not consume; apoptotic cells, which must be consumed without triggering an inflammatory response; and cells undergoing *lysis* (death from some external cause such as injury or disease) that also must be consumed, but which should trigger an inflammatory response [1].

Since cells are constantly undergoing apoptosis throughout the bodies of all multicellular organisms, the mechanism for preventing inflammatory response must work perfectly or disease will result. In humans, serious diseases such as cystic fibrosis, lupus, and some forms of diabetes may be at least partly the result of a breakdown of this mechanism [2].

Phosphatidylserine(PS) is a phospholipid generally found on the cytosolic surface of the cell membrane in most animal cells. Apoptotic cells appear to shift PS rapidly to the exterior of the membrane, at least partly as a signal for phagocytosis. However, it is far from the only signal for this process. What makes PS interesting is the effect PS has on the macrophage. Henson et al. state that the phosphatidylserine receptor (PSR) is ''a crucial molecular switch'' in the control of immune response [3]. Specifically, when PS is expressed in apoptotic cells, it appears to reduce inflammation and other aspects of the immune response by engaging the PSR in macrophages. Cells undergoing lysis do not show this behavior, allowing the full range of immune response, including inflammation, to come into play.

1.2. Computational background

The overall purpose of the project is to analyze microarray data on genes which might be relevant to PSR activity. Below we discuss both the nature of the data and background information on the algorithms used in analyzing the data. Details on the specific methods used in the analysis are found in Section 2.

1.2.1. Data

To ensure that the data were accurate and replicable, three separate and identical experiments (replicants) were performed. In each, microarrays containing 12,488 mouse genes known to be active in mouse macrophages, with each gene being represented by between 5 and 30 RNA probes, were exposed to a single concentration of monoclonal PSR antibody over time. Fluorescent hybridization measurements were taken for RNA at 30 min, 2 h, and 8 h 30 min.

Identical measurements were taken of control samples which were treated with an isotype, i.e., another IgM antibody with better-known effects. Another control sample, of which only one measurement was taken, was not treated. Therefore, seven distinct data points were generated for each probe. These steps were performed using Affymetrix (''Affy'') microarray equipment and the Affy chip MG-U74Av2, a standard chip containing mouse genes.

1.2.2. SiMCAL versus existing methods

The origin of SiMCAL 1 lies in the Jarvis-Patrick (JP) algorithm [4], which has been widely used in fields ranging from astronomy to chemoinformatics, but has so far not been common in bioinformatics applications. JP begins with two user-defined parameters: k and $k_{\min}$. For each data element, a list of that element's k nearest neighbors is calculated. Then the algorithm states that two data elements are in the same cluster if either of the following conditions applies:

- They are within each other's lists of nearest neighbors.
- They have at least $k_{\min}$ nearest neighbors in common.

The advantages of JP are as follows. First, because it allows chains of relationships between