UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Using Machine Learning Algorithms to Detect Cellular Stress of *Listeria monocytogenes* from cDNA Microarray Data

Xiaoji Liu[1], Urmila Basu[1], Petr Miller[1], Nasimeh Asgarian[2], Russell Greiner[2], Lynn M. McMullen[1]

University of Alberta, Department of Agricultural, Food and Nutritional Science, Edmonton, AB, Canada[1], University of Alberta, Department of Computing Science, Edmonton, AB, Canada[2]

Introduction

Listeria monocytogenes is a serious foodborne pathogen that has the ability to form filaments under certain environmental stress such as the presence of antimicrobials. Filament formation is the phenotypical sign of antimicrobial stress of *L. monocytogenes*.

Microarrays are useful tools for measuring gene expression of *L. monocytogenes*, and can be used to determine if a cell population undergoes antimicrobial stress.

Machine learning (ML) algorithms can use a dataset derived from microarrays to learn a classifier that can later identify if a novel cell population is involved in a proposed biological process. While these algorithms [including Bayesian Net, J48 Decision Tree, Random Forest and Support Vector Machine (SVM)] are often used to classify eukaryote microarray experiments, this study focuses on a prokaryotic application using two strains of *L. monocytogenes* as examples.

Objectives

To explore if a machine learning algorithm can learn a classifier that can predict if a population of *L. monocytogenes* is under stress from an antimicrobial:

- to distinguish between cefuroxime treated and untreated *L. monocytogenes* EGE-e, based on the expression level (represented as fluorescence intensity) for each gene from 32 samples [GEO accession GPL14687 (4)];
- to distinguish between *L. monocytogenes* 08-5923 treated with carnocyclin A (cclA) and untreated *L. monocytogenes* 08-5923, based on expression level of 15 selected genes that were ≥ 2-fold up or down-regulated in the presence of cclA. Features were selected using in-fold feature selection (2).

Materials and Methods

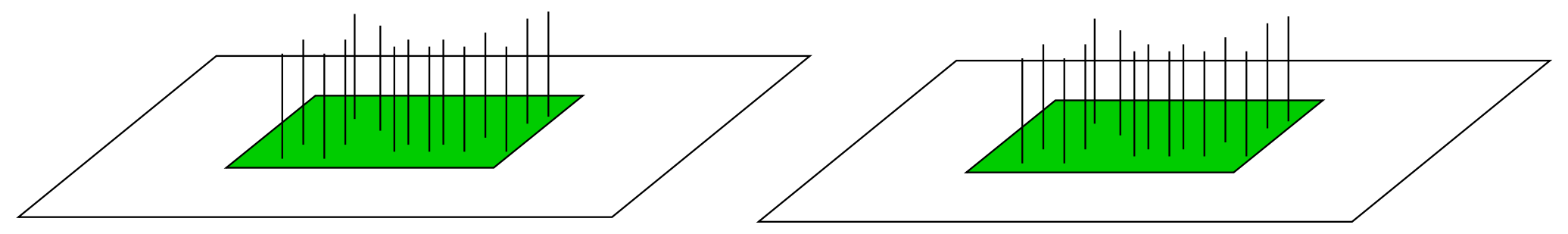
Treat *L. monocytogenes* with cclA

RNA isolation and integrity verification

RNA integrity number >7

Cy3 labelling

Microarray

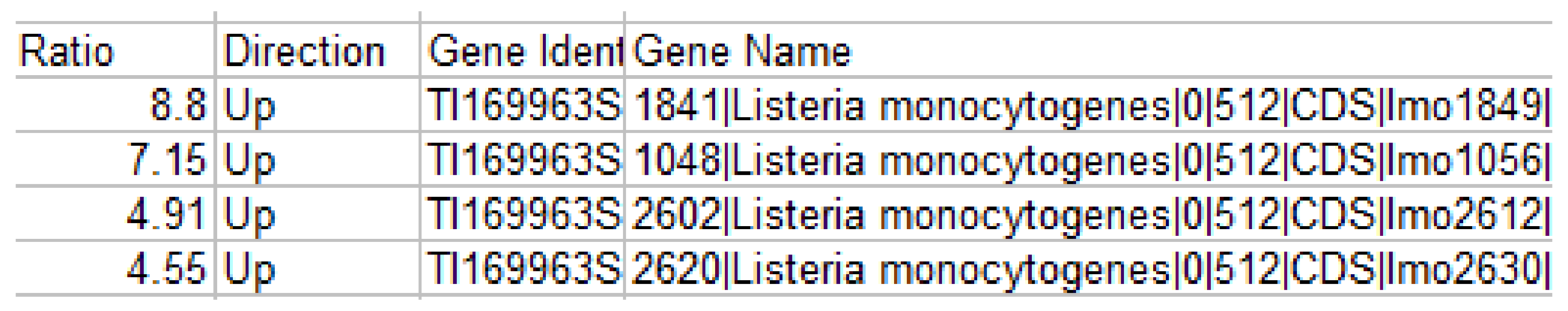


Untreated

Treated

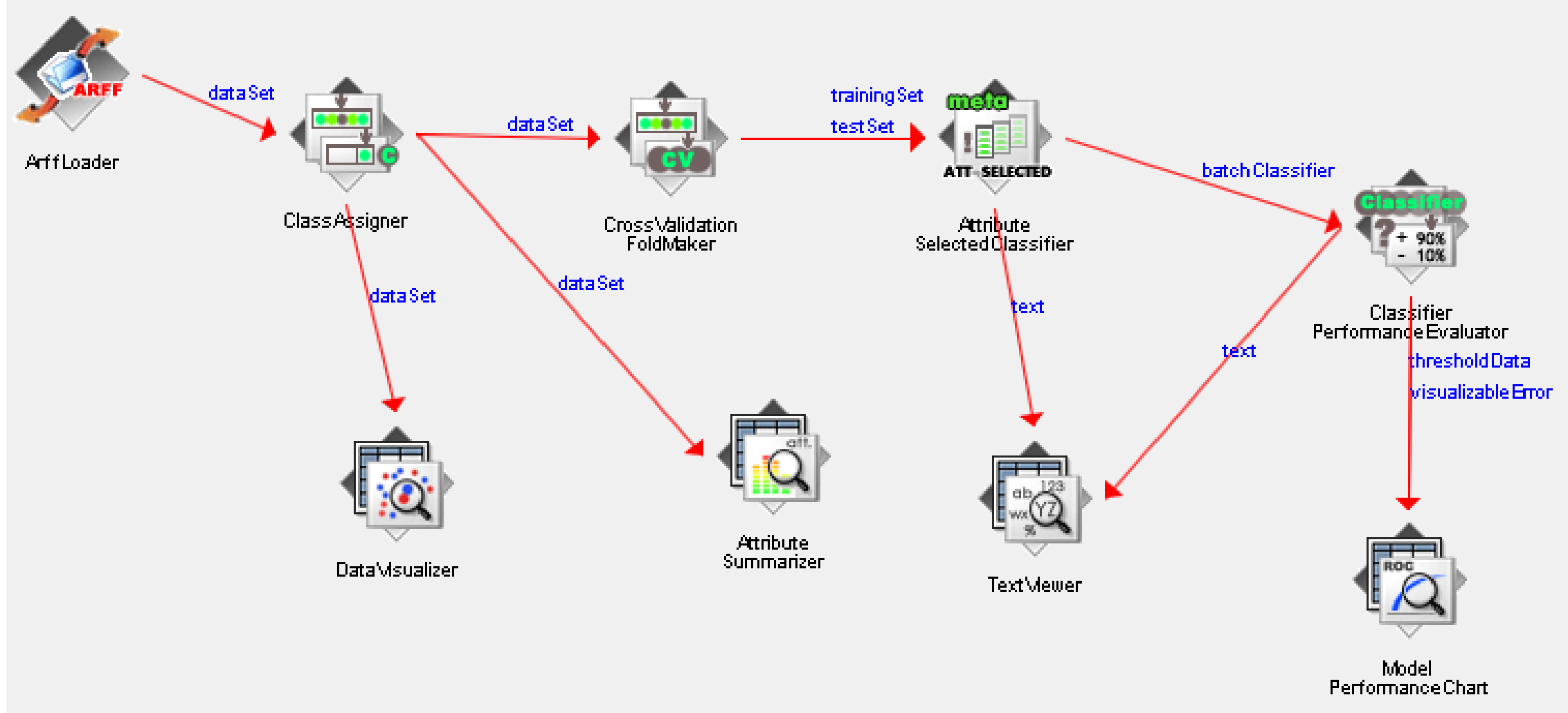
Gene selection

(GeneSifter®)



Ratio	Direction	Gene Ident	Gene Name
8.8	Up	TI169963S	1841|Listeria monocytogenes|0|512|CDS|lmo1849|
7.15	Up	TI169963S	1048|Listeria monocytogenes|0|512|CDS|lmo1056|
4.91	Up	TI169963S	2602|Listeria monocytogenes|0|512|CDS|lmo2612|
4.55	Up	TI169963S	2620|Listeria monocytogenes|0|512|CDS|lmo2630|

Classify based on workflow shown below [WEKA (3)]



Results

Expression levels of genes relevant to cell morphology and death

Table 1: Genes ≥ 2-fold up or downregulated in *L. monocytogenes* 08-5923 when exposed to cclA. The genes from this table, as well as other relevant genes involved in cell division and PTS system (1, 5) such as *lmo2002*, *lmo1973*, *lmo0633*, *lmo1438* and *lmo1892*, were included in the dataset for the subsequent classifying task.

Gene	Function of gene product	Fold change	Differential expression
Cell division protein			
lmo2687	FtsW	2.39	Up
lmo2033	FtsA	2.17	Up
Phosphotransferase (PTS) system			
lmo0096	Mannose-specific	3.60	Up
lmo1035	Beta-glucoside-specific	2.45	Up
lmo1971	Pentitol-specific	2.29	Down
lmo2782	Cellobiose-specific	2.23	Down
lmo0023	Fructose-specific	2.14	Down
lmo2097	Galactitol-specific	2.03	Down
lmo0503	Fructose-specific	2.01	Down

Performance of ML algorithms

Table 2: the accuracy of various algorithms in predicting if a population of *L. monocytogenes* was under stress.

Task	Algorithm	Test mode	Accuracy	Accuracy (in fold cross validation)
CclA-stress	J48	5-fold cross-validation	90%	90%
	Bayes Network		90%	90%
	Random Forest		50%	90%
	SMO		70%	60%
	Naiive Bayes		20%	90%
Cefuroxime-stress	J48	10-fold cross-validation	90.63%	N/A
	J48	32-fold cross-validation	96.88%	N/A

Conclusions

- J48 Decision Tree was the most accurate algorithm for predicting cefuroxime stress (96.9% accuracy with leave-one-out cross validation)
- Both the J48 Decision Tree and Bayesian Network were equally effective for predicting whether *L. monocytogenes* was under stress from carnocyclin A (90.0% accuracy with 5-fold cross validation)
- Bayesian Nets and J48 Decision Tree could be applied to detect the presence of cellular stress in prokaryotes using data from DNA microarrays

Future Work

- Use J48 and Bayes Networks with in fold cross validation to analyze microarray data from the cefuroxime-stress study
- Examine the consistency of the performance of these algorithms in all the biological replicates of the microarray experiments
- Test the performance of the algorithms with various datasets containing expression values of genes from different signalling pathways

Acknowledgements

This project was supported by funding from the Alberta Livestock and Meat Agency.

References

1. Bergholz TM, B Bowen, M Wiedmann and KJ Boor (2012) Appl. Environ. Microbiol. 78:2602-2612.
2. Dupuy A and RM Simon (2007) J. Natl. Cancer Inst. 99:147 – 57.
3. Hall M, E Frank, G Holmes, B Pfahringer, P Reutemann and IH Witten (2009) The WEKA Data Mining Software: An Update. 11(1).
4. Nielsen PK, AZ Andersen, M Mols, S van der Veen,T Abee and BH Kallipolitis (2012) Microbiol. 158(4):963-974.
5. Stasiewicz MJ, M Wiedmann, TM Bergholz (2011) Appl. Environ. Microbiol. 77:5294–5306.