

Using Decision Forest to Classify Prostate
Cancer Samples Based on SELDI-TOF MS
Data – Assessing Chance Correlation and
Prediction Confidence

Weida Tong, Qian Xie, Huixiao Hong, Hong Fang,
Leming Shi, Roger Perkins, Emanuel F. Petricoin
doi:10.1289/txg.7109 (available at <http://dx.doi.org/>)
Online 5 August 2004



The National Institute of Environmental Health Sciences
National Institutes of Health
U.S. Department of Health and Human Services

**Using Decision Forest to Classify Prostate Cancer Samples Based on
SELDI-TOF MS Data – Assessing Chance Correlation and Prediction
Confidence**

Weida Tong^{1*}, Qian Xie², Huixiao Hong², Hong Fang², Leming Shi¹, Roger Perkins²,
Emanuel F. Petricoin³

¹Center for Toxicoinformatics, Division of Biometry and Risk Assessment, National
Center for Toxicological Research (NCTR), FDA, Jefferson, Arkansas 72079

²Bioinformatics Group, NCTR, Jefferson, Arkansas 72079

³NCI-FDA Clinical Proteomics Program, Center for Biologics Evaluation and Research,
FDA, Bethesda, Maryland 20892

*Address correspondence to Dr. Weida Tong, Center for Toxicoinformatics, Division of
Biometry and Risk Assessment, National Center for Toxicological Research (NCTR),
3900 NCTR Rd., HFT020, Jefferson, Arkansas 72079. Telephone: (870) 543-7142. Fax:
(870) 543-7662. E-mail: wtong@nctr.fda.gov

Running Title: Decision Forest for prediction of prostate cancer

Key words: Decision Forest, Class prediction, classification, Bioinformatics, SELDI-TOF, Prediction confidence, Chance correlation, Prostate cancer, Proteomics

Abbreviations:

DF – Decision Forest

DT – Decision Tree

SELDI – Surface Enhanced Laser Deposition/Ionization

TOF – Time-of-flight mass spectrometry

MS – Mass spectrometry

LOO – Leave-one-out cross-validation

L100 – Leave-10-out cross-validation

LNO – Leave-*N*-out cross-validation

SVMs – Support Vector Machines

PCA – Prostate cancer

BPH – Benign prostatic hyperplasia

Outline of section headers:

- Abstract
- Introduction
- Decision Forest
 - General consideration
 - Model development
 - Randomization test for chance correlation
 - Model validation
- Results
 - Assessment of chance correlation
 - Assessment of prediction confidence
 - Comparison of Decision Forest with Decision Tree
 - Biomarker Identification
- Discussion

Abstract

Class prediction using -omics data is playing an increasing role in toxicogenomics, diagnosis/prognosis and risk assessment. Omics data are usually noisy and represented by relatively few samples and a very large number of predictor variables (e.g., genes of DNA microarray data or m/z peaks of mass spectrometry data). These characteristics manifest the importance of assessing potential random correlation and overfitting of noise for a classification model based on -omics data. We present a novel classification method, Decision Forest (DF), for class prediction using -omics data. DF combines the results of multiple heterogeneous but comparable Decision Tree (DT) models to produce a consensus prediction. The method is less prone to overfitting of noise and chance correlation. A DF model was developed to predict presence of prostate cancer using a proteomic dataset generated from surface enhanced laser deposition/ionization time-of-flight mass spectrometry (SELDI-TOF MS). The degree of chance correlation and prediction confidence of the model was rigorously assessed by using an extensive cross-validation and randomization testing. Comparison of model prediction with imposed random correlation demonstrated biological relevance of the model and the reduction of overfitting in DF. Furthermore, two confidence levels (high and low confidences)) were assigned to each prediction, where the majority of misclassifications were associated with the low confidence region. For the high confidence prediction, the model achieved 99.2% sensitivity and 98.2% specificity. The model also identified a list of significant peaks that could be useful for biomarker identification. DF should be equally applicable to other -omics data, such as gene expression data or metabonomic data. The DF algorithm is available upon request.