EPA Analysis for MAQC Toxicogenomics Datasets

Fathi Elloumi(1), Zhen Li(2), Richard Judson(1)
(1)NCCT/ORD, USEPA, RTP NC, USA, (2) Dept. Biostatistics, UNC, Chapel Hill, USA

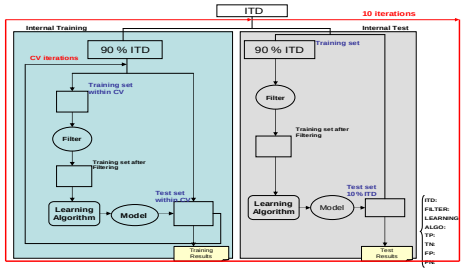
Iconix Data Analysis - Multi-factorial Approach

24 input training data sets (ITD) generated from the raw data using different methods for

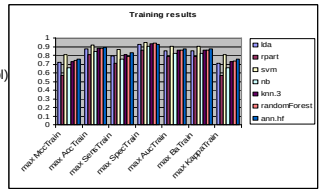
- **Probe selection:** None, DISCOVERY
- **Normalization:** Quantile, Invariant set, Qspline, Cyclic Lowess
- **Differentially Expressed Genes (DEG) for at least one chemical:** Samr parametric t-test , Samr parametric Wilcoxon test, Limma Bayes test
- ratio value (per animal, feature) = norm. signal-treated-avg (norm. signal-control)

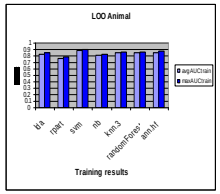
63 learning methods that took in account the following methods:

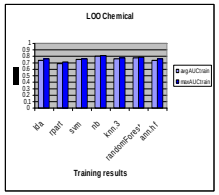
- **Learners:** LDA, RPART, SVM, Naïve Bayes, KNN(k=3), Random Forest, ANN
- **Feature selection (best 50 features):** T-test, Wilcoxon, Bayes
- **Cross validation mode:** LOO animal, 10F-CV animals, LOO chemical

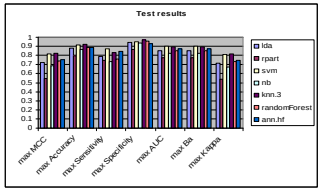


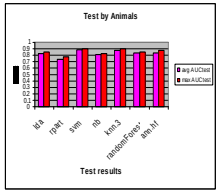
Max values for all Classifiers

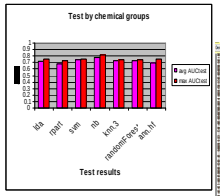








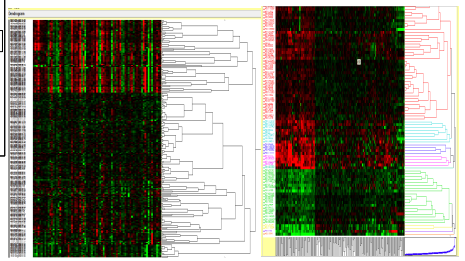




Iconix Best Model

Learner:	SVM	
Probe selection:	None	
Normalization:	Cyclic Lowess	
DEG:	Samr parametric t-test	
Feature selection:	Wilcoxon	
Scores:		
Animals		
AUC LOO:	0.89	+/- 0.019
AUC test:	0.90	+/- 0.052
Animals grouped by chemical		
AUC LOG:	0.75	+/- 0.020
AUC test:	0.74	+/- 0.139

Chemicals clustering



Functional annotation: 81 genes from best model

	#genes	P.value
-BIOLOGICAL_PROCESS		
• regulation of signal transduction	4	6.80E-02
• positive regulation of physiological process	7	5.55E-02
• positive regulation of cellular process	9	7.82E-03
• macromolecule metabolism	19	6.06E-02
• cell death	9	1.89E-03
• establishment of localization	17	3.33E-02
• primary metabolism	30	6.18E-03
• response to oxidative stress	3	3.12E-02
• cellular metabolism	31	5.22E-03
• regulation of cellular physiological process	16	3.12E-02
• response to hypoxia	3	8.88E-03
• transport	14	8.54E-02
• cell organization and biogenesis	11	6.93E-02
-CELLULAR_COMPONENT		
• intracellular organelle	27	1.65E-02
• outer membrane	4	1.16E-03
• intracellular membrane-bound organelle	26	4.63E-03
• organelle outer membrane	4	4.05E-04
• organelle membrane	9	1.28E-03
• organelle envelope	8	7.30E-04
• cytoplasm	22	1.04E-02
• mitochondrial envelope	8	9.13E-05
• nuclear lumen	5	9.58E-02
-MOLECULAR_FUNCTION		
• Transferase activity, transferring acyl grps	4	8.40E-03
• metal ion binding	11	9.40E-02
-KEGG_PATHWAY		
• ADIPOCYTOKINE SIGNALING PATHWAY	3	8.30E-02

Signature prediction

	TRUE	FALSE	
Cluster1	55	11	66
Cluster2	18	132	150
	73	143	216

Hamner Data Analysis

Goal: to predict potential toxicity for the 13 chemicals

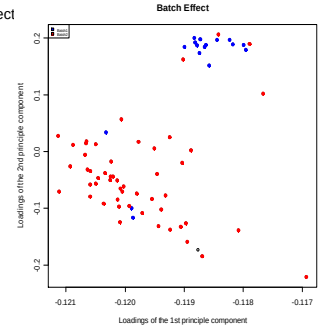
Chemicals (13 in total and single dose):
1,5-Naphthalenediamine, 2,3-benzofuran, 4-Nitroanthranilic,N-(1-naphthyl)ethylenediamine dihydrochloride, benzene, coumarin, pentachloronitrobenzene, 2,2-bis(bromomethyl)-1,3- propanediol, 1,2-dibromoethane, 2-chloromethylpyridine hydrochloride, N-methylolacrylamide, diazinon and malathion.

Controls: corn oil, water, rodent chow

Endpoints:increase in lung tumor incidence; 7 positive and 6 negative.

Replicates: 3 or 4

Other issue: Batch Effect



Hamner Approach

Predict chemical toxicity

Subsequently predict the endpoint of each mouse.

Steps:

1. RMA normalization was done on each chemical vs. its corresponding controls.
2. SAM penalized method was used to filter genes
3. Among the genes declared significantly differentially expressed, identify those that appear for at least k chemicals.
4. Create a binary data-set: gene is / is not significantly differentially expressed
5. Test multiple classifiers
6. Use k-fold cross validation for each classifier (k=6)

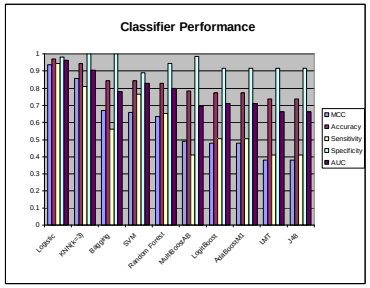
Advantages

- Reduce batch effect by summarizing the data in the least biased way by using binary data derived by penalized methods.
- Use of binary derived data reduces the dependence on the specific method of significance testing and array platform, and therefore, serves as a powerful tool for data integration.
- The binary structure simplifies the process of feature selection through combination of data into a signature gene index.

Predictive Affymetrix Probesets

Affymetrix probeset	Gene Symbol
1416416_x_at	Gstm1
1423627_at	Nqo1
1424266_s_at	AU018778
1424783_a_at	Ugt1a9
1425099_a_at	Arntl
1426261_s_at	Ugt1a6a
1434735_at	Hlf
1439332_at	Ddit4l
1448330_at	Gstm1
1449279_at	Gpx2
1460242_at	Cd55
1426260_a_at	Ugt1a6a
1422438_at	Ephx1
1449486_at	Ces1

14 predictive Affymetrix probesets were identified, which shows up for at least 5 chemicals and represent 12 unique genes. Biological studies have reported that Gstm1, Nqo1 and Ephx1 are associated with lung cancer.



Hamner Top Models

	Logistic	KNN	SVM
MCC	0.94	0.86	0.6578
Accuracy	0.97	0.94	0.8423
Sensitivity	0.95	0.81	0.7669
Specificity	0.98	1	0.8887
AUC	0.96	0.90	0.8278

This poster does has been reviewed by EPA and approved for presentation but does not necessarily reflect EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

United States Environmental Protection Agency

COMPUTATIONAL TOXICOLOGY