NTP
National Toxicology Program

Predicting the hepatocarcinogenic potential of (alkoxy)propenyl benzene derivatives using toxicogenomics

Scott S. Auerbach, Ph.D.

NIEHS/NTP

NTP Board of Scientific Counselors Meeting

November 20-21, 2008

Hypothesis

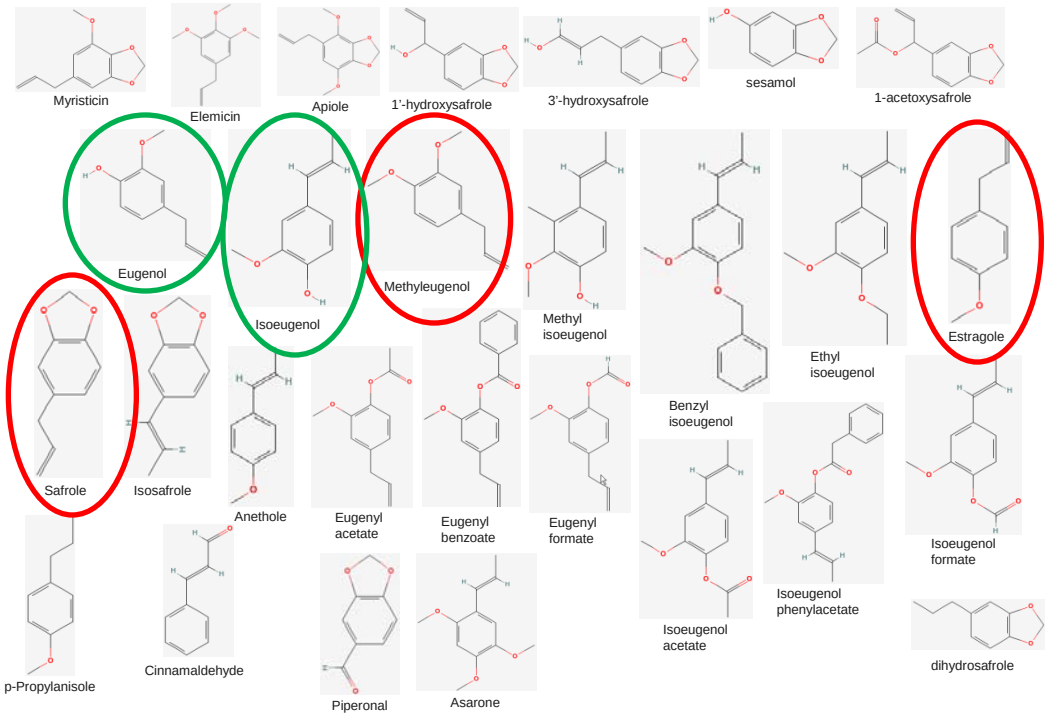
Pattern recognition models trained on hepatic gene expression induced by hepatocarcinogens and non-carcinogens can identify (alkoxy)propenyl benzene derivatives that pose a significant hepatocarcinogenic hazard

Definition of terms

- Supervised machine learning
  - computational methods used to generate pattern recognition models
  - employ prior knowledge about the samples in order to search for genes that correlate with a disease state
- Training data
  - mRNA expression data used to train the pattern recognition models
- Test data
  - mRNA expression data NOT used for training the models which is used to independently evaluate the performance of the models
- Cross-validation
  - classify samples that were used to train the model
- Independent validation
  - classify samples that were NOT used to train the model
- Optimal model
  - a pattern recognition model that achieves 0% (or as close to 0% as possible) cross-validation error with a minimum number of genes

(Alkoxy)propenyl benzene derivatives

- A large class of chemicals that are used in fragrances and/or flavorings agents
- There are naturally occurring and synthetic sources
- Significant fraction are approved for direct addition to food for human consumption
- Limited number have been studied in a carcinogenicity bioassay and some produced increases in hepatic cancer in male rats
- Problem: too many to test
  - Need to prioritize
  - Prioritize based on hepatocarcinogenic potential



Study design

Structurally diverse training data

Male F344 rats dosed for 2, 14 or 90 days	Ames
1 ppm aflatoxin B1	+
5000 ppm 1-amino-2,4-dibromoanthraquinone	?
5 ppm N-nitrosodimethylamine	+
150 mg/kg/day methyleugenol	-
2500 ppm acetaminophen	-
25000 ppm ascorbic acid	-
25000 ppm tryptophan	-
Dose water control	
Dose feed control	
Gavage control (methylcellulose)	

Measure hepatic mRNA levels using Agilent 4X44k microarrays

Use a supervised machine learning method to create and optimize carcinogenicity prediction models based on either a single or a combination of exposure durations. Evaluate models using m-fold cross-validation

Predict (Alkoxy)propenyl benzene test data using the optimized models

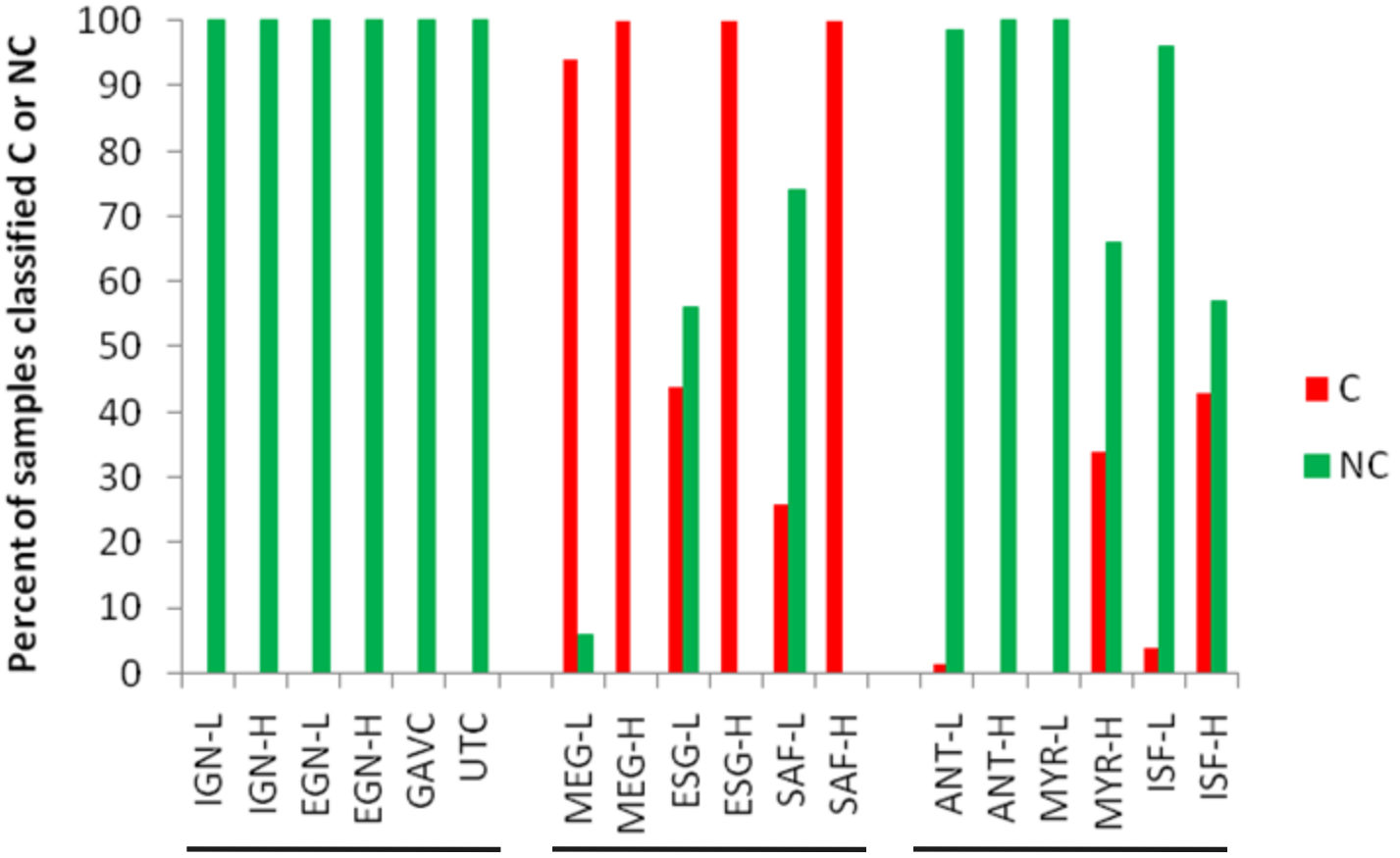
(Alkoxy)propenyl benzene test data

Male F344 rats gavage dosed with 0.2 (L) or 2.0 (H) mmoles/kg/day for 2, 14 or 90 days	Ames
Methyleugenol (MEG)	-
Estragole (ESG)	-
Safrole (SAF)	-
Eugenol (EGN)	-
Isoeugenol (IGN)	-
Gavage control (corn oil) (GAVC)	
Untreated control (UTC)	
Anethole (ANT)	-
Isosafrole (ISF)	-
Myristicin (MYR)	-

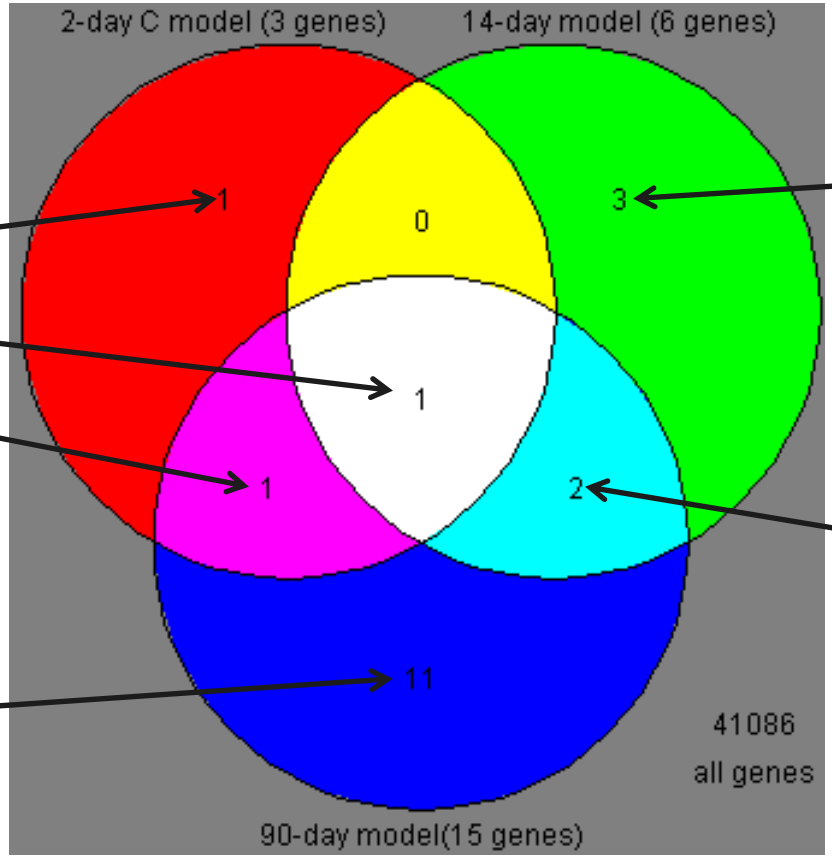
Characteristics of the optimal pattern recognition models

- 7 optimal pattern recognition models were identified using either single or multiple exposure duration training data
  - 2 day, 14 day, 90 day, 2+14 day, 2+90 day, 14+90 day, 2+14+90 day
- All optimal models with the exception of the 2+14 day model achieved 0% error by cross-validation
- The number of features per optimal model ranged from 3 to 59
- Evaluation of the (alkoxy)propenyl benzene derivatives
  - All optimal models classified 90 day test data with the higher accuracy than the 2 or 14 day test data
  - All optimal models classified the 90 day test data with near equal accuracy, therefore we summed the classification results of all the models

Cumulative classification results of the 90-day (alkoxy)propenyl benzene test data



Features (genes) informative to the individual day optimal models



AhR
- Aldh3a1

DNA damage response
- Wwox
- Fhit

Wnt signaling
- Cnsk1a1 (AT)
- C14orf83 (H)
- C8orf46 (H)
- Gbf1(I)

Cell migration
- Unc5d
- Mybl2

Cell cycle
- Sox30
- Abcb1
- Tsx
- Rhbg
- Fundc1

- Cyp26a1
- Dab1(AT)
- CB545387

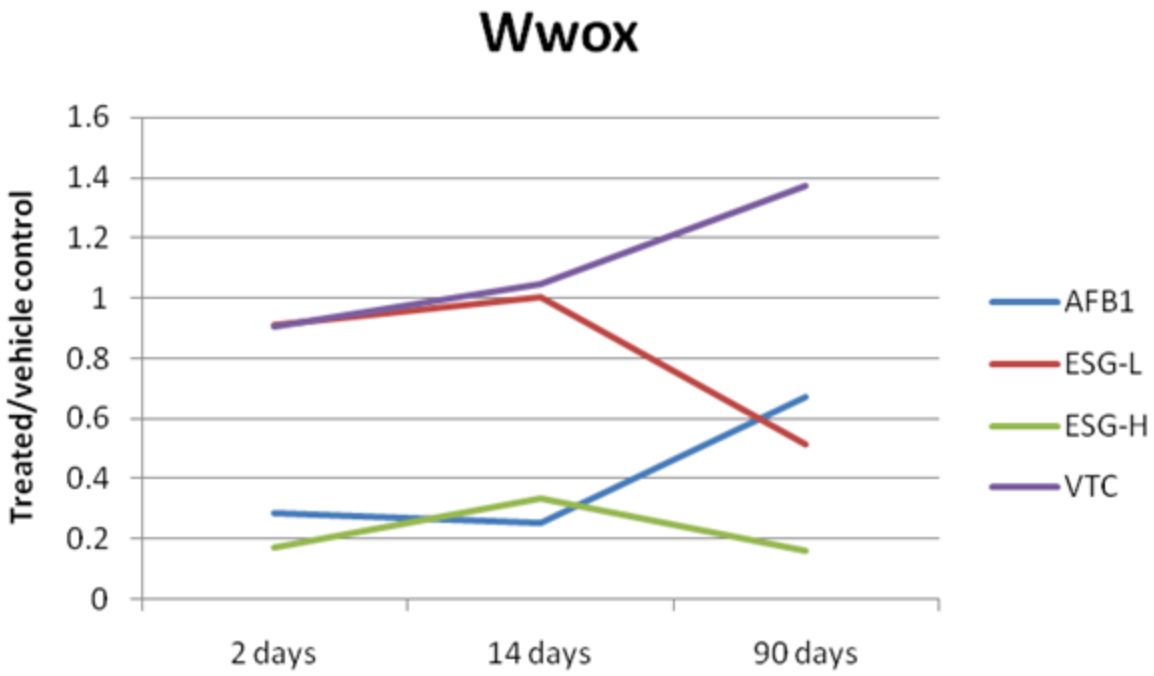
Cell migration
- Adam8

AhR
- Cyp1a1

H = human gene
AT = alternative transcript
I= intronic sequence

Why did the accuracy of the test data prediction get better with increasing exposure duration?

- Short durations of exposure (2 or 14 days) to weak carcinogen/dose combinations failed to induce gene expression changes reflective of carcinogenic activity



Conclusions

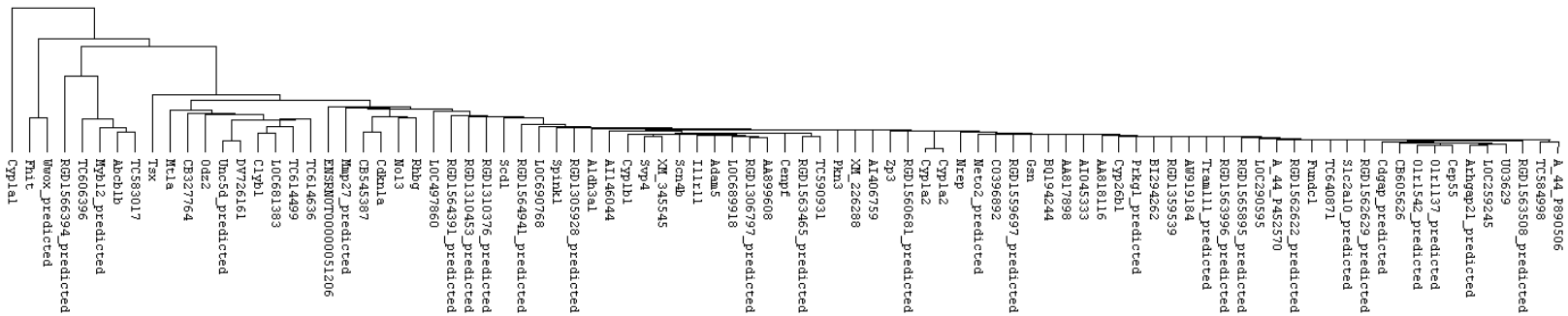
- Myristicin and isosafrole should be given higher priority relative to other members of this class for testing in the carcinogenicity bioassay
- We predict that isosafrole and myristicin, if tested at 2 mmoles/kg/day by corn oil gavage in male F344 rats, would produce significant increases in hepatic cancer
- Highly accurate hepatocarcinogenicity prediction models can be generated from hepatic gene expression changes gleaned from rats exposed for as little as 2 days to highly carcinogenic chemical/dose combinations
  - Models built on 2-day exposure data are equally as accurate as models based on 90-day data
- Genes informing the optimal models reflect pathways known to play a role in rat liver carcinogenesis
- Weakly carcinogenic chemical/dose combinations require longer exposure durations to manifest genomic changes indicative of carcinogenic activity
  - RECOMMENDATION: When performing gene expression-based classification of chemicals with unknown carcinogenic potency one should employ data from longer exposure durations (90 days or greater) in order to avoid false negative predictions

Points to address in future studies

- The chemicals used in the training data act by a limited number of mechanisms (DNA reactive, AhR activation) increasing the chance that some agents, acting by different mechanisms (PPAR activators), may be misclassified as non-carcinogenic
  - Study more chemicals with varied mechanisms of action
  - Alternative: 90 days of exposure may be enough time to produce gene expression changes that are more universally related to carcinogenesis
    - i.e. genes related to tissue remodeling and cell cycle
- The models currently do not address potency or dose-response
- The predictions are limited to male F344/N rat liver
  - The models presented here need to be validated across sexes, strains and species
  - More models need to be created using gene expression from other common target organ systems

Questions?

Hierarchical clustering of FA samples using 89 genes informative to models



Independent validation of the minimum feature models

Test data: 2 dose levels

Carcinogens – safrole*, estragole, methyleugenol

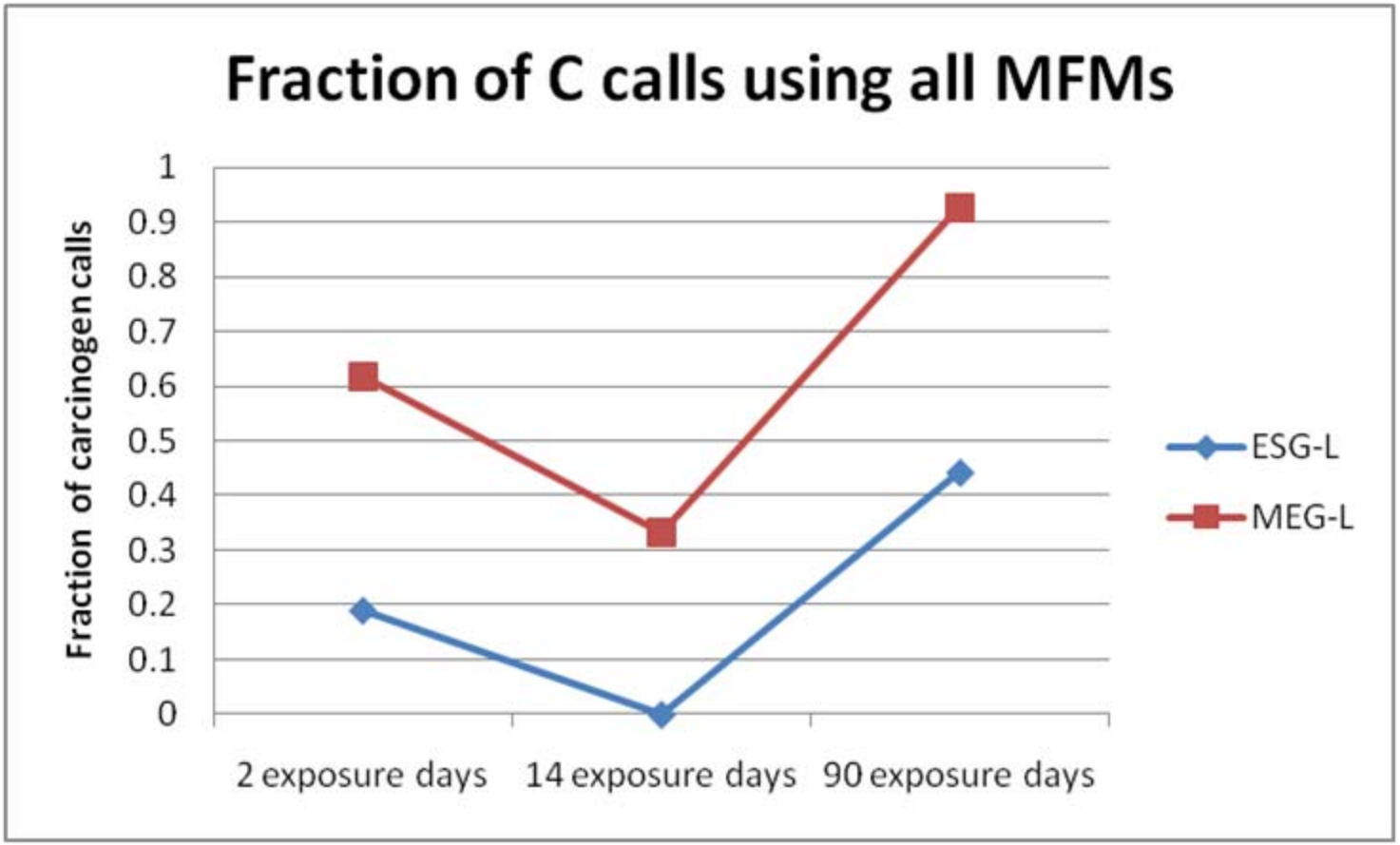
Non-carcinogens – eugenol, isoeugenol, gavage control, untreated control

*high dose only

Model	2-day test data (% error)	14-day test data (% error)	90-day test data (% error)	All test data (% error)
2 day	22	20	5	14
14 day	52	20	4	21
90 day	27	18	6	15
2+14 day	12	13	4	9
2+90 day	20	18	7	14
14+90 day	13	13	6	9
2+14+90 day	10	13	6	9
Average error (all models)	22	16	5	13

- 90 day test data yields the lowest overall error rate

Exposure duration and the identification of weak hepatocarcinogens



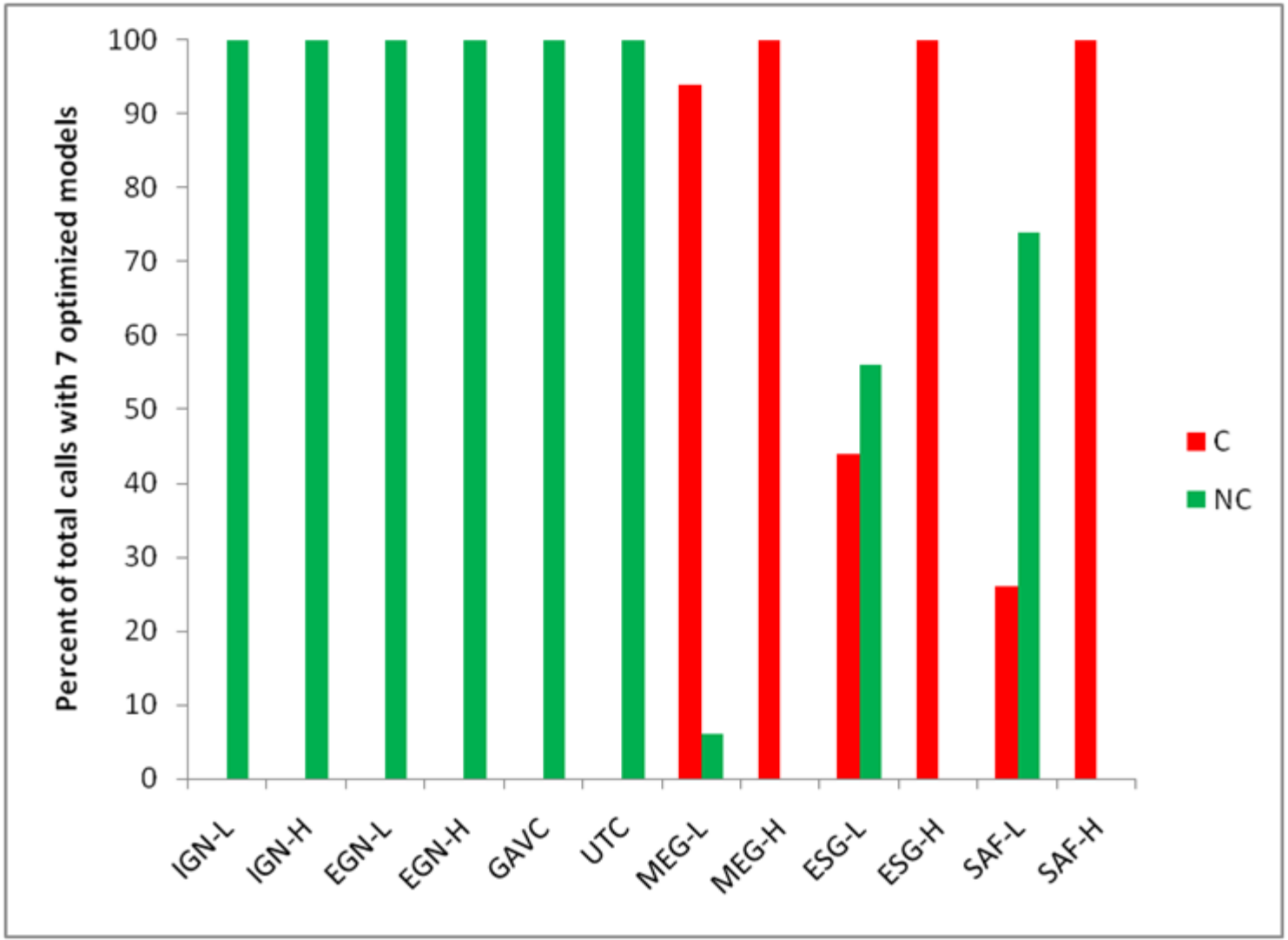
Prediction of the tested (alkoxy)propenyl benzene derivatives using the minimum feature models

	MEG(L)	MEG(H)	ESG(L)	ESG(H)	SAF(H)	EGN(H)	IGN(H)
2 day model	100(10/10)	100(10/10)	50(5/10)	100(10/10)	100(10/10)	0/10	0/10
14 day model	70(7/10)	100(10/10)	60(6/10)	100(10/10)	100(10/10)	0/10	0/10
90 day model	90(9/10)	100(10/10)	40(4/10)	100(10/10)	100(10/10)	0/10	0/10
2+14 day model	90(9/10)	100(10/10)	30(3/10)	100(10/10)	100(10/10)	0/10	0/10
2+90 day model	100(10/10)	100(10/10)	30(3/10)	100(10/10)	100(10/10)	0/10	0/10
14+90 day model	100(10/10)	100(10/10)	40(4/10)	100(10/10)	100(10/10)	0/10	0/10
2+14+ 90 day model	100(10/10)	100(10/10)	60(6/10)	100(10/10)	100(10/10)	0/10	0/10
Total C calls	94(66/70)	100 (70/70)	44(31/70)	100(70/70)	100(70/70)	0/70	0/70

Percent of 90 day exposure samples from the untested chemical group demonstrating a signature of carcinogenicity

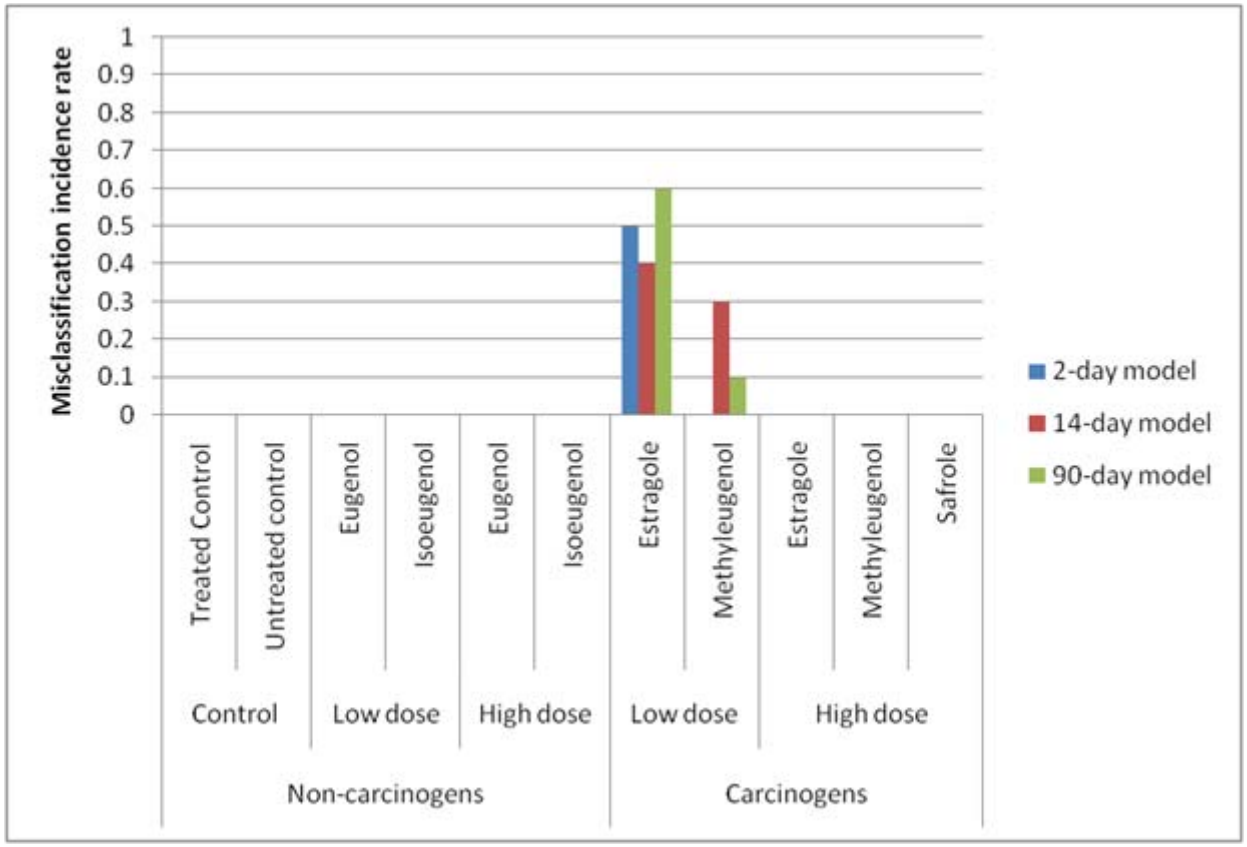
	Anethole(L)	Anethole(H)	Myristicin(L)	Myristicin(H)	Isosafrole(L)	Isosafrole(H)	Safrole(L)
2 day model	10 (1/10)	0	0	0	10 (10/10)	0	0
14 day model	0	0	0	80 (8/10)	0	70(7/10)	0
90 day model	0	0	0	90(9/10)	20(2/10)	90(9/10)	30(3/10)
2+14 day model	0	0	0	0	0	20(2/10)	0
2+90 day model	0	0	0	20(2/10)	0	40(4/10)	0
14+90 day model	0	0	0	0	0	20(2/10)	50(5/10)
2+14+ 90 day model	0	0	0	50(5/10)	0	60(6/10)	10(10/10)
Total C calls	1.5(1/70)	0/70	0/70	34(24/70)	4(3/70)	43(30/70)	26(18/70)

Meta-model predictions of the tested (alkoxy)propenyl benzene derivatives using 90 day exposure data



What caused the residual error when classifying the 90 day test samples?

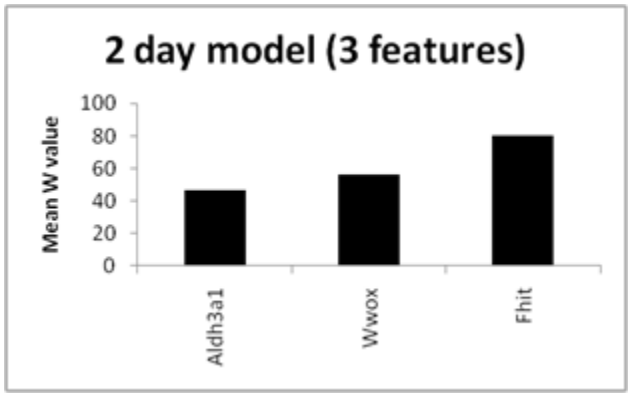
- Evan after 90 days of exposure some of the animals treated with low doses of carcinogens failed to exhibit changes in gene expression reflective carcinogenic activity

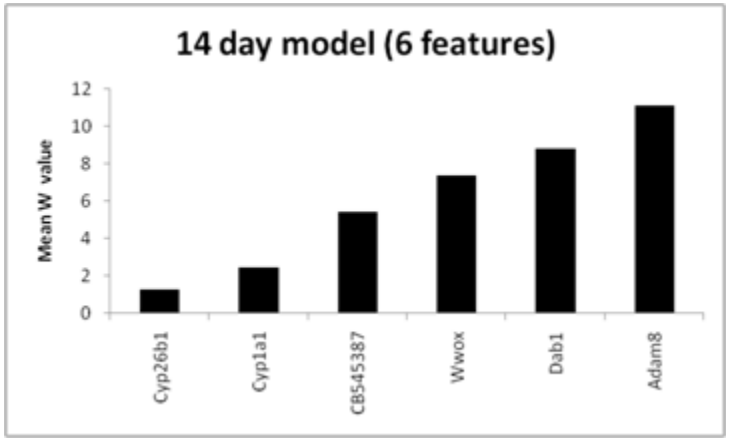


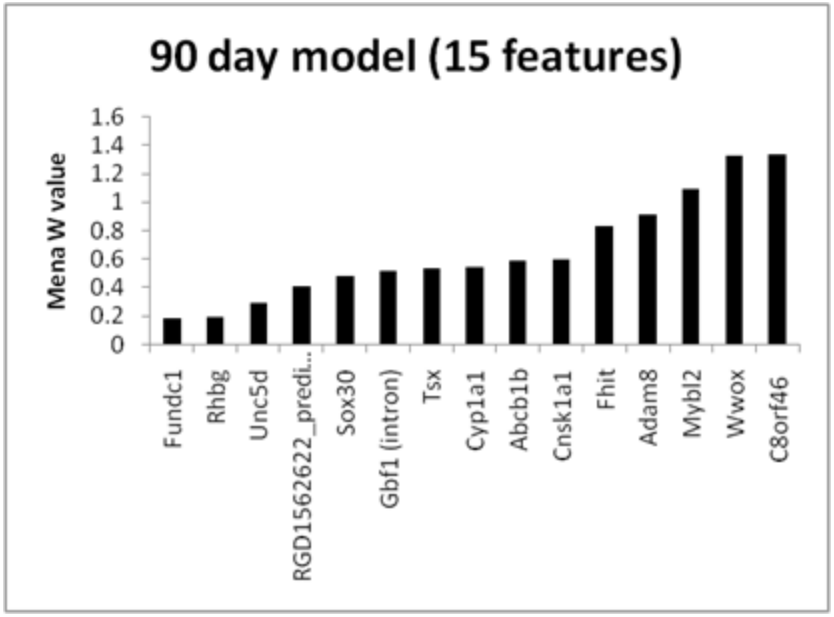
Percent of 90 day exposure samples from the untested chemical group demonstrating a signature of carcinogenicity

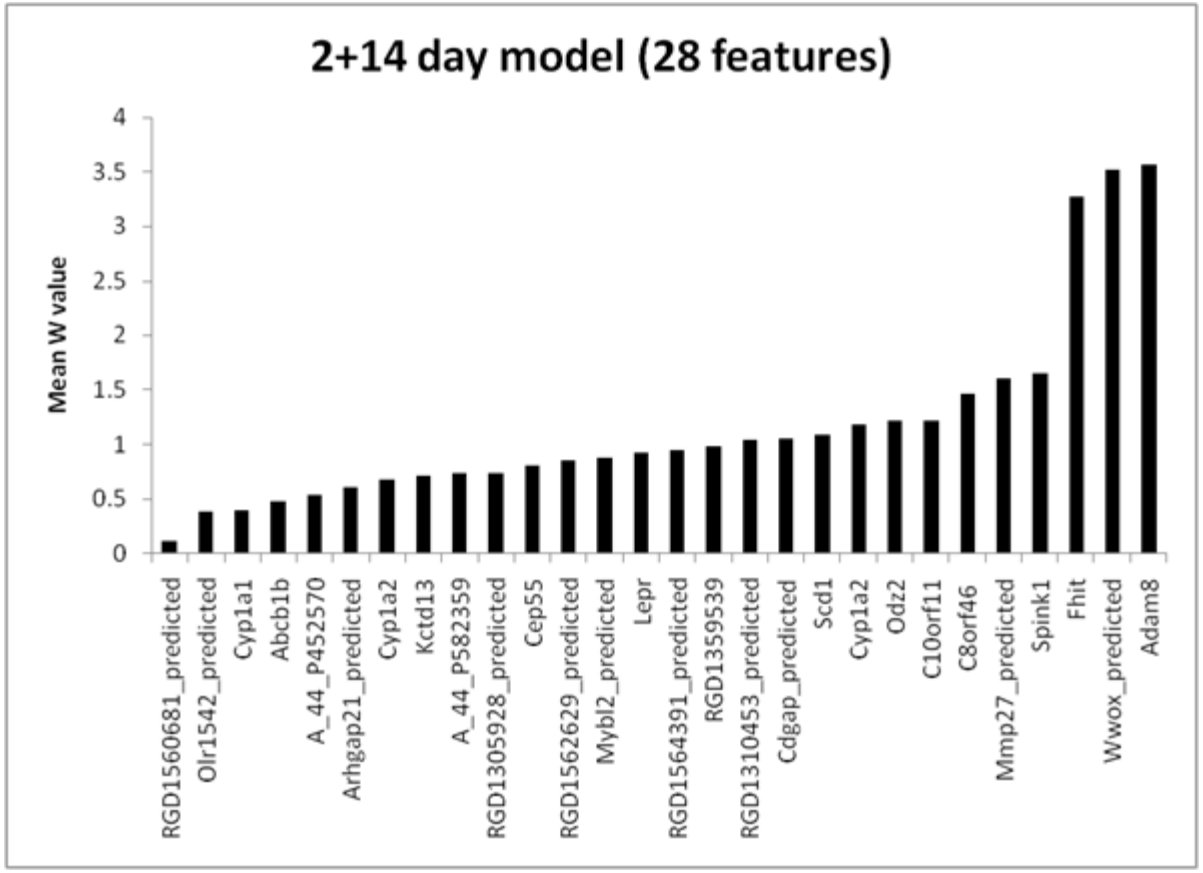
	Anethole(L)	Anethole(H)	Myristicin(L)	Myristicin(H)	Isosafrole(L)	Isosafrole(H)
2 day model	10	0	0	0	10	0
14 day model	0	0	0	80	0	70
90 day model	0	0	0	90	20	90

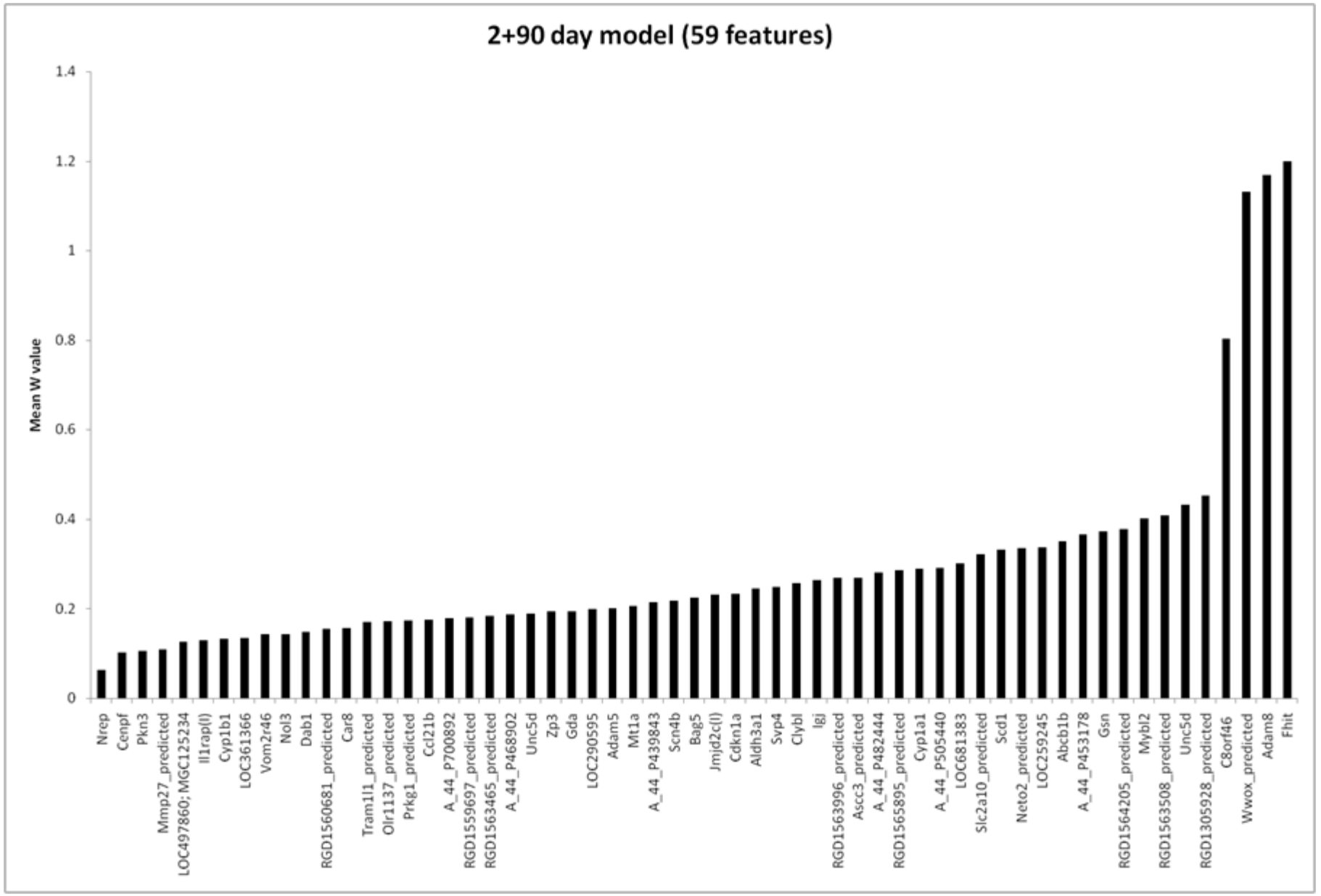
* Myristicin and isosafrole are Cyp1a1 inducers

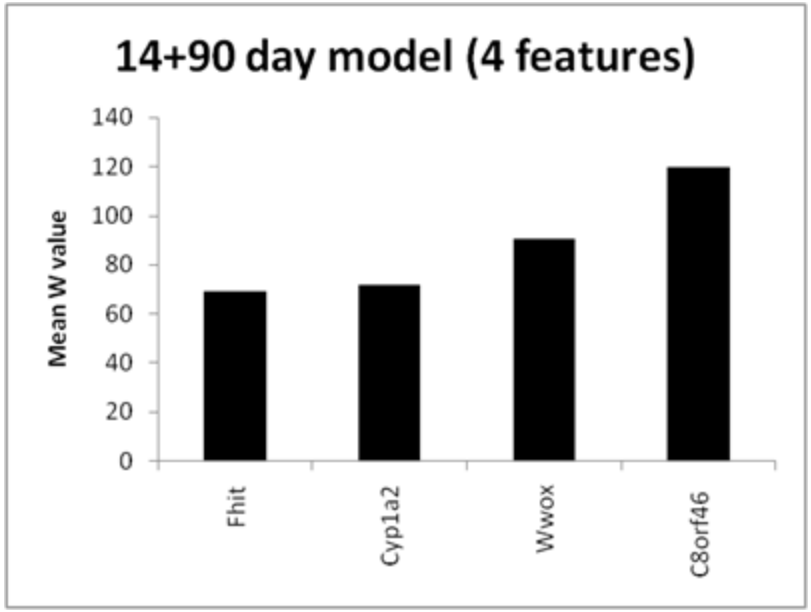

2 day model (3 features)
Mean W value
100
80
60
40
20
0
Aldh3a1
Wwox
Fhit

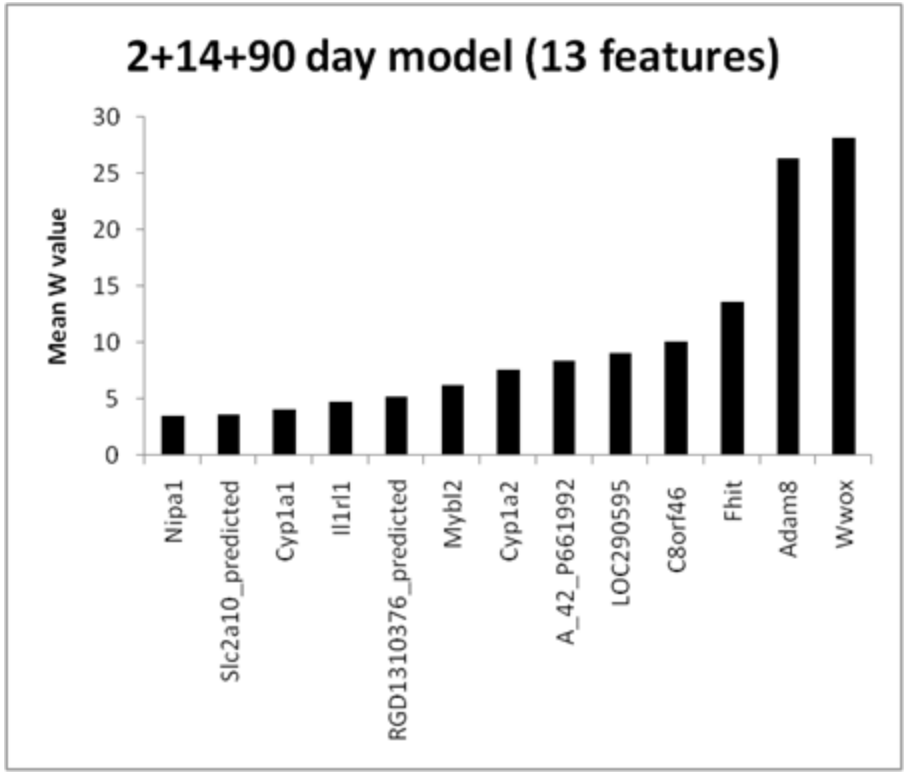

14 day model (6 features)
Mean W value
12
10
8
6
4
2
0
Cyp26b1
Cyp1a1
CB545387
Wwox
Dab1
Adam8


90 day model (15 features)
Mena W value
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Fundc1
Rhbg
Unc5d
RGD1562622_predi...
Sox30
Gbf1 (intron)
Tsx
Cyp1a1
Abcb1b
Cnsk1a1
Fhit
Adam8
Mybl2
Wwox
C8orf46


2+14 day model (28 features)
Mean W value
4
3.5
3
2.5
2
1.5
1
0.5
0
RGD1560681_predicted
Olr1542_predicted
Cyp1a1
Abcb1b
A_44_P452570
Arhgap21_predicted
Cyp1a2
Kctd13
A_44_P582359
RGD1305928_predicted
Cep55
RGD1562629_predicted
Mybl2_predicted
Lepr
RGD1564391_predicted
RGD1359539
RGD1310453_predicted
Cdgap_predicted
Scd1
Cyp1a2
Odz2
C10orf11
C8orf46
Mmp27_predicted
Spink1
Fhit
Wwox_predicted
Adam8


2+90 day model (59 features)
Mean W value
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Nrep
Cenpf
Pkn3
Mmp27_predicted
LOC497860; MGC125234
Il1rap(l)
Cyp1b1
LOC361366
Vom2r46
Nol3
Dab1
RGD1560681_predicted
Car8
Tram1l1_predicted
Olr1137_predicted
Prkg1_predicted
Ccl21b
A_44_P700892
RGD1559697_predicted
RGD1563465_predicted
A_44_P468902
Unc5d
Zp3
Gda
LOC290595
Adam5
Mt1a
A_44_P439843
Scn4b
Bag5
Jmjd2c(l)
Cdkn1a
Aldh3a1
Svp4
Clybl
Igj
RGD1563996_predicted
Ascc3_predicted
A_44_P482444
RGD1565895_predicted
Cyp1a1
A_44_P505440
LOC681383
Slc2a10_predicted
Scd1
Neto2_predicted
LOC259245
Abcb1b
A_44_P453178
Gsn
RGD1564205_predicted
Mybl2
RGD1563508_predicted
Unc5d
RGD1305928_predicted
C8orf46
Wwox_predicted
Adam8
Fhit


14+90 day model (4 features)
Mean W value
140
120
100
80
60
40
20
0
Fhit
Cyp1a2
Wwox
C8orf46


2+14+90 day model (13 features)
Mean W value
30
25
20
15
10
5
0
Nipa1
Slc2a10_predicted
Cyp1a1
Il1rl1
RGD1310376_predicted
Mybl2
Cyp1a2
A_42_P661992
LOC290595
C8orf46
Fhit
Adam8
Wwox

Cross and independent validation of the minimum feature models

Test data: 2 dose levels
Carcinogens – safrole*, estragole, methyleugenol
Non-carcinogens – eugenol, isoeugenol, gavage control, untreated control

Model	# of features	Cross-validation error	All test data (% error)	2-day test data (% error)	14-day test data (% error)	90-day test data (% error)
2 day	3	0%	14	22	20	5
14 day	6	0%	21	52	20	4
90 day	15	0%	15	27	18	6
2+14 day	28	1%	12	12	17	7
2+90 day	59	0%	14	20	18	7
14+90 day	4	0%	9	13	13	6
2+14+90 day	13	0%	9	10	13	6

* 90 day test data yields the lowest error rate

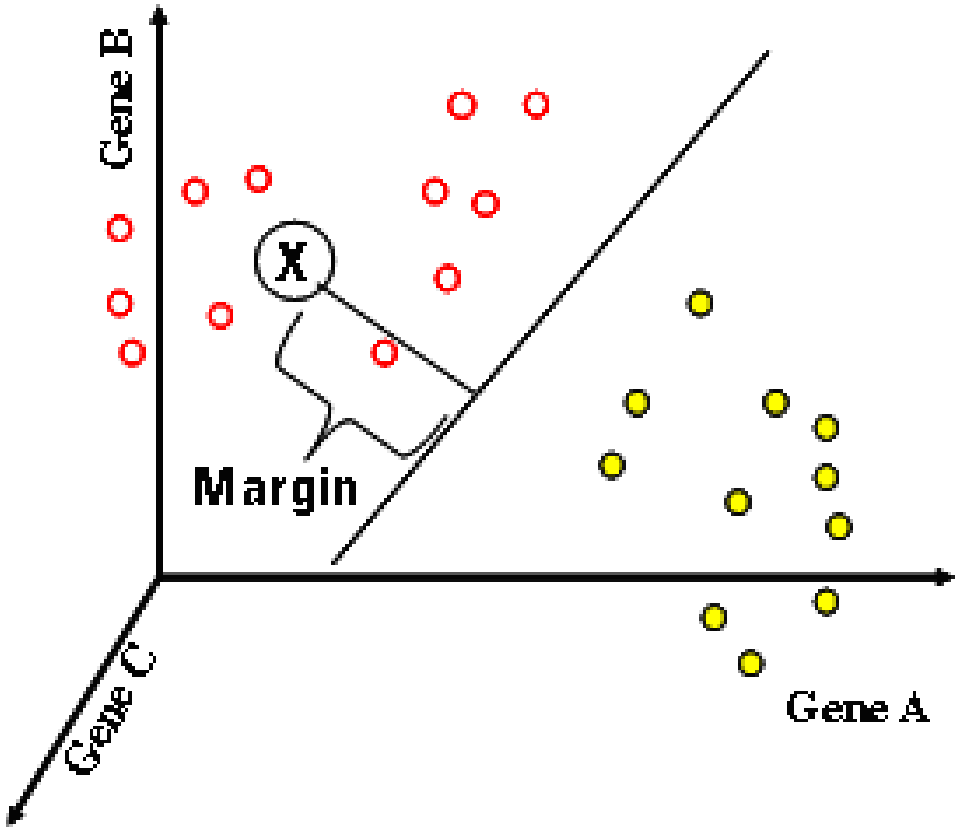
Support Vector Machines (1)

- Supervised machine learning technique
- Plot each training set sample according to its expression intensity for the selected predictor genes.
  - The space in which the samples reside is termed input space of n-dimensions, where n equals the number of predictor genes specified for the analysis
- SVM algorithm then attempts to locate a linear hyperplane that will separate the samples of the two classes
  - If multiple classes are being discriminated, a linear hyperplane is drawn for each class, in a one class-versus-rest approach
- Samples not separable in input space can eventually be made separable by mapping the samples to a higher dimensional feature space
- The SVM algorithm is able to circumvent the problem of working in higher-dimensional space by using a kernel function to define a linear separating hyperplane without explicitly mapping the samples into feature space

Support Vector Machines (2)

- Once the hyperplane has been defined, each test sample is plotted according to its expression intensity for the selected predictor genes, and the distance between each test sample and the hyperplane is calculated into a margin score
- A margin score for each test sample is calculated for each class
- A test sample will have a positive margin score for the class if it is on the same side of the hyperplane as the training samples representing that class, and a negative margin score if it is on the opposite side of the hyperplane as the training samples representing that class
- The magnitude of the margin score also indicates the degree of confidence in that prediction
  - A margin score of +1 or greater indicates that the algorithm has high confidence that the sample belongs to that class, and a score of -1 or less reflects a high confidence that the sample does not belong to that class

Support Vector Machines (3)



Procedure for model creation and refinement