# Signalling circuit activities as mechanism-based features to predict mode of action of chemicals.

Cankut Cubuk[1], Marta R. Hidalgo[1], Jose Carbonell-Caballero[1], and Joaquín Dopazo[1,2,3]

1. Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain.
2. Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, Spain.
3. Functional Genomics Node, (INB) at CIPF, Valencia, Spain.

## Abstract

Most biological phenotypes are too complex to be described as consequences of the activities of individual genes but rather as a complex interaction among these. Here we propose the transformation of individual gene expression data into numerical descriptors of signalling pathway activities and its use to predict the mode of action (MOA) of chemicals. Here we addressed the Challenge 1a: SEQC Rat TGx - rat liver response to chemicals data, Topics 1 and 2. Our results show how the performance of the transformed values is quite good and how the predictions derived from RNA-seq seem to be better that the ones derived from microarrays.

## Introduction

Many complex traits, as drug response, are associated with complex changes in biological pathways rather than being the direct consequence of single gene alterations. Actually, the idea of using the information contained in different biological pathways to understand complex traits, such as disease or drug mode of action is gaining acceptance [1]. Signaling pathways provide a formal representation of the processes by which the cell triggers actions in response to particular stimulus through a network of intermediate gene products. In particular, specific sub-networks (or circuits) that connect stimulus reception proteins to proteins that produce the consequent cell response can directly be related to cell functionalities. Recently some methods have developed that focus particularly on the estimation of the activity of these stimulus-response signaling circuits from gene expression data [2, 3].

## Method:

We obtained *Rattus norvegicus* (rat) signalling pathway information from KEGG database. A total of 23 signalling pathways were examined. Each pathway was split up into their elementary signalling circuits, as described previously [3]. Activation-inactivation relationships between nodes (proteins) along the circuits enabled us to use

a graph traversal methodology for updating signal intensity at each visited node and finally compute a global value of signal transduction for the circuit, that we call signalling circuit activity therein. Unlike in previous methods [2, 3], the algorithm used here for the calculation of these signalling circuit activities is platform independent and can use gene expression data either from microarrays or from RNA-seq.

The microarray and RNAseq datasets (GSE55347,GSE47792) were downloaded from the GEO database. The raw microarray data were normalized by RMA method. The probe IDs were converted into Entrez Gene IDs. The probe expression values were summarized into gene expression values by 90 quantile.

The RNAseq data were already normalized, as provided by the MAGIC pipeline and we used them directly, and annotated with Entrez gene IDs (duplicated gene IDs were excluded).

In total, 1334 genes were used to calculate signalling circuit activities for the 867 sub-pathways that compose the 23 signalling pathways studied here. These signalling circuit activities and normalized gene expression values were used to compare their respective prediction accuracies.

ANOVA was used to detect the differential expressed genes and signalling circuits. All training and test set groups were used together for ANOVA.

For the prediction, support vector machine (SVM) with radial basis function (RBF) kernel was used [4]. Two parameters for an RBF kernel were used: cost and sigma. Best sigma and cost parameters were selected among different values tested. The model optimized with 10 fold cross validation.

(MOAs were used as endpoints for training the model as follows:

Training Set:

"PPARA", "CAR/PXR", "CONTROL", "UNKNOWN"(AhR, Cytotoxoc, DNA Damage)

Test Set:

"PPARA", "CAR/PXR", "CONTROL","UNKNOWN"(ER, HMGCOA)

## Results and discussion

For both platforms the prediction accuracy obtained using signalling circuit activities as classification variables was reasonable and better than the corresponding accuracy obtained when using genes alone (see Figure 1).

It must be taken into account that not all the chemicals studied are acting at the level of the signalling pathways and therefore some MOAs will probably be deficiently predicted using only information on signalling. For example, HMGCOA (all) and AHR (LEFLUNOMIDE) MOAs are known to act at the level of metabolic pathways.

**RNA-seq**

Signalling Circuits

```
Confusion Matrix and Statistics

          true
pred       CAR_PXR PPARA UNKNOWN VEHICLE
  CAR_PXR        2     2       3       0
  PPARA          0     3       0       0
  UNKNOWN        7     2      15       0
  VEHICLE        0     2       0       6

Overall Statistics

               Accuracy : 0.619
                 95% CI : (0.4564, 0.7643)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 0.01

                  Kappa : 0.4372
 Mcnemar's Test P-Value : NA
```

Model accuracy= 72.7%

Gene expression

```
Confusion Matrix and Statistics

          true
pred       CAR_PXR PPARA UNKNOWN VEHICLE
  CAR_PXR        0     0       0       0
  PPARA          0     0       0       0
  UNKNOWN        0     0       0       0
  VEHICLE        9     9      18       6

Overall Statistics

               Accuracy : 0.1429
                 95% CI : (0.0543, 0.2854)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 1

                  Kappa : 0
 Mcnemar's Test P-Value : NA
```

Model accuracy = 40.90909

**Microarray**

Signalling Circuits

```
Confusion Matrix and Statistics

          true
pred       CAR/PXR Control PPARA UNKNOWN
  CAR/PXR        1       1     4       3
  Control        4       4     5       8
  PPARA          3       0     0       4
  UNKNOWN        1       1     0       3

Overall Statistics

               Accuracy : 0.1905
                 95% CI : (0.086, 0.3412)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 0.99970

                  Kappa : -0.0171
 Mcnemar's Test P-Value : 0.00796
```

Model accuracy = 60.0%

Gene expression

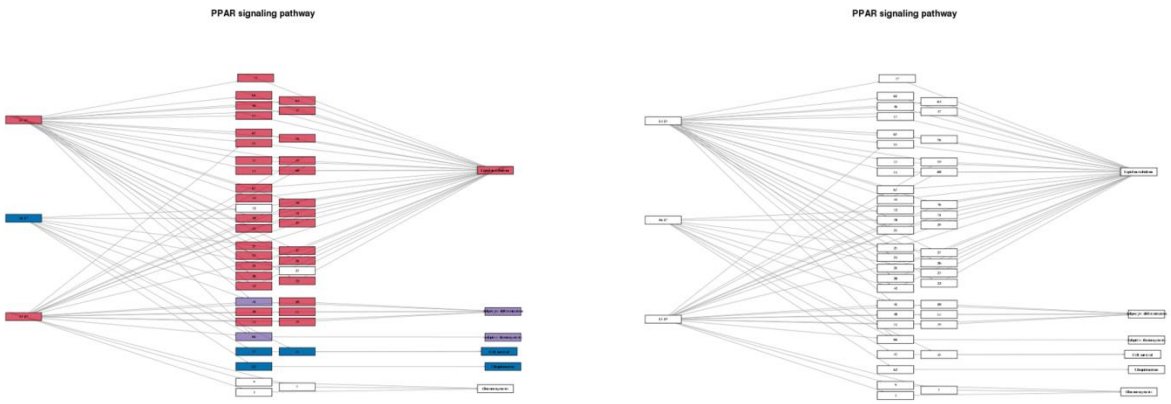
```
Confusion Matrix and Statistics

          true
pred       CAR/PXR Control PPARA UNKNOWN
  CAR/PXR        0       1     1       0
  Control        5       5     8      16
  PPARA          4       0     0       2
  UNKNOWN        0       0     0       0

Overall Statistics

               Accuracy : 0.119
                 95% CI : (0.0398, 0.2563)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0444
 Mcnemar's Test P-Value : NA
```

Model accuracy=48.9%

***Figure 1**. Prediction accuracy obtained using signalling circuit activities or gene expression values as classification variables obtained for RNA-seq and microarray data.*

An example in which the different effect of chemicals over pathways is obvious is depicted in Figure 2. It presents the analysis results of the PPARA signalling pathway

among the two MOA groups. Common MOAs groups of the test and training sets were merged for this analysis.
In the PPARA group (the group which exposed to PPARA agonists) the PPAS signalling pathway present a clear alteration in the lipid metabolism, while the AHR (and actually the other MOA groups, data not shown) have the pathway unaltered. These analyses were carried out for both, RNAseq and microarray data, rendering highly correlated results.




PPARA group AHRgroup

***Figure 2.*** *Differential activities in the PPAR signalling pathway in presence of chemicals belonging to the PPAR (left) and AHR (right) groups.*

## Conclusions

The method presented here shows that transforming the gene expression data into mechanism-based biomarkers within the context of signalling pathways is useful to predict molecular phenotypes that are controlled by signalling pathways. In addition, we were able to distinguish different phenotypes using signalling circuit activities.
We propose that approaches that model cell functionalities will be not only more accurate in predicting phenotypic traits, such as the drug response, but will also provide insights into the molecular mechanisms that account for such phenotype.

## References

1. Davis MJ, Ragan MA: **Understanding cellular function and disease with comparative pathway analysis.** *Genome Med* 2013, **5:**64.

2. Sebastian-Leon P, Carbonell J, Salavert F, Sanchez R, Medina I, Dopazo J: **Inferring the functional effect of gene expression changes in signaling pathways.** *Nucleic Acids Res* 2013, **41:**W213-217.
3. Sebastian-Leon P, Vidal E, Minguez P, Conesa A, Tarazona S, Amadoz A, Armero C, Salavert F, Vidal-Puig A, Montaner D, Dopazo J: **Understanding disease mechanisms with models of signaling pathway activities.** *BMC Syst Biol* 2014, **8:**121.
4. Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, et al: **The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance.** *Nat Biotech* 2014, **32:**926-932.