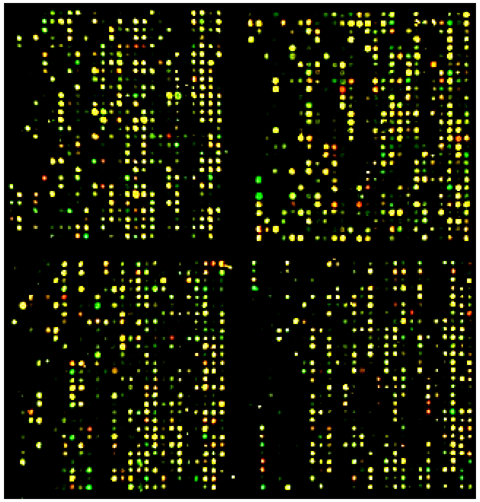
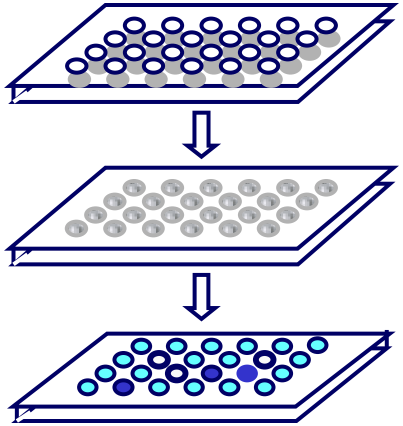
KDD Cup Task 2

Mark Craven
Department of Biostatistics & Medical Informatics
Department of Computer Sciences
University of Wisconsin
craven@biostat.wisc.edu
www.biostat.wisc.edu/~craven

Task Motivation

- molecular biology has entered a new era in which experimentation can be done in a high-throughput manner
  - *microarrays* can simultaneously measure the "activity" of thousands of genes under some set of conditions
  - *yeast deletion arrays* can measure the activity of some "reporter" system when each of ~5k genes is knocked out





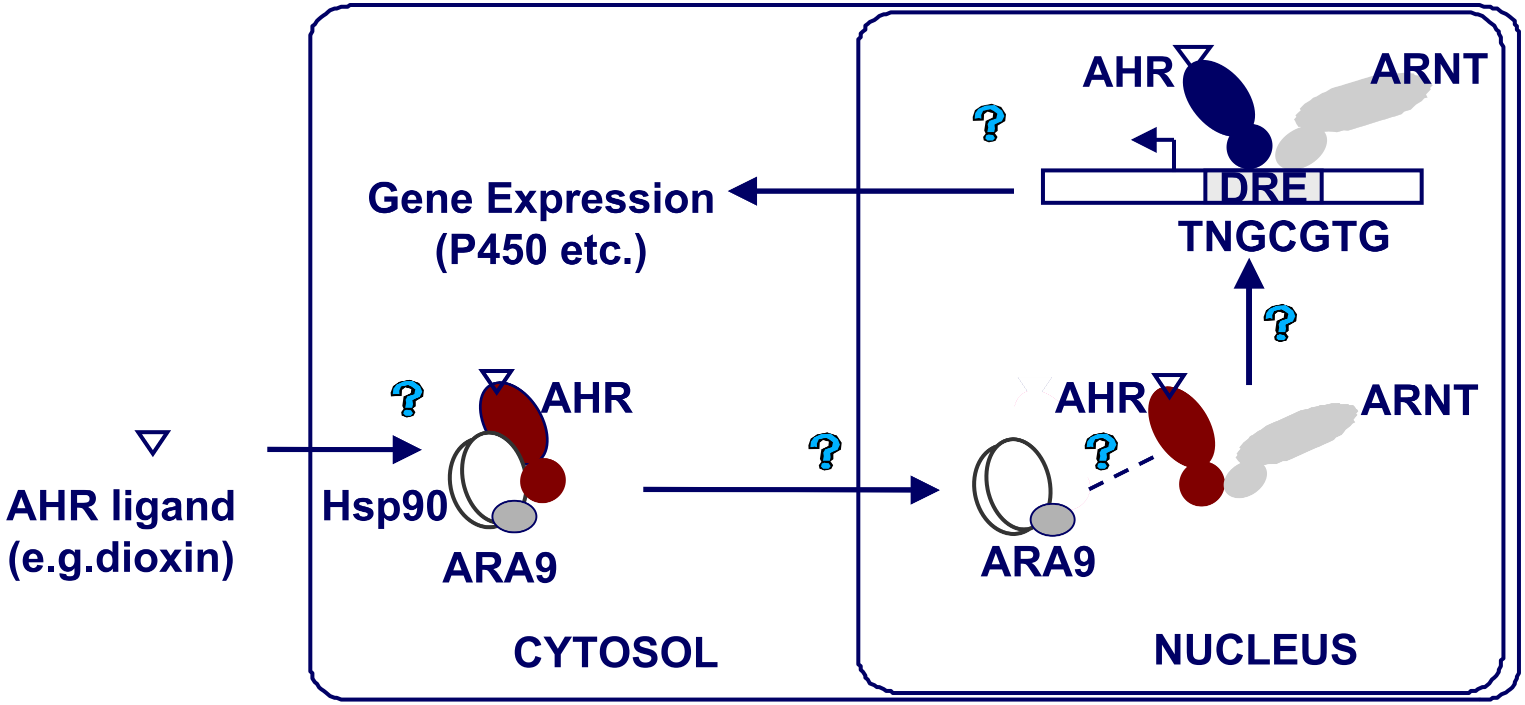
- **key problem**: it is difficult for biologists to assimilate and interpret thousands of measurements per experiment

The Problem Domain: Characterizing the *Regulatome* of the AHR Signaling Pathway

- experimental data kindly provided by

  Guang Yao and Prof. Chris Bradfield
  McArdle Laboratory for Cancer Research
  University of Wisconsin
- the *Aryl Hydrocarbon Receptor (AHR)* is a member of the protein family that mediates the biological response to dioxin, hypoxia, circadian rhythm, etc.
- focus of project: determine which proteins affect the activity of AHR

The AHR Signaling Pathway



- when a cell is exposed to say, dioxin, AHR acts to turn on/off various genes
- experiment motivation: which proteins (gene products) in the cell regulate how AHR does this?

Characterizing the *Regulatome* of the AHR Signaling Pathway

- a high-throughput experiment using the *Yeast Deletion Array* (~5k strains of yeast, each with a specified gene knocked out)
- for each strain
  - insert a specially engineered AHR gene
  - insert a “reporter” system that is activated by AHR signaling
  - prod the AHR signaling pathway with a dose of agonist
  - see if the reporter lights up
- result: we can see which genes encode proteins that affect AHR signaling

The KDD Cup Task

- key computational task : help annotate/explain the results of the experiment, using available data sources

- a proxy task for KDD Cup: develop models that can predict the experimental result for a given gene from available data sources

- rationale:
  - annotation/explanation task not amenable to objective evaluation
  - prediction task, like annotation/explanation task, involves eliciting patterns from available data that explain why individual genes behave as they do in the experiment

The KDD Cup Task

- **given**: data describing a gene
  - hierarchical (functional/localization annotation)
  - relational (protein-protein interactions)
  - text (scientific abstracts from MEDLINE)
- **do**: predict if knocking out the gene will have a significant effect on AHR signaling

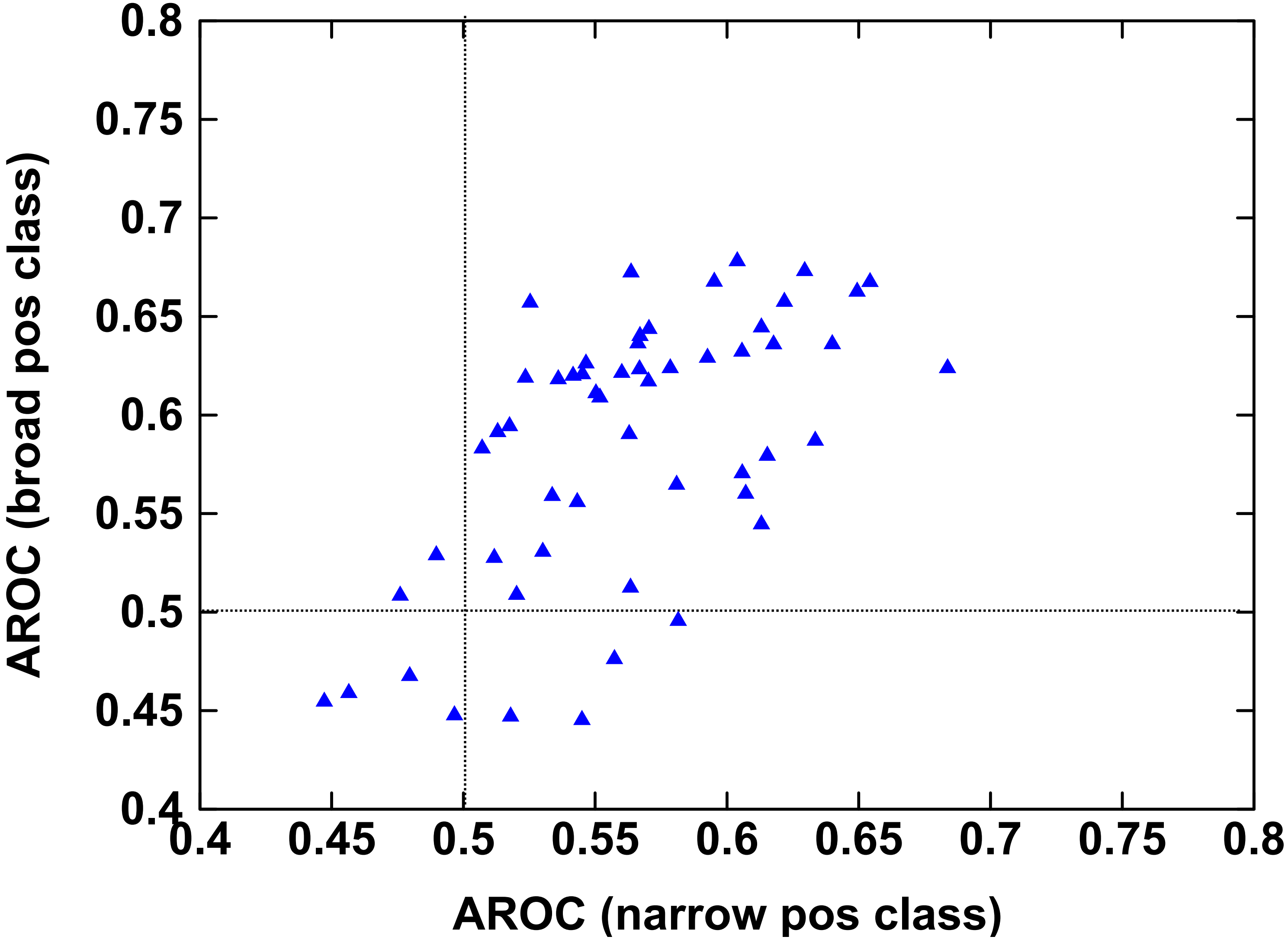
Characteristics of the Problem

- rich data sources
- much missing data
  - function/localization annotations
  - protein-protein interactions
  - abstracts
- few positive instances (127 pos, 4380 neg)
- very “disjunctive”

Task Evaluation

- evaluated as a two-class problem
  - positive: knockout has significant effect on AHR signaling
- but two different definitions of positive class
  - narrow: knockout has an AHR-specific effect
  - broad: knockout also affects a control pathway
- the scoring metric was the sum of the *area under the ROC curve* (AROC) for the two class partitions

AROC Scores for All Teams



Task 2 Winning Teams

- winner
  - ✯ Adam Kowalczyk and Bhavani Raskutti
    Telstra Research Laboratories

- honorable mention
  - ✯ David Vogel and Randy Axelrod
    A.I. Insight Inc. and *Sentara Healthcare*
  - ✯ Marcus Denecke, Mark-A. Krogel, Marco Landwehr and Tobias Scheffer
    Magdeburg University
  - ✯ George Forman
    Hewlett Packard Labs
  - ✯ Amal Perera, Bill Jockheck, Willy Valdivia Granda, Anne Denton, Pratap Kotala and William Perrizo
    North Dakota State University

Current and Future Activity

- figure out what lessons have been learned
  - value of text?
  - which algorithms learned most accurate models?
  - etc.
- determine if learned models can provide insight into the domain
- write articles (task overview, descriptions of winning teams’ methods) for *SIGKDD Explorations*
- maintain public access to data set (do Google search on KDD Cup)

Acknowledgements

- the experimental data was generated by
  Guang Yao and Prof. Chris Bradfield
  McArdle Laboratory for Cancer Research
  University of Wisconsin