

# Applying Text Mining to Aid in Understanding Gene Regulation

**Mark Craven**

Department of Biostatistics & Medical Informatics

Department of Computer Sciences

University of Wisconsin.

[craven@biostat.wisc.edu](mailto:craven@biostat.wisc.edu)

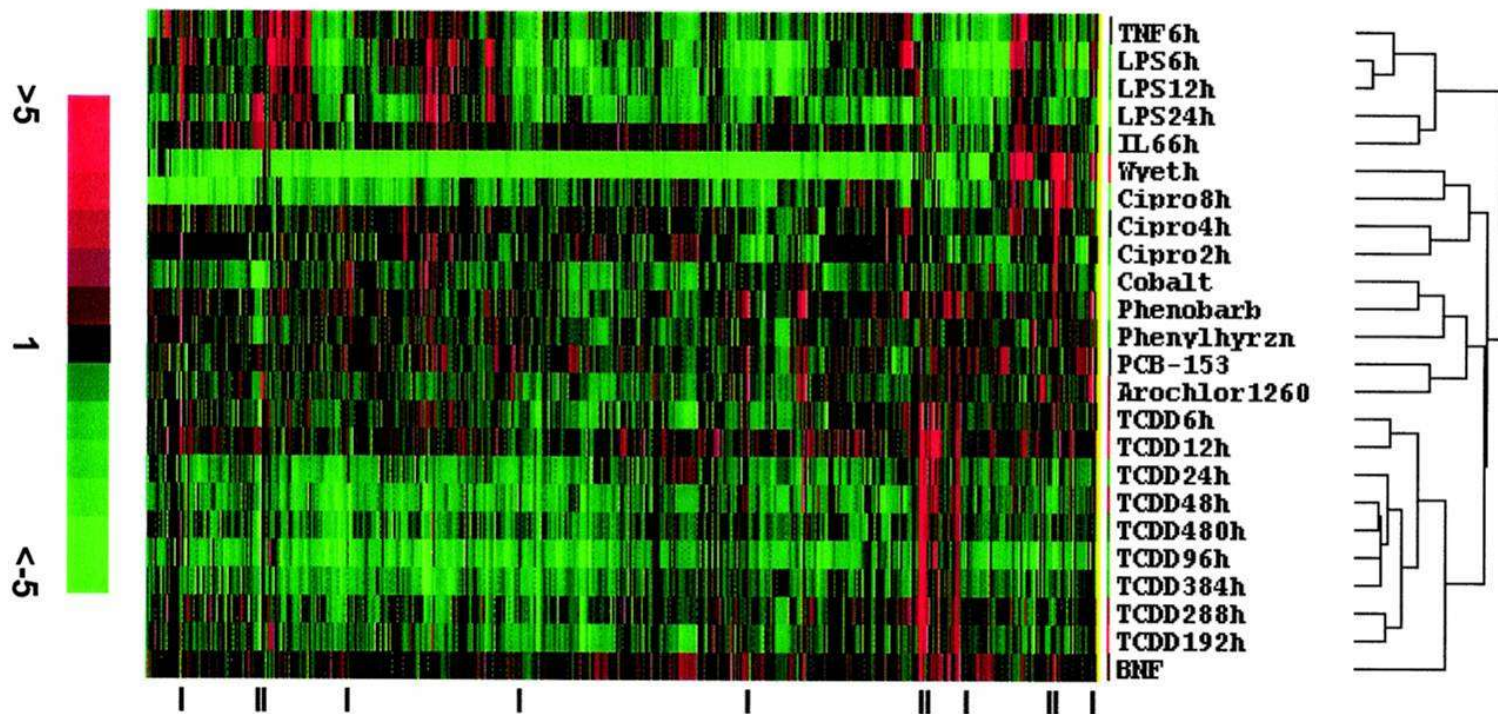
[www.biostat.wisc.edu/~craven](http://www.biostat.wisc.edu/~craven)

# Outline

- two motivating tasks
- a few important issues
  - identity uncertainty
  - how much to return
  - heterogeneity
  - data integration
  - different amounts of literature for each gene

# Thomas et al., *Molecular Pharmacology* 2002

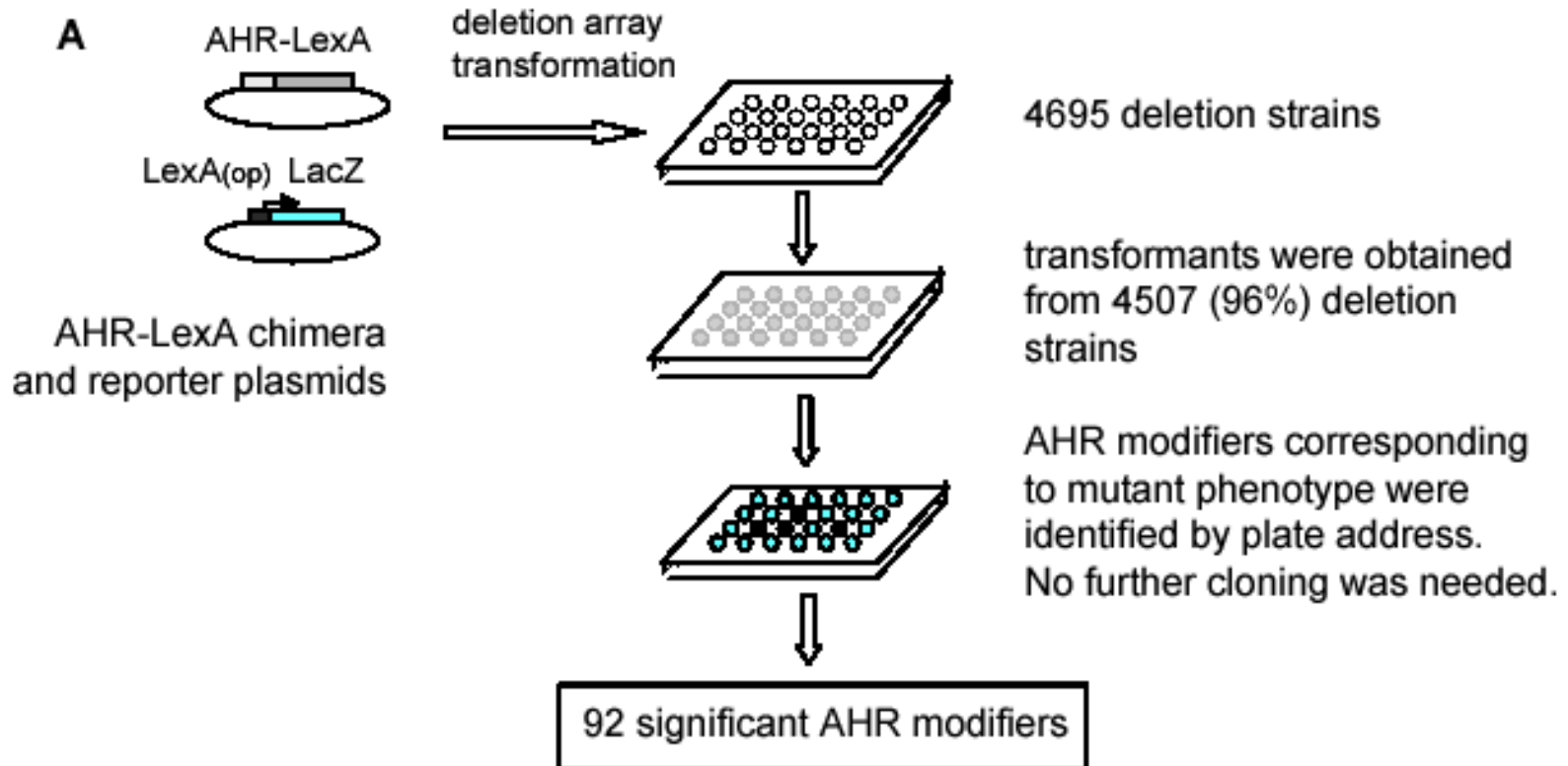
Genes



- in initial experiments, a mysterious set of genes that were upregulated in all treatments

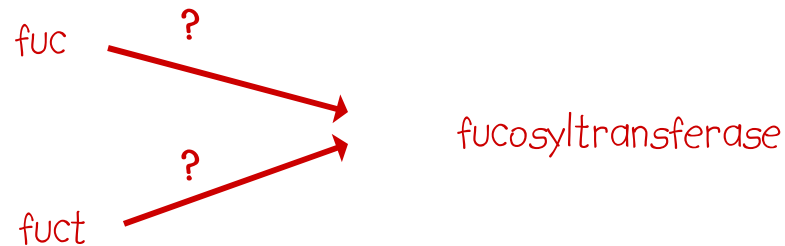
# Yao et al., *PLoS Biology* 2004

- a high-throughput deletion-array screen for knockouts that modify AHR signaling



# Issue: Identity Uncertainty

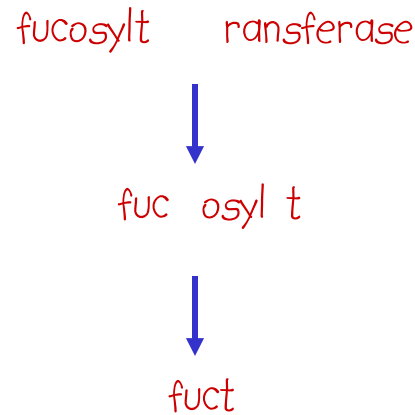
- issue: determine which sets of mentions refer to the same object



- a.k.a. *coreference resolution, de-duping, record linkage*

# Issue: Identity Uncertainty

- one approach: learning edit-distance functions



- edit operations and costs need to be sensitive to context/morphology

# Issue: What/How Much to Return?

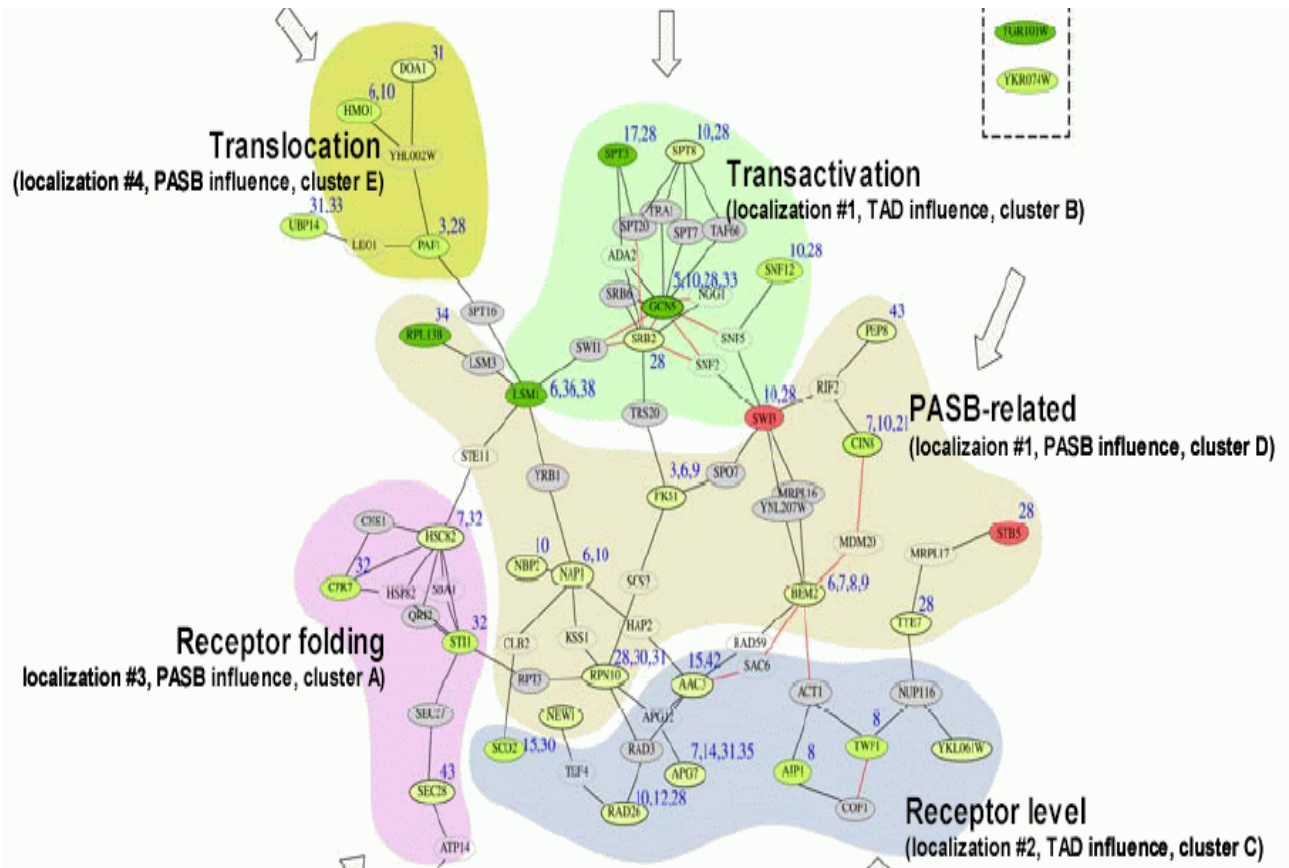
Keyphrases ranked according to "association" with genes involved in viral replication

| Number | Keyphrase                              | InCluster(Abstracts:Genes) | OutOfCluster(Abstracts:Genes) | Score  | Probability            |
|--------|--|----------------------------|-------------------------------|--------|------------------------|
| 1      | <a href="#">decapping</a>              | 17:4                       | 10:9                          | 221.24 | 1.0842928548219345E-47 |
| 2      | <a href="#">tyrosyl-tma</a>            | 8:2                        | 1:1                           | 217.06 | 8.689802866527016E-47  |
| 3      | <a href="#">decapping enzym</a>        | 12:2                       | 2:2                           | 188.44 | 1.3232829028312333E-40 |
| 4      | <a href="#">tyrosyl-tma synthetase</a> | 7:2                        | 1:1                           | 186.12 | 4.2027594882441745E-40 |
| 5      | <a href="#">mma decapping</a>          | 9:3                        | 2:2                           | 162.97 | 4.196212573556378E-35  |
| 6      | <a href="#">patlp</a>                  | 6:3                        | 2:2                           | 157.5  | 6.353683296010251E-34  |
| 7      | <a href="#">lsm protein</a>            | 5:3                        | 1:1                           | 147.69 | 8.28250521212477E-32   |
| 8      | <a href="#">eif-4e</a>                 | 5:2                        | 4:4                           | 119.11 | 1.202477065402657E-25  |
| 9      | <a href="#">lsm</a>                    | 5:3                        | 3:3                           | 109.39 | 1.4850205170281127E-23 |
| 10     | <a href="#">aminoacylation</a>         | 6:2                        | 7:7                           | 94.19  | 2.756746440978038E-20  |
| 11     | <a href="#">la protein</a>             | 5:3                        | 5:5                           | 88.02  | 5.834740450201008E-19  |
| 12     | <a href="#">cap-binding</a>            | 7:5                        | 18:15                         | 83.28  | 6.074427407019047E-18  |
| 13     | <a href="#">cap-binding protein</a>    | 5:4                        | 10:10                         | 83.23  | 6.213453212370963E-18  |
| 14     | <a href="#">ef-1 alpha</a>             | 7:2                        | 5:3                           | 78.23  | 7.368936124659748E-17  |
| 15     | <a href="#">ef-1</a>                   | 7:2                        | 7:5                           | 66.08  | 2.940926675347068E-14  |

- return a *sufficient* explanation or all significant terms?
- the latter raises the *multiple comparisons* problem; but perhaps new methods for bounding FDR could be applied [e.g. Storey & Tibshirani]

# Issue: Heterogeneity

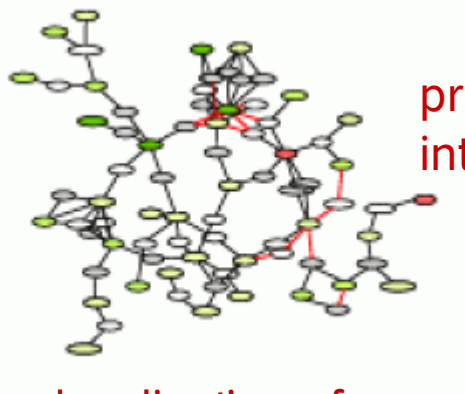
- genes may exhibit the same behavior for different reasons



- need for disjunctive explanations
- need for integration of various data sources

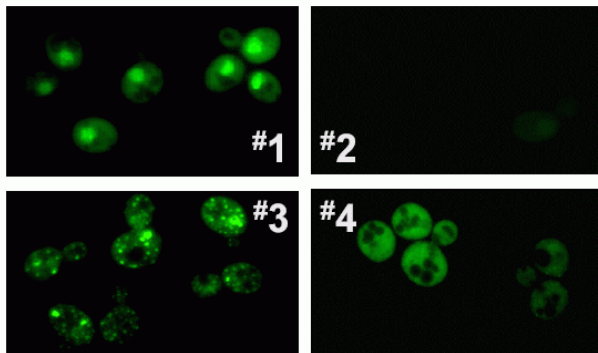
# Issue: Data Integration

- to understand and organize the set of AHR modifiers, Yao et al. considered other sources of data

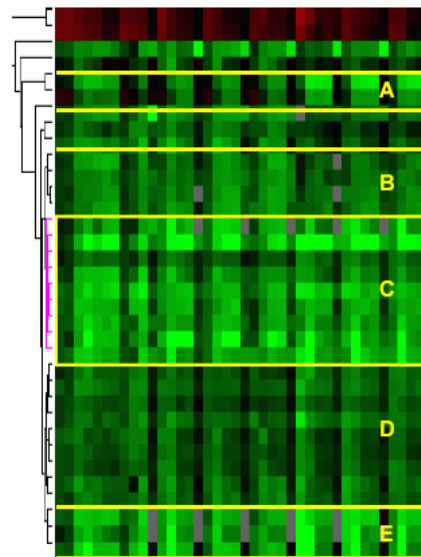
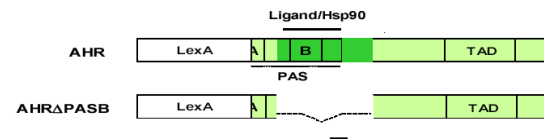


protein-protein interactions

localization of AHR-GFP in knockouts



AHR vs. AHR $\Delta$ PASB signaling in knockouts



clustering analysis of AHR pharmacology in knockouts

# Issue: Significant Variance in Amount of Literature Per Gene

- number of abstracts referencing specific yeast genes ranges from 0 to 1,000+
- how should we treat genes that have little/no text annotation?
- one approach: empirical Bayesian method

The diagram illustrates the empirical Bayesian method for estimating the number of abstracts for a gene  $g$ . The formula is:

$$\frac{a_{k,g} + M \left( \frac{a_k}{a} \right)}{a_g + M} \times A$$

Annotations with red arrows:

- $a_{k,g}$ : # abstracts  $k$  occurs in for gene  $g$
- $a_k$ : # abstracts  $k$  occurs in
- $a$ : total # abstracts
- $a_g + M$ : # abstracts for gene  $g$

# Conclusions

- text mining important for aiding in the annotation of high-throughput experiments
- a few important issues
  - identity uncertainty
  - what to return
  - heterogeneity
  - data integration
  - different amounts of literature for each gene

# Acknowledgments

NSF CAREER grant IIS-0093016

NIH/NLM grant 1R01 LM07050-01