

Coupled analysis of mRNA and protein profiles

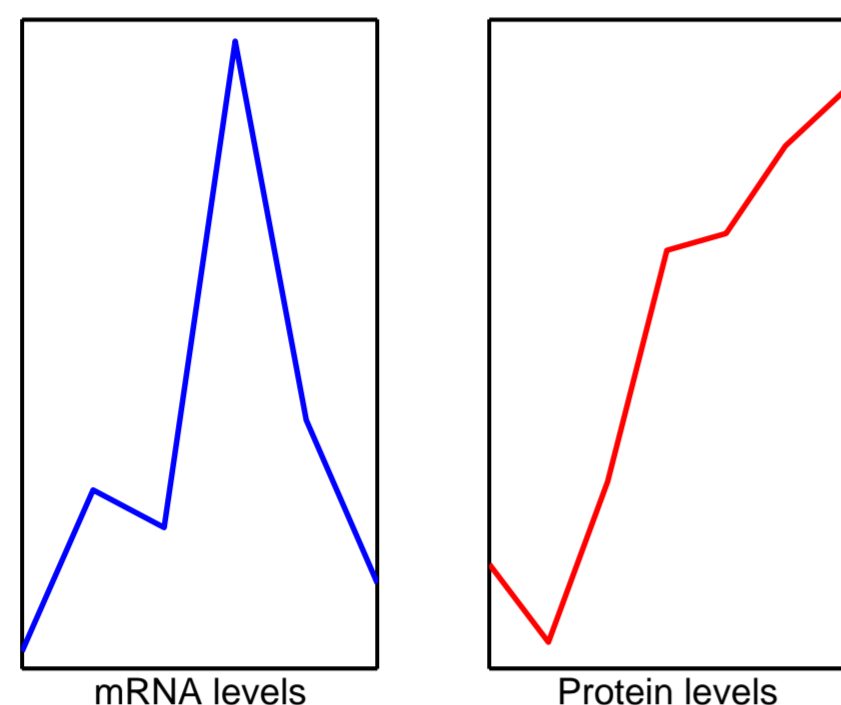
Simon Rogers and Mark Girolami: Computing Science, University of Glasgow

Walter Kolch: Beatson institute for cancer research, Glasgow

Katrina M. Waters, Tao Lui, Brian Thrall and H. Steven Wiley: Pacific Northwest Laboratory, USA

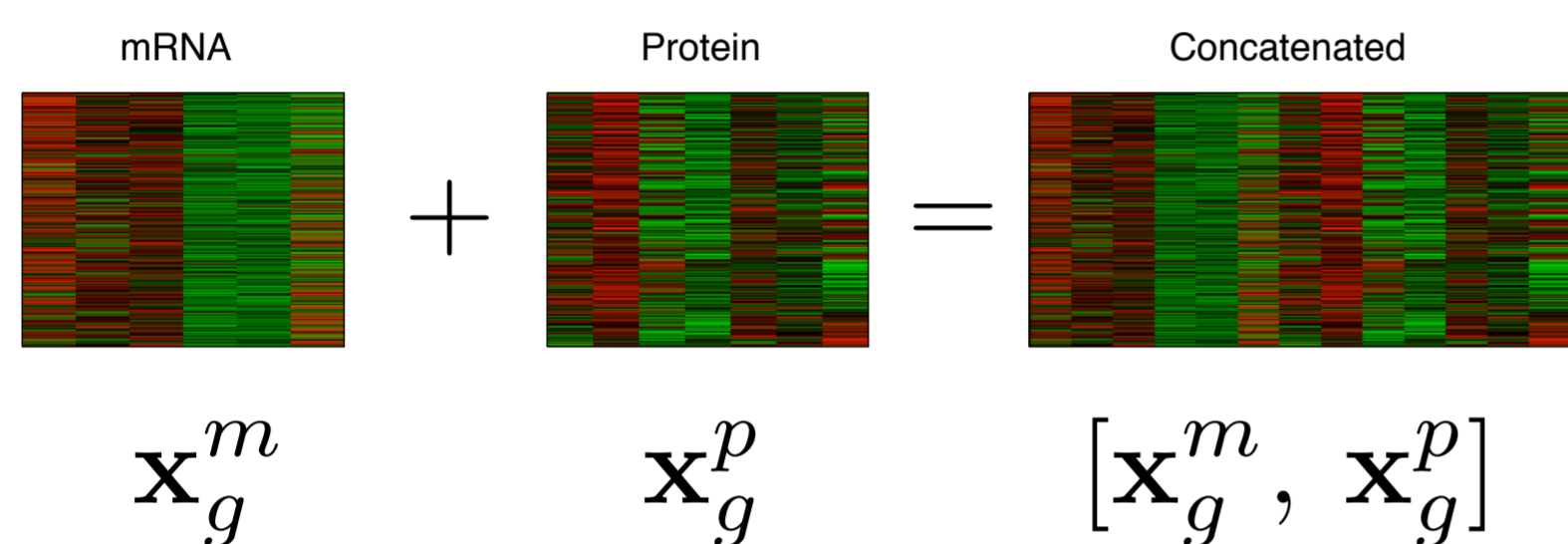
The data

- mRNA and protein profiles for ~ 500 human genes
- Measurements taken at various times after stimulation with EGF
- e.g. TLN1

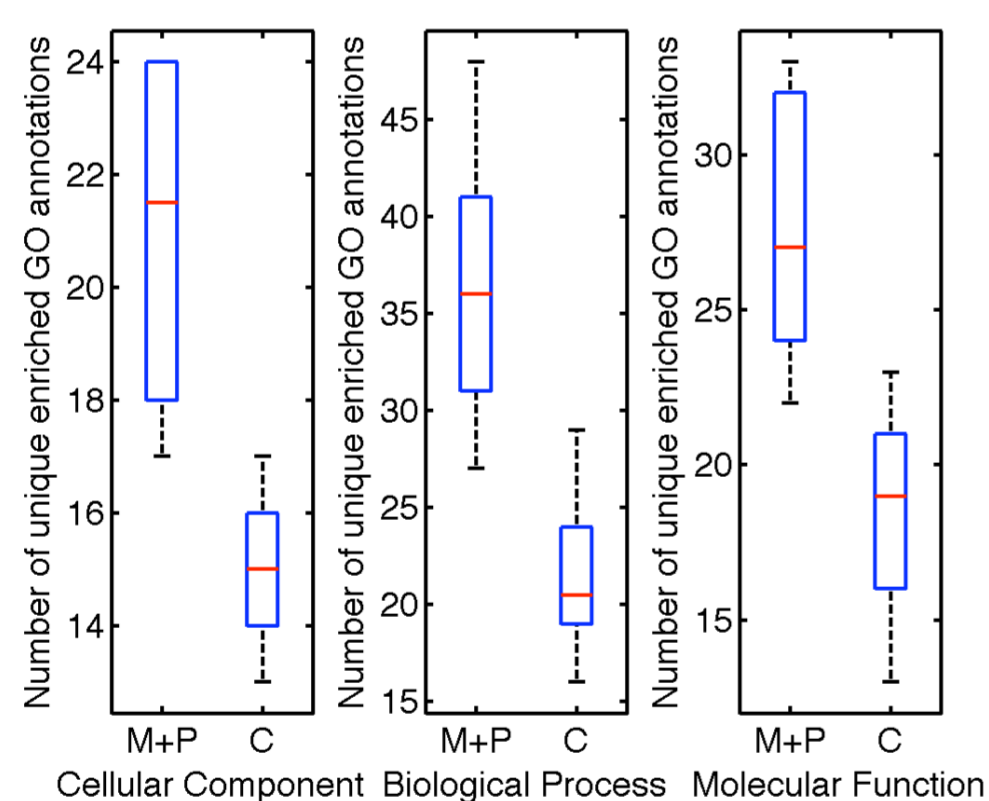


The problem

- Cluster analysis is a useful exploratory tool
- Clustering one dataset is easy - coupled datasets are harder
- Simple option is concatenation:



- But concatenation loses biological richness - fewer enriched Gene Ontology (GO) terms are found when clustering the concatenated data ('C' in figure below) than are found when clustering the two data types independently ('M+P').



The solution

We propose a pair of coupled K (mRNA) and J (protein) component Gaussian mixtures where the coupling is through parameterisation of the joint prior distribution on components $p(k, j)$. Particularly, we factorise the joint distribution as

$$p(k, j) = p(k)p(j|k)$$

where the $p(k)$ and $p(j|k)$ are model parameters to be inferred.

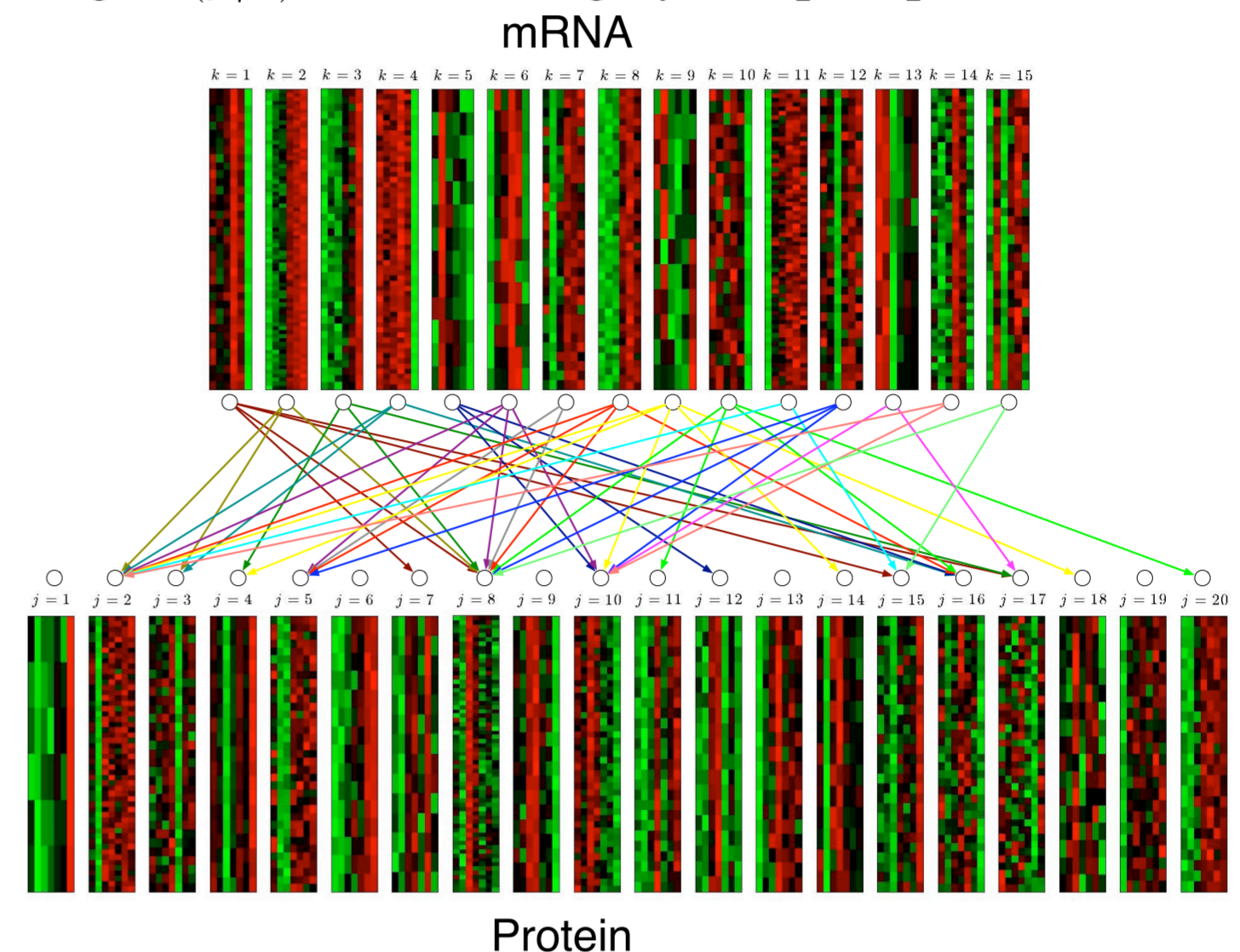
Likelihood function:

$$p(\mathbf{x}_n, \mathbf{y}_n | \dots) = \sum_k p(k) \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) \sum_j p(j|k) \mathcal{N}(\mathbf{y} | \mu_j, \Sigma_j)$$

- EM algorithm can be used to learn model parameters and mRNA and protein assignments
- Use multiple restarts to attempt to overcome local optima
- Note: special cases: Independent clustering, $p(j|k) = p(j)$ and concatenated $p(j|k) = \delta(j = k)$

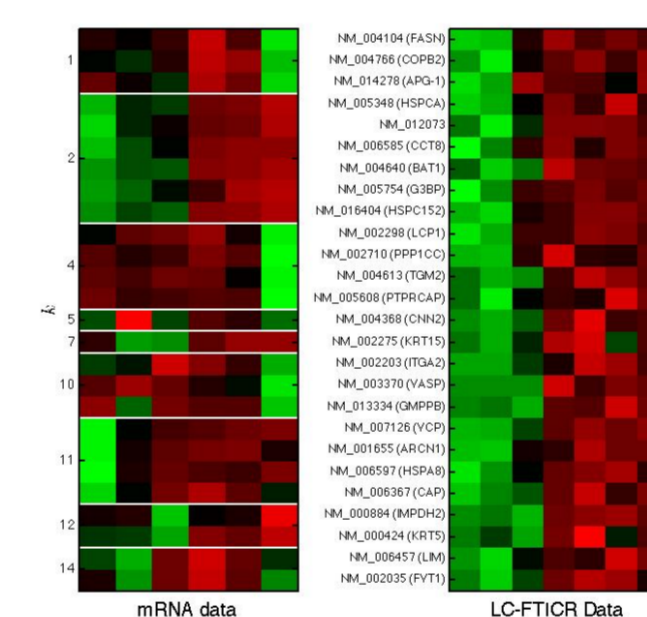
High level results

Visualisation of connectivity between mRNA and protein clusters (through $p(j|k)$) reveals a highly complex picture.

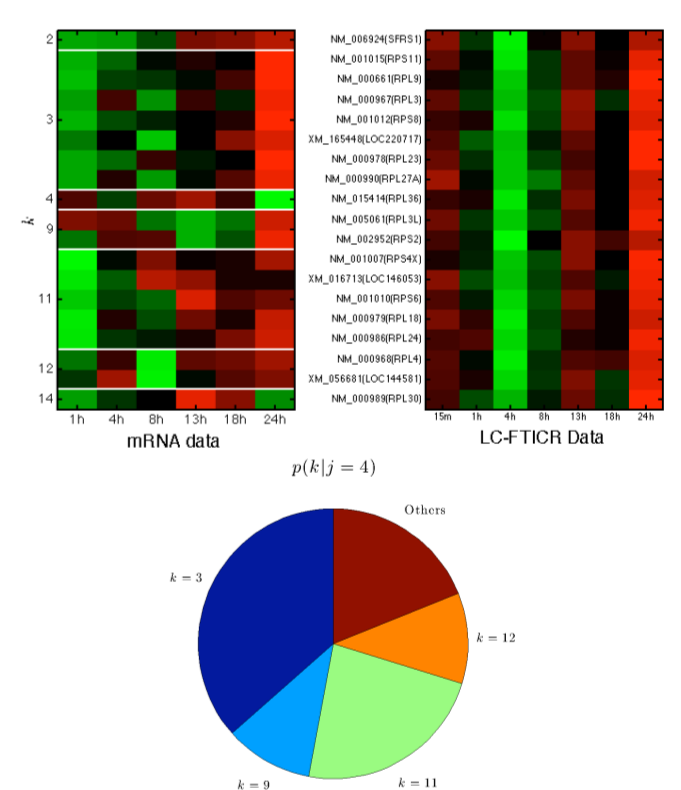


Low level results

Many small modules of genes with homogenous mRNA and protein profiles. For example, looking at one protein cluster we see small groups of very different mRNA profiles



One notable exception are the ribosomal proteins - a large group exhibiting homogenous expression at both mRNA and protein level. Such control is not surprising for a complex requiring such coordination.



Conclusions

- Relationship between mRNA and protein levels appears (in this dataset at least) to be highly complex/
- The coupled mixture model allows us to gain insight into the mRNA and protein levels individually (via the marginal clusterings) as well as the relationships between them.
- It is more flexible than resorting to concatenation and has uncovered many interesting biological phenomenon from this data alone.

Future work

- Further mining these results for interesting biology.
- More sophisticated mixture components.
- Alternative methods for parameterising $p(j, k)$.

e.g. currently developing model based on

$$p(k, j) = \sum_i p(i)p(j|i)p(k|i)$$

with infinite (DP) priors over i, j and k .