
Computational functional genomics

(Spring 2002: Lecture 18)

David K. Gifford

(Adapted from a lecture by Tommi S. Jaakkola)

MIT LCS and AI Lab

gifford@mit.edu

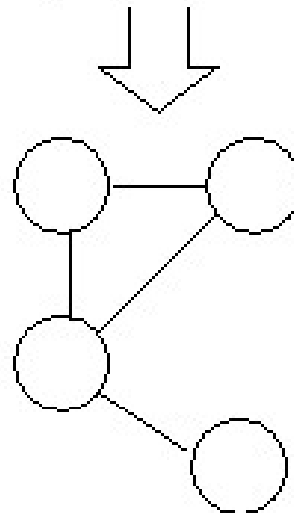
Information fusion

- Information fusion refers to a combination of multiple sources of information
- Possible settings:
 - expression + location analysis
 - expression + sequence analysis
 - expression + location + sequence analysis
 - across species comparison
etc.
- Multiple **independent constraints** can dramatically increase the significance of otherwise elusive effects
- How can/should we combine the sources?

Example

- Fusion via graph representations

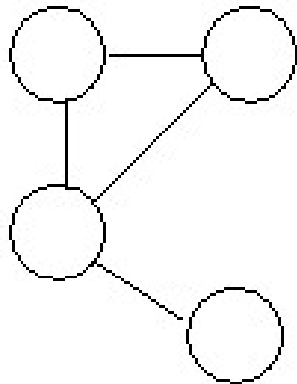
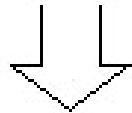
Expression data



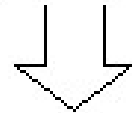
Example

- Fusion via graph representations

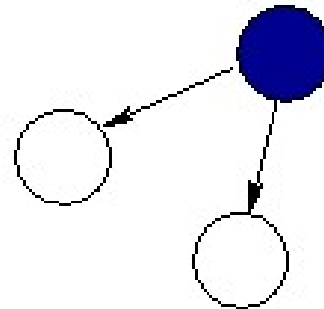
Expression data



Sequence/location data

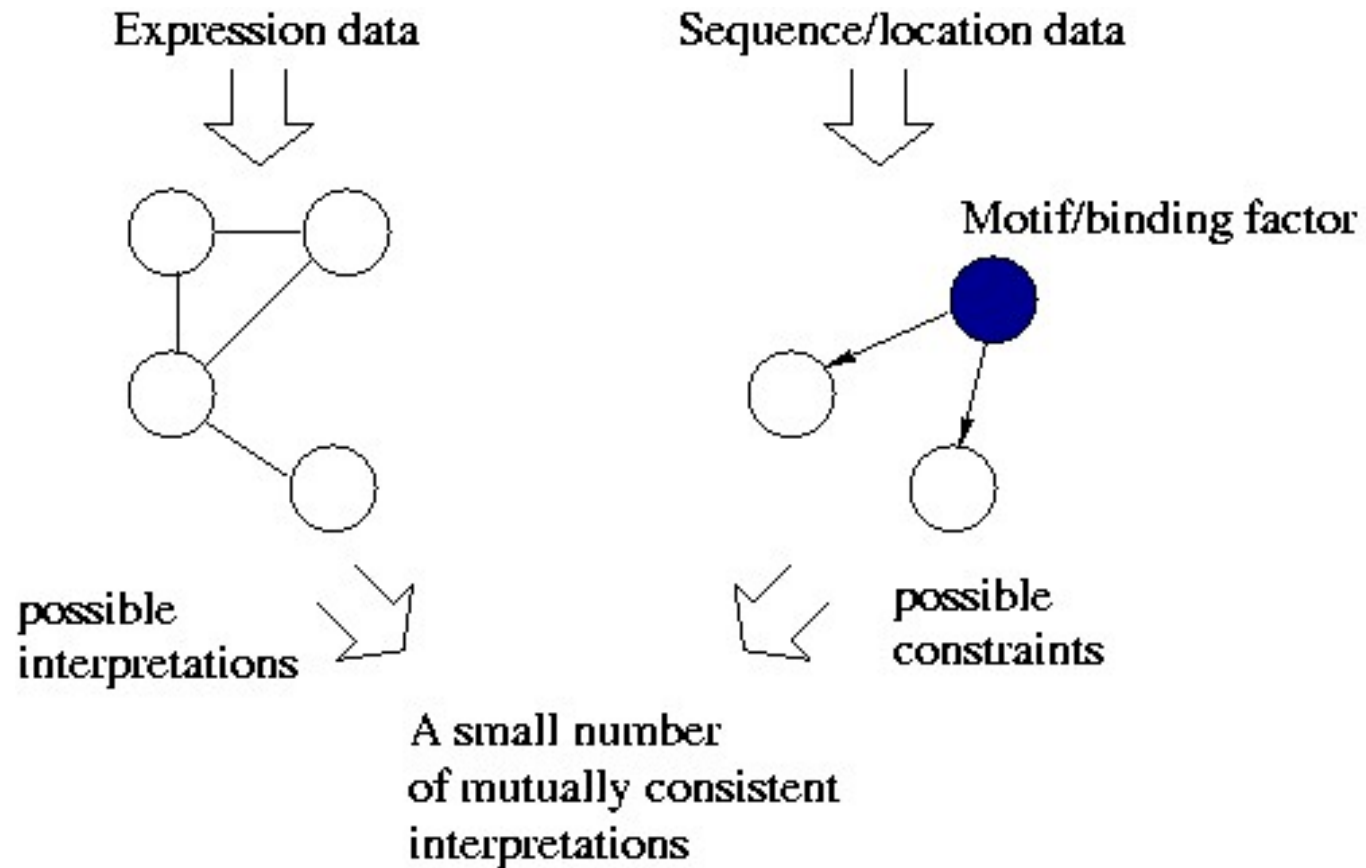


Motif/binding factor



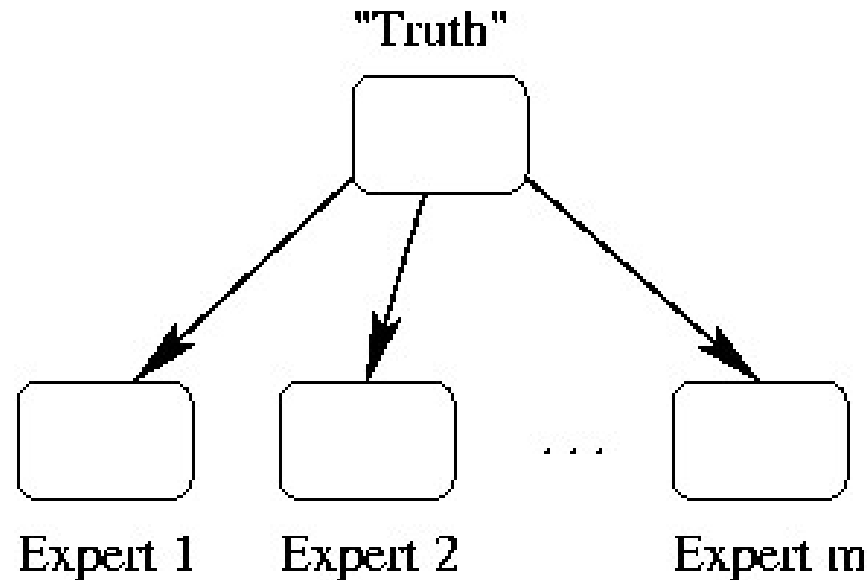
Example

- Fusion via graph representations



Model based information fusion

- Multiple “experts”



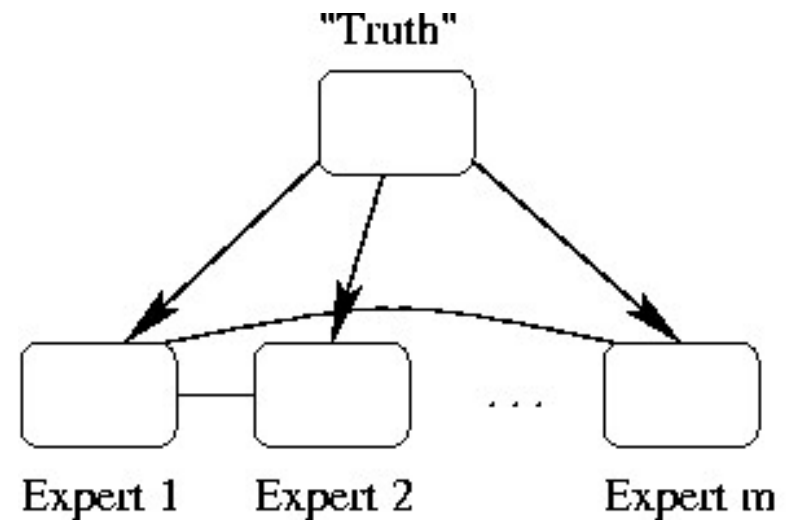
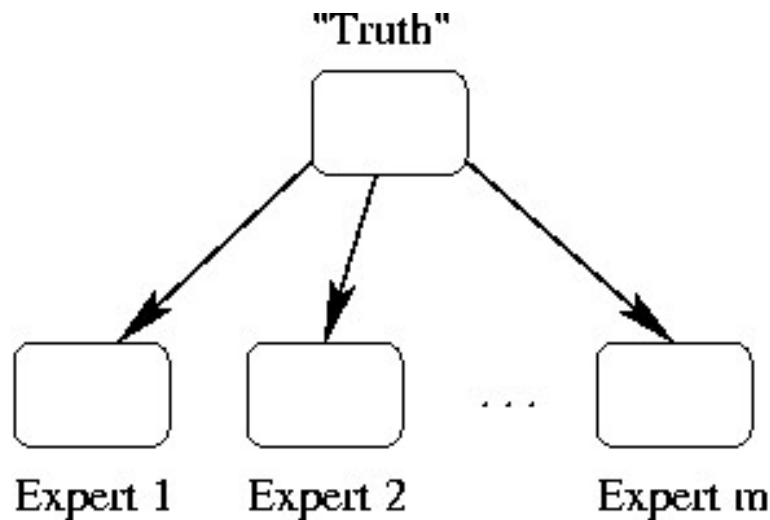
possible experts: sequence motif model, model of the gene expression, etc.

possible “truths”: regulated/not, binding/not, etc.

- The decision maker must be able to model what the experts say given the “truth”

Model based information fusion

- Multiple “experts”



- The experts may be dependent or independent of each other
 - mutually dependent experts makes it necessary to build a joint model of expert opinions
 - what might cause dependence?
- Simple setting: **independent** experts

Model based information fusion

- Example: independent* “experts”

Suppose we wish to deduce whether gene 1 regulates genes 2

Expert 1: p-value (p_1) for the comparison of $P(x_2|x_1)$ vs. $P(x_2)$ based on expression data

Expert 2: p-value (p_2) for the event that the promoter of gene 2 has a sequence motif corresponding to gene 1

- Assuming these “experts” are independent, combining their predictions can substantially improve the p-value... but how exactly?

$$p^* = p_1 \times p_2 ? \quad (1)$$

NO!!!

Combined p-value

- Simple example: testing the same thing (H_0 vs. H_1) in two different ways using two **independent** datasets

test statistic

dataset 1: $T_1(X) \sim \chi_{\nu_1}^2$

dataset 2: $T_2(Y) \sim \chi_{\nu_2}^2$

What is the combined p-value?

$$T_1(X) + T_2(X) \sim \chi_{\nu_1 + \nu_2}^2 \quad (2)$$

and so we can evaluate the p-value as follows (in MATLAB)

$$p = 1 - \text{chi2cdf}(\hat{T}_1 + \hat{T}_2, \nu_1 + \nu_2) \quad (3)$$

For example: if $p_1 = p_2 = 0.1$ for the two tests when carried out independently (with $\nu_1 = \nu_2 = 1$), then $p = 0.0668$.