Methods of Genomic Data Fusion

An Overview

Konstantin Tretyakov (kt@ut.ee)

May 4, 2005

Motivation: Why combine data?

- "High-throughput" methods are increasingly popular.
- They produce lots of *indirect*, *generic* and *partial* data.
- The data produced is often *very noisy*.
- Combination of *different kinds of evidence* might improve statistical significance and reduce the noise.

What is Data Fusion

- Many kinds of analyses obtain results by combining *certain* different datasets in one way or another.
- We are, however, interested in approaches that can *scale* to integrating different and *nearly arbitrary* kinds of data.

Outline

- Data fusion "in general"
- Case study: Bayes nets
- Case study: kernel methods

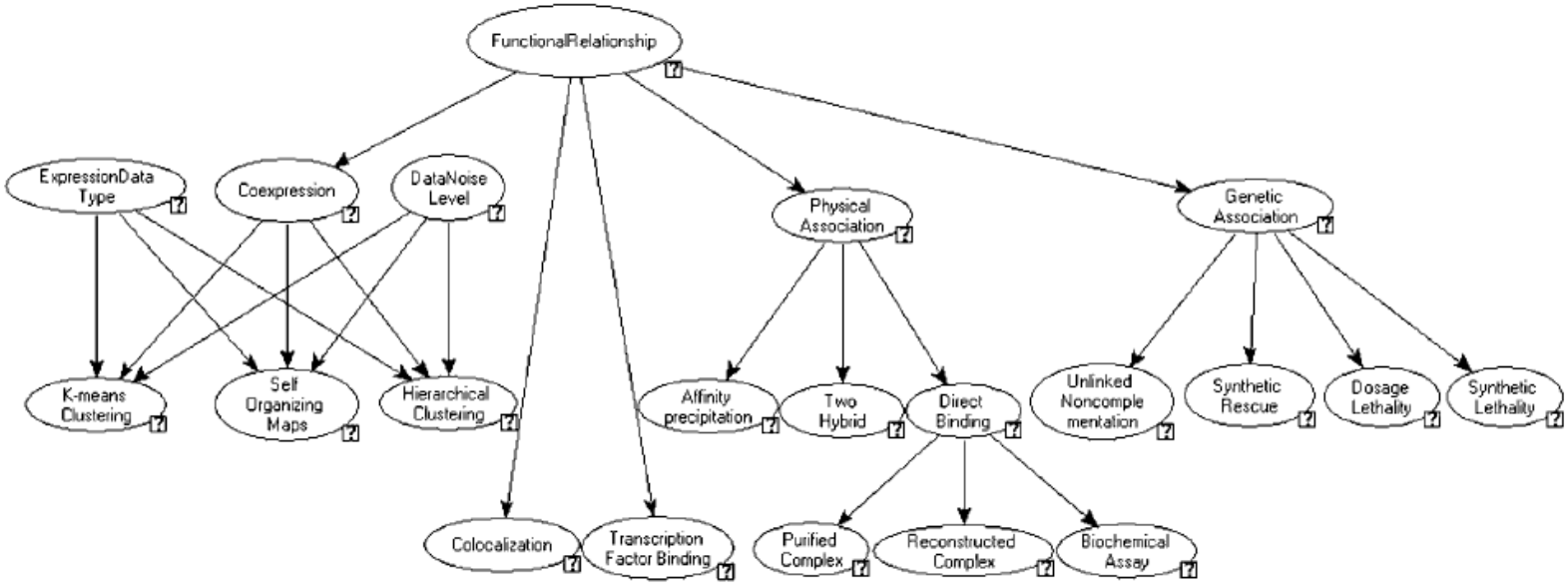
A General Data Analysis Strategy

- Any data analysis looks like that:
  1. Represent the data in an appropriate form
  2. Perform a certain statistical procedure
  3. Convert the results to a useful form
- Data fusion is easiest at step 1 (*early integration*) or 3 (*late integration*).

Common Representation

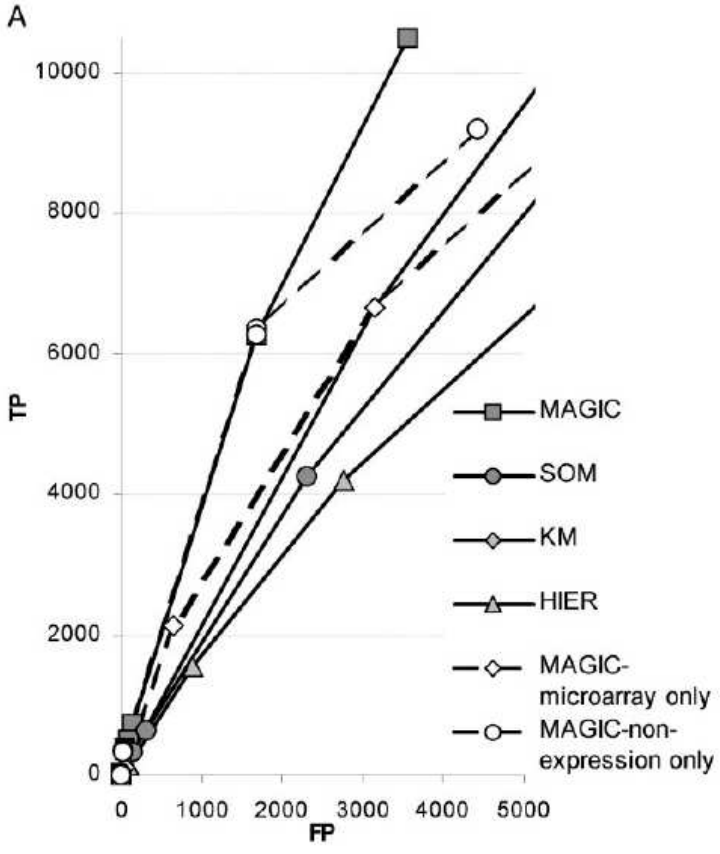
- In most cases the choice of common data representation pretty much defines the algorithm to be used:
  - Data vectors/matrices
  - Distance metrics/Kernel matrices
  - Graphs/binary predicates

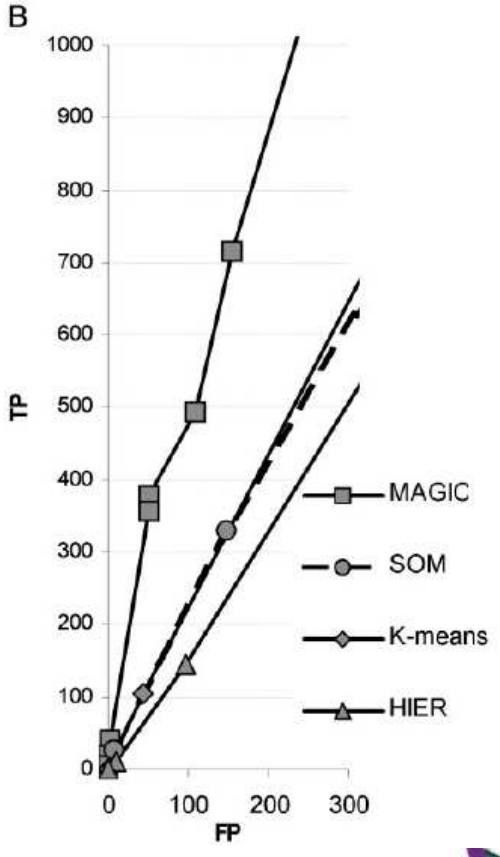
Example: Bayes Nets



A Bayesian framework for combining heterogeneous data sources for gene function prediction. Troyanskaya et al. PNAS, 2003.

Example: Bayes Nets





Example: Kernel Methods

- Often *pairwise inner products* suffice for analysis.
- $\Rightarrow$ the *kernel matrix* as a common representation.
- A variety of kernels exist for all kinds of data.
- Different kernels can be combined by summing.

Kernels

- Protein sequence kernels
- Protein interaction kernels
- Gene expression kernels

Kernel combination

- Different kernels can be weighed and summed:

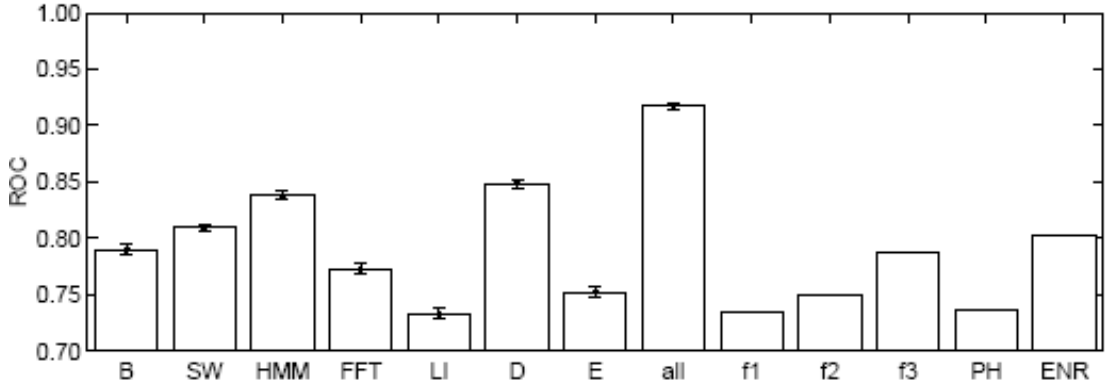
$$\mathbf{K} = \lambda_1 \mathbf{K}_1 + \lambda_2 \mathbf{K}_2 + \cdots + \lambda_n \mathbf{K}_n$$

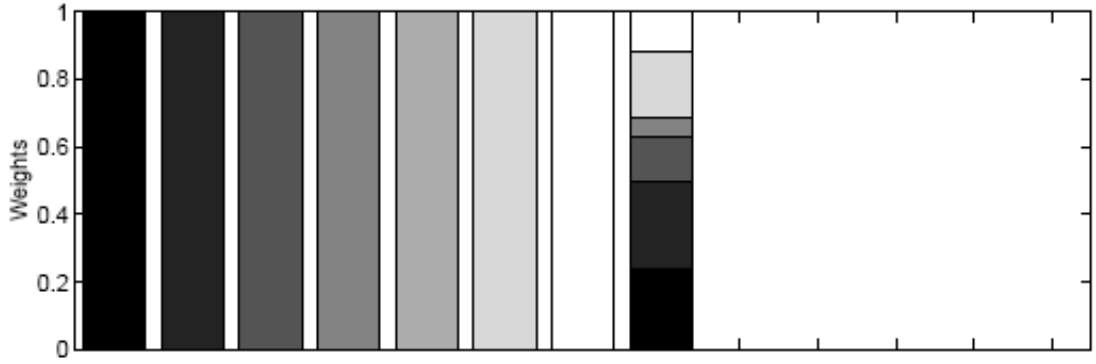
- *Lancriet et al.* present an SVM classification algorithm that can figure out the weights automatically.

Example: Membrane proteins

- Input: the collection of different kernel matrices + CYGD annotations.
- Procedure: split the data into training/test sets, predict values for test set.
- Evaluation: compare the performance to alternative methods.

Example: Membrane proteins





More applications

- Yeast protein function prediction
- Biological pathway prediction
- Protein-protein interaction prediction
- Clustering & search for motifs

Summary

- Although a recent development, there are already several frameworks for data fusion
- The major approaches are *clustering*, *Bayesian inference* and *kernel methods*.

Hope you don't have any questions...