Computational modeling of transcriptional regulation using multiple heterogeneous data sources

Harri Lähdesmäki

Computational Systems Biology Group
Department of Signal Processing
Tampere University of Technology

August 16, 2008

Overview of research topics

- learning regulatory mechanisms from gene expression and protein level data
  - transcriptional regulatory networks
  - signaling pathways
- biological sequence analysis ★
  - modeling/predicting binding of transcription factors (or other molecules)
  - analysis of mutations (e.g. SNPs) and their effect on transcriptional regulation
- statistical methods for biological data fusion ★
- modeling high-throughput measurements

Joint work with collaborators

- TUT
  - **Xiaofeng Dai**
  - **Kirsti Laurila**
- ISB, Seattle
  - **Alistair Rust**
  - **Ilya Shmulevich**
  - Matti Nykter
  - Vesteinn Throsson

1. Motivation

2. Probabilistic TF binding prediction & data fusion
   - Computational methods
   - Results
   - Recent improvements

3. Effects of regulatory mutations on TF binding
   - Disease related mutations
   - Simultaneous and competitive binding of multiple TFs

4. Incorporating expression data into TF binding prediction
   - BCL6 case study

Motivation: TF binding

- transcriptional regulation (TR) is a central control mechanism
- TR involves transcription factors (TFs) that control gene expression by binding regulatory DNA in a sequence specific manner

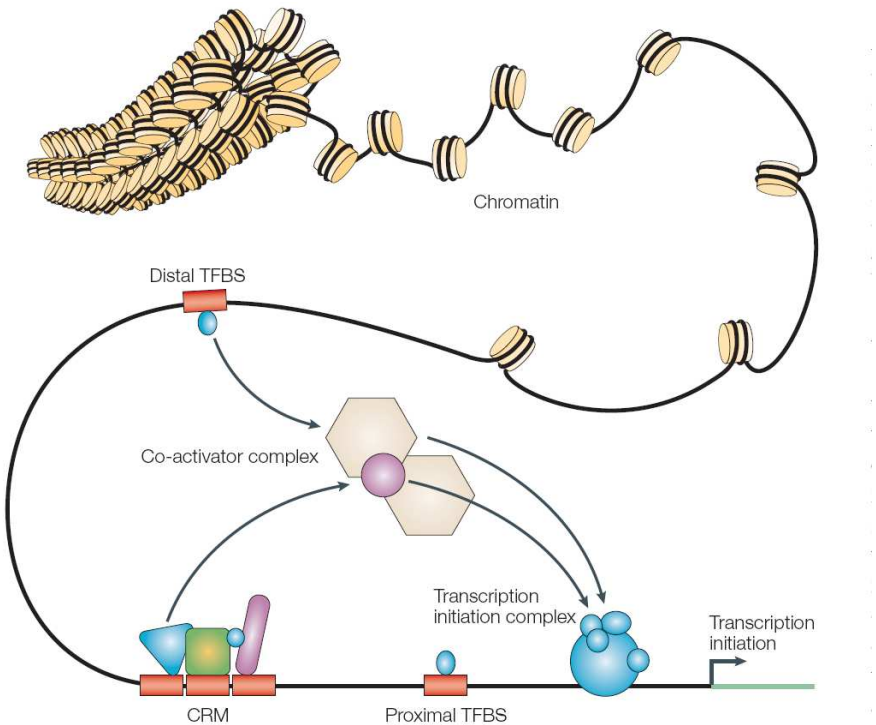


Figure from (Wasserman and Sandelin, 2004)

Motivation: TF binding prediction

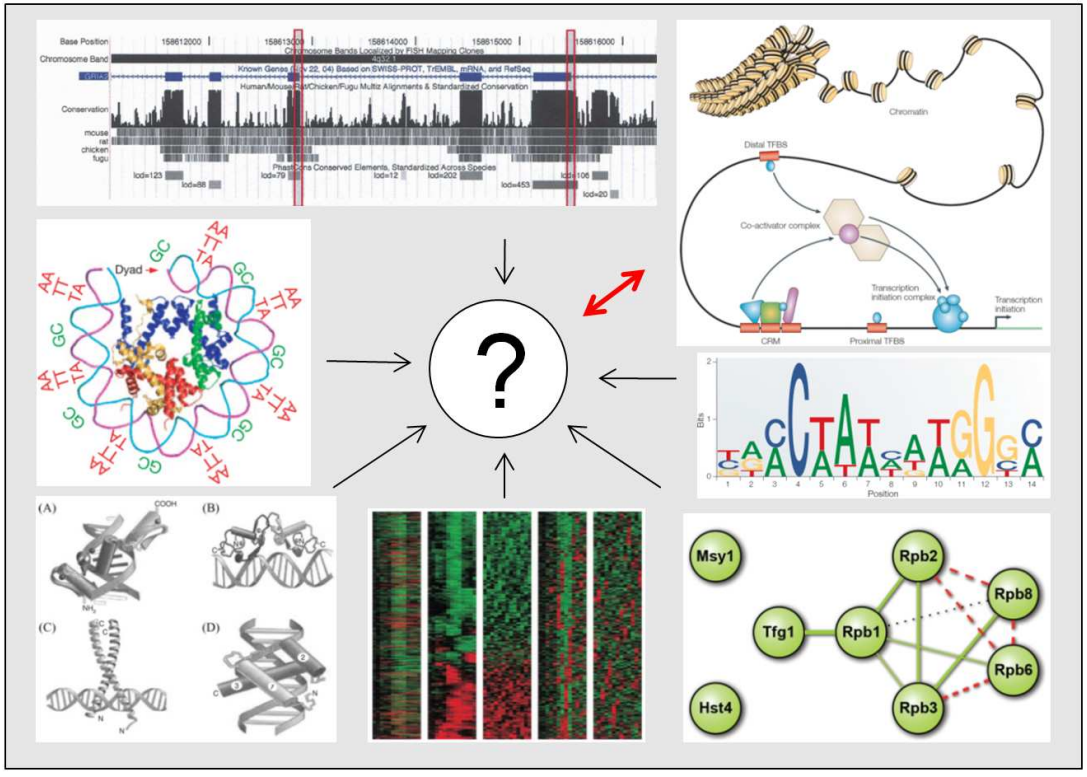
- revealing regulatory mechanisms is one of the key problems in understanding genome-wide transcriptional regulation
  - regulatory code
- genome-wide binding of transcription factors (TF) to gene promoters is largely unknown

$\rightsquigarrow$ computational analysis of protein-DNA binding is important

  - motif discovery
  - TF binding prediction
- prediction problem is becoming increasingly important

Motivation: data fusion

- each information source gives only a *partial* and *noisy* view of a biological system
- probabilistic analysis of multiple heterogeneous information sources



Part I:

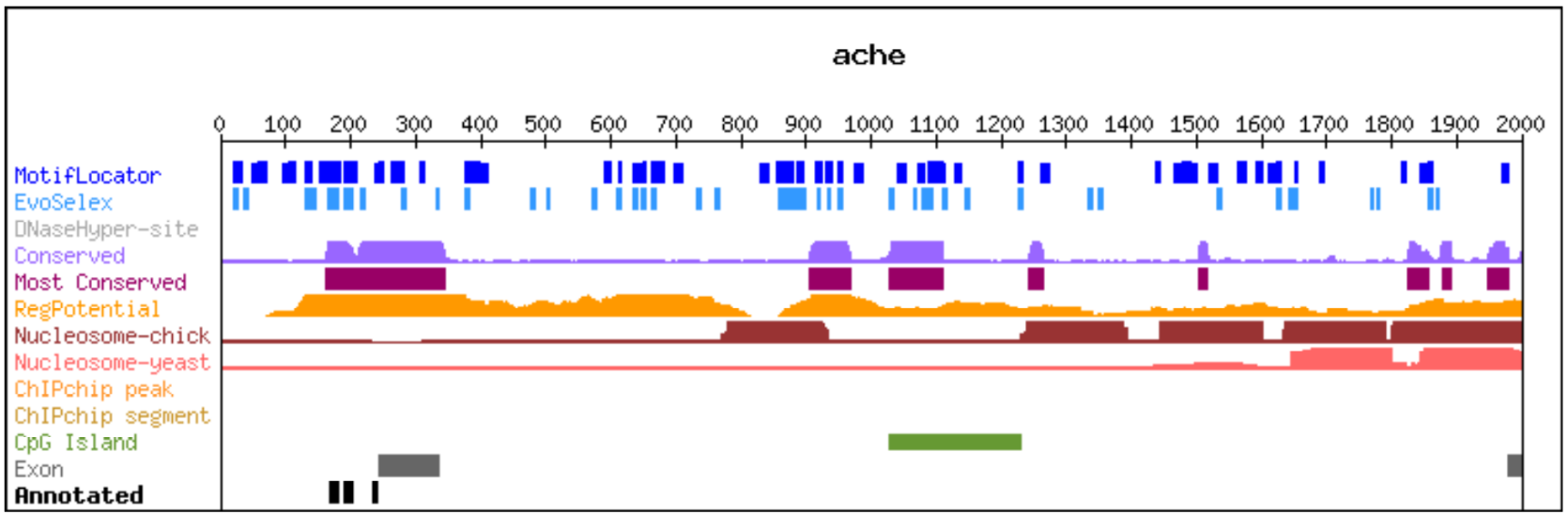
Probabilistic TF binding prediction & data fusion

TF binding prediction: our approach

- *in silico* genome-wide TF binding prediction
- the standard binding site prediction tools are based on hypothesis testing (i.e., *p*-values and "present" / "absent" calls) and have high false positive rate
- **our goal:** to develop a method that
  1. is probabilistic in nature (outputs probabilities)
  2. analyzes each gene promoter as a whole: probability that a TF binds to
     - a regulatory region of a gene
     - a particular location in a promoter
  3. provides a principled way of combining multiple data sources
     - evolutionary conservation, nucleosome positioning, CpG islands, DNase hypersensitive sites, ChIP-chip data, DNA duplex stability, regulatory potential, etc.
- (Lähdesmäki H, Rust AG & Shmulevich I, *PLoS ONE*, 2008)

Test set

- a merger of annotated TF binding sites for mouse from the ABS and ORegAnno databases (47 promoters)

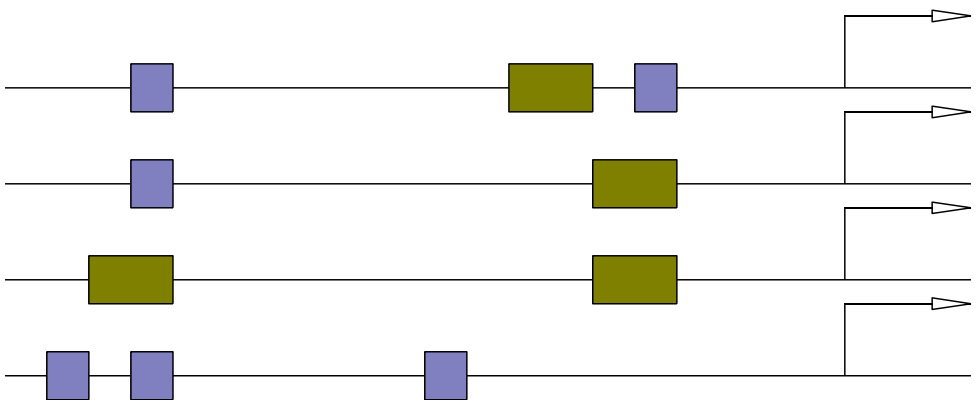


- performance evaluation using receiver operating characteristics (ROC)
  - true positive rate versus false positive rate
  - area under the ROC

Modeling framework

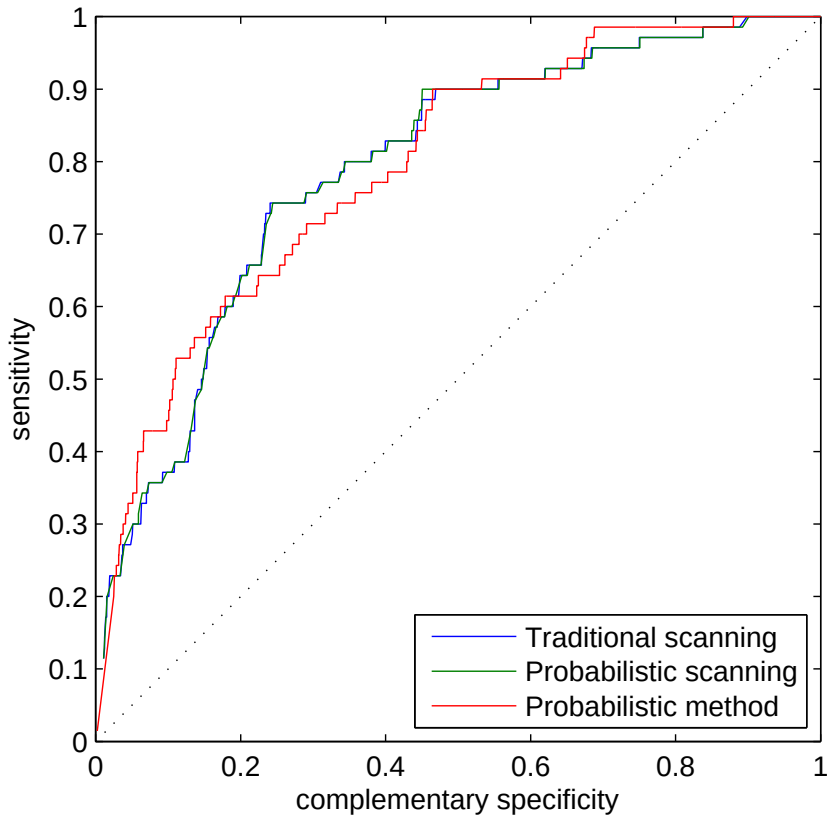
- standard building blocks (choice is arbitrary though)
  - position specific frequency matrix (PSFM)
  - Markovian background model
- a key unknown quantity is the number of binding sites Q
  - if $Q = 0$, then "no binding"
  - if $Q > 0$, then "binding"
- probabilistic modeling

$$P(Q|S, \Theta, \phi) \propto P(S|Q, \Theta, \phi)P(Q)$$



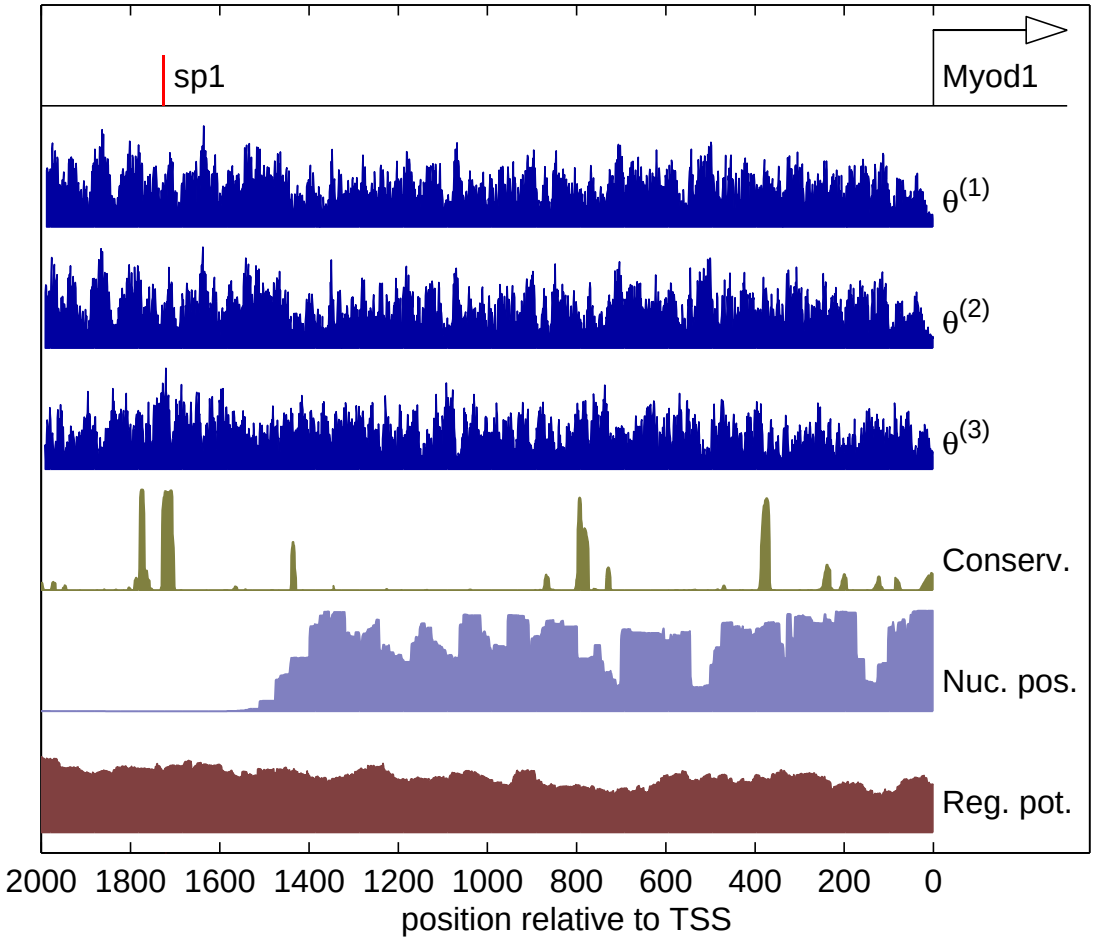
Results: without data fusion

- Performance assessment using the mouse test set
- Comparison with traditional motif scanning



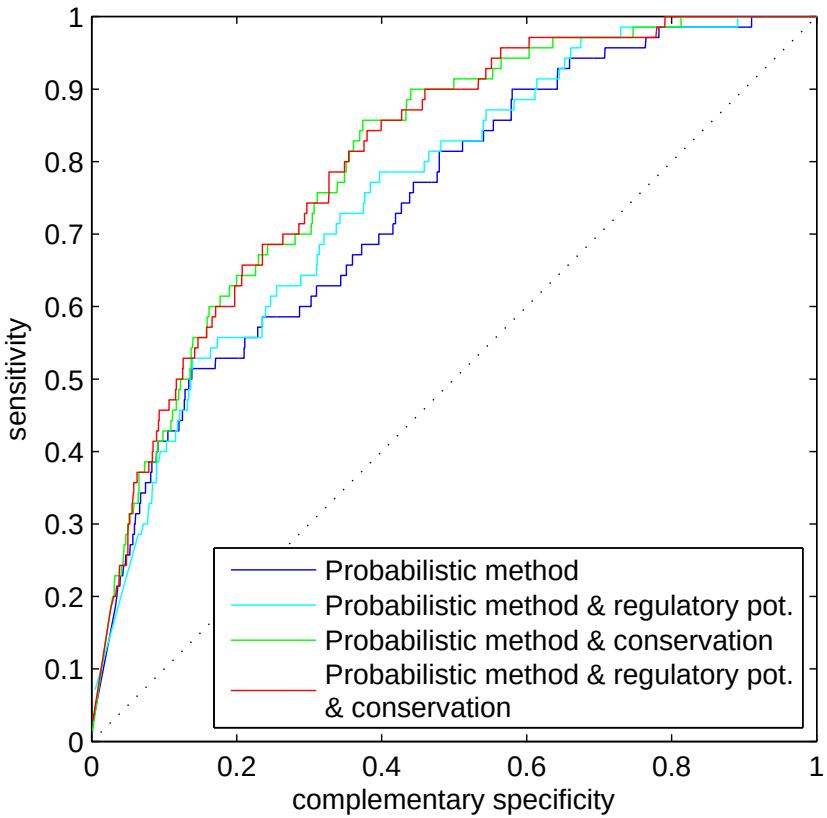
Data fusion example

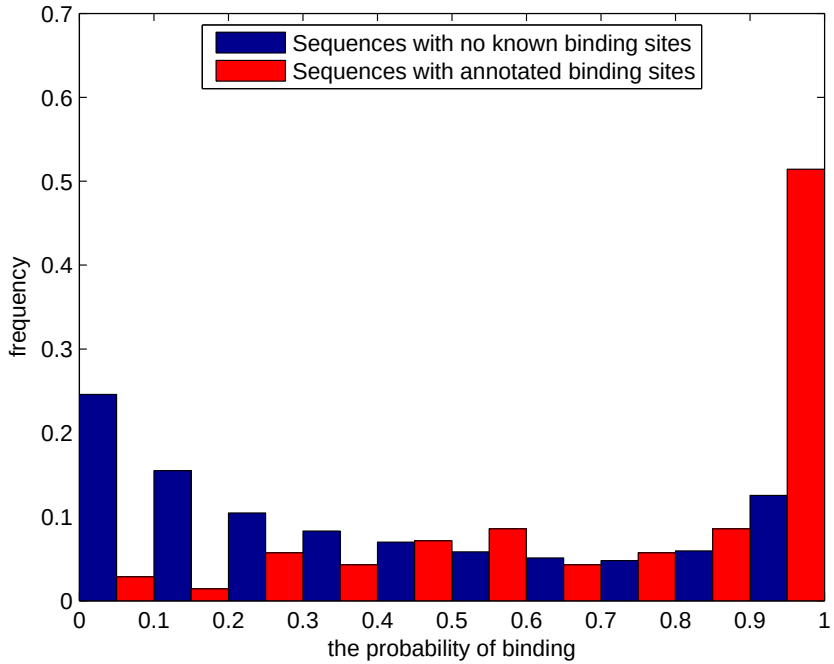
- an illustration



Data fusion results

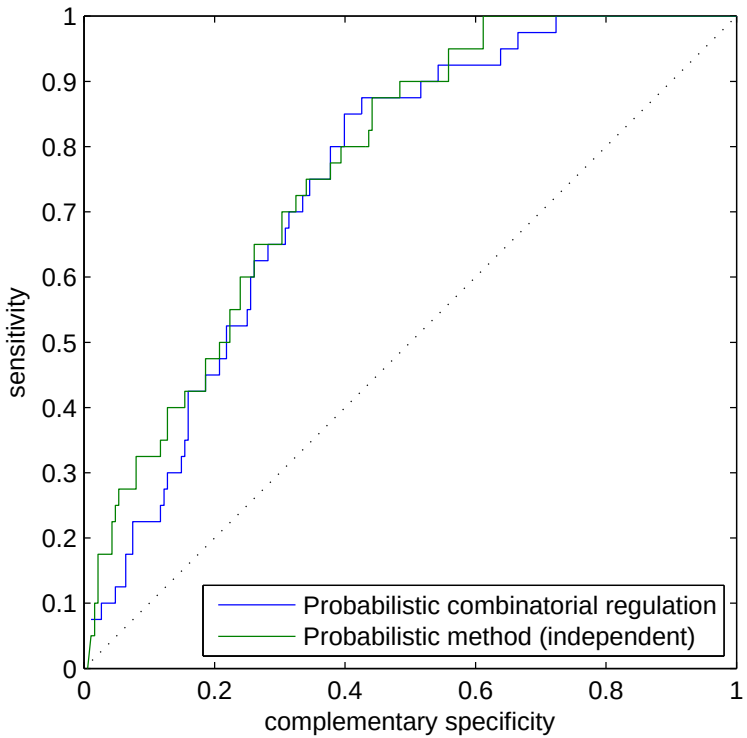
- combining multiple information sources improves binding predictions

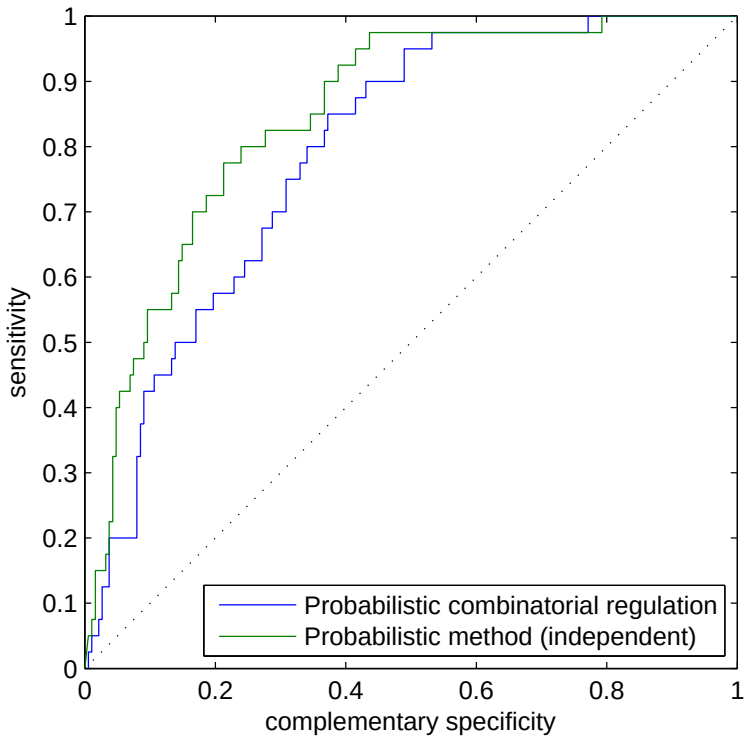




Data fusion results: combinatorial regulation

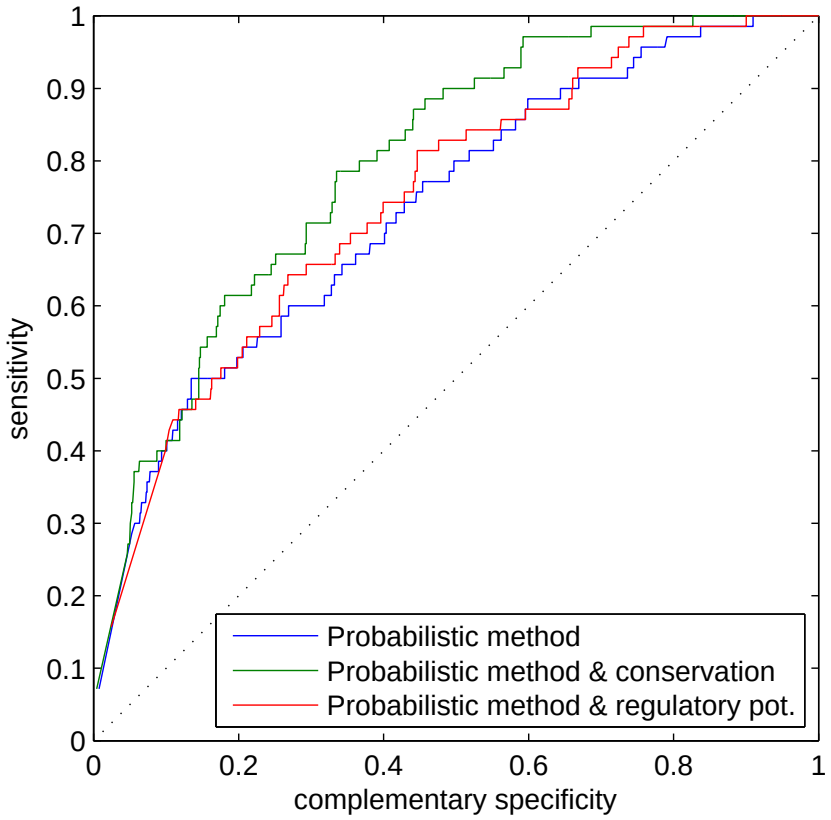
- TR is typically controlled combinatorially by multiple TFs
- example on pairwise regulation:
  - probability that both TFs have at least one binding site





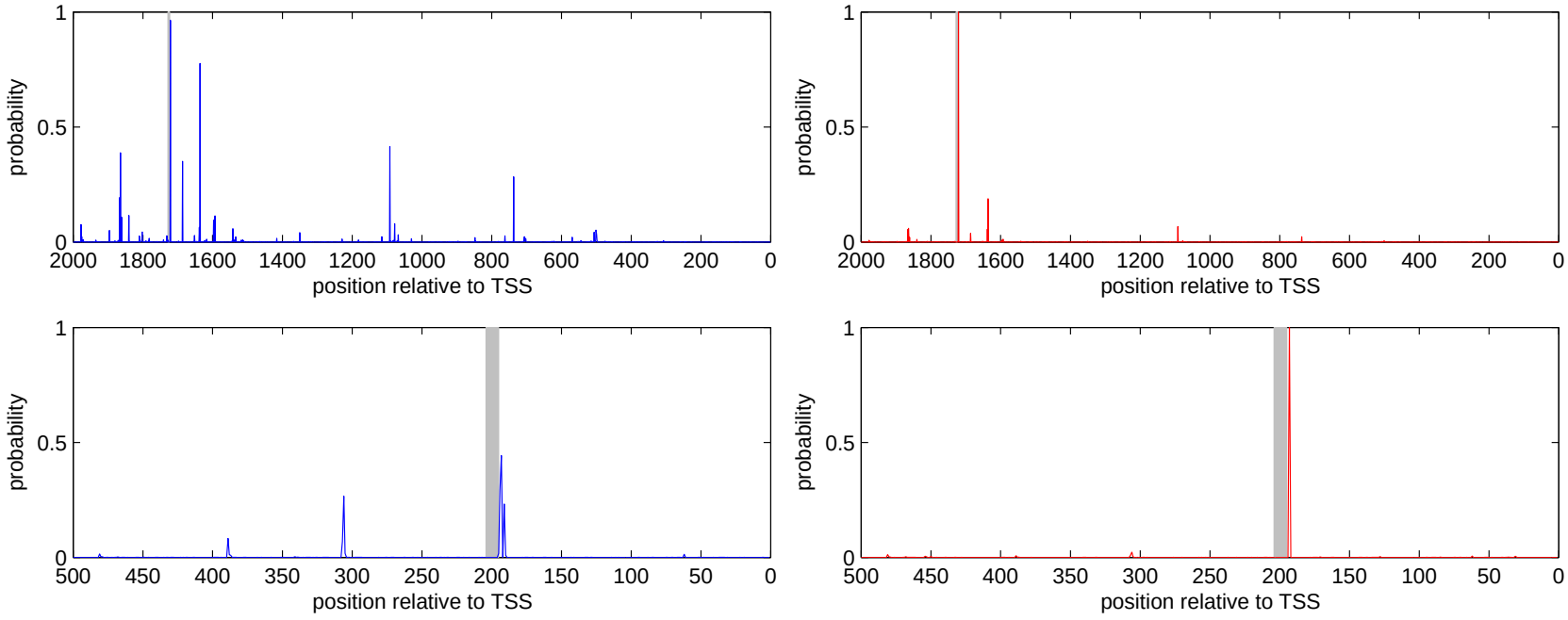
Data fusion results: double stranded DNA

- method can equally well use both strands of DNA



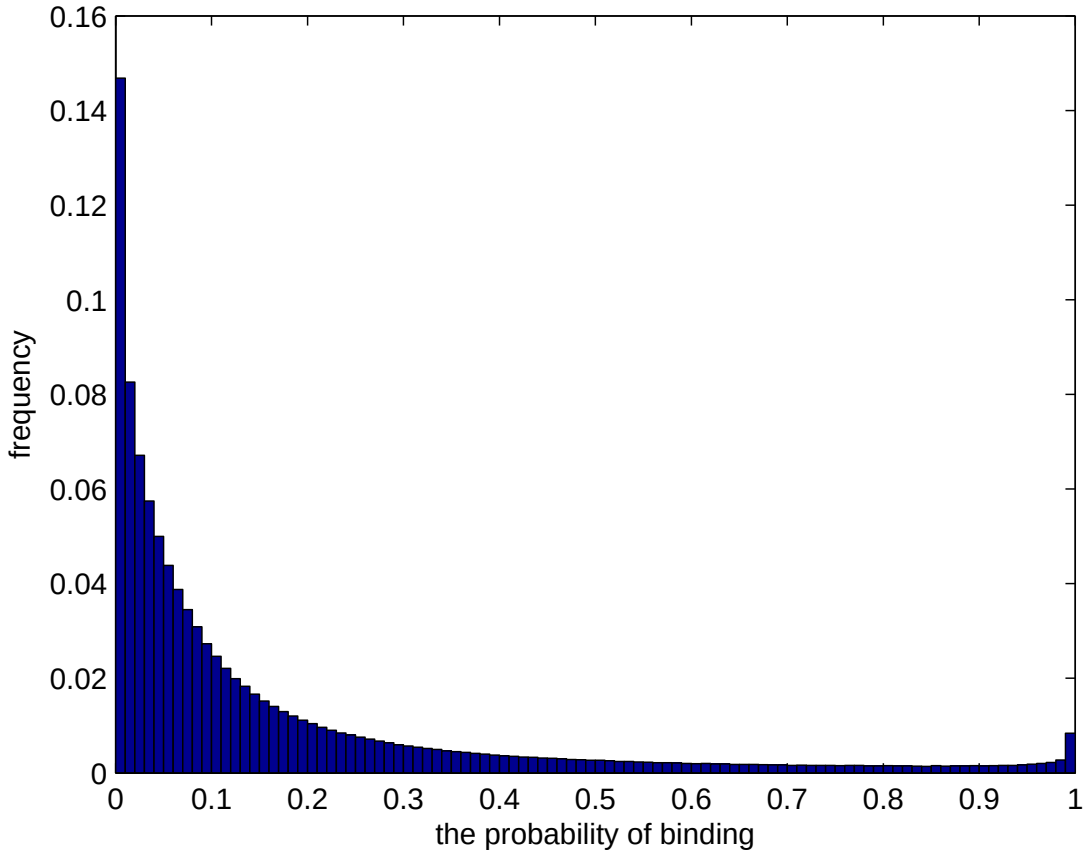
Data fusion results: higher resolution

- predict binding positions at single base pair resolution
- estimate the probability of binding at a location i



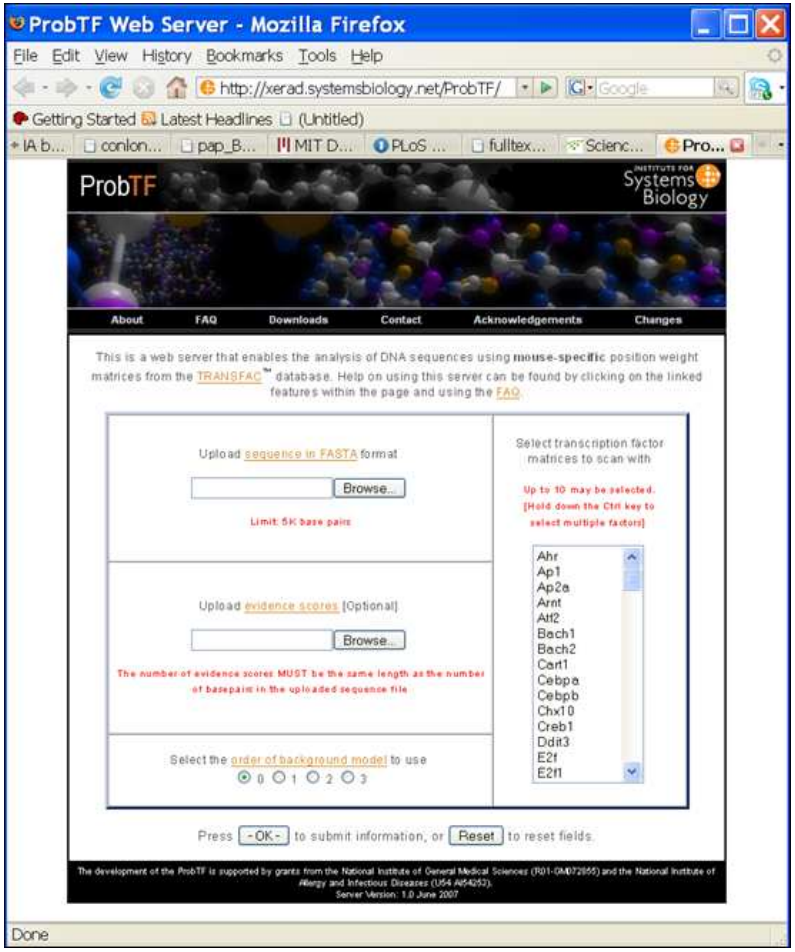
Genome-wide analysis

- binding predictions for all mouse promoters ($\sim$ 20000) and all mouse TFs in TRANSFAC (266) combined with conservation data



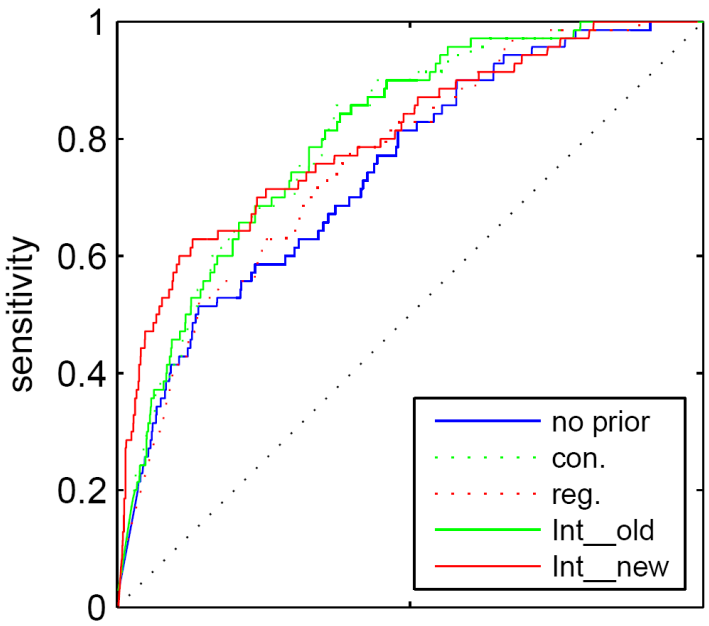
Web tool

- open source and accessibility (web tool)



Summary & improvements

- data fusion improves binding prediction at least moderately
  - our statistical data fusion is not optimal?
  - data is not good?
- (Dai, X. and Lähdesmäki, H., manuscript)
- an improved data fusion method



Better nucleosome data

- e.g. better nucleosome data (=predictions) are now available

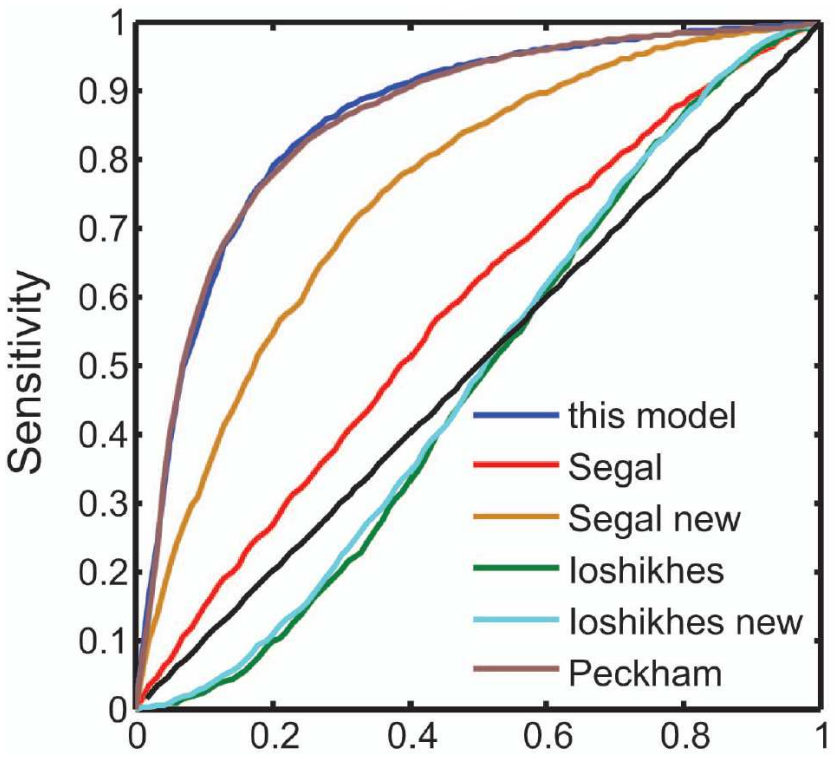
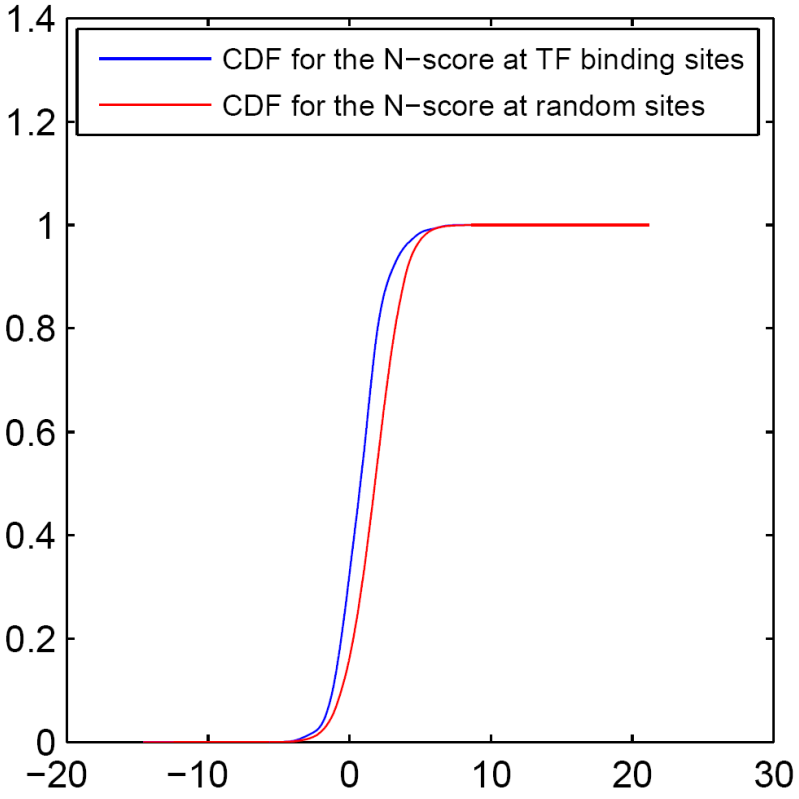
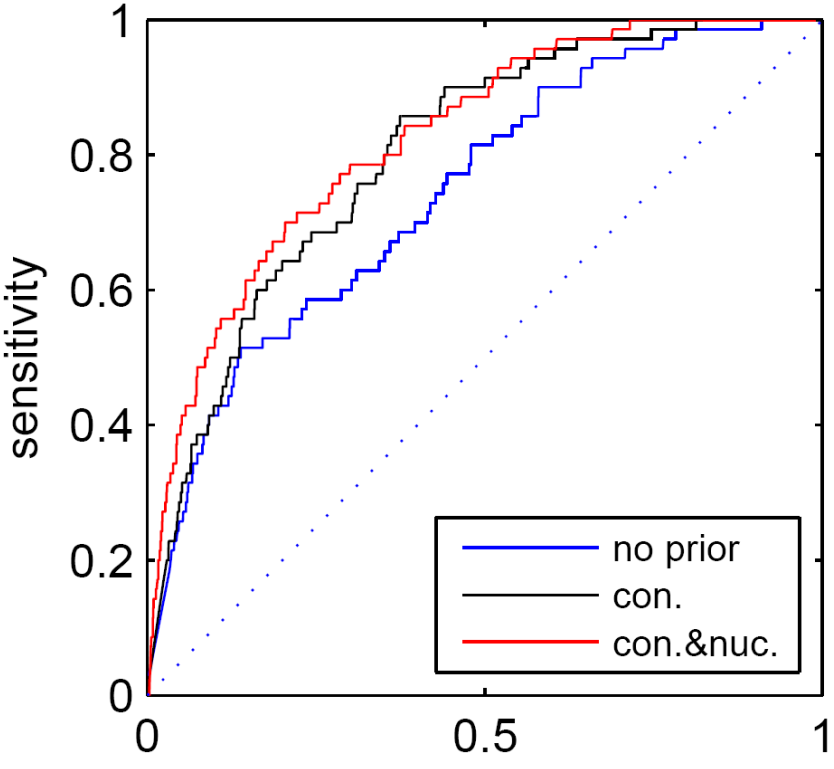


Figure from (Yuan & Liu, 2008)

Better nucleosome data

- better nucleosome data gives better results





DNA duplex stability data

- evidence that TFs bind in
  - single-strand manner (e.g. FBP protein in human)
  - double-strand manner (crystal structures have revealed interactions with both strands)
- A computational tool for predicting DNA duplex destabilization energy (Benham, 1992; Bi & Benham, 2004)

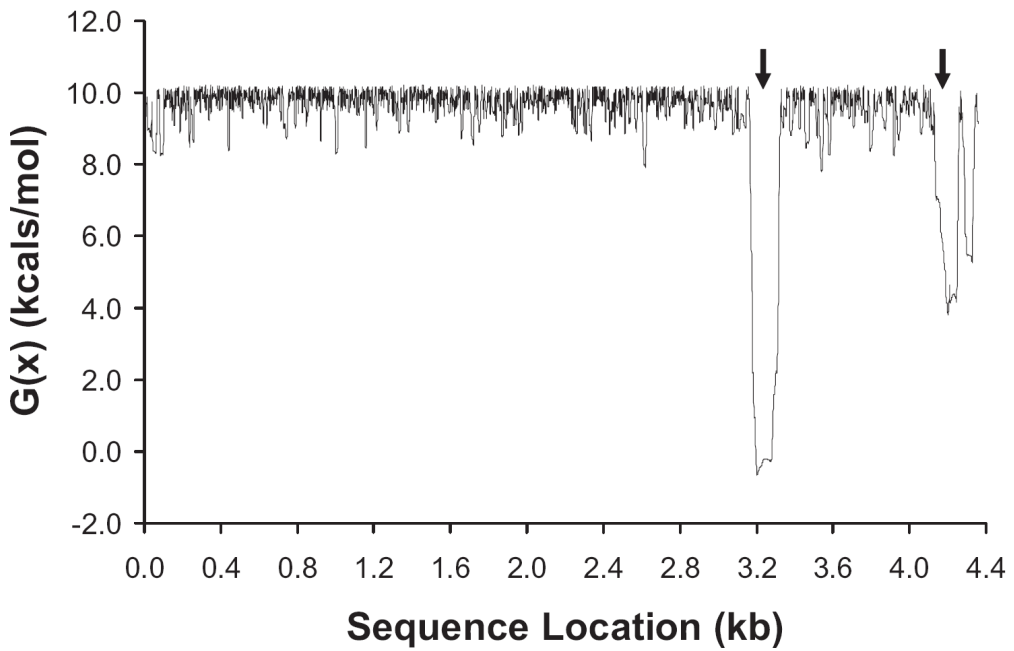
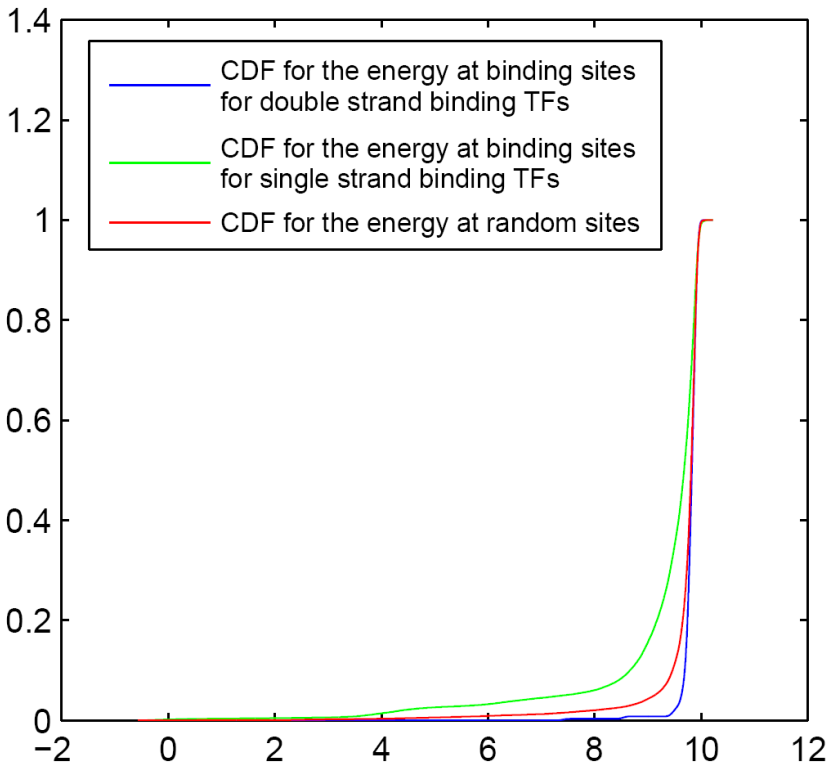
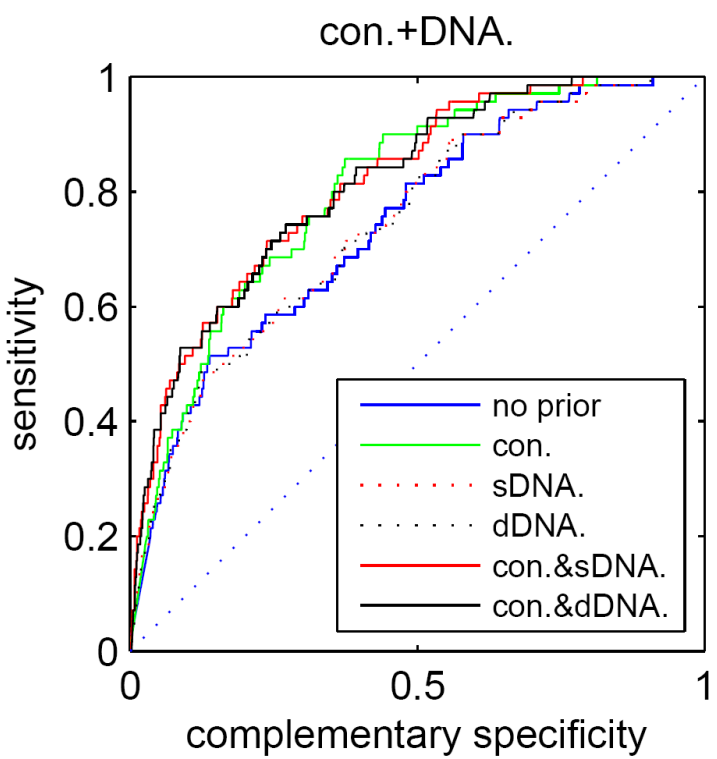


Figure from (Bi & Benham, 2004)

DNA duplex stability data

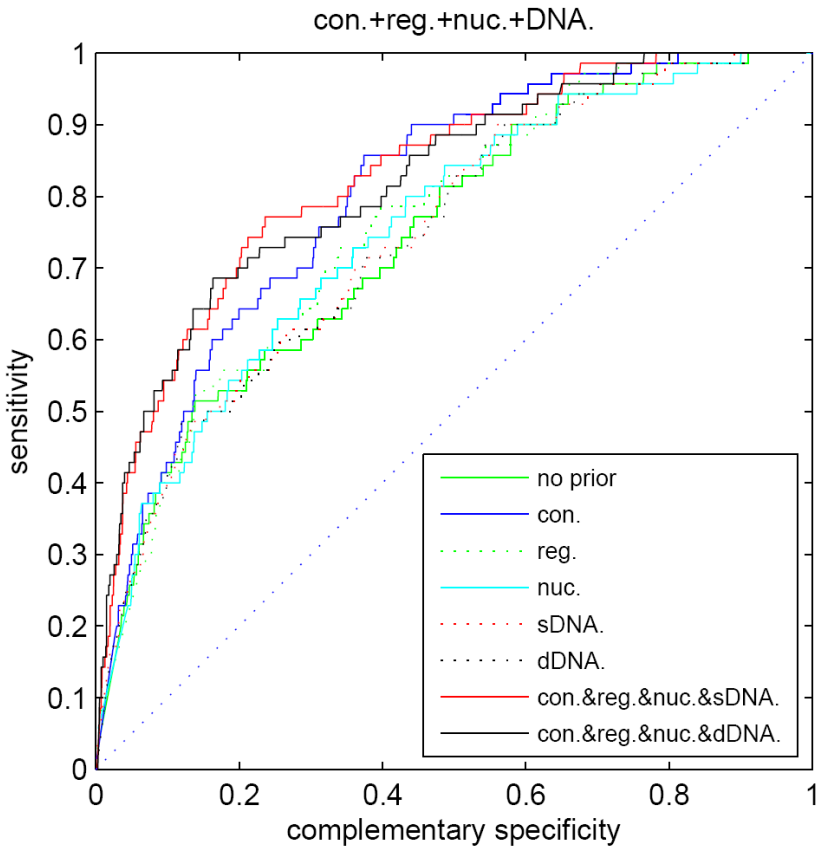
- DNA duplex stability data is discriminative





Summary of part I

- combining multiple information sources helps
- the best results we've gotten so far



Part II:

Computational analysis of effects of regulatory mutations on TF binding

Part III:

Incorporating expression data into TF binding prediction

Thanks!