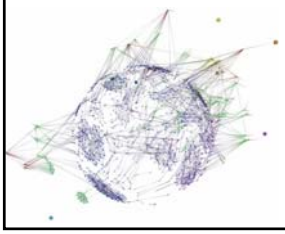
Computational Genomics

10-810/02-710, Spring 2009

Physical Molecular Networks and Network algorithms



Eric Xing

Lecture 28, April 29, 2009

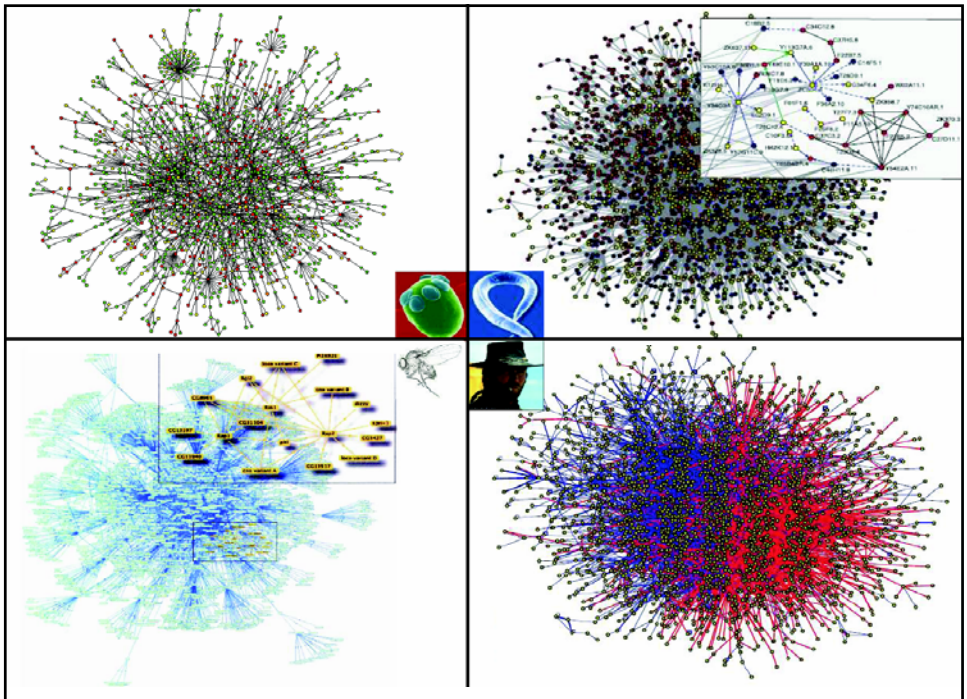
Reading: handouts

Molecular Networks

- Inferred molecular networks:
  - Gene correlation networks (lecture 26)
  - Module networks (lecture 27)
  - …
- Physical molecular networks:
  - Protein-protein interaction (PPI) networks
  - Protein-DNA interaction (PPI) networks --- transcription regulation networks

Protein-Protein Interactions (PPI)

- Protein-protein interactions involve the association of protein molecules
- Eg. signals from the exterior of a cell are mediated to the inside of that cell by protein-protein interactions
- Eg. form protein complex, such as nuclear pore, that carries another protein from cytoplasm to nucleus

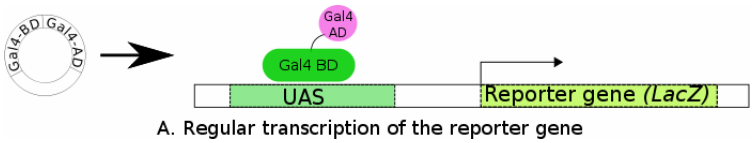


Yeast Two-Hybrid System (Y2H)

- A molecular biology technique used to discover protein-protein interactions.
- It tests physical interactions (such as binding) between two proteins
- Key: the activation of downstream reporter gene(s) by the binding of a transcription factor onto an upstream activating sequence (UAS)

Y2H: I

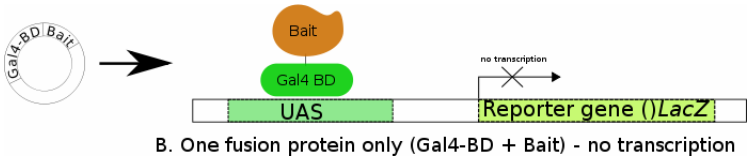
- Gal4 transcription factor gene produces two domain protein (BD and AD) which is essential for transcription of the reporter gene (LacZ).
- BD is responsible for binding to the UAS
- AD is responsible for activation of transcription



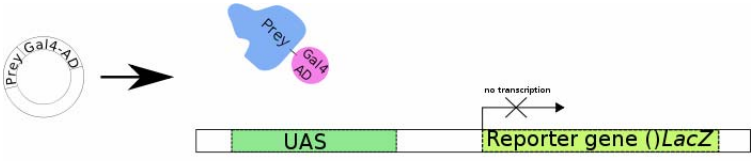
A. Regular transcription of the reporter gene

Y2H: II

- Two fusion proteins are prepared: Gal4BD+Bait and Gal4AD+Prey. None of them is usually sufficient to initiate the transcription (of the reporter gene) alone.



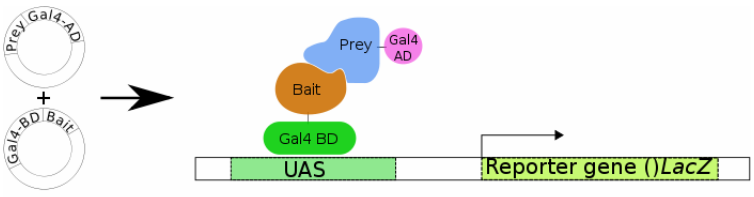
B. One fusion protein only (Gal4-BD + Bait) - no transcription



C. One fusion protein only (Gal4-AD + Prey) - no transcription

Y2H: III

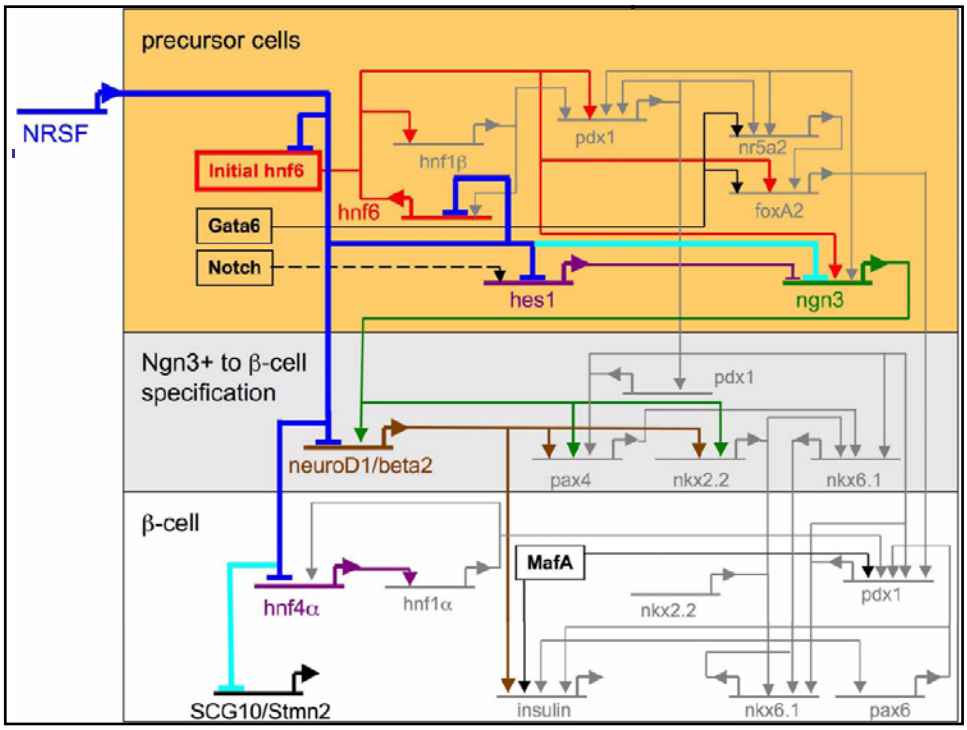
- When both fusion proteins are produced and Bait part of the first interact with Prey part of the second, transcription of the reporter gene occurs.



D. Two fusion proteins with interacting Bait and Prey

Protein-DNA Interactions

- Find DNA binding target seq for each transcription factor
- Understand the regulatory relations between genes
- System biology: build gene regulatory networks

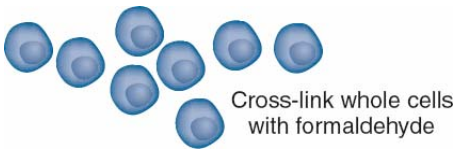


ChIP-Sequencing (ChIP-Seq)

- A molecular biology technique used to analyze protein interactions with DNA.
- It combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify binding sites of DNA-associated proteins
- It can be used to precisely map global binding sites for any protein of interest (more accurate than ChIP-chip).

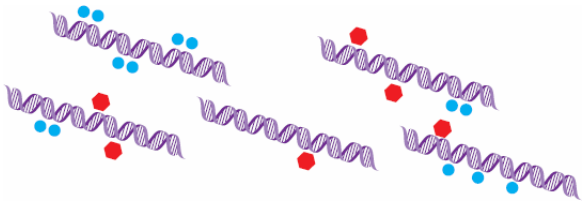
ChIP-Seq: I

- Covalent cross-links between proteins and DNA are formed, typically by treating cells with formaldehyde or another chemical reagent.



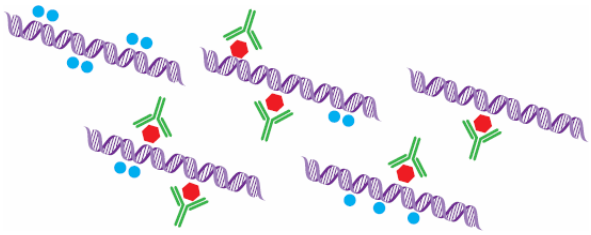
ChIP-Seq: II

- Isolate genomic DNA
- Sonicate DNA to produce sheared, soluble chromatin



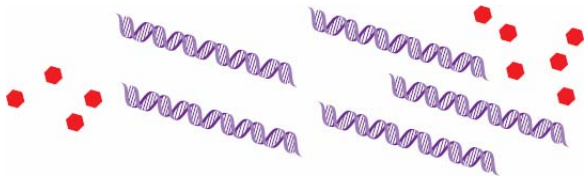
ChIP-Seq: III

- An antibody specific to the protein of interest is used to selectively coimmunoprecipitate the protein-bound DNA fragments that were covalently cross-linked.



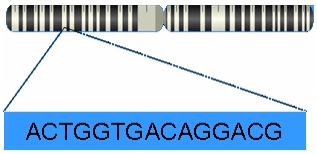
ChIP-Seq: IV

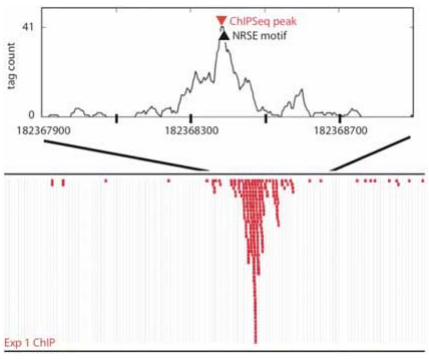
- Reverse cross-links, purify DNA and prepare for sequencing



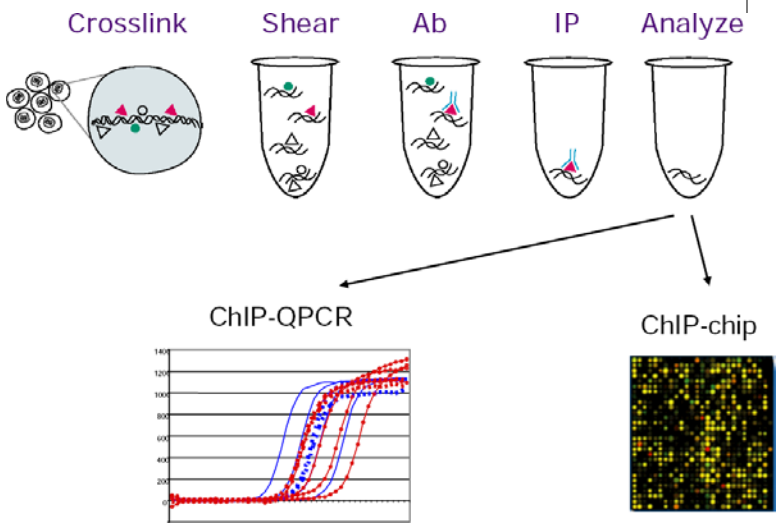
ChIP-Seq: V

- Map the resulting sequences back to the reference genome, whereby the most frequently sequenced fragments formed peaks at specific genomic regions.





Other Related Techniques



Mining and analyzing networks

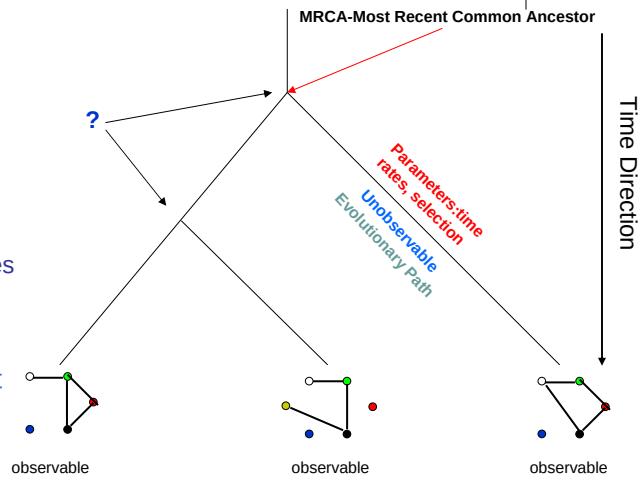
- Identifying Signaling Pathways
  - color-coding technique (Alon, Yuster and Zwick. 1995) and generalizations (Scott et al. RECOMB 2005)
- Identifying Interaction Complexes (clique-like structures)
  - Statistical subgraph scoring (Sharan et al. RECOMB 2004)
- Network alignment
  - PathBLAST: identify conserved *pathways* (Kelley et al 2003)
  - MaWISh: identify conserved *multi-protein complexes* (Koyuturk et al 2004)
  - Nuke: Scalable and General Pairwise and Multiple Network Alignment (Flannick, Novak, Srinivasan, McAdams, Batzoglou 2005)
- Network Dynamics
  - Sandy: backtracking to find active sub-network (Luscombe et al, Nature 2005)
- Node function inference
  - Stochastic block models (Aroldi et al, 2006)
  - Latent space models (Hoff, 2004)
- Link prediction
  - Naïve Bayes classifier, Bayesian network
  - MRF

Network evolution

3 Problems:

1. Test all possible relationships.
2. Examine unknown internal states.
3. Explore unknown paths between states at nodes.

→ Network alignment

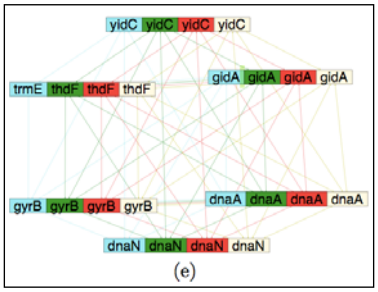


Motivation

- Sequence alignment seeks to identify conserved DNA or protein sequence
  - Intuition: conservation implies functionality

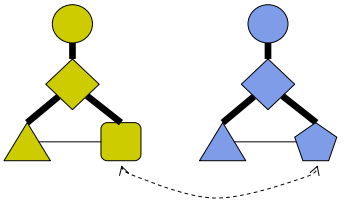
```
EFTPPVQAAYQKVVAGV   (human)
DFNPNVQAAFQKVVAGV   (pig)
EFTPPVQAAYQKVVAGV   (rabbit)
```

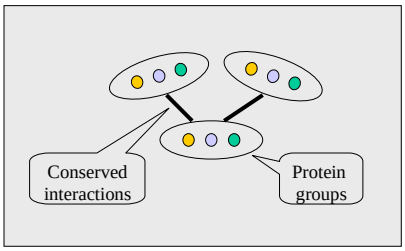
- By similar intuition, subnetworks conserved across species are likely functional modules



Network Alignment

- "Conserved" means two subgraphs contain proteins having **homologous** sequences, serving **similar** functions, having **similar** interaction profiles
  - Key word is similar, not identical





- Product graph:
  - Nodes: groups of sequence-similar proteins, one per species.
  - Edges: conserved interactions.

Scoring Scheme

- Given two protein subsets, one in each species, with a many-to-many correspondence between them, we wish:
  - Each subset induces a dense subgraph.
  - Matched protein pairs are sequence-similar.
- Two hypothesis:
  - **Conserved complex model**: matched pairs are similar.
  - **Random model**: matched pairs are randomly chosen.

$$L(C,C') = L(C)/L(C') \times \prod_{u,v \text{ matched}} \frac{\Pr(S_{u,v} \mid \text{similar})}{\Pr(S_{u,v} \mid \text{random})}$$

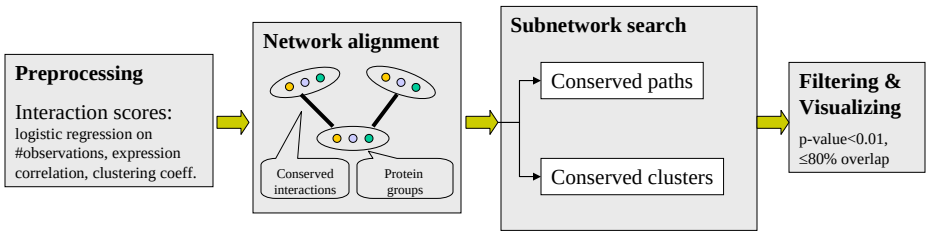
Similarity (BLAST E-value)

Scoring Scheme cont.

- For multiple networks: run into problem of scoring a multiple sequence alignment.
- Need to balance edge and vertex terms.

- Practical solution:
  - Sensible threshold for sequence similarity.
  - Nodes in alignment graph are filtered accordingly.
  - Node terms are removed from score.

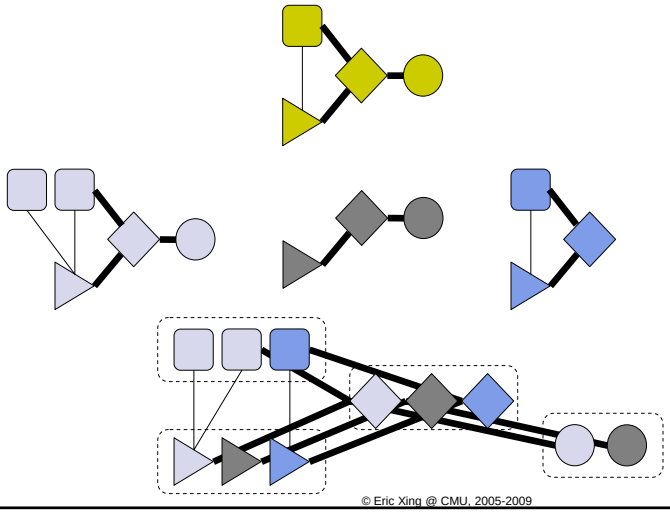
Multiple Network Alignment



- Two recent algorithms:
  - ???, Sharan et al. PNAS 2005
  - Nuke: Flannick, Novak, Srinivasan, McAdams, Batzoglou 2005

Nuke: the model

- Example:



hypothetical ancestral module

descendants

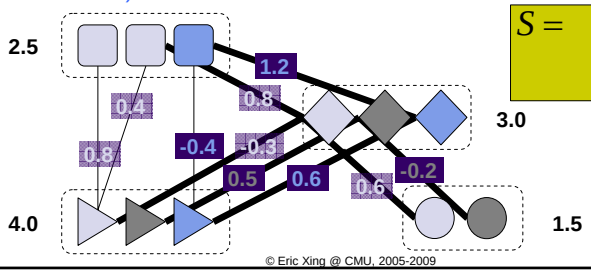
equivalence classes

Nuke: Scoring

- Probabilistic scoring of alignments:

$$\log\frac{P(nodes\,|\,M)}{P(nodes\,|\,R)}+\log\frac{P(edges\,|\,M)}{P(edges\,|\,R)}$$

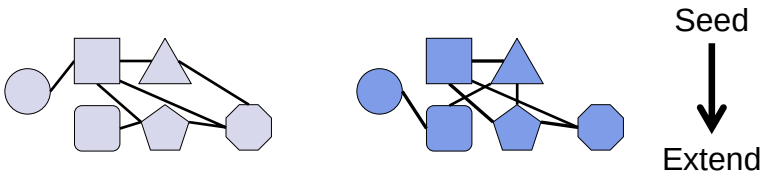
  - *M* : **Alignment model** (network evolved from a common ancestor)
  - *R* : **Random model** (nodes and edges picked at random)
  - Nodes and edges scored independently: How? This is hot research issue! (not covered here)



$$S = S_N + S_E$$

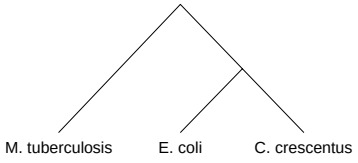
A General Network Aligner: Algorithm

- Given this model of network alignment and scoring framework, how to efficiently find alignments between a pair of networks (N1, N2)?
- Constructing every possible set of equivalence classes clearly prohibitive
- Idea: seeded alignment
  - Inspired by seeded sequence alignment (BLAST)
  - Identify regions of network in which "good" alignments likely to be found
    - MaWISh does this, using high-degree nodes for seeds
    - Can we avoid such strong topological constraints?



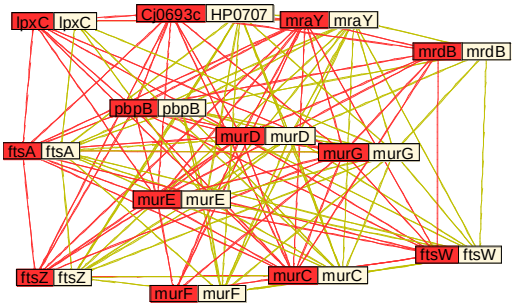
Multiple Alignment

- Progressive alignment technique
  - Used by most multiple sequence aligners

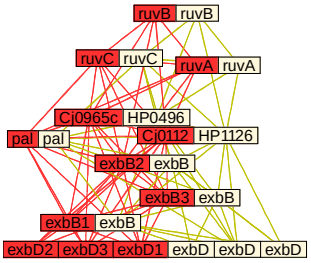


- Simple modification of implementation to align *alignments* rather than *networks*
  - Node scoring already uses weighted SOP
  - Edge scoring remains unchanged

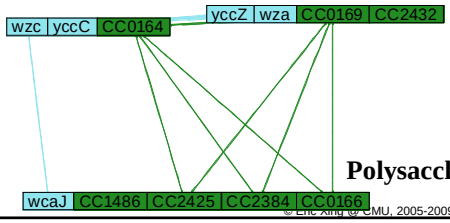
Pairwise alignments



Cell division

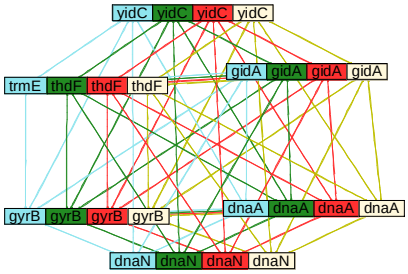


DNA uptake

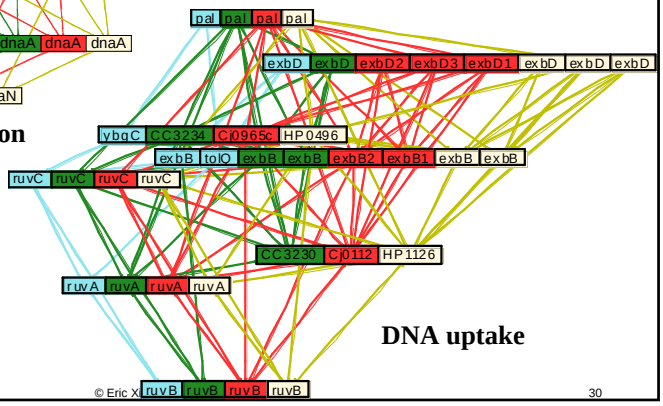


Polysaccharide transport

Multiple alignments

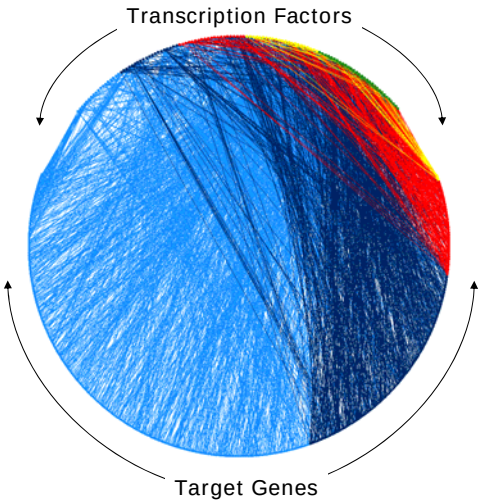


DNA replication



DNA uptake

Dynamic Yeast TF network



- Analyzed network as a static entity
- But network is *dynamic*
  - Different sections of the network are active under different cellular conditions
- Integrate gene expression data

[Luscombe et al, *Nature*]

Gene expression data

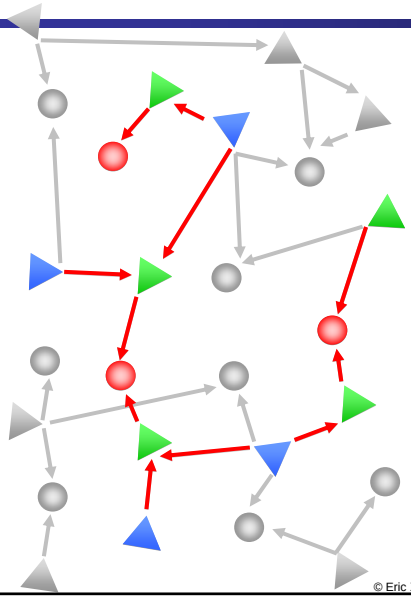
- Genes that are differentially expressed under five cellular conditions

Cellular condition	No. genes
Cell cycle	437
Sporulation	876
Diauxic shift	1,876
DNA damage	1,715
Stress response	1,385

- Assume these genes undergo transcription regulation

[Luscombe et al, *Nature*]

Backtracking to find active sub-network



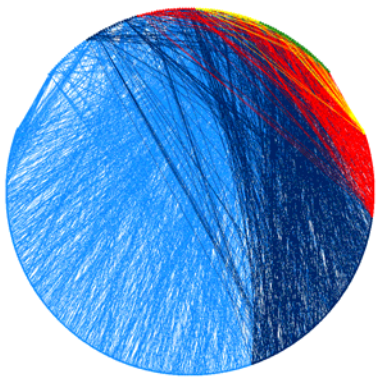
- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network

[Luscombe et al, *Nature*]

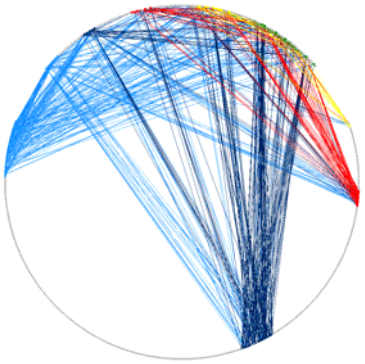
Network usage under different conditions

static



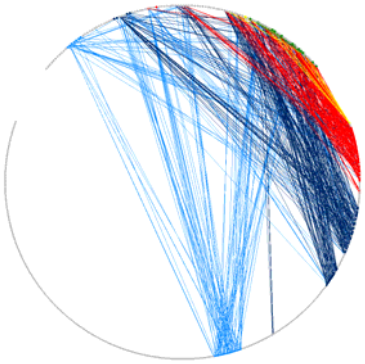
Network usage under different conditions

cell cycle



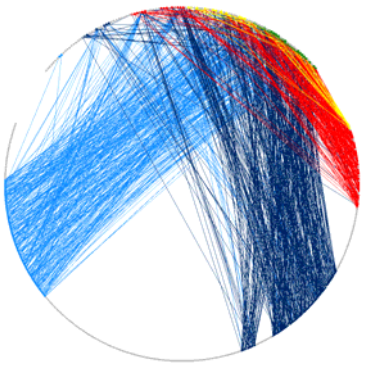
Network usage under different conditions

sporulation



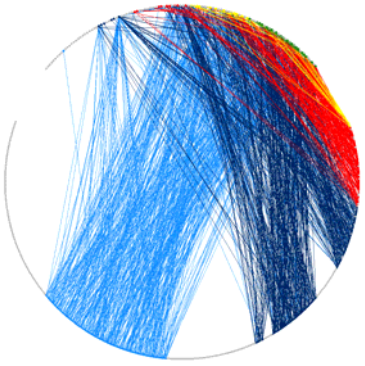
Network usage under different conditions

diauxic shift



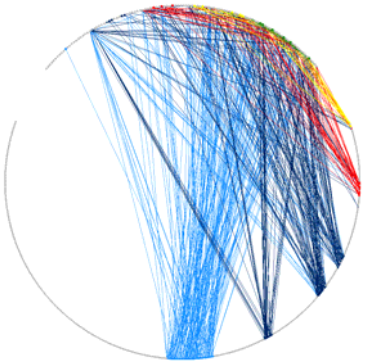
Network usage under different conditions

DNA damage



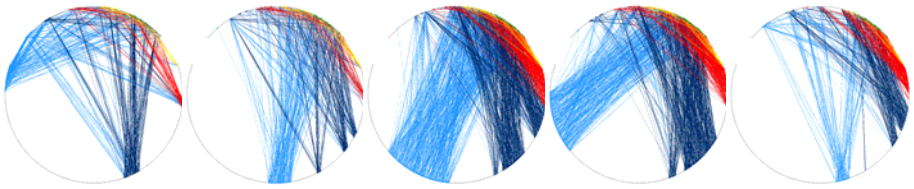
Network usage under different conditions

stress response



Network usage under different conditions

Cell cycle **Sporulation** **Diauxic shift** **DNA damage** **Stress**

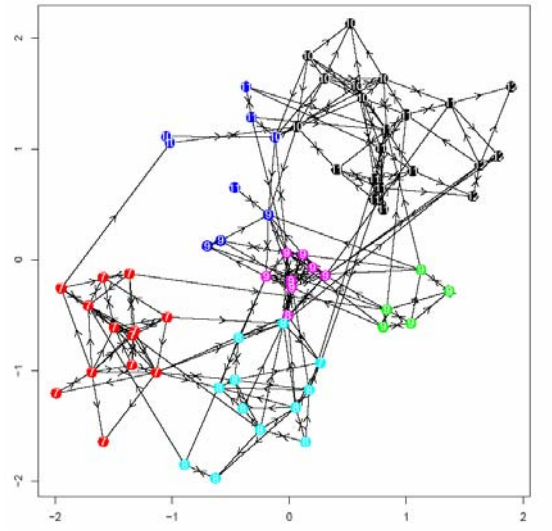


How to model the networks change?

--- an open problem

[Luscombe et al, *Nature*]

Network Tomography: functional analysis

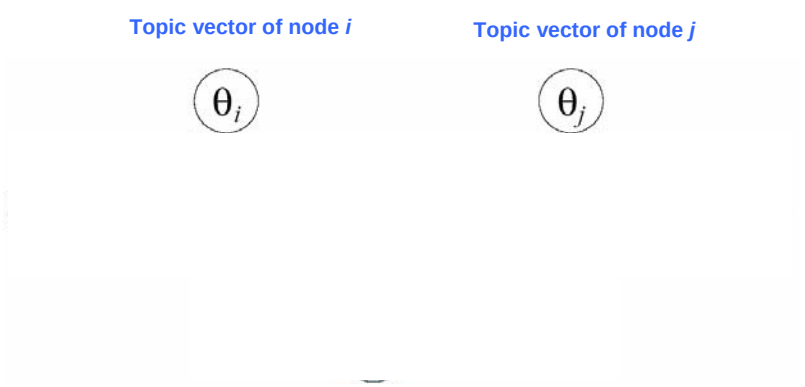


A Latent Mixture Membership Model

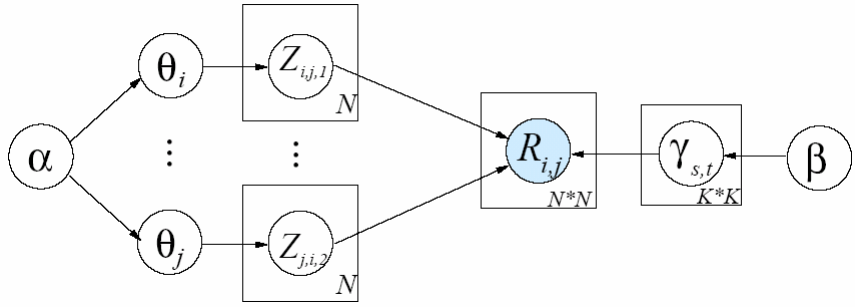
Motivation

- In many networks (e.g., biological network, citation networks), each node may be "multiple-class", i.e., has multiple functional/topical aspects.
- The interaction of a node (e.g., a protein) with different nodes (partners) may be under different function context.
- Prior knowledge of group interaction may be available.

A Mixture Membership Stochastic Blockmodel (MMSB) Airoldi, Blei, Fienberg, and Xing, 2008



Hierarchical Bayesian MMSB



For each object i=1,…,N:

$\theta_i \sim \text{Dirichlet}(\alpha)$

For each topic-pair (*s*,*t*):

$\gamma_{s,t} \sim \text{Beta}(\beta)$

For each pair of object (*i*,*j*)

$Z_{i,j,1} \sim \text{Multi}(\theta_i)$

$Z_{i,j,2} \sim \text{Multi}(\theta_j)$

$R_{i,j} \sim \text{Bernoulli}(\rho\gamma_{z_{i,j,1},z_{i,j,2}} + (1-\rho)\delta_0)$

Variational Inference

- The Joint likelihood:

$$p(r,z,\theta,\gamma)=\prod_i \theta_i^{\sum_j z_{i,j,1}+z_{i,j,2}+\alpha-1}\times\gamma_{m,n}^{\sum_{i,j} r_{i,j} z^m_{i,j,1} z^n_{i,j,2}+\beta_1-1}(1-\gamma_{m,n})^{\sum_{i,j}(1-r_{i,j}) z^m_{i,j,1} z^n_{i,j,2}+\beta_2-1}$$

- GMF approximation:

$$q(\mathbf{r},\mathbf{z},\theta,\gamma\mid\alpha,\beta)=\left(\prod_{i=1}^{N} q(\theta_i\mid\mu_i)\right)\times\left(\prod_{s=1,t=1}^{K} q(\gamma_{s,t}\mid\nu_{s,t})\right)\times\left(\prod_{i=1,j=1}^{N} q(z_{i,j,1},z_{i,j,2},r_{i,j}\mid\varphi_{i,j})\right)$$

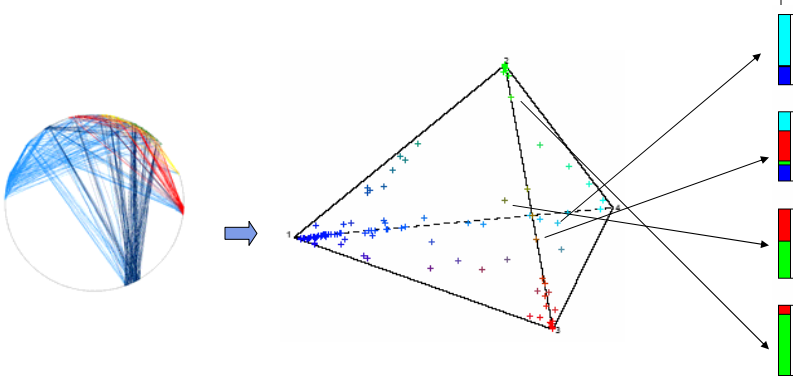
$$\mu_i=\alpha+\sum_j\langle z_{i,j,1}\rangle+\sum_j\langle z_{i,j,2}\rangle$$

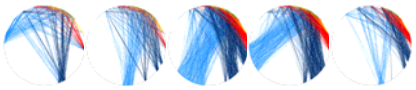
$$\nu_{s,t}=\beta+\sum_{i,j} r_{s,t}\langle z_{i,j,1}z_{i,j,2}\rangle$$

- MF approximation: ...

$$q(\mathbf{r},\mathbf{z},\theta,\gamma\mid\alpha,\beta)=\left(\prod_{i=1}^{N} q(\theta_i\mid\mu_i)\right)\times\left(\prod_{s=1,t=1}^{K} q(\gamma_{s,t}\mid\nu_{s,t})\right)\times\left(\prod_{i=1,j=1}^{N} q(z_{i,j,1}\mid\phi_{i,j,1})q(z_{i,j,2}\mid\phi_{i,j,1})q(r_{i,j}\mid\varphi_{i,j})\right)$$

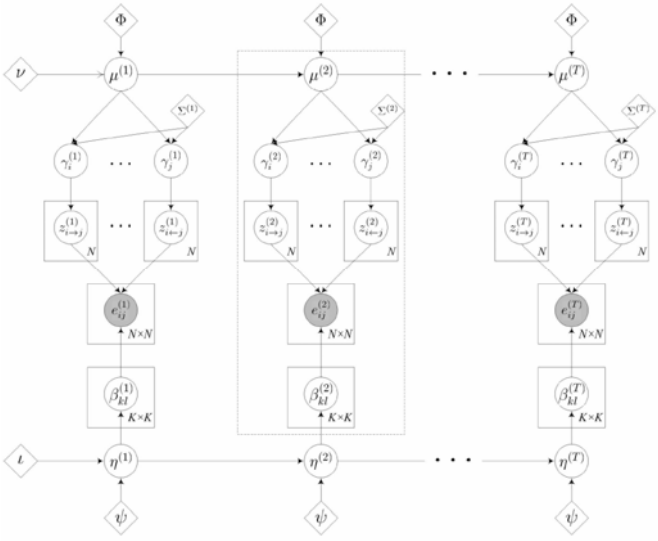
Inferred Mixed membership Network Tomography

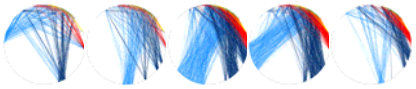


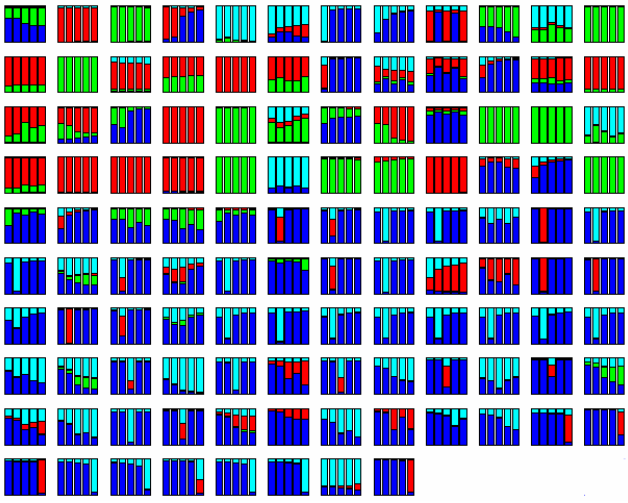


Dynamic MMSB

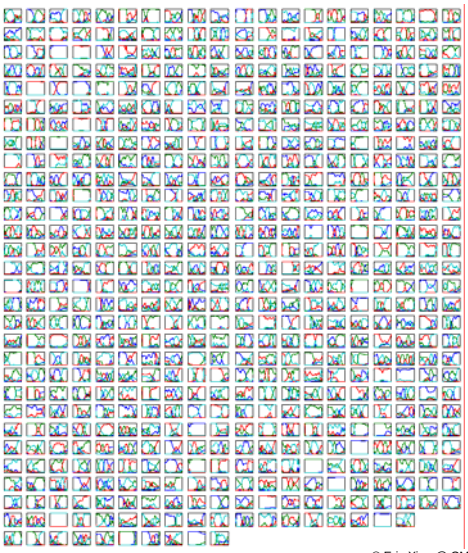
Fu, Song, and Xing, 2009

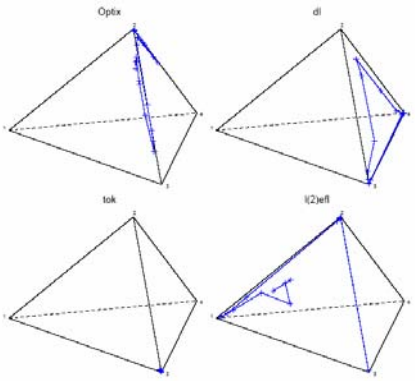






Trajectory of MM of genes during Drosophila life cycle

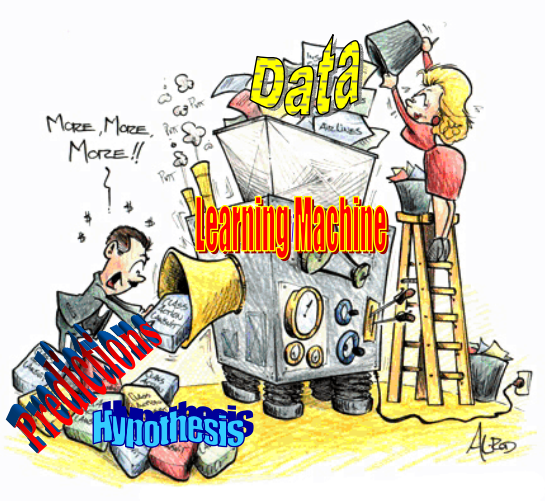




Summary of MMSB

- A stochastic block model
- Each node can play "multiple roles", and its ties with other nodes can be explained by different roles
- Hierarchical Bayesian formalism
- Dynamic tomography
- Efficient variational inference

Computational Molecular Biology



Using
mathematical models
and
computational reasoning
to pursue
predictive understanding
of life

Research in Computer Science

- Computer science is a ``science of the artificial.''
- Problems are precisely stated and are often generic rather than application-specific.
- The quality of an algorithm is measured by its worst-case time bound.
- Mathematical elegance is just as important as relevance to applications.

Research in Computational Biology

- The goal is to understand ground truth.
- Problem statements are often fuzzy.
- Problems are often application-specific, and problem formulations must be faithful to those applications.
- The quality of an algorithm is measured by its performance on real data.
- Biological findings are more important than computational methods.

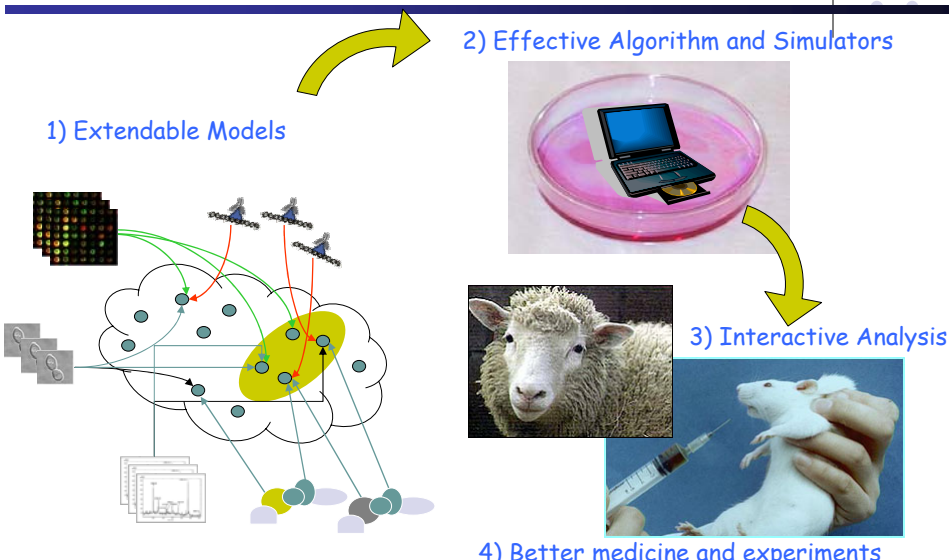
Adapting to Computational Biology

- Choose problems that are fundamental, timely and relevant.
- Mathematical depth and elegance are highly desirable, but often simple mathematics, artfully applied, is the key to success.
- Avoid problems that will change when technology changes.
- Learn the biological background of your problem, the available sources of data and their noise characteristics.
- Work with an application-oriented team and don't get typecast as an algorithms specialist or just "play with numbers."
- Benchmark your algorithms on real data, establish a user community and make your software available and easy to use.

Computational Biology can Benefit from Research in Machine Learning

- Biological processes are stochastic and partially observed
  - probabilistic models and statistical inference/learning algorithms
- Biological data are usually non-linear and high dimensional
  - kernel methods and convex optimization
- Biological systems are complex and usually intractable
  - efficient representation and approximation techniques
- Biological prior knowledge provide crucial model constrains and biological subjects can be studied from different angles
  - Bayesian approach and data fusion methods

Conclusion



Reference

- **Deng et al.** ***Assessment of the reliability of protein-protein interactions and protein function prediction.*** Proc. PSB, 140-151 (2003).
- **Bader et al.** ***Gaining confidence in high-throughput protein interaction networks***. Nat. Biotechnol., 78-85 (2004).
- **Kelley et al.** ***PathBLAST: a tool for alignment of protein interaction networks***. Nucl. Acids Res. **32**, W83-8 (2004).
- **Kelley et al.** ***Conserved pathways within bacteria and yeast as revealed by global protein network alignment***. PNAS **100**, 11394-9 (2003).
- **Sharan et al.** ***Conserved patterns of protein interaction in multiple species.*** PNAS **102**, 1974-9 (2005).
- **Sharan et al.** ***Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data.*** J. Comp. Biol. In press (2005).
- **Scott et al.** ***Efficient algorithms for detecting signaling pathways in protein interaction networks***. Proc. RECOMB, 1-13 (2005).
- E. Airoldi, D. Blei, S. Fienberg, and E. P. Xing, Mixed Membership Stochastic Blockmodel**,** *Journal of Machine Learning Research, 9(Sep):1981--2014, 2008.*
- W. Fu, L. Song, and E. P. Xing, A State-Space Mixed Membership Blockmodel for Dynamic Network Tomography**,** *Manuscript, arXiv:0901.0135.*