Belief Propagation Algorithms: From Matching Problems to Network Discovery in Cancer Genomics

Jennifer Chayes
Microsoft Research New England
Microsoft Research New York City

Outline

1. Graphical Models and Belief Propagation
2. A Simple Example: Matching
3. The Steiner Tree Problem
4. Application to Networks in Systems Biology

1. Graphical Models & Belief Propagation

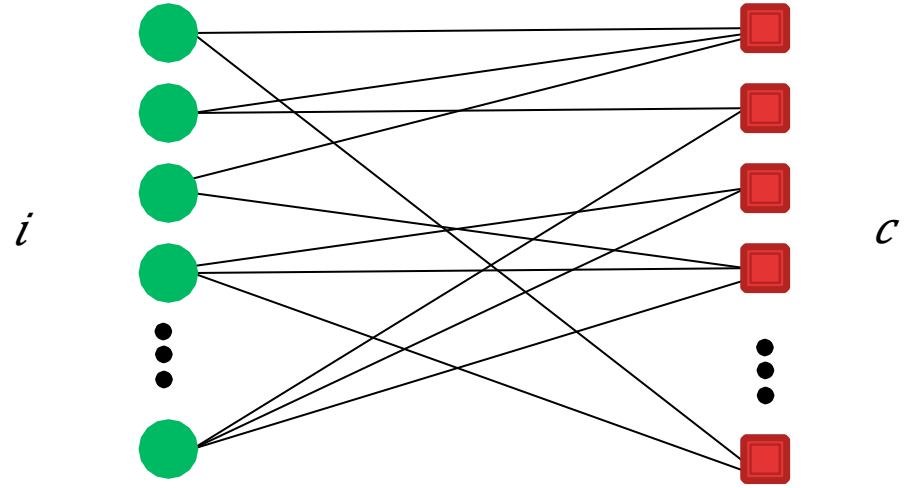
- (Hyper)Graphical model: Representation of dependency structure of a collection of random variables with local constraints

$G=(V, E)$

- Each node $i \in V$ has random variable σ_i with *a priori* distribution φ_i
- Each hyperedge $c \in E$ has (hard or soft) constraint ψ_c
- Probability distribution of the set of variables $\sigma_V = \{\sigma_i\}_{i \in V}$:

$$\mu(\sigma_V) = 1/Z \prod_{i \in V} \varphi_i(\sigma_i) \prod_{c \in E} \psi_c(\sigma_c)$$

- Visualize dependency structure: Factor Graph F



ic is an edge of F
if
i is constrained by c

- Interested in calculating/estimating:
  - Marginals $\mu\downarrow i$ of $\sigma\downarrow i$

$$\mu\downarrow i\,(\sigma\downarrow i)=\sum\sigma\downarrow j\in\sigma\downarrow V\backslash \mathrm{i}\uparrow\blacksquare\mu(\sigma\downarrow V\,)$$

  - Modes (configurations of maximal weight)

$$\sigma\downarrow max=\text{argmax }\mu$$

Belief Propagation

- Iterative method for approximating marginals and modes, exact if the factor graph is a tree
- In general, 2 sets of equations* relating:
  - "message from i to c":

    $\mu_{i \to c}$ = marginal i would have if it ignored constraint c
  - "message from c to i":

    $\mu_{c \to i}$ = marginal i would have if it were only constrained through c (and had uniform prior)

*Note: There are simplifications in problems in which the variables or constraints have only degree 2 in the factor graph

General Belief Propagation Equations

- Fixed-Point Equations (exact on trees):

$$\mu_{i \to c}(\sigma_i) \propto \varphi_i(\sigma_i) \prod_{c' \neq c} \mu_{c' \to i}(\sigma_i)$$

$$\mu_{c \to i}(\sigma_i) \propto \sum_{\sigma_k \in \sigma_c \setminus i} \psi_c(\sigma_k) \prod_{j \in c, j \neq i} \mu_{j \to c}(\sigma_j)$$

- Easy to implement corresponding update equations
- Often work well in practice
- Question: When does the solution converge to the right answer?

Rigorous results on BP: Convergence and correctness

- Maximum weight matching
  - Bipartite graph (when solution is unique):
    - Bayati, Shah, Sharma ('08)
  - General graph, b-matching (when corresponding LP is tight):
    - Bayati, Borgs, Chayes, Zecchina ('09)
    - Sanghavi, Shah, Willsky ('09)
- Nash bargaining on networks (when corresponding MWM LP is tight):
  - Bayati, Borgs, Chayes, Kanoria, Montanari ('11)
- Min-cost network flow:
  - Garmanik, Shah, Wei ('11)

2. A Simple Example of BP: Matching

- The model and graphical representation
- Derivation of BP for (max) weighted matchings
- LP and statement of BP results

Perfect b-Matching Problem

- Given
  - Graph $G=(V, E)$
  - Degree sequence $\{b_i\}_{i \in V}$, $b_i = 1, 2, \ldots, |V|$
  - Weights $\{w_{ij}\}_{ij \in E}$
- Perfect b-matching M

$$M \subseteq E \text{ s.t. } \forall\, i \in V \quad |\{e \in M \mid e \ni i\}| = b_i$$

Ex: $b \equiv 1$ perfect matching

$b \equiv 2$ 2-factor

- Max-weight b-matching problem: Find

M_{max} s.t. $W(M_{max}) = \sum_{ij \in M_{max}} w_{ij}$ is maximal

Graphical Model for Perfect b-Matching

- Here the variables sit on the edges and the constraints on the sites of the graph $G=(V,E)$
  - Variables: $\forall\, ij \in E,\ x_{ij} = \begin{cases} 0 \text{ if vacant} \\ 1 \text{ if occupied} \end{cases}$
  - Constraints: $\forall\, i \in V,\ \sum_{j \in N(i)} x_{ij} = b_i$

  M $\leftrightarrow$ edge variables $x_E = \{x_{ij}\}$ with $x_{ij} = \begin{cases} 1 \text{ if } ij \in M \\ 0 \text{ if } ij \notin M \end{cases}$
- Probability distribution of x_E at "temperature" β:

$$\mu(x_E) = 1/Z \prod_{ij \in E} e^{\beta w_{ij} x_{ij}} \prod_{i \in V} \mathbb{I}\left(\sum_{j \in N(i)} x_{ij} = b_i\right)$$

Derivation: BP Matching Equations on Trees

- Simplifications:
  - Consider only $b \equiv 1$ (perfect matchings)
  - Notational: leave out constraint in equations, and enforce constraints implicitly

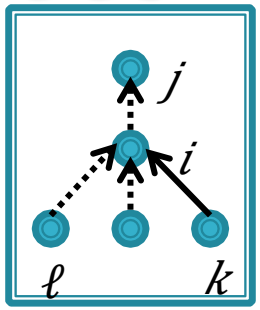
$$\mu(x_E) = 1/Z \prod_{ij \in E} e^{\beta w_{ij}} x_{ij}$$

- Messages:
  - Since variables have only degree 2 in the factor graph, we need only one set of equations, e.g. for $\mu_{\{i,j\} \to j}$ = marginal at ij if constraint at j is ignored, which we'll just call $\mu_{i \to j} = \mu_{i \to j}(x_{ij})$.
  - Also, instead of taking just $\mu_{i \to j}(1)$ or $\mu_{i \to j}(0)$, as the message, try the log-ratio $m_{i \to j}$ defined by

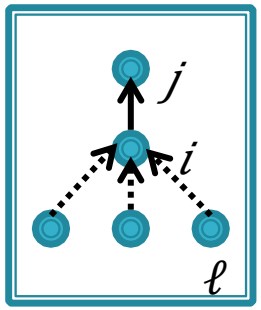
$$e^{\beta m_{i \to j}} = \mu_{i \to j}(1) / \mu_{i \to j}(0)$$

Iterative Calculations on Trees

- $\mu_{i\to j}(0)$
$= 1/Z_{ij} \sum_{k\in N(i)\setminus j} \mu_{k\to i}(1) \prod_{\ell\in N(i)\setminus\{j,k\}} \mu_{\ell\to j}(0)$



- $\mu_{i\to j}(1)$
$= e^{\beta w_{ij}} / Z_{ij} \prod_{\ell\in N(i)\setminus j} \mu_{\ell\to j}(0)$



$$\Rightarrow e^{-\beta m_{i\to j}} = \mu_{i\to j}(0)/\mu_{i\to j}(1) = \sum_{k\in N(i)\setminus j} e^{-\beta(w_{ij} - m_{k\to i})}$$

As $\beta\to\infty$

$$m_{i\to j} = w_{ij} - \max_{k\in N(i)\setminus j} m_{k\to i}$$

To get matching $M\downarrow max$ from messages:

- Similarly, on trees, one can show:

$$\mu(ij\in M) \quad = \quad e\uparrow\beta m\downarrow i\to j \ / \sum k\uparrow\square e\uparrow\beta m\downarrow k\to j$$

$$\to\beta\to\infty\top \quad \{\blacksquare 0 \quad \text{if } m\downarrow i\to j < \max\top k\in N(j)$$

$$m\downarrow k\to j \ 1 \ \text{if } m\downarrow i\to j = \max\top k\in N(j) \ m\downarrow k\to j$$

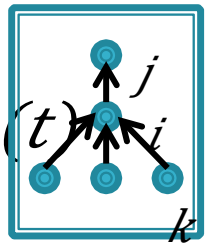
(assuming the max above is not degenerate).

$\Longrightarrow$ For each $i\in V$, algorithm chooses edge ki into i with maximum message $m\downarrow k\to i$

Summary: BP for Perfect Matching

- Define "message" $m\downarrow i\to j$ on directed edge $i\to j$ by

$$m\downarrow i\to j\,(0)= w\downarrow ij$$
$$m\downarrow i\to j\,(t+1)= w\downarrow ij - \max\nolimits_{-} k\in N(i)\backslash \mathrm{j}\; m\downarrow k\to i\,(t)$$



- To estimate $M\downarrow max$ at time t, $M(t)$:

For each site i choose as the candidate edge into i the edge $i\ell$ such that

$$m\downarrow \ell\to i\,(t)= \max\nolimits_{-} k\in N(i)\; m\downarrow k\to i\,(t)$$

and add this maximum message edge to the candidate "matching" $M(t)$. (Note $M(t)$ may not be a b-matching.)

- Note: This is exact on trees.
- Question: Can we determine when else it converges to the correct answer, and how fast?

Rigorous Result on BP for b-Matching

- Consider the corresponding LP relaxation and its dual:
  - LP: max $\sum_{ij\in E} w_{ij} x_{ij}$
    subj. to $0 \le x_{ij} \le 1$
    $\sum_{j\in N(i)} x_{ij} = b_i$
  - Dual: min $\sum_{ij\in E} \lambda_{ij} - \sum_{i\in V} b_i y_i$
    subj. to $\lambda_{ij} \ge 0$
    $\lambda_{ij} \ge w_{ij} + y_i + y_j$

- Theorem (Bayati, Borgs, Chayes, Zecchina '09): If the LP has a unique optimum which is integer, then $M(t)$ converges to the correct solution M_{max}. In particular $M(t) = M_{max}$ for

$$t \ge 2|V|/\epsilon \ \max_i |y_i^* - 1/2| \ ,$$

where y^* is an optimal solution of the dual LP and $\epsilon = \min_{ij}\{|\ w_{ij} + y_i^* - 1/2 + y_j^* - 1/2\ | > 0\}$.

3. The Steiner Tree Problem

- Given
  - Graph $G=(V, E)$
  - Costs $\{c_{ij}\}_{ij \in E}$, $c_{ij} \geq 0$
  - Set of "terminals" $S \subseteq V$
- Problem: Find a tree $T \subseteq G$ containing all terminals, i.e. all nodes in S, which minimizes the cost:

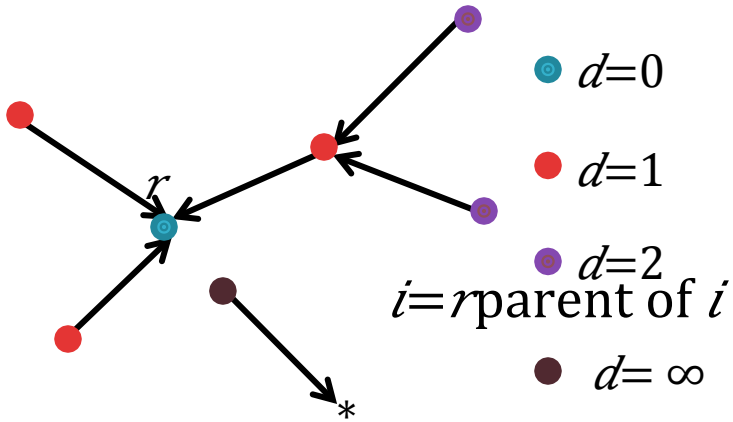
$$C(T) = \sum_{ij \in E(T)} c_{ij}$$

- Difficulty: Want to do BP on this, but don't have a local way to enforce the global constraint of a (connected) tree
- Solution: Introduce a new representation

New Representation

Bayati, Borgs, Braunstein, Chayes, Ramezanpour, Zecchina ('08)

- Designate one terminal $r \in S$ as root and set $c_{rr} = 0$
- $\forall i \in V$, introduce two variables
  - Distance: $d_i \in \{0, 1, \ldots, |V|-1\}$
  - Parent: $p_i \in N(i) \cup \{*\}$
- If T is a Steiner tree, set
  - $d_i = \mathrm{dist}_T(i,r) \quad \forall i \in V(T)$
  - $p_i = \begin{cases} * & \text{if } i \neq V(T) \\ \blacksquare & i \text{ if } i = r \\ \text{in } T & \text{otherwise} \end{cases}$ parent of i



- Cost of the tree: $C(T) = \sum_{i \in V(G)} c_{i p_i} \mathbb{I}(p_i \neq *)$
- Constraints:
  - $p_i \neq * \quad \forall i \in S$
  - If $p_k = j \notin \{*, r\}$, then $p_j \neq *$ and $d_j = d_k - 1$

Graphical Model

- Define interactions enforcing these constraints (and including the weights):

$$\psi_{jk} = [1-\mathbb{I}(p_k = j)\mathbb{I}(d_j \neq d_k - 1)][1-\mathbb{I}(p_k = j)\mathbb{I}(p_j = *)]$$

and

$$\varphi_i = [1-\mathbb{I}(i \in S)\mathbb{I}(p_i = *)] \exp[-\beta c_i \mathbb{I}(p_i \neq *)]$$

- Then the probability distribution is

$$\mu(\{d_i, p_i\}) = 1/z \prod_{i \in V} \varphi_i \prod_{i,j \in V;\ ij \in E} \psi_{ij}$$

- Variants:
  - Bounded diameter D tree: Take $d_i \in \{0,1, \ldots, D\}$
  - Prize-collecting Steiner tree: Replace φ_i by soft constraints, removing $\mathbb{I}(i \in S)$ and adding "prizes" to cost function

See Angel, Flaxman, Wilson ('08 –'12)

BP Results on the Steiner Tree

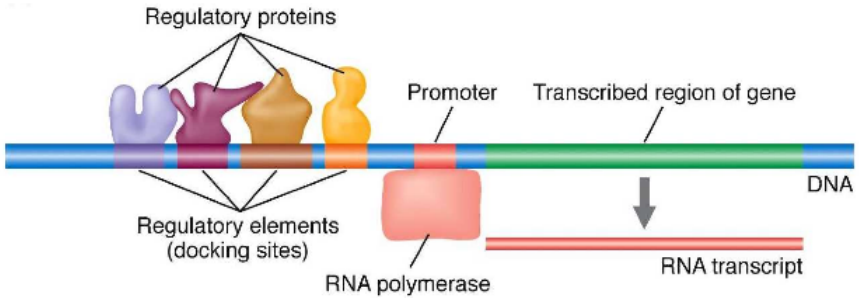
- Rigorous Results: Minimum spanning tree
  - If BP converges, then it converges to the correct solution (Bayati, Braunstein and Zecchina '08)
- Non-Rigorous Results: Minimum Steiner tree
  - Tests of our BP algorithm vs. LP algorithms for a benchmark library of several dozen Steiner tree instances (SteinLib), show that our algorithm is *much faster*. Also, it gets better optima in all but two (very small) instances (Bailley-Bechet, Borgs, Braunstein, Chayes, Dagkessamanskaia, Francois, Zecchina '11)
  - On biological data sets in the Fraenkel Lab at MIT, the LP algorithms were too slow to give any results on human data
- Open Problem: Find sufficient conditions for BP for the MWST to converge to the correct solution, or at least to a solution within ϵ of an optimizer.

4. Applications to Networks in Systems Biology

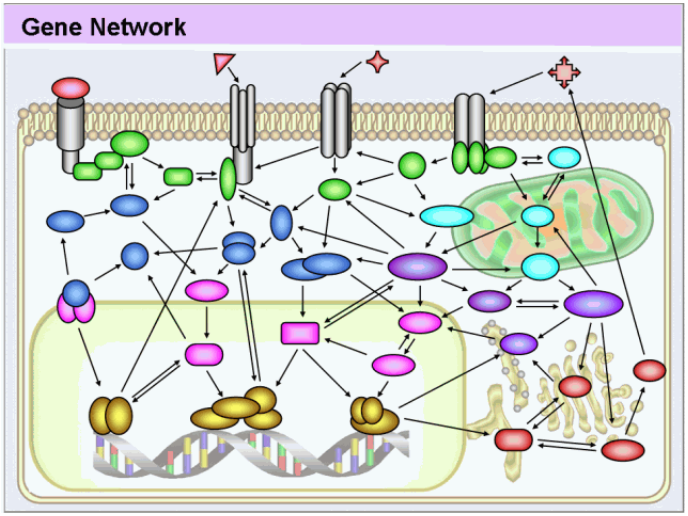
- The Biological Problem
- Formulation of the Algorithmic Problem: The Prize-Collecting Steiner Tree (PCST)
- Biological Applications of the PCST
- A Variant Algorithmic Problem: The Prize-Collecting Steiner Forest

The Biological Problem

- Standard Dogma: DNA → RNA → Proteins



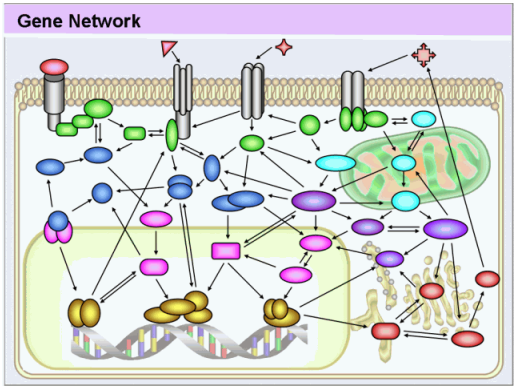
⇒ Gene Regulatory Network



Protein Interactome

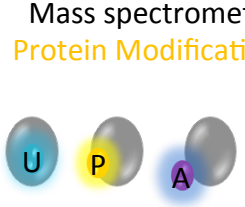
Gene Regulation and Disease

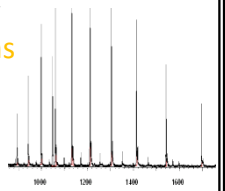
- Problems with the gene regulatory network are the sources of many diseases
- How do we infer the network structure from partial data?
- Can we identify particular nodes on the network responsible for dysregulation in certain diseases and individuals?
- Are one or more nodes in combination viable drug targets?



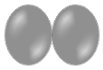
Drug Discovery Paradigm

Mass spectrometry
Protein Modifications

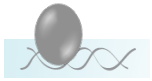




Yeast two-hybrid
Affinity capture mass-spec
Protein-protein interactions



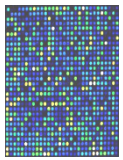
ChIP-Seq, Dnase-Seq, …
Protein-DNA interactions



Genetic/Chemical
Screens

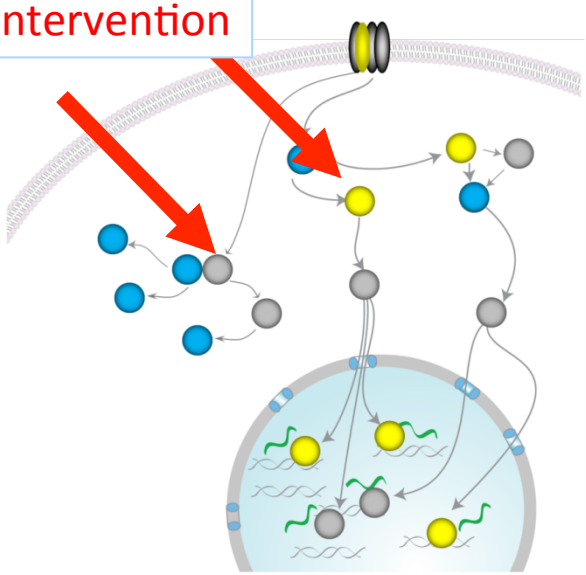
Microarrays
RNA-Seq
mRNA

```
AAATAGCCATTATACGTA
CCTAATACTGAAGAGTCA
TTCCTAGTAAAGCATGCT
ACTTTTCAGTATATTCCA
TTATATTTTTAACTACAA
GCGGCGCAGAAACCAGAG
```

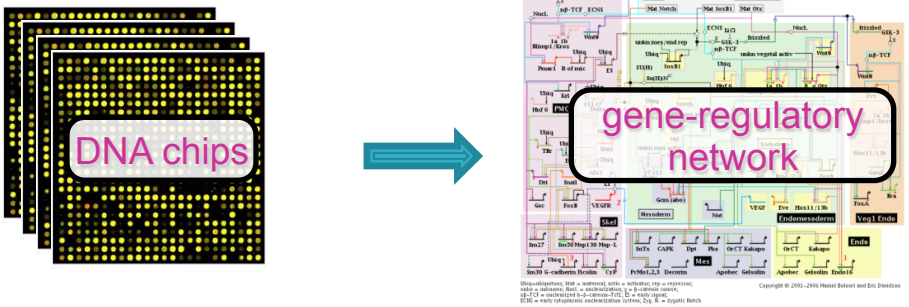


Computational
Models

Points of
Intervention



Gene Expression Data



- Microarrays tell us which gene is expressed in the presence of which other gene under a particular set of conditions
- From the differential expression of a particular gene, we infer the node weight of the corresponding transcription factor protein (prize in the PCST)
- To get edge weights between two proteins, we use the probability of interaction of these two proteins inferred from (properly weighted) databases of known interactions for the given organism

Question: How do we determine the network most likely to have produced this data?

Formulation of the Problem: The Prize-Collecting Steiner Tree

- Given
  - Graph $G=(V, E)$
  - Costs $\{c_{ij}\}_{ij \in E}$, $c_{ij} \geq 0$
  - Set of "prize terminals" $S \subseteq V$ with prizes $\{\pi_i\}_{i \in S}$, $\pi_i > 0$
  - Parameter $\lambda > 0$
- Problem: Find a tree $T \subseteq G$ which minimizes the cost:

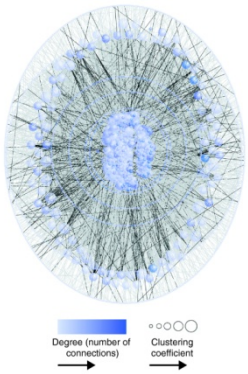
$$C(T) = \sum_{ij \in E(T)} c_{ij} - \lambda \sum_{i \in V(T)} \pi_i$$

- Note: As $\lambda \to \infty$, this turns into the standard Steiner tree problem with terminals $S = i \pi_i > 0$.

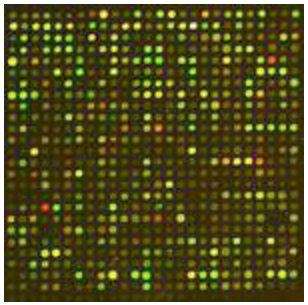
Mapping to Biological Data

- Find the tree which minimizes

$$C(T) = \sum_{ij \in E(T)} c_{ij} - \lambda \sum_{i \in V(T)} \pi_i$$



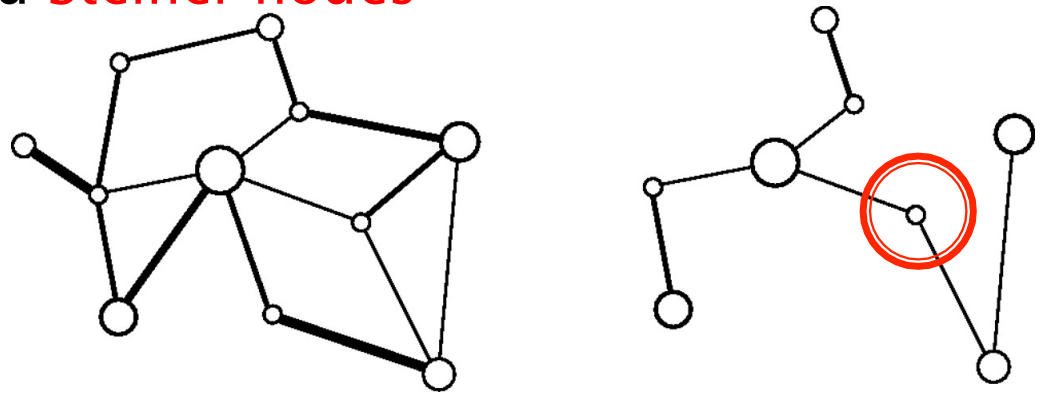
$c_{ij} = -\log \text{prob}(ij \text{ exists})$
where $\text{prob}(ij \text{ exists})$ is the probability that proteins i and j interact in the given organism (from databases)



$\pi_i = -\log p\text{-value}(i)$
where $p\text{-value}(i)$ is the p-value of the differential expression of the gene corresponding to protein i, in the given experiment

Steiner Nodes

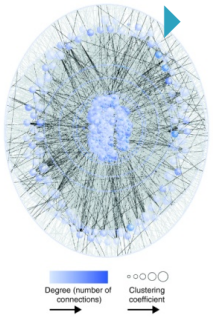
- In the standard Steiner tree problem, nodes which are included in the minimizing solution but which are not terminals, i.e. not in the set S, are called Steiner nodes
- Similarly, in the PCST, nodes which have zero (or low) prizes but which are included in the minimizing solution are called Steiner nodes



- In the context of the gene regulatory networks, Steiner nodes correspond to proteins whose genes which are not differentially expressed a lot, but which nevertheless seem likely to participate in the network ⇒ identification of proteins not previously know to participate in the pathway

Example 1: Yeast Pheromone Response Pathway

(Bailley-Bechet, Borgs, Braunstein, Chayes, Dagkessamanskaia, Francois, Zecchina: PNAS '11)

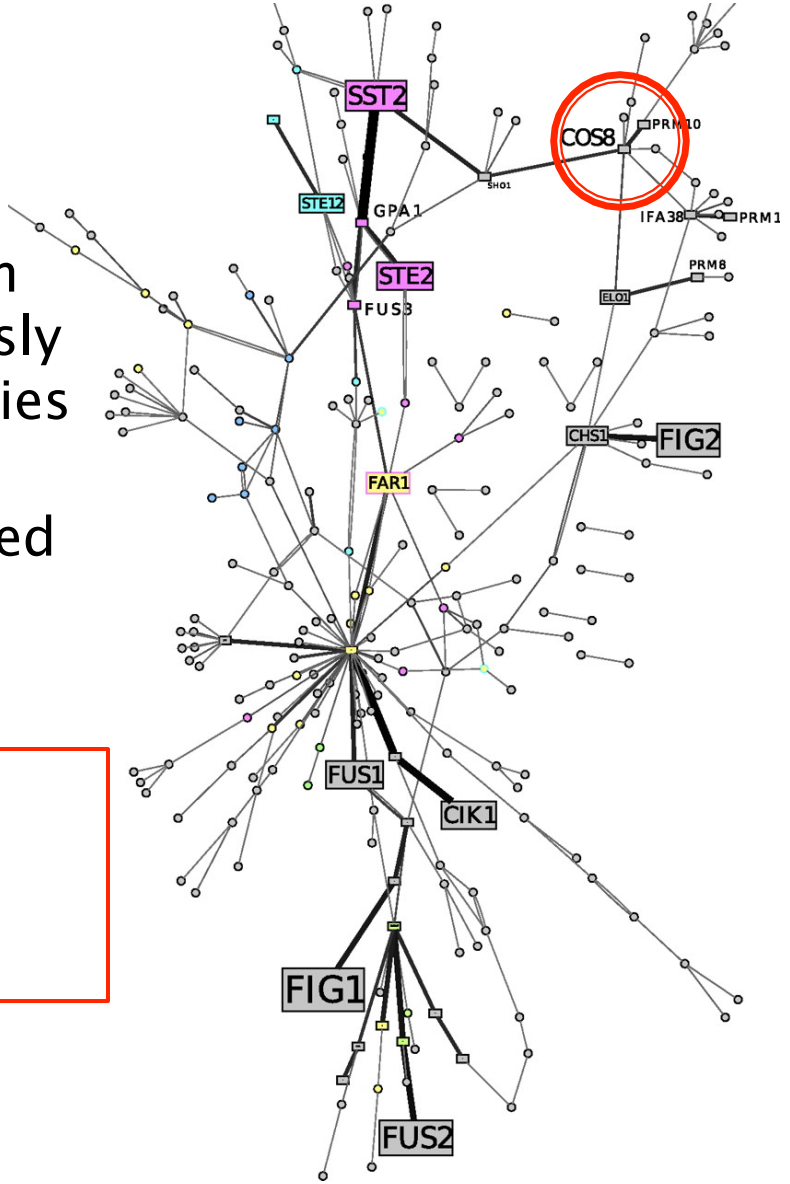


- Yeast protein signal transduction network:
  - 4689 Proteins
  - 14928 Protein-Protein interactions
  - Gives set of weights $\{c\downarrow ij\}$ for relevant proteins in pheromone response pathway
- Considered 56 large-scale gene expression data sets used to reconstruct the yeast pheromone pathway. For each data set
  - Get set of prizes $\{\pi\downarrow i\}$
- Construct 56 solutions to bounded-D PCST problem
- "Merge solutions" to get one network

Results: Pathway identified

- Two types of proteins on network
  - Proteins differentially expressed in pheromone response and previously discovered by transcriptomic studies (terminals)
  - Proteins not differentially expressed but bridging between different subnetworks ("Steiner proteins")

Question: Are the Steiner proteins important in the pheromone response pathway?



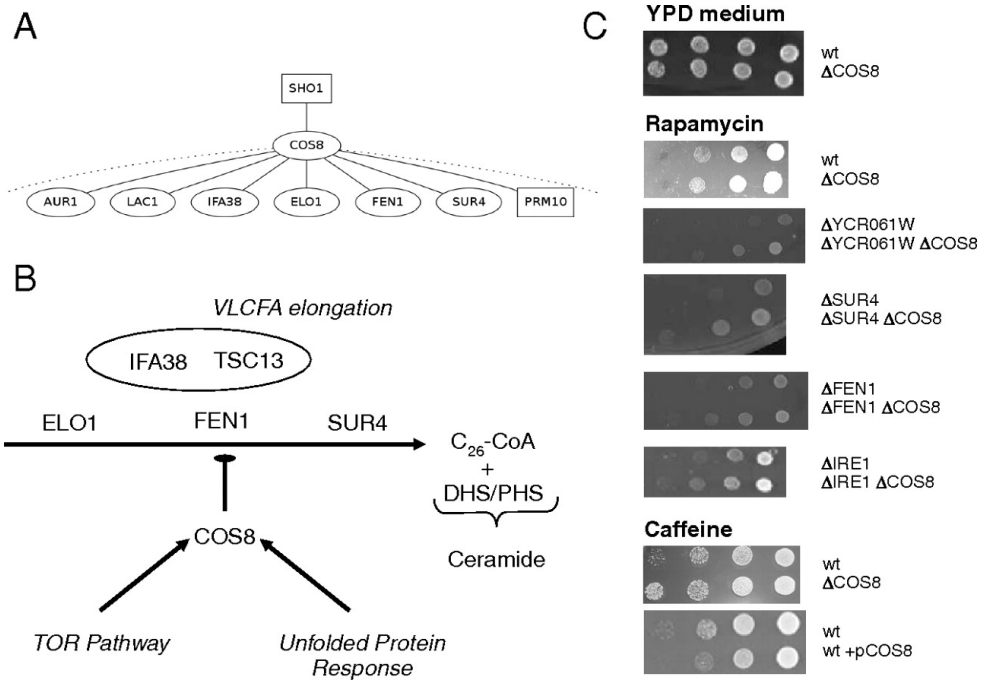
Testing a Steiner Node

- Did an experiment to knock out the gene corresponding to COS8

$\Rightarrow$

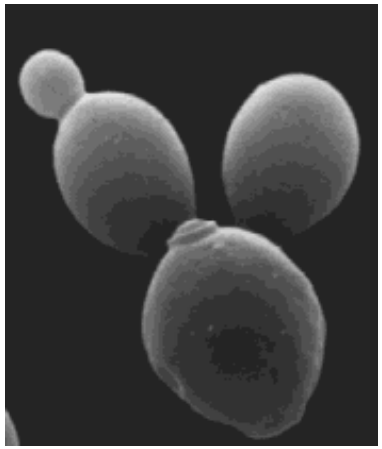
Pheromone response pathway failed.

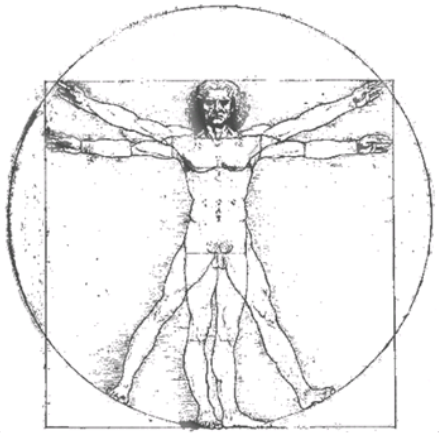
"Experimental proof" of the importance of the Steiner node



From Yeast to Mammals

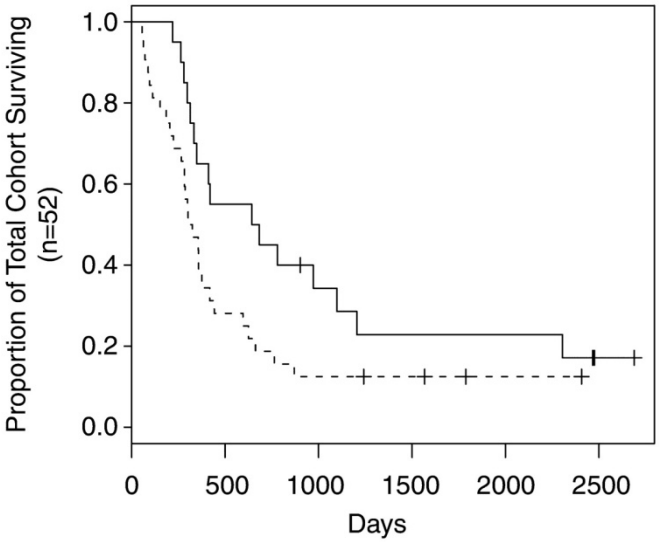
- Problems (mammals relative to yeast):
  - Incomplete interactome data
  - Ten times as many transcription factors
  - Huge intergenic regions
- Need fast algorithms



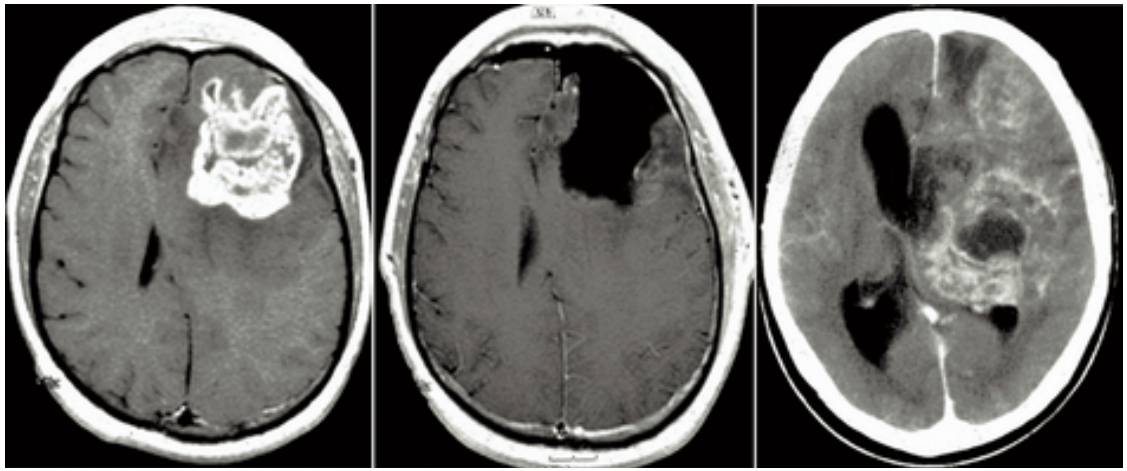


Example 2: Glioblastoma Pathways

- Glioblastoma:
  - particular form of brain cancer
  - the human cancer with the worst outcome
  - much more common in men than women



Pope W B et al. Radiology 2008;249:268-277

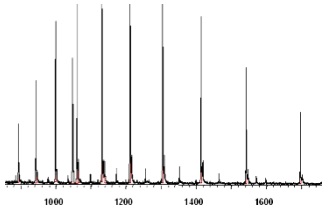


Presentation Post-op Recurrence

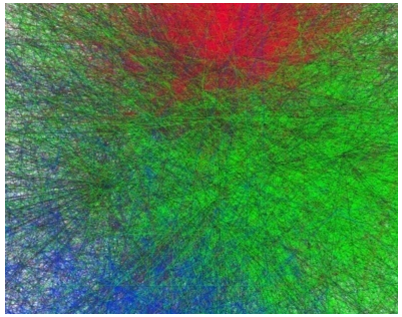
Weil RJ (2006) PLoS Med 3(1): e31.

Can we find GBM pathways using the PCST?

(Fraenkel Lab, MIT, work in progress using our PCST algorithm)

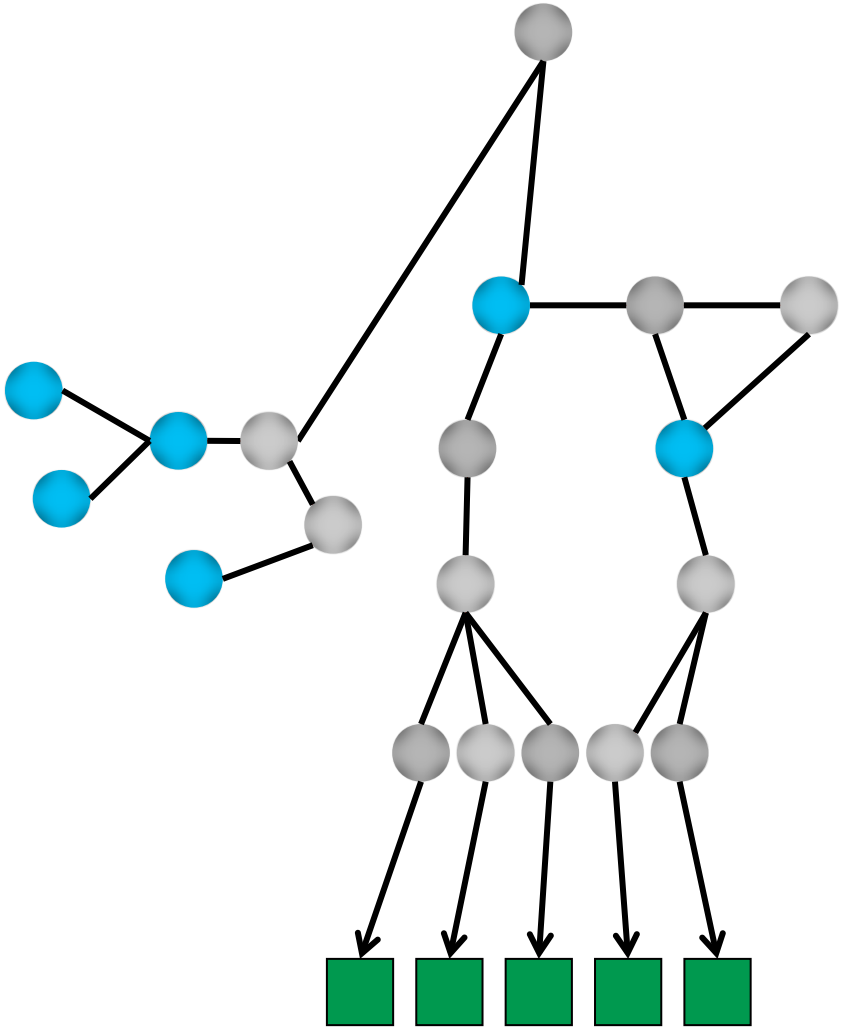


Mass spectrometry



Interactome

Expression/Epigenomics

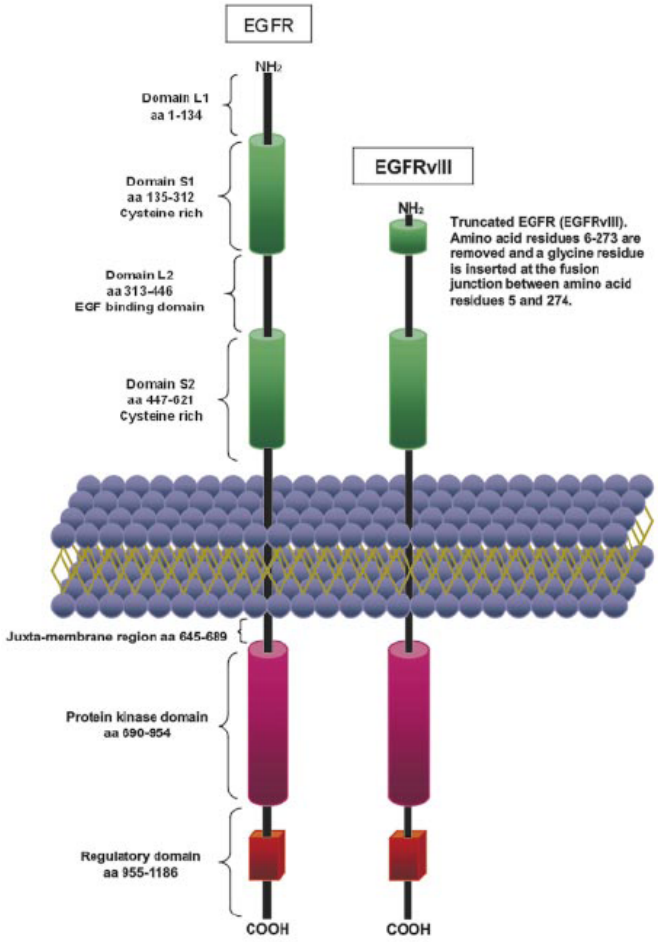


How to choose the root of the PCST?

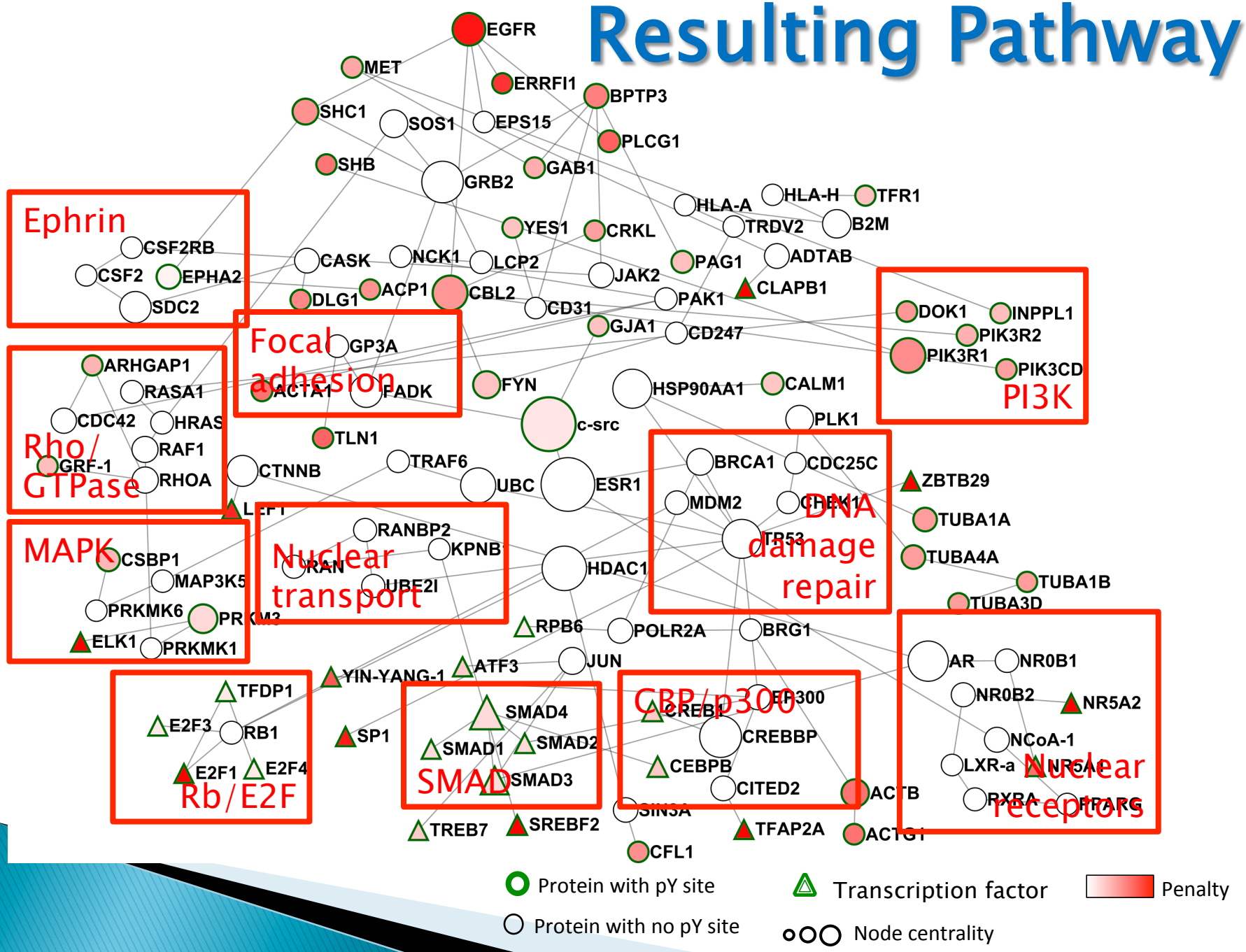
Always good to choose receptor proteins since these often begin signaling pathways

Try EGFR

- EGFR variant III mutation is most common EGFR mutation in human cancer
- Present in 60% of GBMs
- EGFRvIII expression correlates with shorter life expectancies



Resulting Pathway



Identify interesting Steiner nodes

- Top 5 Nodes ranked by betweeness centrality*:
  SRC, ESR1, HDAC1, CREBBP, GRB2
- SRC well-known to be active in many types of cancer, and had relatively large "prize"
- What about ESR1?
  - No "prize" and not previously identified for Glioblastoma
  - What is ESR1?
  - This is the Estrogen Receptor
- **First pathway link between glioblastoma and gender!**
- Experimental test: EGFR inhibitor and Estrodiol together inhibit the growth of GBM cells in culture better than the EGFR inhibitor alone

  ⇒ possible drug therapy for glioblastoma

*Relative percentage of shortest paths in graph through given node

Multiple Signaling Pathways

(Tuncbag, Braunstein, Pagnani, Huang, Chayes, Borgs, Zecchina, Frankel; RECOMB '12)

- How do we explain multiple disjoint signaling pathways altered in a particular condition?
- Use Prize-Collecting Steiner Forest:
- Just like prize-collecting Steiner tree, but now we also specify that there be k disjoint trees* (= forest F) as the minimizing solution of

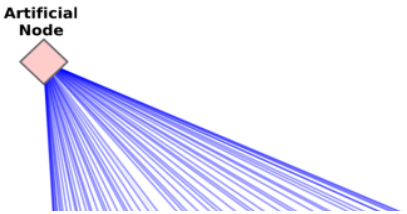
$$C(F) = \sum_{ij \in E(F)} c_{ij} - \lambda \sum_{i \in V(F)} \pi_i$$

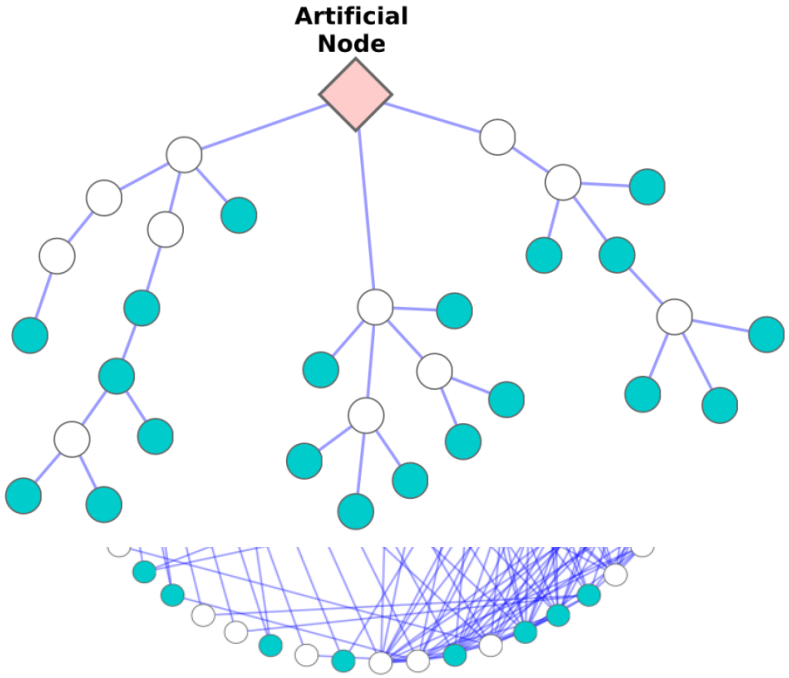
- To implement PCSF, just add an "artificial node" A, connect every node i to A with strength c_{iA} $\Rightarrow$ new PCST with 1 more node and $|V|$ more edges

*Or let k vary by adding another term to C

Method

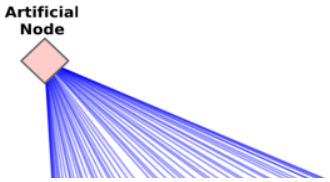
Prize Collecting Steiner Forest



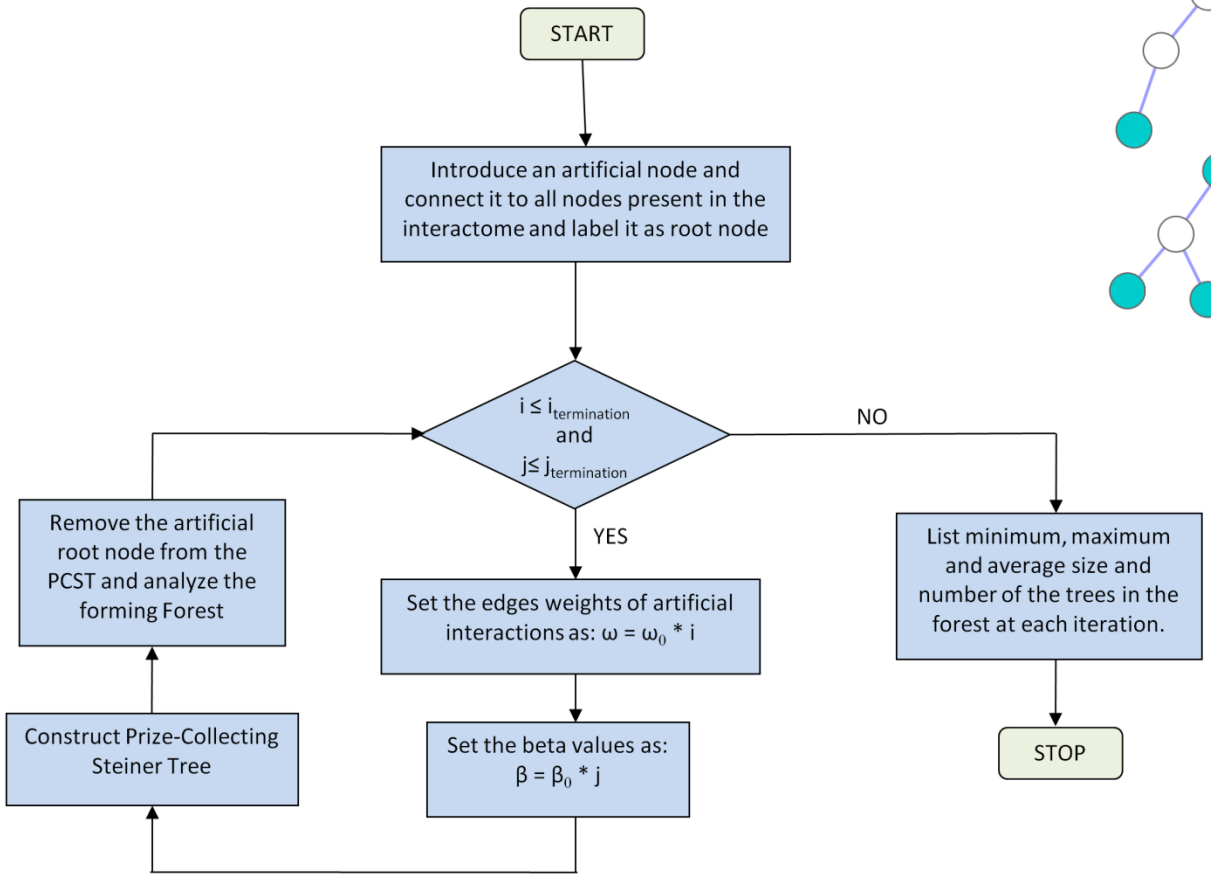


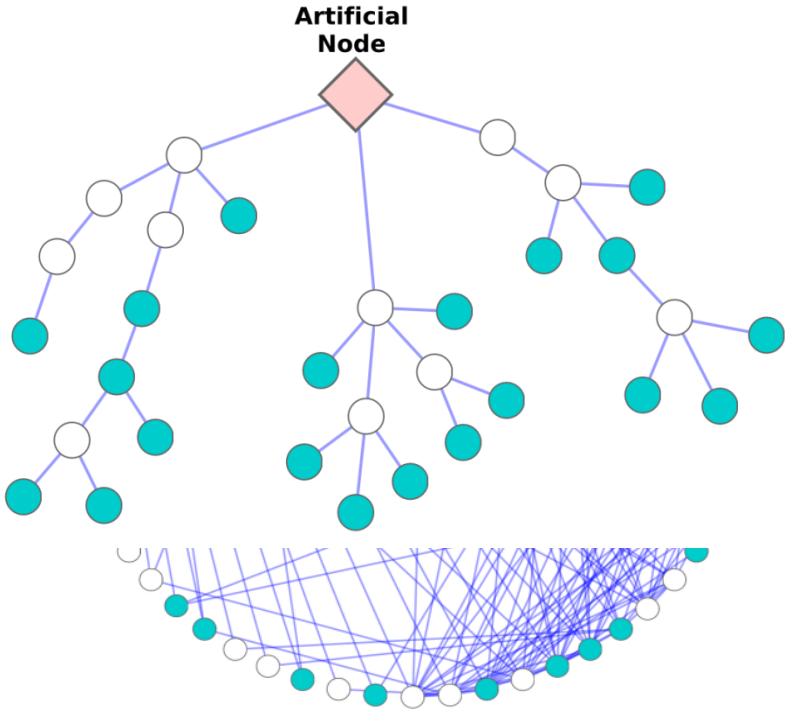
Reveals parallel working pathways, in addition to "hidden" (Steiner) individual proteins or genes

Method



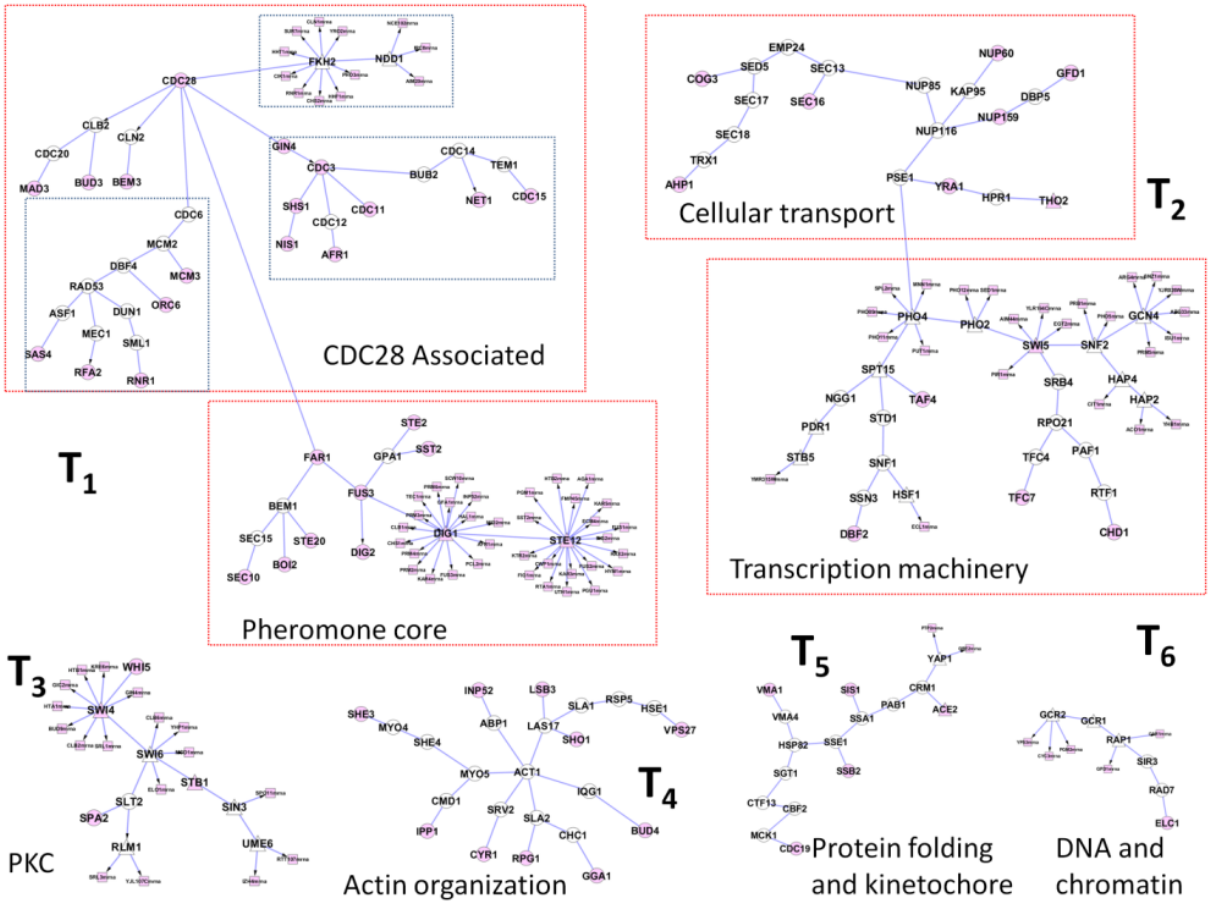
Prize Collecting Steiner Fores



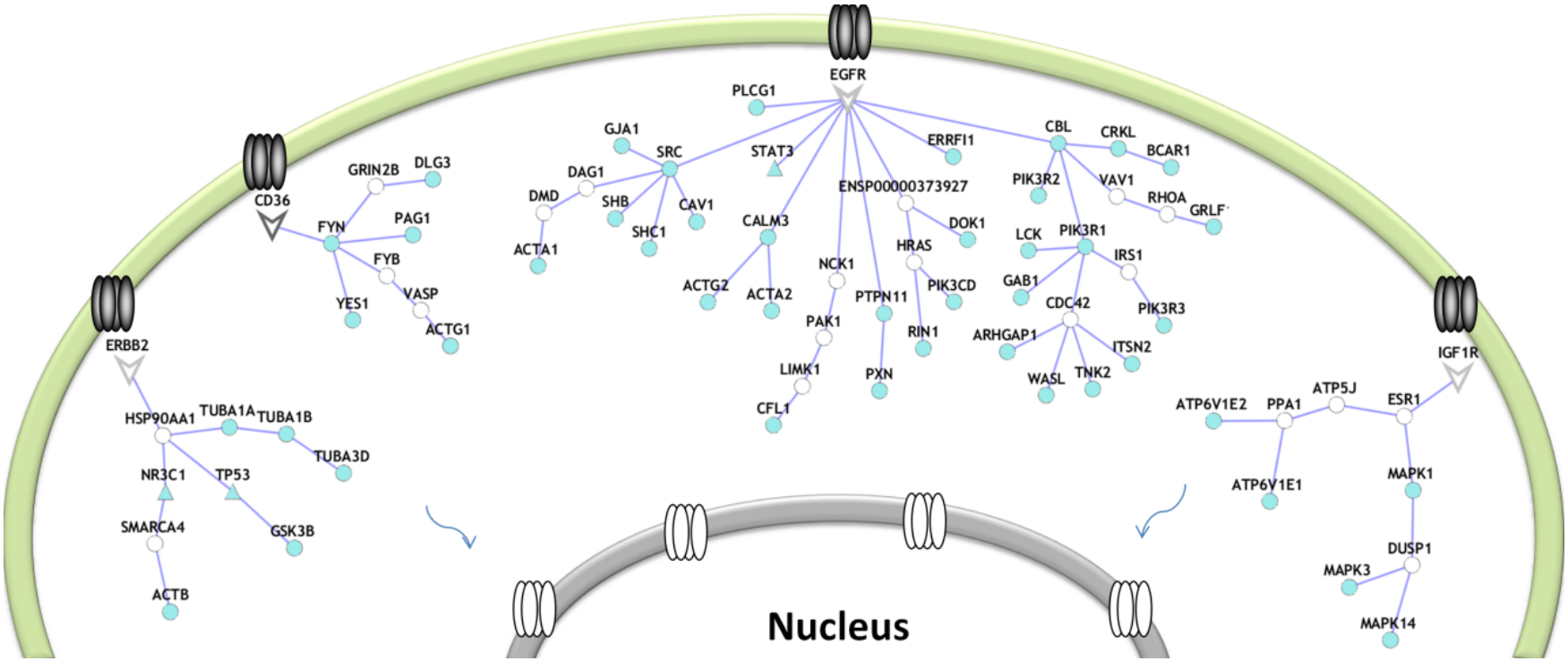


Reveals parallel working pathways, in addition to "hidden" individual proteins or genes

Derived Forest: Yeast Pheromone Response Network



Derived Forest: Human Glioblastoma Data Set



Summary

- Graphical models give us succinct representations for capturing local dependencies among random variables, and (with the right representation) even some global dependencies, e.g., the prize-collecting Steiner tree
- Belief propagation give us a way of approximiating marginals and modes of graphical models
  - Rigorously can be proved to converge quickly to the correct solution in particular cases (e.g., b-matching when LP has only integral optima)
  - In practice converges to near optimal solutions very rapidly on known benchmarks and new biological data sets
- There is biological evidence that BP algorithms do very well in identifying signaling and regulatory pathways among proteins, and also identify "Steiner proteins", suggesting drug targets for human disease

Open Question

- Find conditions under which these new BP algorithms (for the Steiner tree, the prize-collecting Steiner tree or forest, or even the minimum spanning tree) converge to either the correct solution or at least to a solution within ϵ of an optimizer.
- Get bounds on the rate of convergence.

Thanks for your attention