计算系统生物学

王 勇

中国科学院数学与系统科学研究院

Conditional specific pathway or subnetwork identification

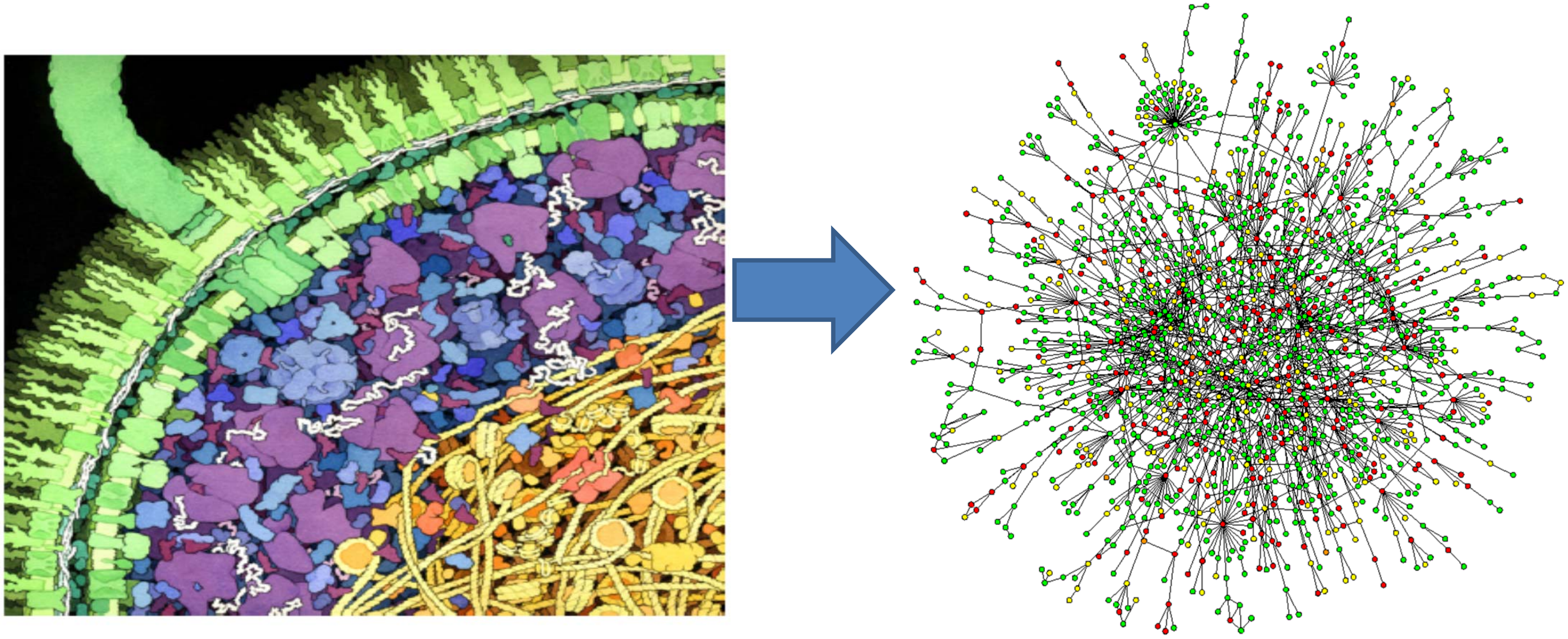
Yong Wang

Academy of Mathematics & Systems Science

http://zhangroup.aporc.org

Chinese Academy of Sciences

Background: Network biology



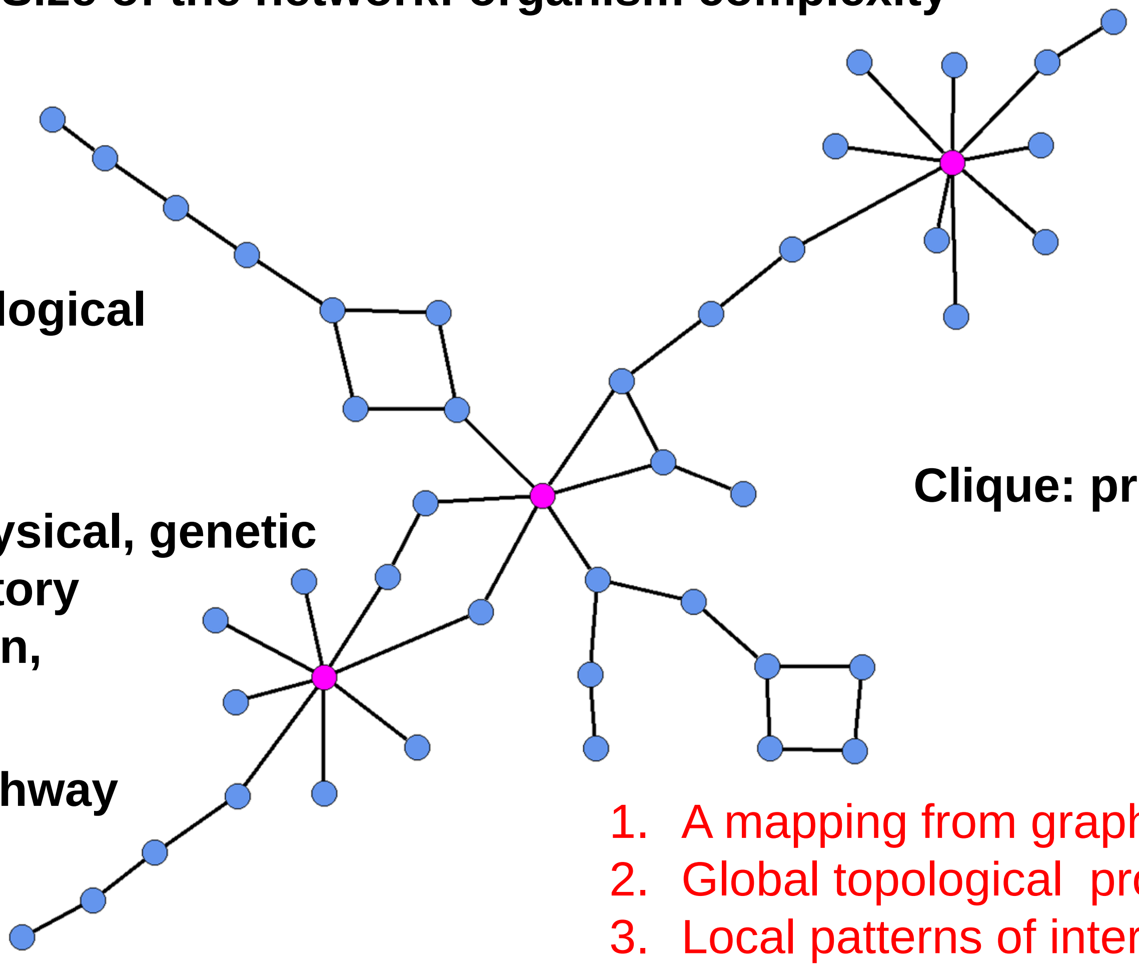
- Usually graphs are used to represent these complex biological systems

•1D Vs 3D: 2D representation

- Nodes denote biological molecules and edges denote their relationships

A quick view of Network biology

Size of the network: organism complexity



Hub: essential or Evolutionarily conserved molecules

Node: biological molecule

Clique: protein complex

Edge: physical, genetic or regulatory interaction,

Path: pathway

1. A mapping from graph theory to biology
2. Global topological properties
3. Local patterns of interaction

Huge successes

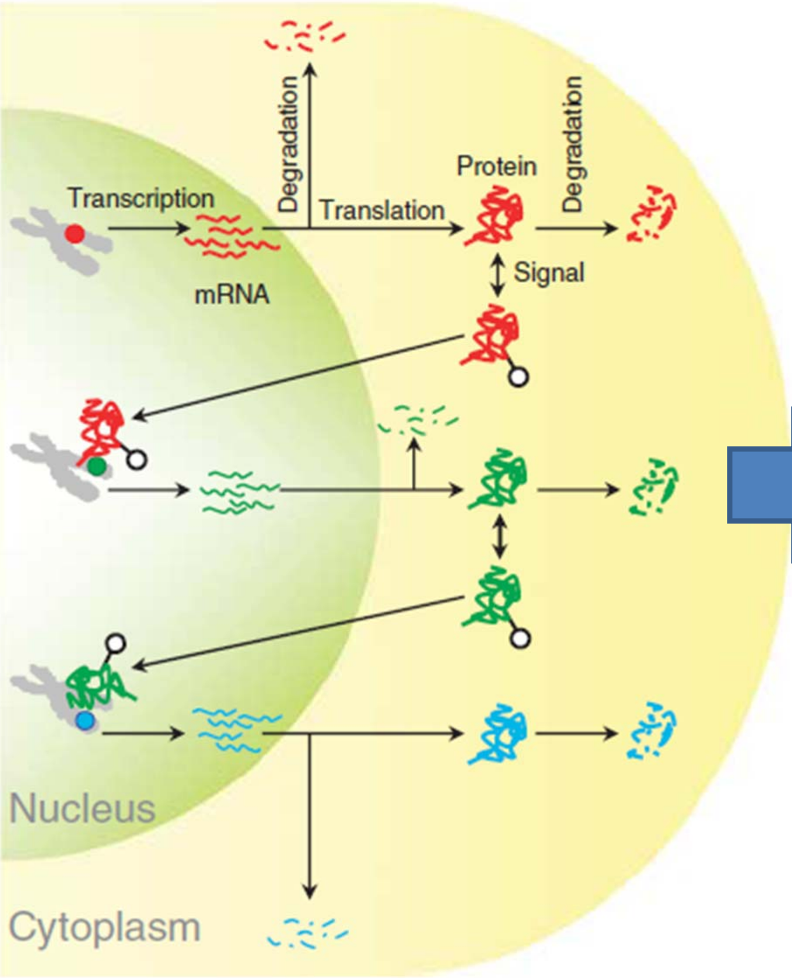
Revealing the large scale organization and evolutionary principles of a cell

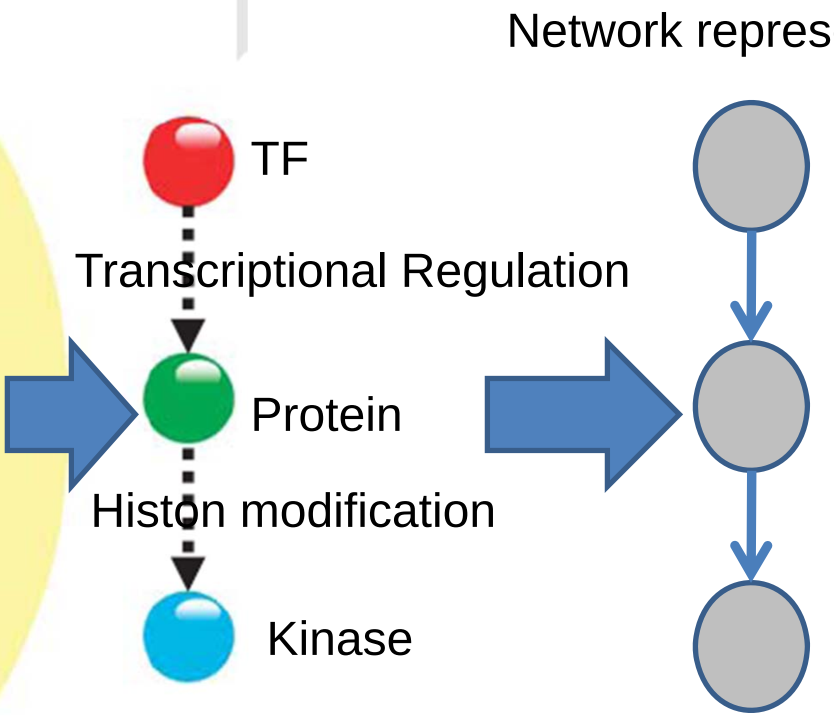
- ***Cellular networks are scale-free***
- ***High clustering in cellular networks***
- ***Motifs are elementary units of cellular networks***
- ***Hierarchy organization of topological modules***
- ***Topological, functional and dynamic robustness***
- ***…***

NATURE REVIEWS | GENETICS VOLUME 5 | FEBRUARY 2004 | **101**

Is it enough to study the whole network?

- Observation: Although protein-protein interactions are conveniently represented as nodes and edges in a network, it is important to note that each node in the network represents several entities (proteins in different tissues) and events (transcription, translation, degradation, etc) that are compressed in both space and time.




Network representation

TF

Transcriptional Regulation

Protein

Histon modification

Kinase

Although a series of regulatory events can be conveniently represented as a node in the network, the dynamics of the entities and the biological processes that make up the node are not captured.

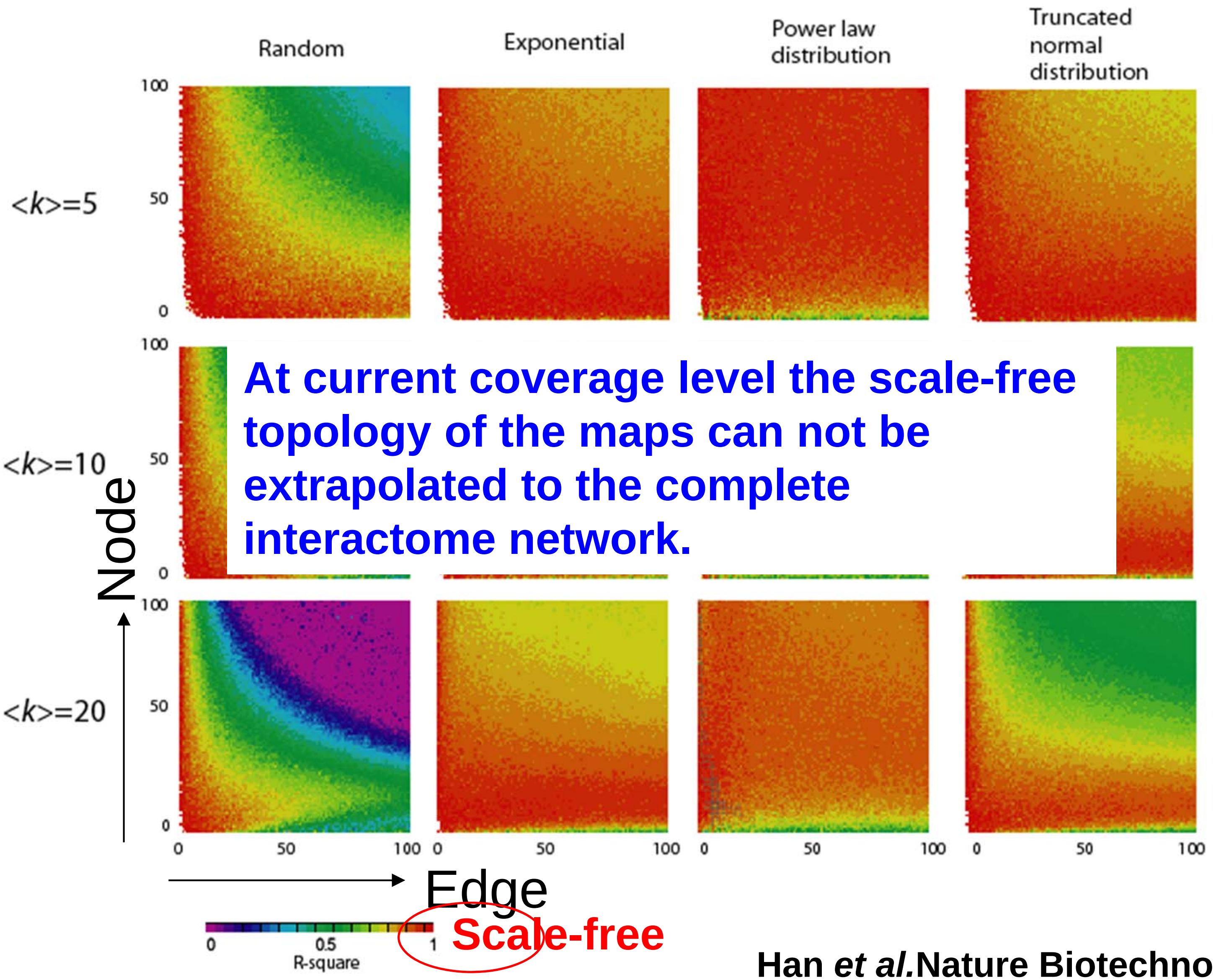
Molecular Systems Biology 5:294

Subnetwork VS whole network?

- Observation: Genome-wide network and subnetwork can be very different
- An example:

1. The current interactome maps cover only a small fraction of the total interactome (3-15%).
2. Basic observation: the current interactome is scale free
3. Question: can we infer that the topology of complete interactome networks is scale free?

The answer is: No



At current coverage level the scale-free topology of the maps can not be extrapolated to the complete interactome network.

Scale-free

Han *et al.*Nature Biotechnology, 2005

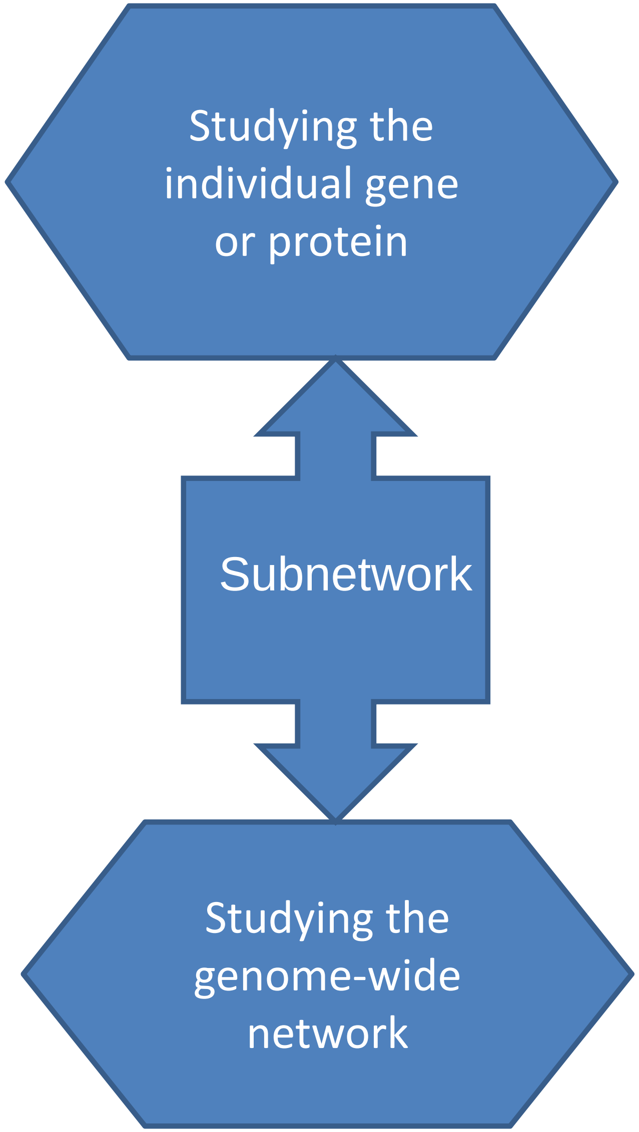
Subnetwork?

- Many network-based studies focus on graph theoretical analysis of nodes and edges within a single, global biomolecular network. However, there exists a high level of chemical and functional heterogeneity within the underlying biomolecules, biomolecular interactions, and interactome subnetworks.
- It remains an open question whether or not the global properties of the full interactome extend to these subnetworks.
- In addition, subnetworks may exhibit unique, emergent properties that are absent in the conglomeration of the full interactome.

Studying subnetwork is important

- Studying a group of condition specific genes or proteins and their relationships.

- The concept of subnetwork is very important and extensively applied in different contexts.



Studying the individual gene or protein

Subnetwork

Studying the genome-wide network

Subnetwork

- Subnetworks can reveal the complex patterns of the whole-genome network

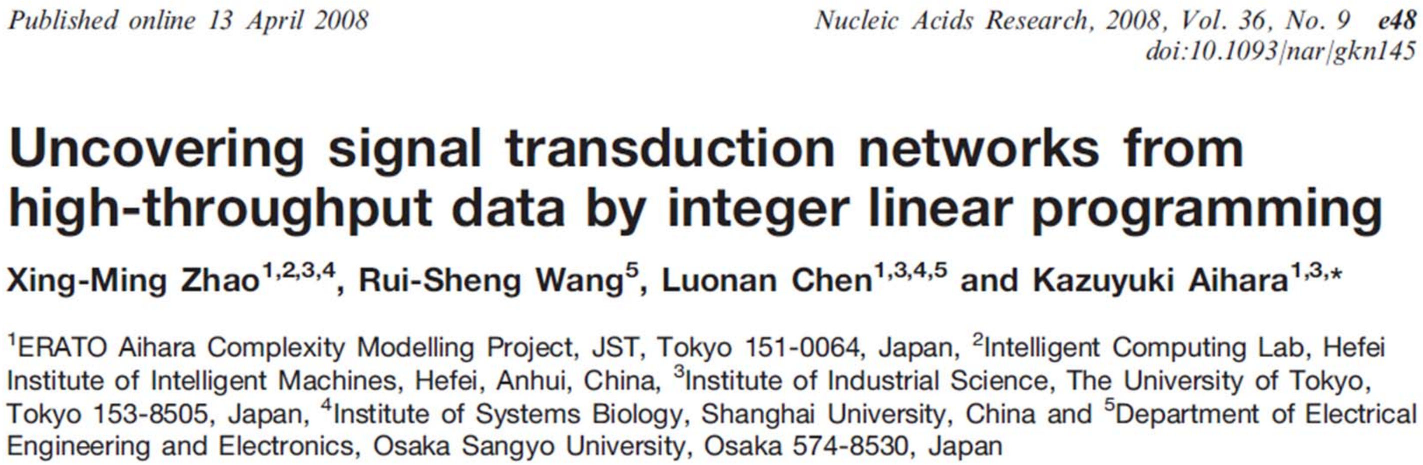
Temporal: The evolutionarily conserved subnetworks

Spatial: Protein complexes depending on the sub-cellular localization

Condition specific context: Subnetwork biomarker for diseases

- Novel subnetwork identification methods that are flexible and efficient are still much needed.

Automatic modeling of signaling pathways from protein-protein interaction networks



Published online 13 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 9 ***e48***
doi:10.1093/nar/gkn145

Uncovering signal transduction networks from high-throughput data by integer linear programming

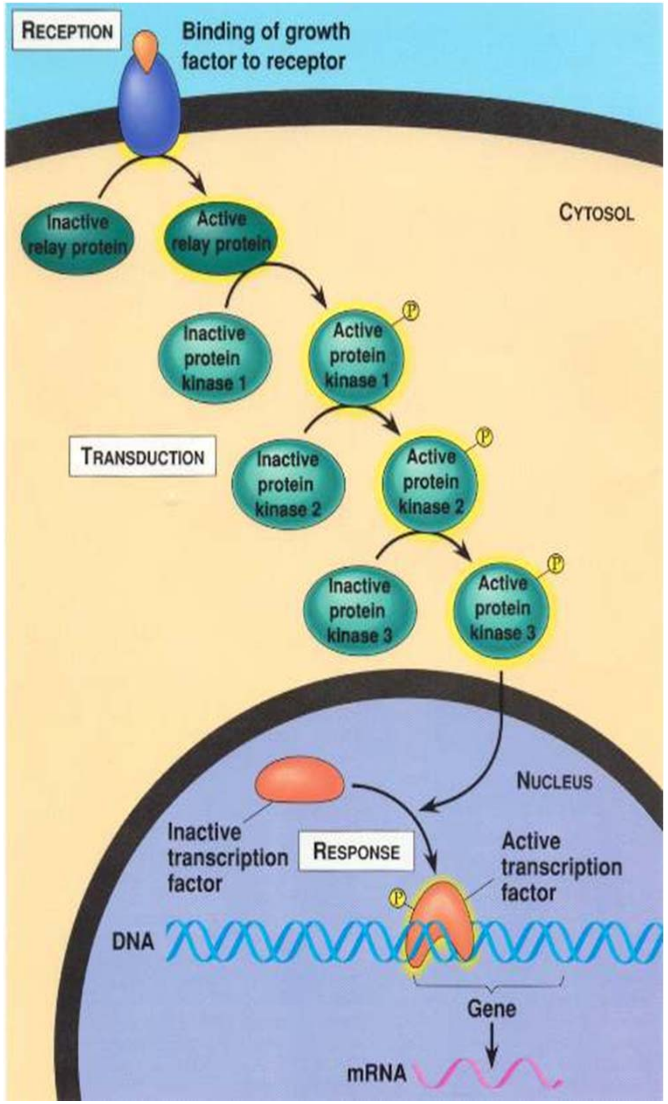
Xing-Ming Zhao[1,2,3,4], Rui-Sheng Wang[5], Luonan Chen[1,3,4,5] and Kazuyuki Aihara[1,3,*]

[1]ERATO Aihara Complexity Modelling Project, JST, Tokyo 151-0064, Japan, [2]Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Hefei, Anhui, China, [3]Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan, [4]Institute of Systems Biology, Shanghai University, China and [5]Department of Electrical Engineering and Electronics, Osaka Sangyo University, Osaka 574-8530, Japan

Outline

- Background
- Previous works on this topic
- Signaling network reconstruction by integer linear programming
- Experimental results
- Conclusions

Background


RECEPTION
Binding of growth factor to receptor
CYTOSOL
Inactive relay protein
Active relay protein
Inactive protein kinase 1
Active protein kinase 1
TRANSDUCTION
Inactive protein kinase 2
Active protein kinase 2
Inactive protein kinase 3
Active protein kinase 3
P
NUCLEUS
Inactive transcription factor
RESPONSE
Active transcription factor
DNA
Gene
mRNA

Signal transduction

» Movement of signals from outside the cell to inside; Cells always receive different signals from the physical environment and from other cells.
细菌的群体感应 (quorum sensing)

» Mediate the sensing and processing of stimuli; Many cellular decisions such as proliferation, differentiation, development and other responses to external stimuli are achieved by signal transduction.

» Abnormality in cellular information processing are responsible for diseases such as cancer, heart disease, autoimmunity, and diabetes.

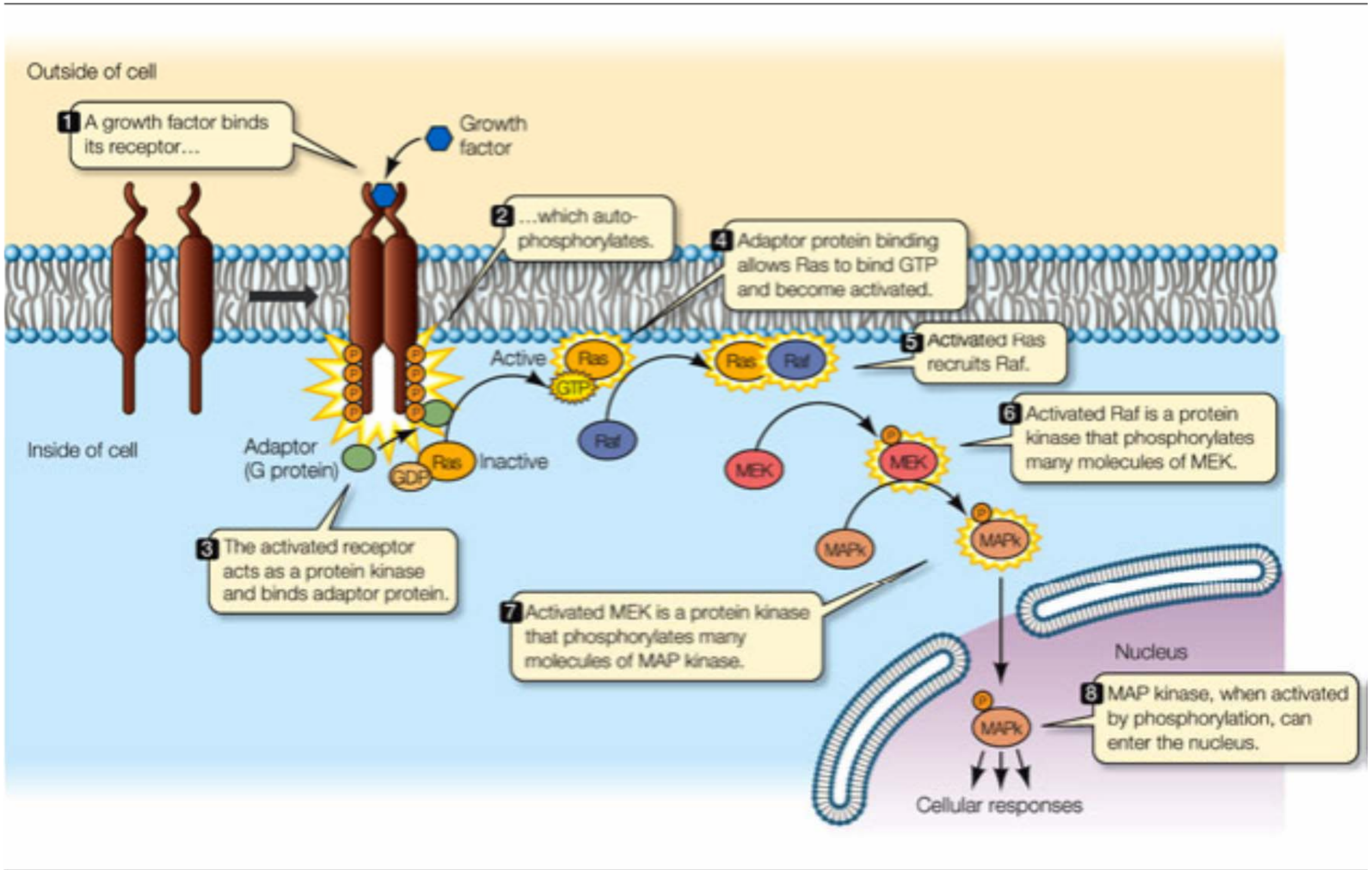
Background (cont.)

- Signal transduction processes are activated by multiple extracellular factors as well as cell membrane receptors to mediate the regulation of target gene expression.

- Cells respond to signals by specific receptor proteins that can bind those signals.

- The ultimate cellular response to a signal may be the opening of ion channels, the alteration of enzyme activities, or changes in gene transcription.

Background (cont.)



1 A growth factor binds its receptor…

2 …which auto-phosphorylates.

3 The activated receptor acts as a protein kinase and binds adaptor protein.

4 Adaptor protein binding allows Ras to bind GTP and become activated.

5 Activated Ras recruits Raf.

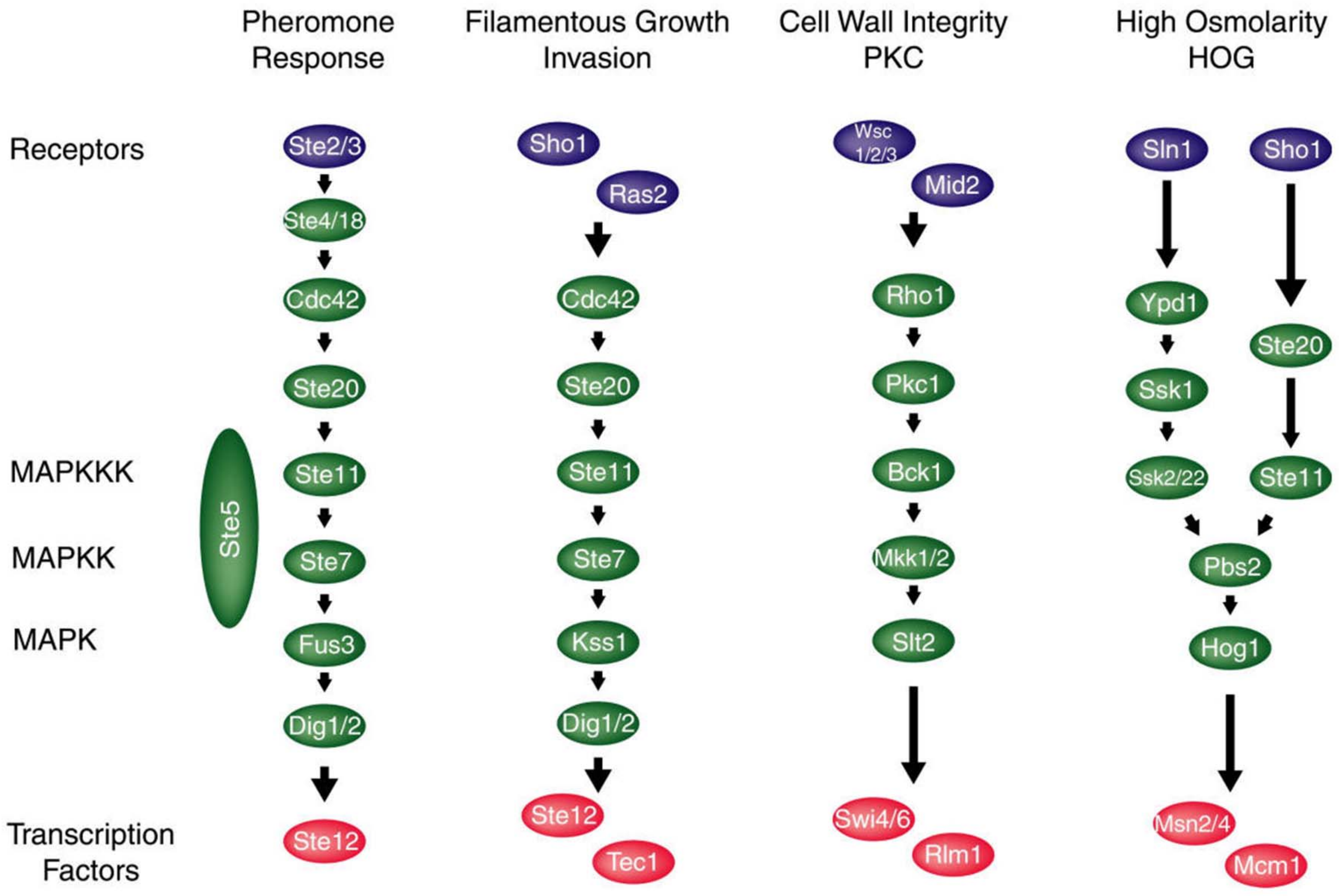
6 Activated Raf is a protein kinase that phosphorylates many molecules of MEK.

7 Activated MEK is a protein kinase that phosphorylates many molecules of MAP kinase.

8 MAP kinase, when activated by phosphorylation, can enter the nucleus.

An illustration of signal transduction from outside to inside of cell

MAPK signal transduction pathways in yeast



Background (cont.)

Methods for detecting components in signaling pathways:

- Experimental methods:
  - Knock out specific genes;
  - Time consuming and expensive;
    - Every reaction and component even in a relatively simple signaling pathway requires a concerted and decades-long effort.
    - Many signaling components and mechanisms are unknown. There is not a lot of kinetic data available with which to create models of pathway component interaction.
- Computational methods
  - Knowledge based methods;
  - Data based methods.

Background (cont.)

- Knowledge based methods:
  - Modeling pathways by ordinary differential equations;
  - Modeling pathways by Petri net
  - Limited by the scale, lack of kinetic coefficients

- Data based (our focus):
  - High-throughput techniques result in large mounts of biological data.
  - Recovering signal transduction pathways and identifying key components from multiple data sources.
    - Large scale.
    - Data dependency.

Previous works

NetSearch algorithm

Steps:

- Potential pathways detected by Depth First Search (DFS) algorithm from PPI network;
- Ranking candidate pathways according to the clustering results on gene expression data.
- The more the elements in candidate pathways overlap with a cluster, the more likely they are true components.

Ref: “Automated modelling of signal transduction networks”, BMC Bioinformatics 2002, 3:34.

Previous works (cont.)

Ordering the signal pathway with score function

Steps:

- Assume the components in a signaling pathway are known. Only the order of the components is unknown
- Find the candidate pathways by using PPIs, i.e. assign each order a score
- Ordering the signal pathways by using gene expression data (pairwise correlation coefficients).

Ref: “A computational approach for ordering signal transduction pathway components from genomics and proteomics data”, *BMC Bioinformatics, 5, 158, 2004*

Previous works (cont.)

Color coding

Given a weighted PPI network

- Find candidate signaling pathways by a variant of color coding algorithm;
- Assemble top-scoring candidate pathways into signaling network.

- Ref: "Efficient Algorithms for Detecting Signaling Pathways in Protein Interaction Networks", Journal of Computational Biology 2006.

Previous works (cont.)

- Problems lying in the previous work:
  - Individual signaling pathways are identified and then heuristically rank and assemble them into a signal transduction network;
  - Multi-stage tends to lead to local optimal solutions.
- A one-stage method with global optimal solutions is needed

Our ideas about recovering signaling networks

- Proteins involving in a same signaling pathway tend to interact with each other

- The model tries to find a subnetwork with highest sum of edge weights (there is a tradeoff between the sum of edge weights and the number of edges) from a membrane protein (receptor) to a transcription factor in a big protein-protein interaction (PPI) network.

- The extraction process is formulated into an integer linear programming model, which will be relaxed into a linear programming in the practical applications

Recovering signaling networks by integer linear programming

$$\text{Min} \quad \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} a_{ij}e_{ij} + \lambda \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} e_{ij}$$

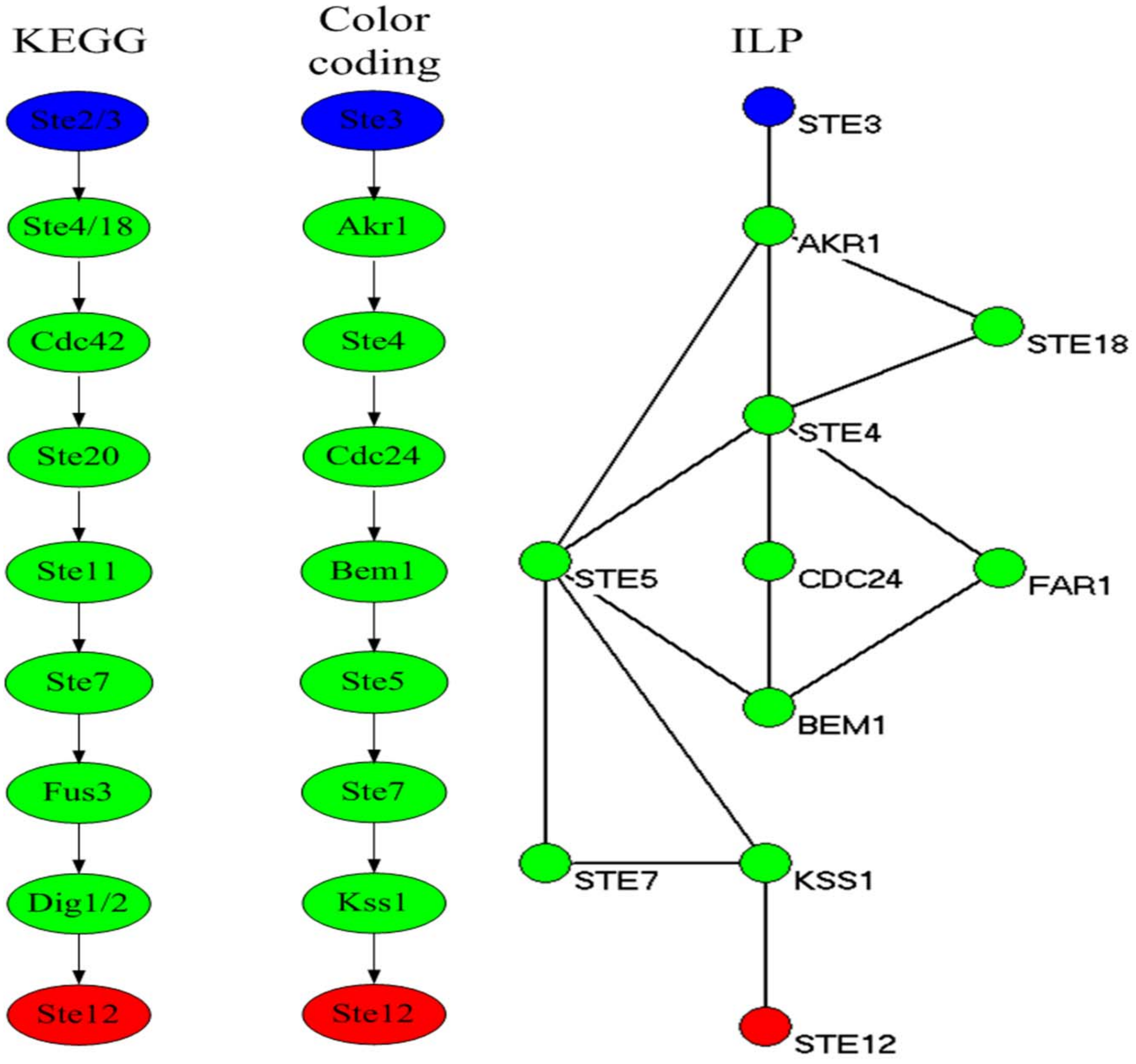
$$\begin{aligned}
\textbf{s.t.} \quad & e_{ij} \le x_i \\
& e_{ij} \le x_j \\
& \sum_{j} e_{ij} \ge 1, \quad \text{if } i \text{ is a membrane protein or TF} \\
& \sum_{j} e_{ij} \ge 2x_i, \text{if } i \text{ is not a membrane protein or TF} \\
& x_i = 1 \text{ , if } i \text{ is a membrane protein or TF} \\
& x_i \in \{0, 1\}, \ i = 1, 2, \cdots, |V| \\
& e_{ij} \in \{0, 1\}, \ i, j = 1, 2, \cdots, |V|
\end{aligned}$$

- a_{ij} – PPI strength
- x_i – binary variable for protein *i*
- e_{ij} – binary variable for protein interaction *(i,j)*
- λ – penalty parameter
- One step and global model !!!

Experimental results

- Experimental data:
  - Yeast protein interaction network with ∼4,500 nodes and ∼14,500 edges.
- Pre-process:
  - Find the paths of length 6-8 from the PPI network using the Depth-first search;
  - The reduced network consist of all possible candidate pathways.

Pheromone response (linear path)

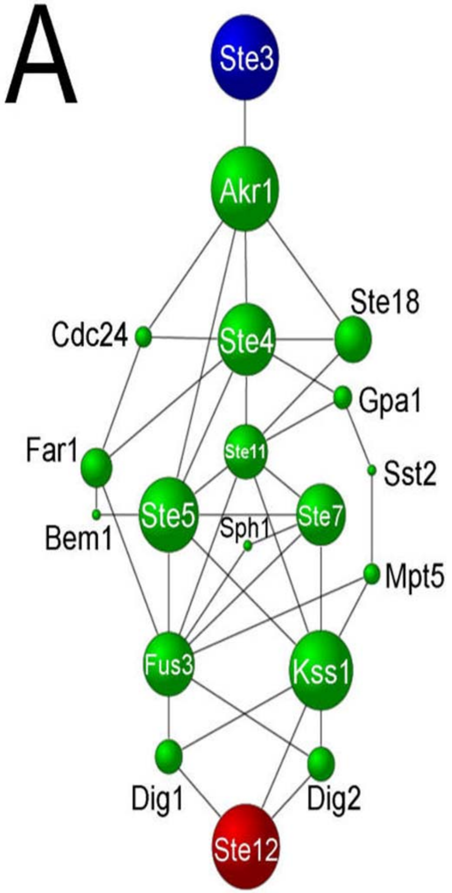


信息素响应

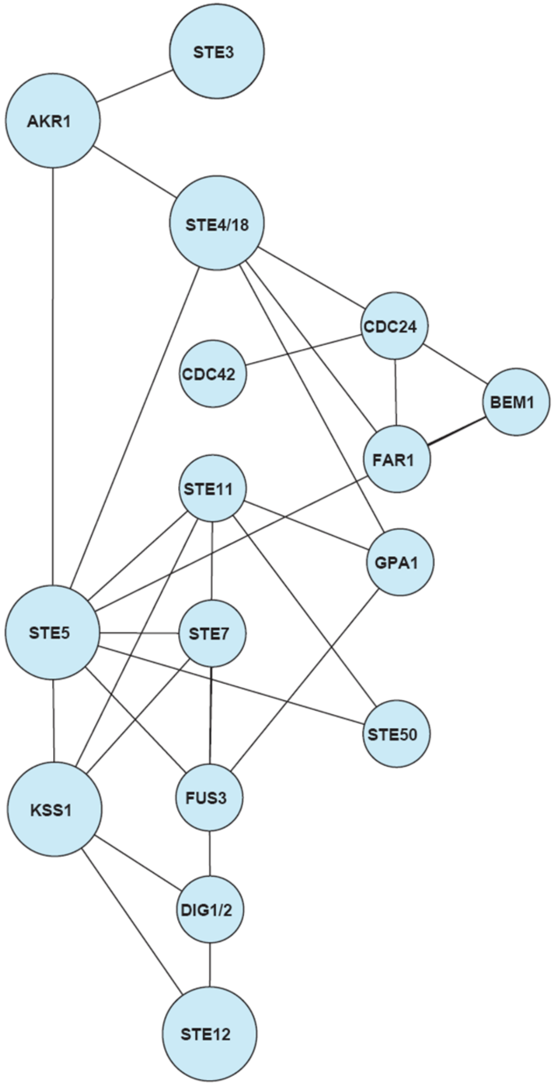
We find additional components. Such redundant mechanisms can compensate single protein disruptions and maintain signal transduction unblocked

- pheromone response (signaling network)

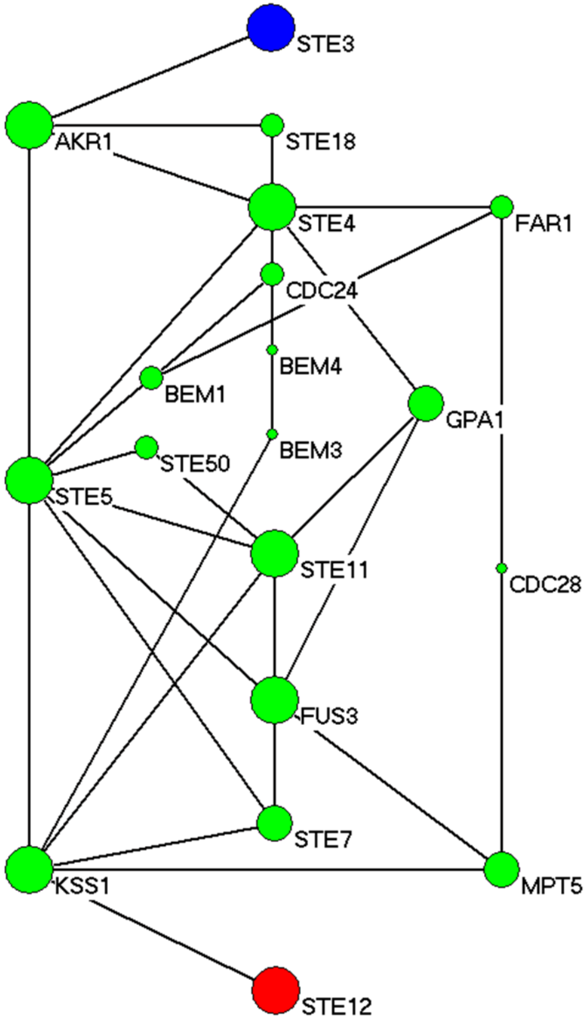
Results by Netsearch



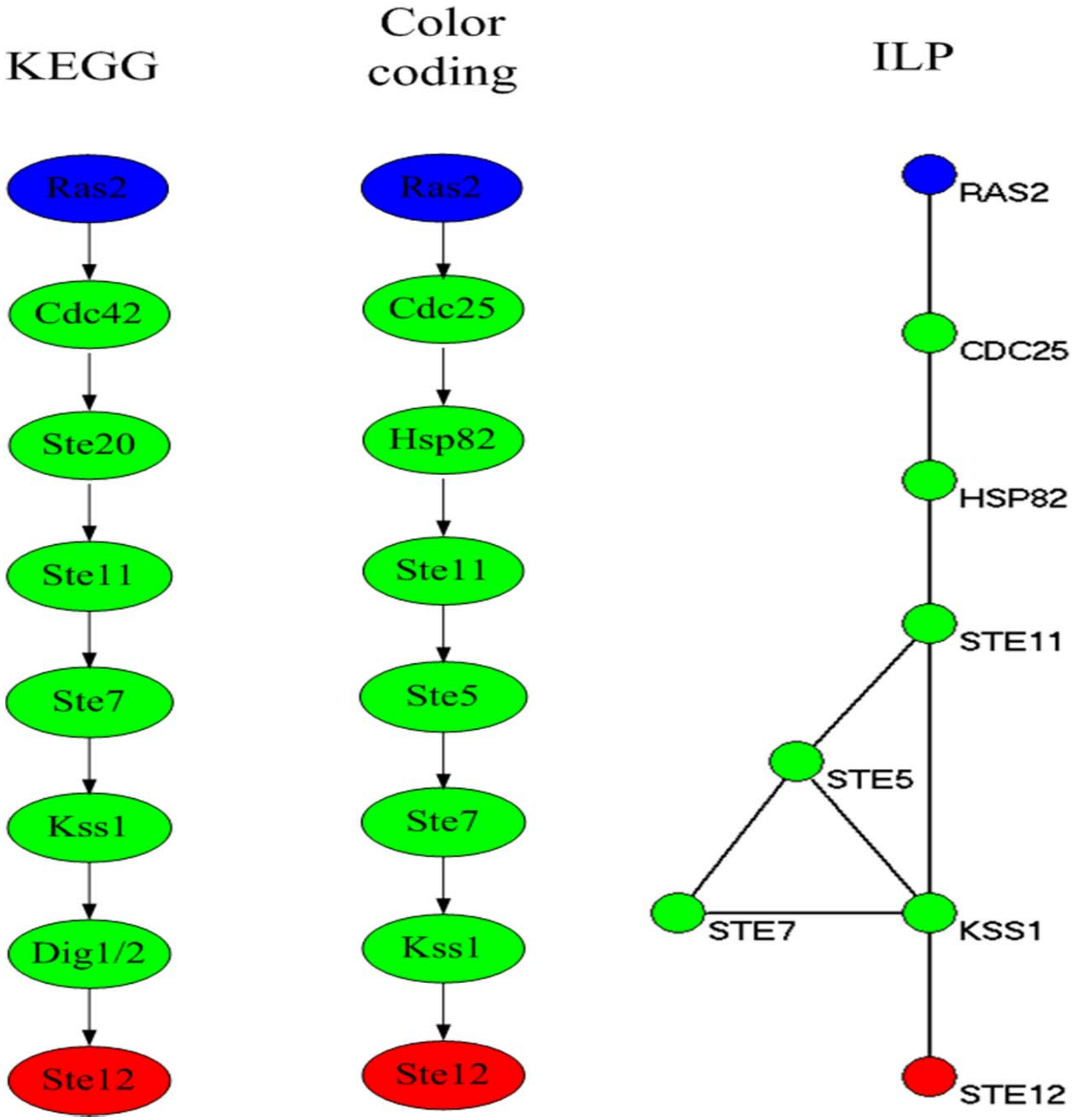
results by color coding



results by ILP



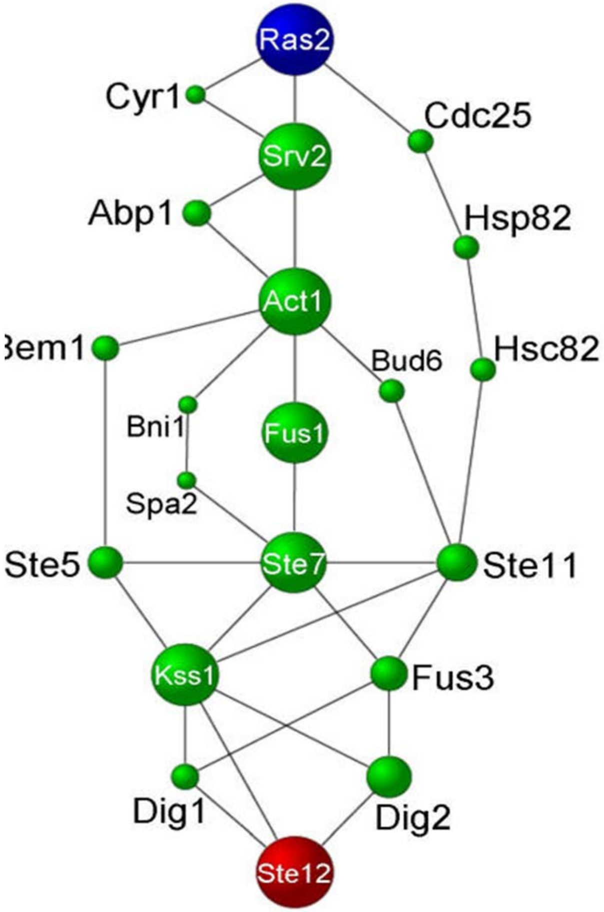
- Filamentation pathways (linear path)



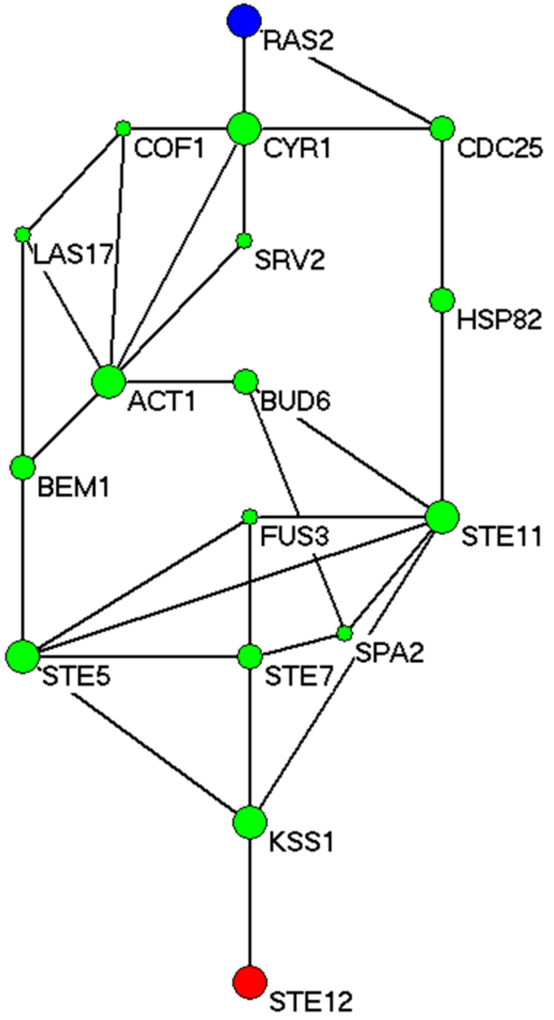
细菌成丝

Filamentation pathways(signaling network)

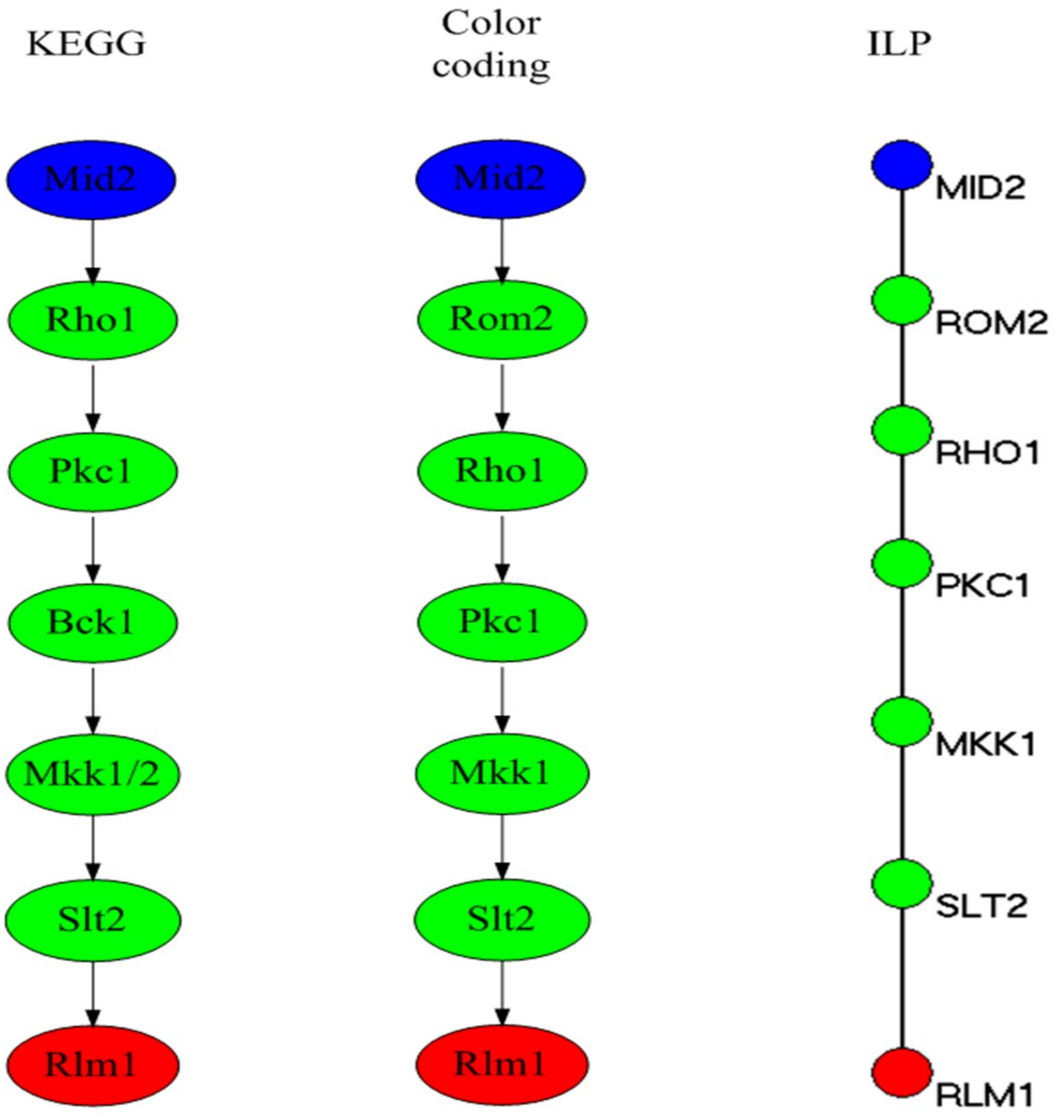
Results by Netsearch



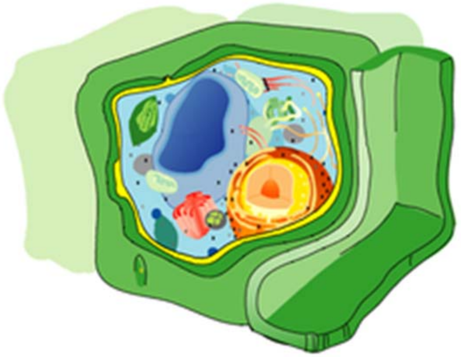
results by ILP



Cell wall integrity (linear path)



细胞壁



Our method can detect the exact pathway that other algorithms found

- These results on known yeast MAPK signaling pathways demonstrate that the ILP model can recover the known signaling pathways, and the reconstructed STNs match most parts of those published results

- Compared with existing methods, our method is much simpler in both algorithm and computation because it can detect the signaling networks from protein interaction data directly in an integrated and accurate manner

- Our method can handle a large scale system without numerical difficulty due to the LP algorithm.

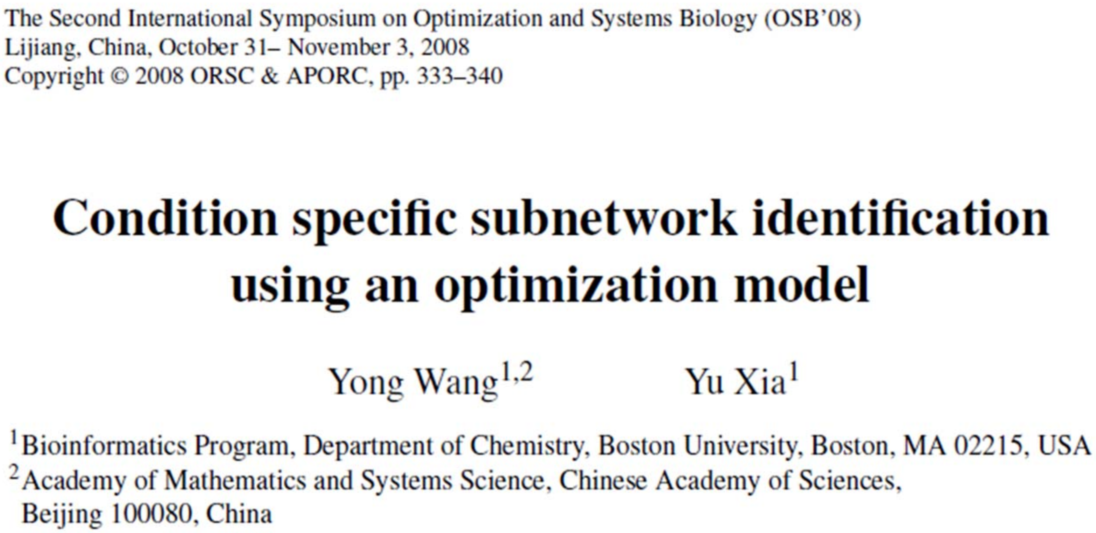
Conclusion and future work

- Proposed LP algorithm is effective for inferring the signaling network; It is a one-stage method and does not need heuristic ranking and assembling

- Protein interactions have no timing information. In the future, we will integrate PPIs with gene expression data for signaling network detection, which will make the detection more realistic

- We will also explore the further application of the method to other signaling networks except MAPK pathways.

Optimization model for condition specific subnetwork identification



The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
Copyright © 2008 ORSC & APORC, pp. 333–340

Condition specific subnetwork identification using an optimization model

Yong Wang[1,2] Yu Xia[1]

[1]Bioinformatics Program, Department of Chemistry, Boston University, Boston, MA 02215, USA
[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
Beijing 100080, China

http://zhangroup.aporc.org

Chinese Academy of Sciences

Subnetwork

- Subnetworks can reveal the complex patterns of the whole-genome network

Temporal: The evolutionarily conserved subnetworks

Spatial: Protein complexes depending on the sub-cellular localization

Condition specific context: Subnetwork biomarker for diseases

- Novel subnetwork identification methods that are flexible and efficient are still much needed.

Problem formulation

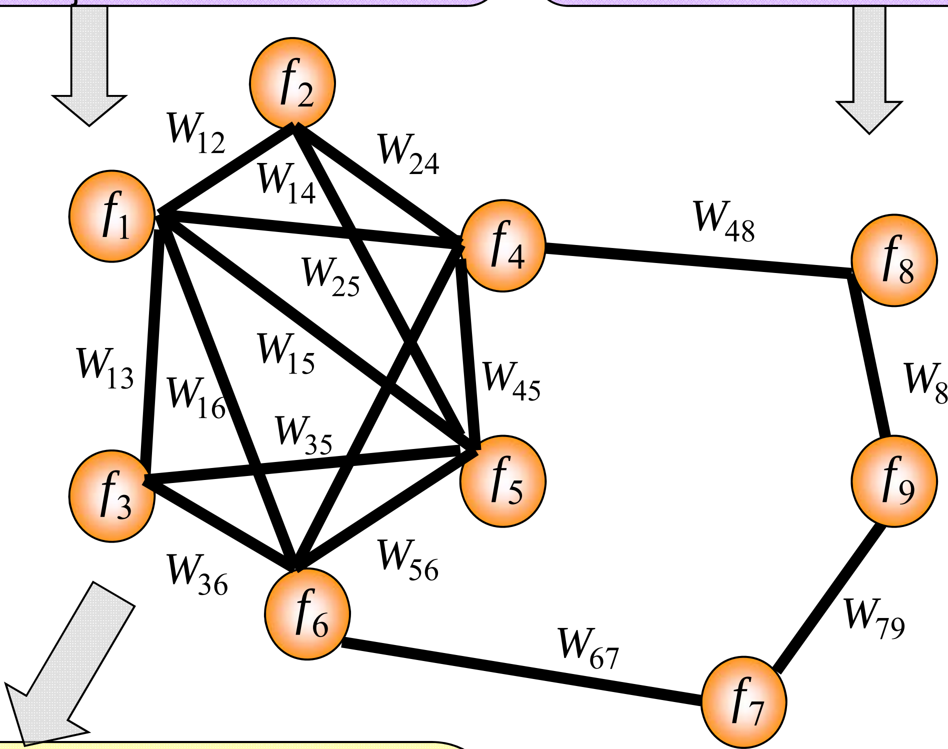
- Input:

G=(V,E) is the network with n nodes V_1,V_2. … V_n. We use a symmetric weight matrix W to quantify the connectivity strength (for example, W can be the edge confidence scores for biomolecular interaction or functional linkage networks). $W_{ij} \geqslant 0$, I, j=1, 2$\cdots$n.

Every node V_i is associated with a profile (for example gene expression data, or other properties related to the nodes). We consider the simplest case (weight f_i).

Constructing molecular network G=(V, E) and edge weight Matrix {W_{ij}, i,j=1,2,…,n }

Assembling condition specific information $f_1, f_2, f_3, \ldots, f_n$ for molecules



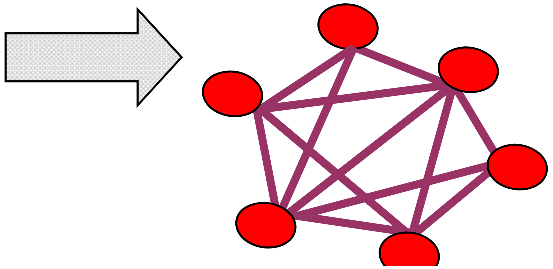
Optimization model

$$\max \quad \sum_{\mathrm{i}} \sum_{\mathrm{j}} W_{ij} x_i x_j + \lambda \sum_{i} f_i x_i$$

$$s.t. \quad x_1^{\beta} + x_2^{\beta} + x_3^{\beta} + \cdots + x_n^{\beta} = 1$$

$$x_{\mathrm{i}} \geq 0 \qquad i = 1, 2, \cdots, n$$

Condition specific subnetwork



- Then we have two objects:

1. Choose as many as possible edges within the subnetwork (maximize the interconnectivity within the subnetwork)
2. Maximize the degree of association between the subnetwork nodes and the specific condition.
3. We introduce a parameter to integrate them.

- We introduce a regularization constraint that limit the number of nodes selected.

1. Parameter β is introduced to adjust the strength of regularization applied to the variable $x=(x^1,x^2,\ldots,x^n)$
2. When β=2, this is a trust region problem which optimizes a quadratic function
3. When β=1, the L1-type constraint will lead to a sparse solution, i.e., many of the entries will be zeros

Computational complexity

- If we focus only on the first term of objective function, our model can be used to find the maximum clique in an weighted graph (the Motzkin-Struss Formalism for computing maximal cliques, Motzkin-Straus Theorem, 1965)
- Both the maximum cardinality and the maximum weight clique problems are NP-hard.
- Biomolecular networks are often large in scale. In yeast the protein-protein interaction network is estimated to have about 6,000 nodes and 50,000 interactions.

A fast algorithm for large-scale problem

The KKT condition is:

$$L = -\sum_{\mathrm{i}}\sum_{\mathrm{j}} W_{ij} x_i x_j - \lambda \sum_i f_i x_i + \alpha(x_1^{\beta} + x_2^{\beta} + x_3^{\beta} + \cdots + x_n^{\beta} - 1) - \sum_i \mu_i x_i$$

$$\frac{\partial L}{\partial x_i} = 0 \Rightarrow \mu_i = -2(WX)_i - \lambda f_i + \alpha\beta x_i^{\beta-1} \qquad i = 1,2,\cdots,n$$

$$\mu_i x_i = 0 \qquad i = 1,2,\cdots,n$$

$$x_{\mathrm{i}} \geq 0, \qquad \mu_i \geq 0 \qquad i = 1,2,\cdots,n$$

$$x_1^{\beta} + x_2^{\beta} + x_3^{\beta} + \cdots + x_n^{\beta} = 1$$

Then we can use the following iterative algorithm to quickly converge to a local minimum satisfying KKT condition:

$$\alpha = {(2X^T WX + \lambda \sum_i f_i x_i)} \Big/ {\beta}$$

$$x_i^{t+1} = (x_i^t \frac{2(WX)_i + \lambda f_i)}{\alpha\beta})^{\frac{1}{\beta}} = (x_i^t \frac{2(WX)_i + \lambda f_i)}{2X^T WX + \lambda \sum_i f_i x_i})^{\frac{1}{\beta}}$$

Proof of Correctness

Lagrangian function

$$L=-\sum_{\mathrm{i}}\sum_{\mathrm{j}}W_{ij}x_ix_j-\lambda\sum_i f_ix_i+\alpha(x_1^\beta+x_2^\beta+x_3^\beta+\cdots+x_n^\beta-1)-\sum_i\mu_ix_i$$

Complementarity Slackness:

$$[2(Wx)_i+\lambda f_i-\alpha\beta x_i^{\beta-1}]x_i=0$$

Lagrangian multipier value: $\alpha=2(x^TWx+\lambda\sum_{i=1}^n f_ix_i)/\beta$

Update rule: $x_i\leftarrow\left(x_i\dfrac{2(Wx)_i+\lambda f_i}{\alpha\beta}\right)^{1/\beta}$

At Convergence $x_i^*=\left(x_i^*\dfrac{2(Wx^*)_i+\lambda f_i}{\alpha\beta}\right)^{1/\beta}$ satisfies KKT condition

Proof of Convergence

Introducing auxiliary function

G(x,x') is an auxiliary function of L(x) if

$$G(x,x') \le L(x), \quad G(x,x) = L(x)$$

set $$x^{(t+1)} = \arg\max_x G(x, x^{(t)})$$

$$L\left(x^{(t)}\right) = G(x^{(t)}, x^{(t)}) \le G(x^{(t+1)}, x^{(t)}) \le L\left(x^{(t+1)}\right)$$

$$L\left(x^{(1)}\right) \le L\left(x^{(2)}\right) \le L\left(x^{(3)}\right) \le \cdots$$

L(x) is monotonically increasing and is bounded from up. Thus the algorithm converges

Proof of Convergence (cont)

Key: (1) find auxiliary function, (2) find global maxima

The auxiliary function is

$$G(x,x') = \sum_{ij} x'_i W_{ij} x'_j (1 + \log \frac{x_i x_j}{x'_i x'_j}) + \lambda \sum_{i=1}^{n} f_i x'_i - \alpha (\sum_i x_i^{\beta} - 1)$$

First order derivative:

$$\frac{\partial G(x,x')}{\partial x_i} = 2\frac{x'_i (Wx')_i}{x_i} + \lambda f_i x'_i - \alpha\beta x_i^{\beta-1}$$

2nd order derivative: is negative definite

$$\frac{\partial^2 G(x,x')}{\partial x_i \partial x_j} = -[2\frac{x'_i (Wx')_i}{x_i^2} + \lambda f_i x'_i + \alpha\beta(\beta-1) x_i^{\beta-2}]\delta_{ij}$$

Thus G(x,x’) is concave in x. we can obtain global maxima.

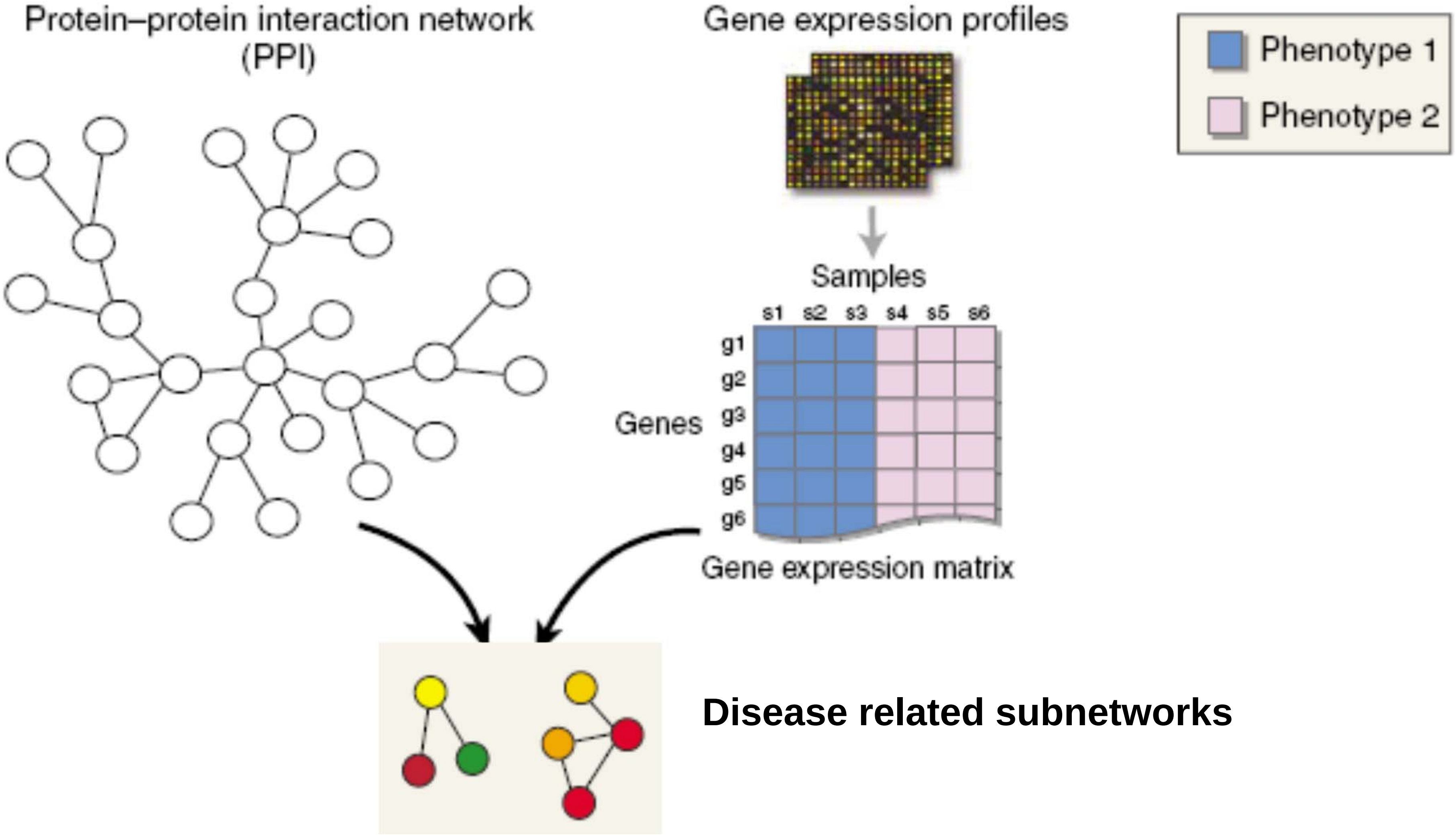
Notes on the model

- To relax the variable from integer to continuous variable in [0,1], we get a quadratic programming problem. The meaning can be the probability of that node to be a biomarker.

- The hardness of this programming depends on the network structure, maybe many local minimums exist. So careful choose of initial solution is necessary.

- We provide a deterministic way to replace the current heuristic based methods for subnetwork identification.

Finding the disease related subnetwork



Disease related subnetworks

Type 2 diabetes related subnetwork

- Type 2 diabetes mellitus is a complex disease with profound impact on health and longevity.
- It is estimated to affect more than 150 million people worldwide by the World Health Organization statistics.

Data integration

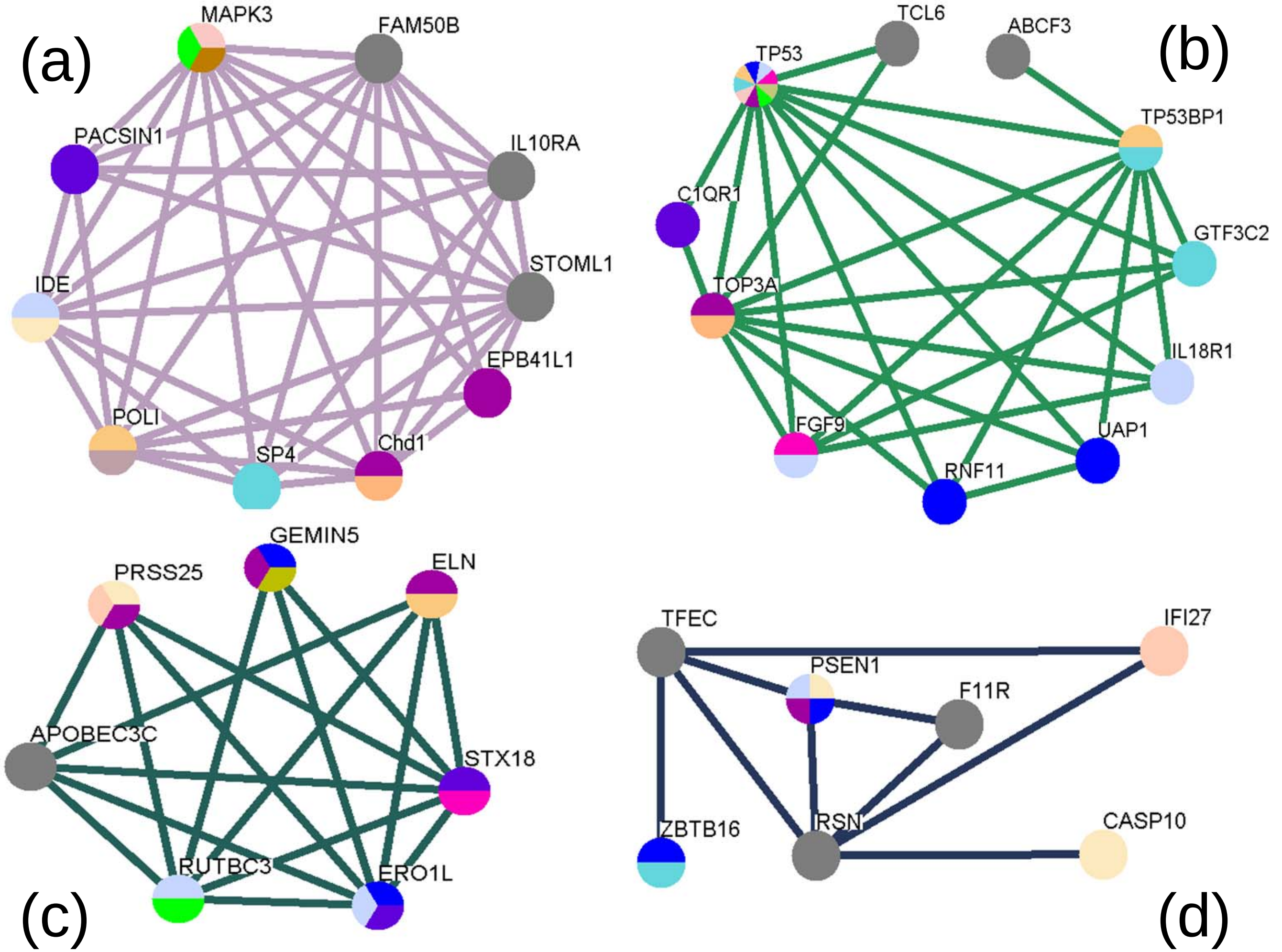
- The basic network is protein interaction network

We assembly the protein-protein interaction data in human have 7,903 proteins and 44,422 interactions. We make the sparse (the percentage of protein pairs that interact is only 0.14%.) denser by considering indirect interaction. In this way, we get a weighted protein-protein interaction network with 724,144 edges (2.3% of all protein pairs, a 16-fold increase in network size).

- Disease related data is confidence of association with T2D

We collected 2503 genes related to T2D and each gene is assigned a confidence score to be T2D candidate gene

Identified subnetworks



They are closely related to insulin-degradation, signal transduction, and metabolism functions.

Why “pilot study”?

- First, the present protein-protein interaction network in human is noisy and far from complete.
- Second, our basic assumption is that subnetworks are better biomarkers than single proteins, which needs further experimental and clinical verification especially for complex diseases such as T2D.
- Further research directions include validation of the effectiveness of subnetwork biomarkers, and improvement of the subnetwork identification algorithm.

Conclusions

- We propose a general framework to integrate two different kind of data.
- To find the disease related subnetwork is only a special case.
- We develop a general methodology to deal with it.

Take-home messages

- Subnetwork concept is very important.
- In essence it provides a efficient way to integrate heterogeneous data sources