パスウェイ解析・システム生物学入門

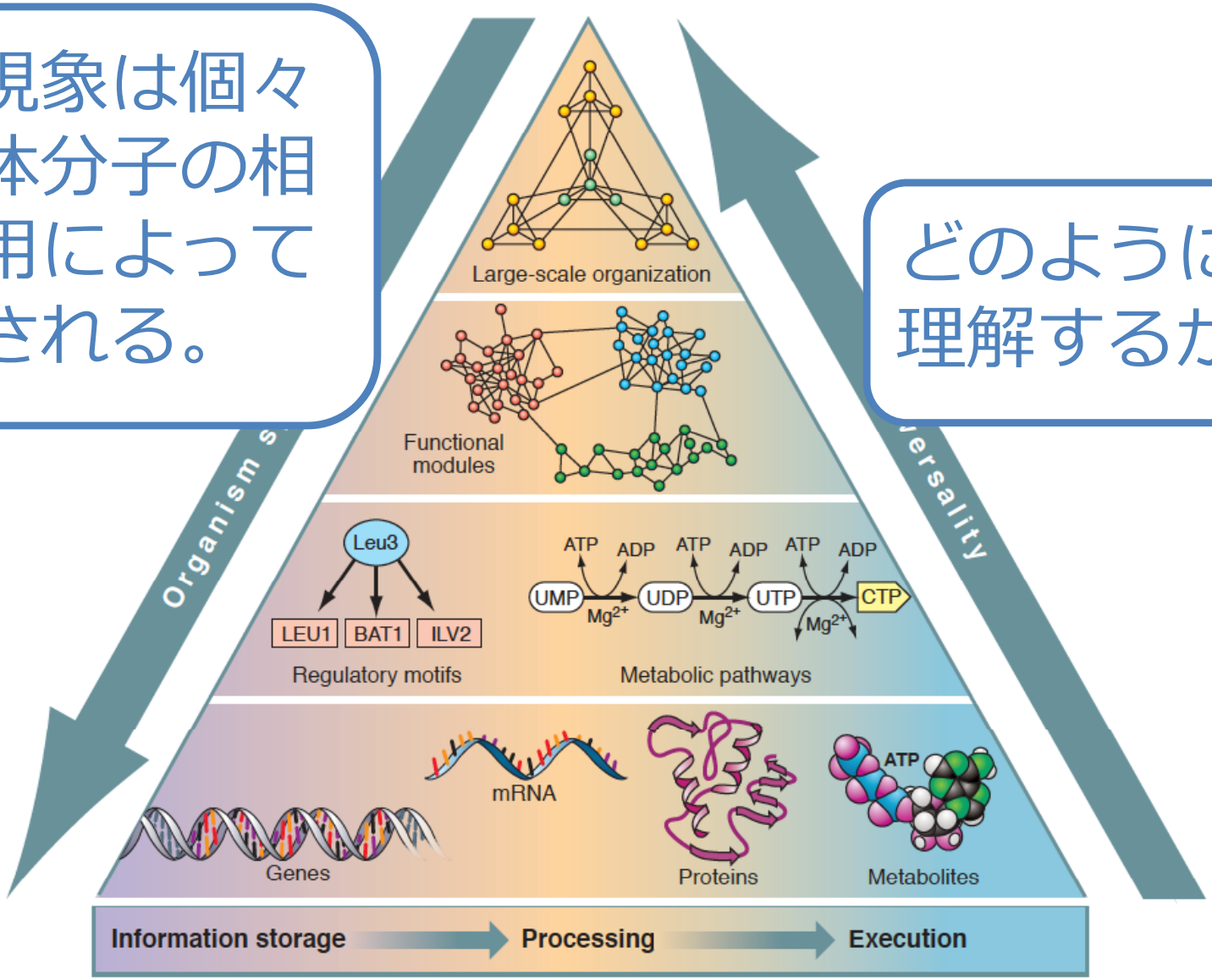
竹本和広（JSTさきがけ）

takemoto@cb.k.u-tokyo.ac.jp

パスウェイ解析・システム生物学の必要性

生命現象は個々の生体分子の相互作用によって記述される。

どのように記述、理解するか



Science, 298, 763 (2002)

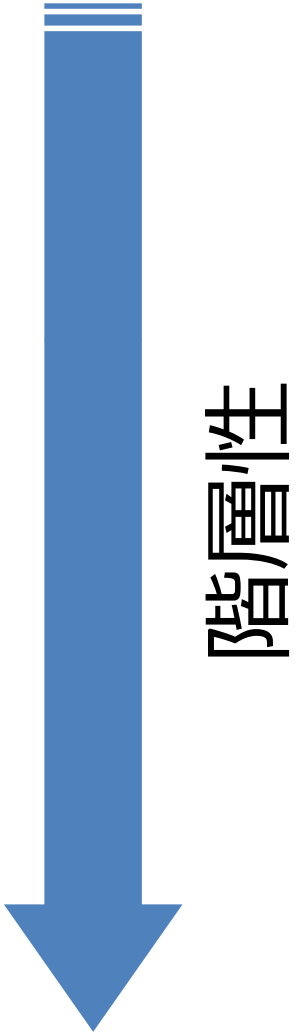
本日の流れ

- パスウェイ／ネットワーク解析
  - 相互作用をどのように表現・解析するか
- システム生物学
  - 上記で表現される相互作用する系のダイナミクスをどのように記述するか
- 演習
  - バイオインフォマティクス技術者認定試験の過去問を解く

パスウェイ／
ネットワーク解析入門

生体分子ネットワークの種類

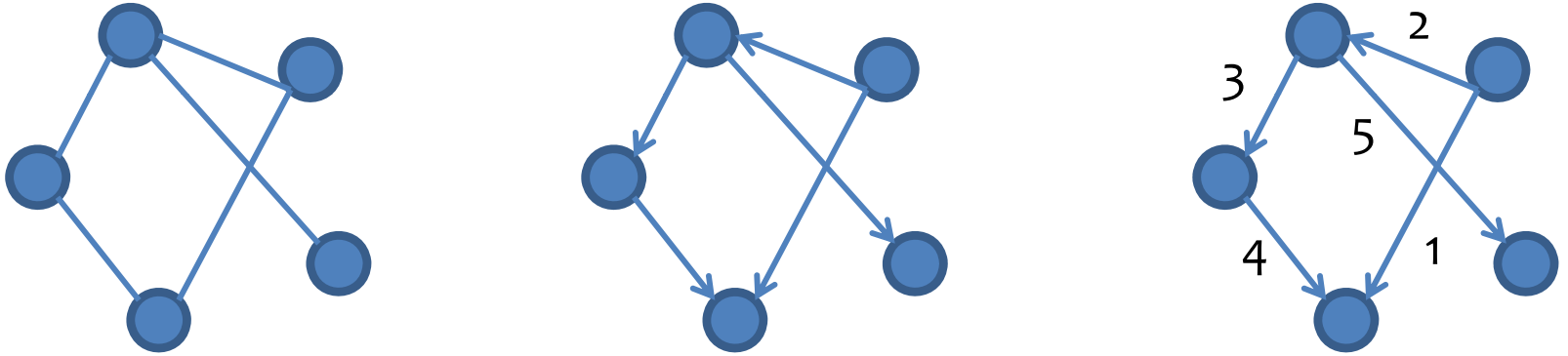
- 遺伝子制御ネットワーク
  - 遺伝子発現
- タンパク質相互作用ネットワーク
  - タンパク質複合体形成
- 代謝ネットワーク
  - 代謝（化合物の変換）



グラフによる相互作用の表現

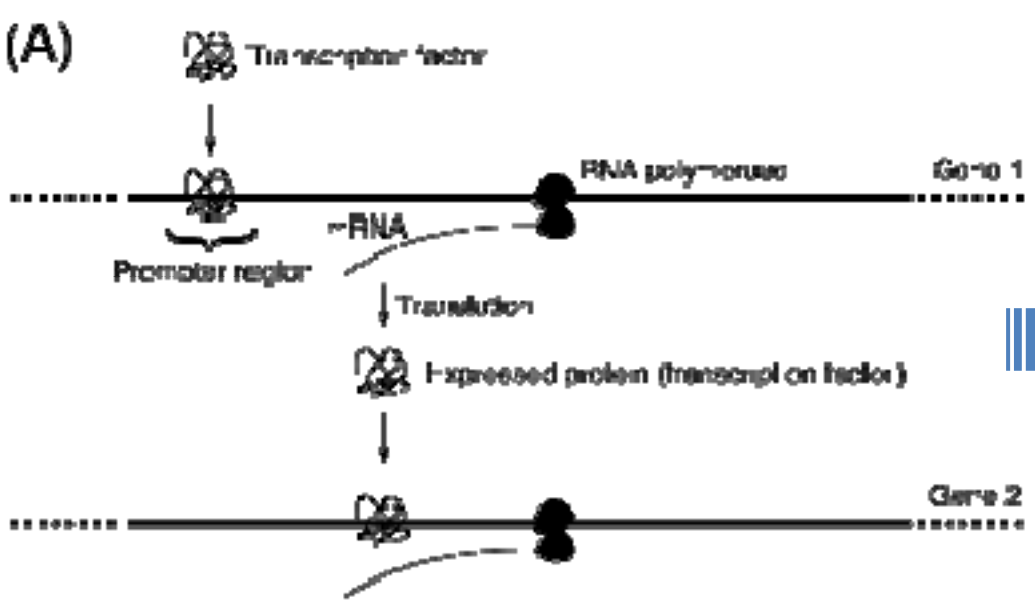
- 点（頂点）と線（辺）の集合

無向グラフ　　有向グラフ　　重み付き有向グラフ

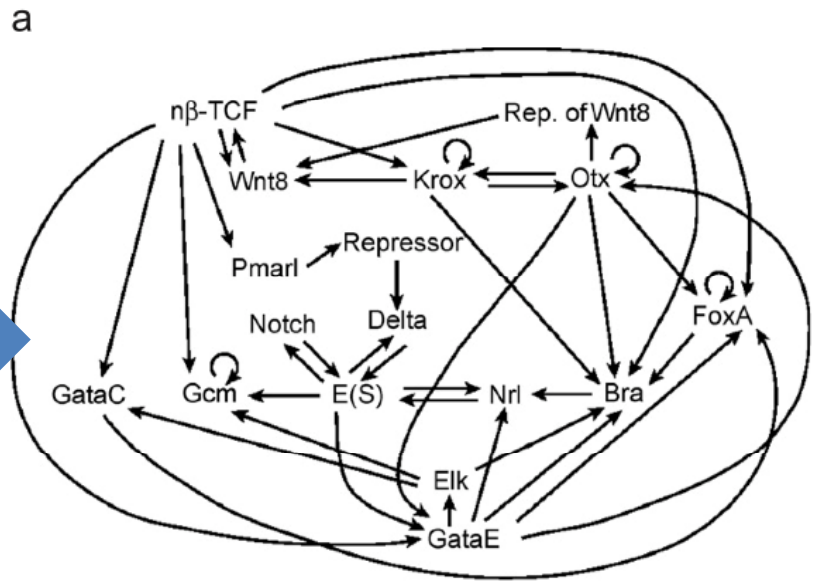


- 数学・情報学の分野におけるグラフ理論が適応できる。
- ただ、**表現力には限界がある**。

転写制御ネットワーク



活性／抑制



J Theor Biol **250,** 307 (2007)

ある遺伝子の転写はたの遺伝子にコードされるタンパク質（転写因子）によって調節される。

RegulonDB

- http://regulondb.ccg.unam.mx/
- 大腸菌の転写制御関係のデータベース

ここからダウンロードできる

転写因子

転写される因子

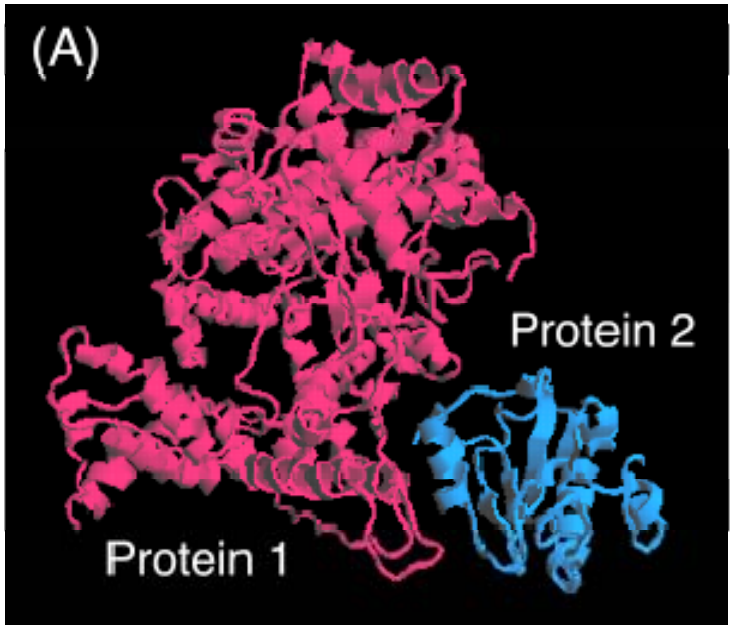
```
b2213   b2213   +-   Site mutation;Human inference based on similarity to c
b3131   b3139   -    Gene expression analysis;Human inference based on simi
b0504   b0515   +    Binding of cellular extracts;Human inference based on
…
```

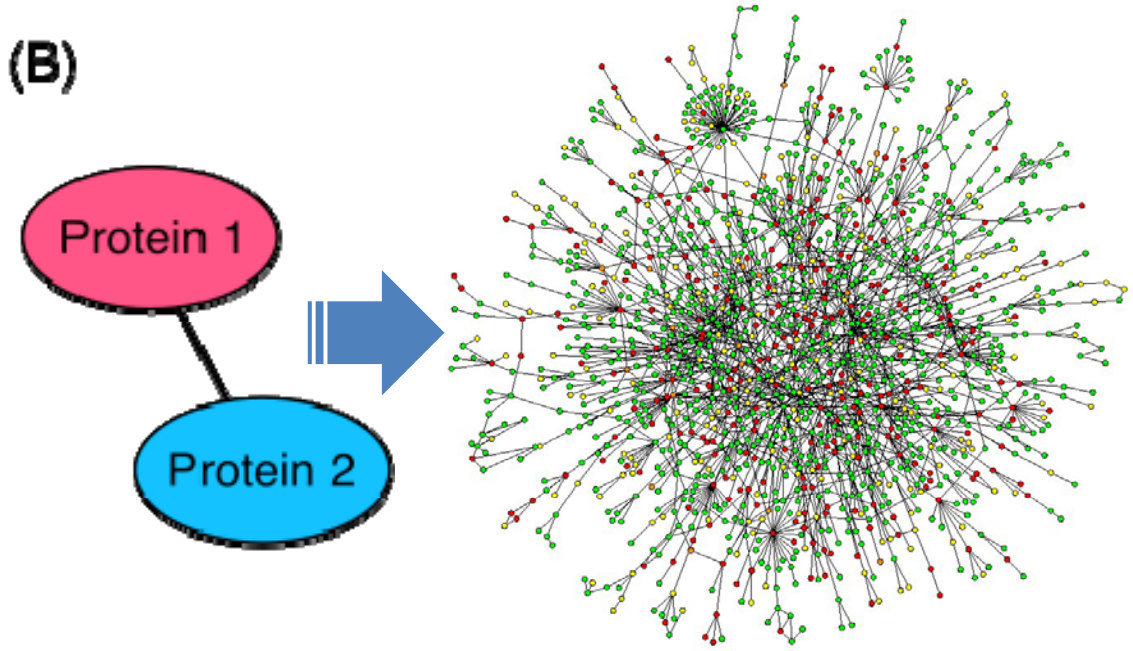
タイプ（活性／抑制）

検証方法

タンパク質間相互作用（PPI）ネットワーク

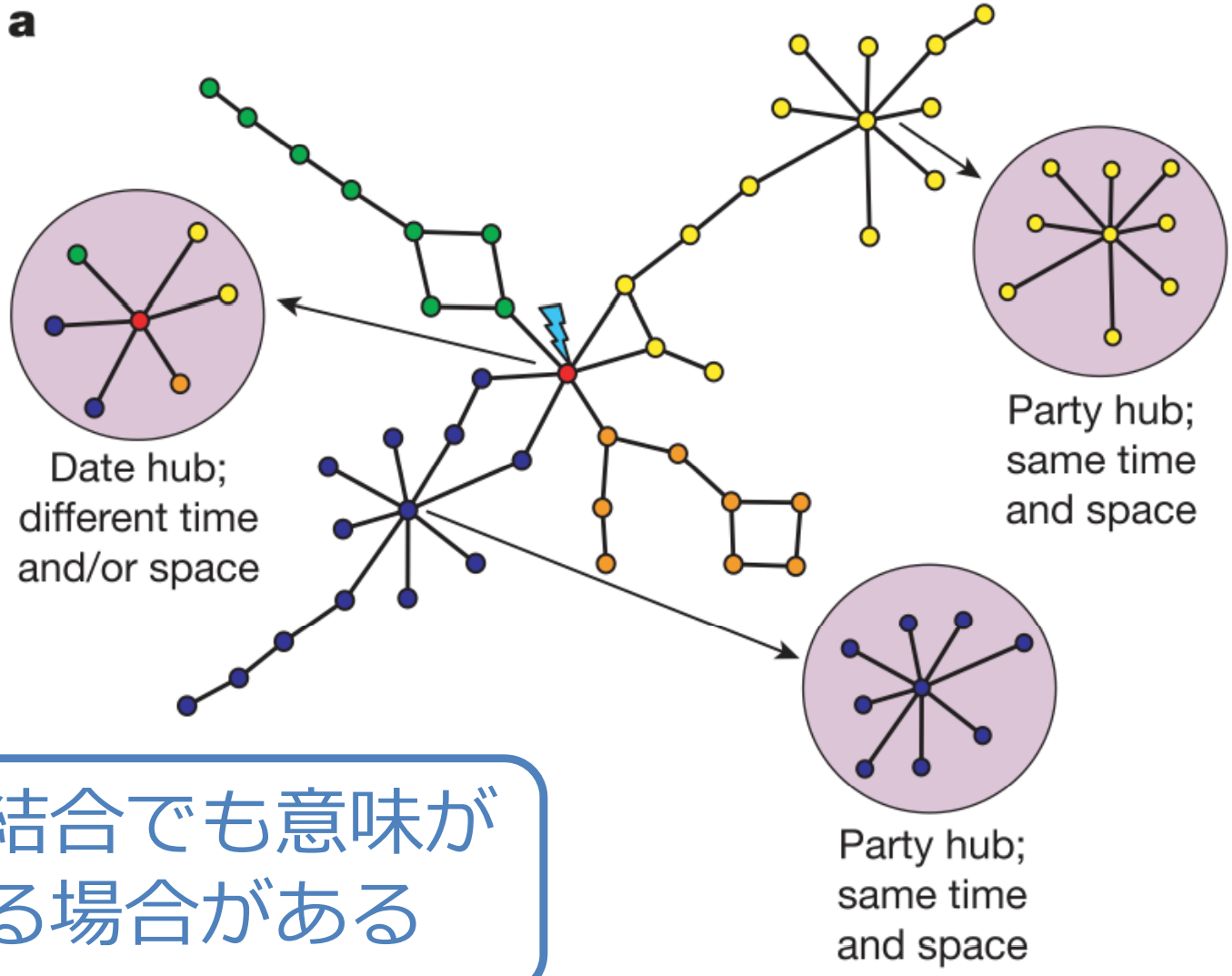
- 発現したタンパクは単一ではなく複数のタンパク質と複合体を形成して機能する。





Nature **411,** 41 (2001)

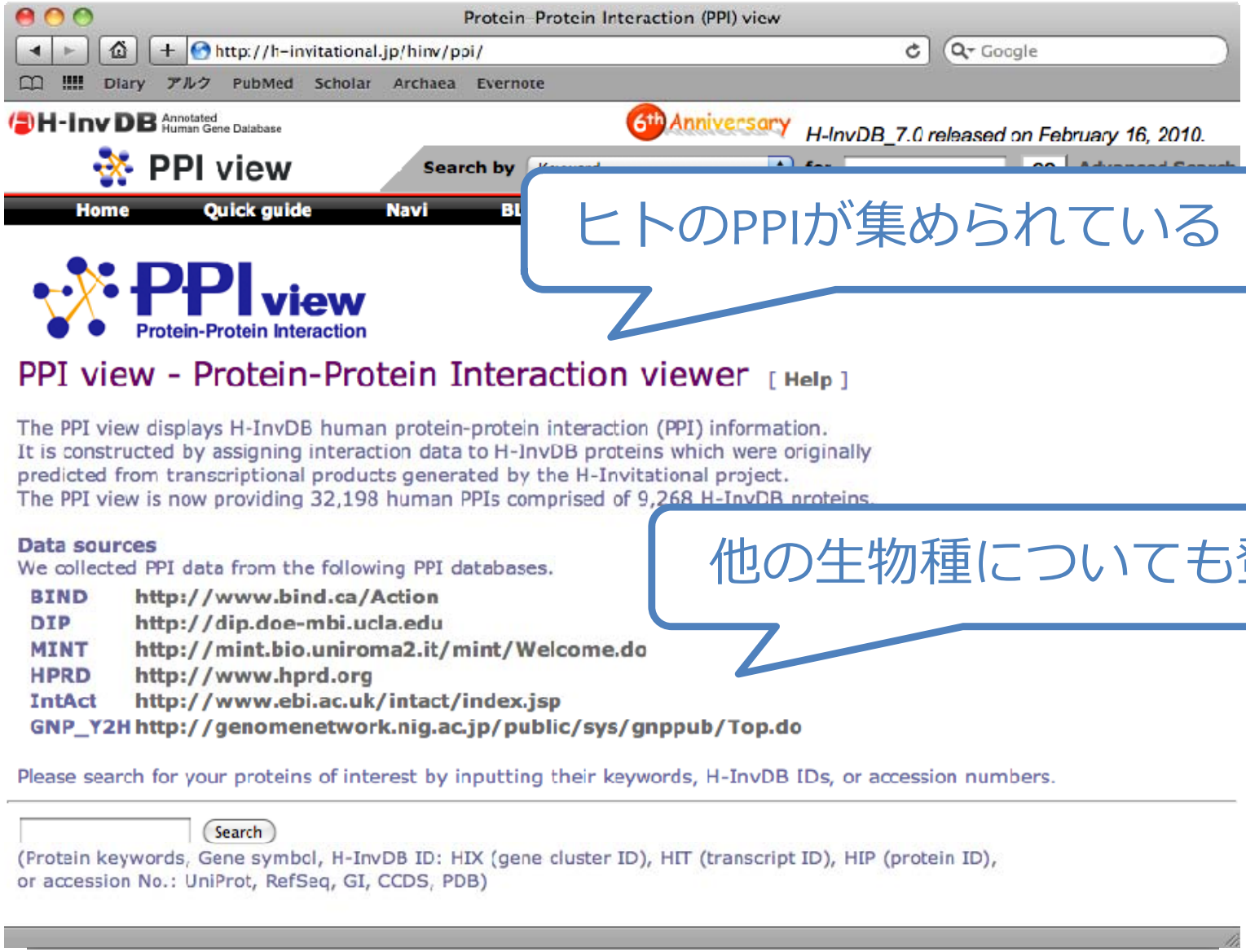
タンパク質間相互作用の種類



同じ結合でも意味が異なる場合がある

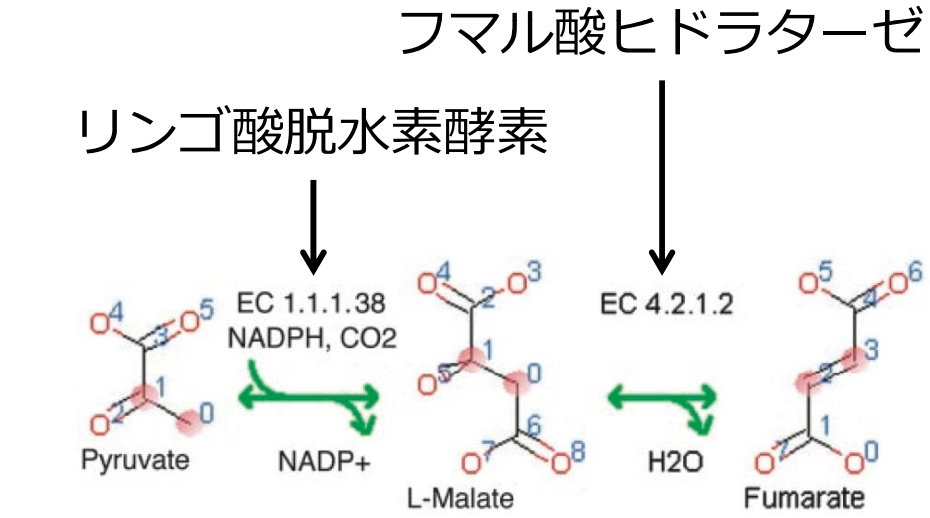
Nature **430,** 88 (2004)

PPIのデータベース

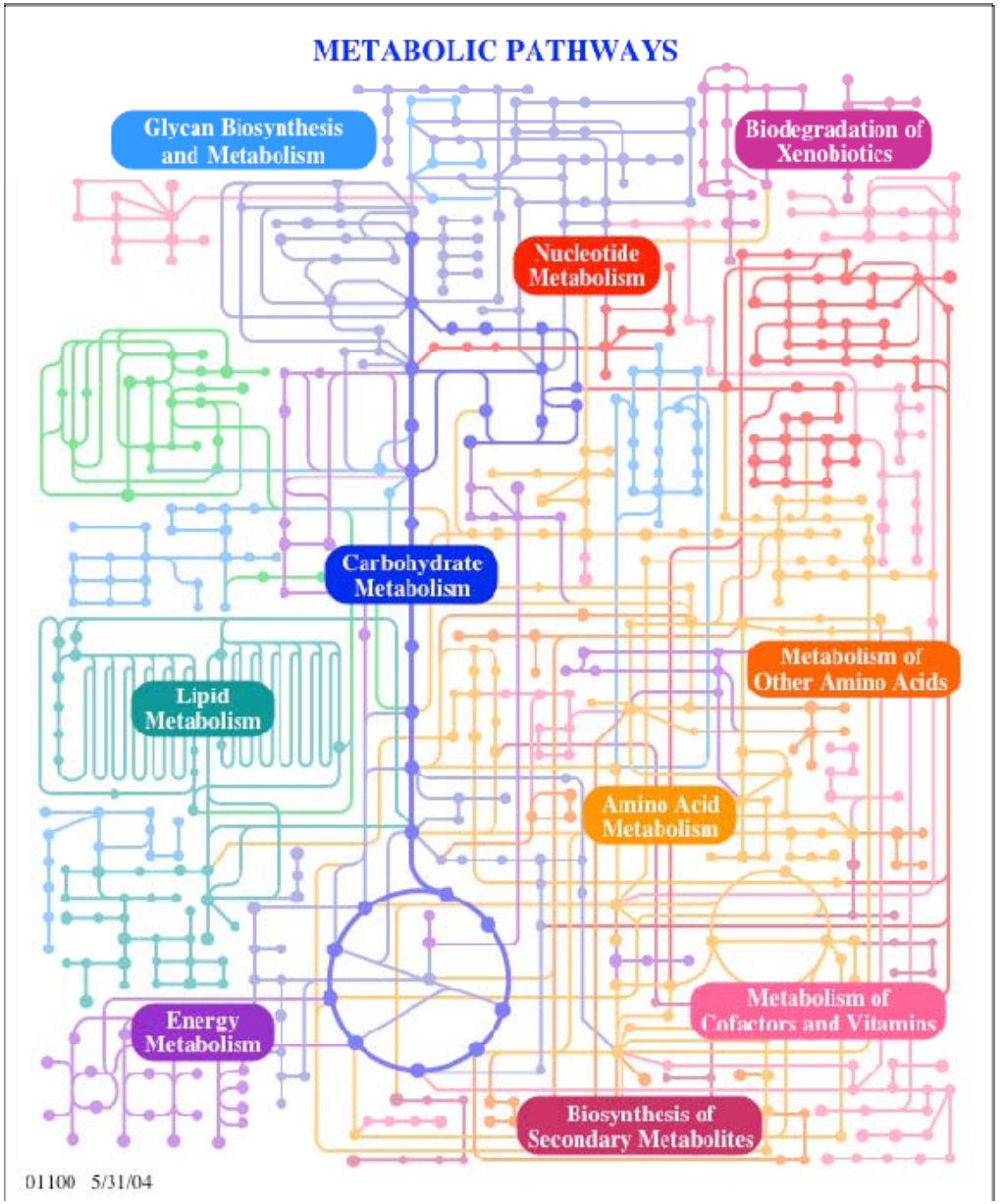

Protein-Protein Interaction (PPI) view
http://h-invitational.jp/hinv/ppi/
Google
Diary
PubMed
Scholar
Archaea
Evernote
H-InvDB Annotated Human Gene Database
6th Anniversary
H-InvDB_7.0 released on February 16, 2010.
PPI view
Search by
Home
Quick guide
Navi
ヒトのPPIが集められている
PPI view
Protein-Protein Interaction
PPI view - Protein-Protein Interaction viewer [Help]
The PPI view displays H-InvDB human protein-protein interaction (PPI) information.
It is constructed by assigning interaction data to H-InvDB proteins which were originally
predicted from transcriptional products generated by the H-Invitational project.
The PPI view is now providing 32,198 human PPIs comprised of 9,268 H-InvDB proteins.
Data sources
We collected PPI data from the following PPI databases.
BIND http://www.bind.ca/Action
DIP http://dip.doe-mbi.ucla.edu
MINT http://mint.bio.uniroma2.it/mint/Welcome.do
HPRD http://www.hprd.org
IntAct http://www.ebi.ac.uk/intact/index.jsp
GNP_Y2H http://genomenetwork.nig.ac.jp/public/sys/gnppub/Top.do
他の生物種についても登録
Please search for your proteins of interest by inputting their keywords, H-InvDB IDs, or accession numbers.
Search
(Protein keywords, Gene symbol, H-InvDB ID: HIX (gene cluster ID), HIT (transcript ID), HIP (protein ID),
or accession No.: UniProt, RefSeq, GI, CCDS, PDB)

代謝パスウェイ（ネットワーク）

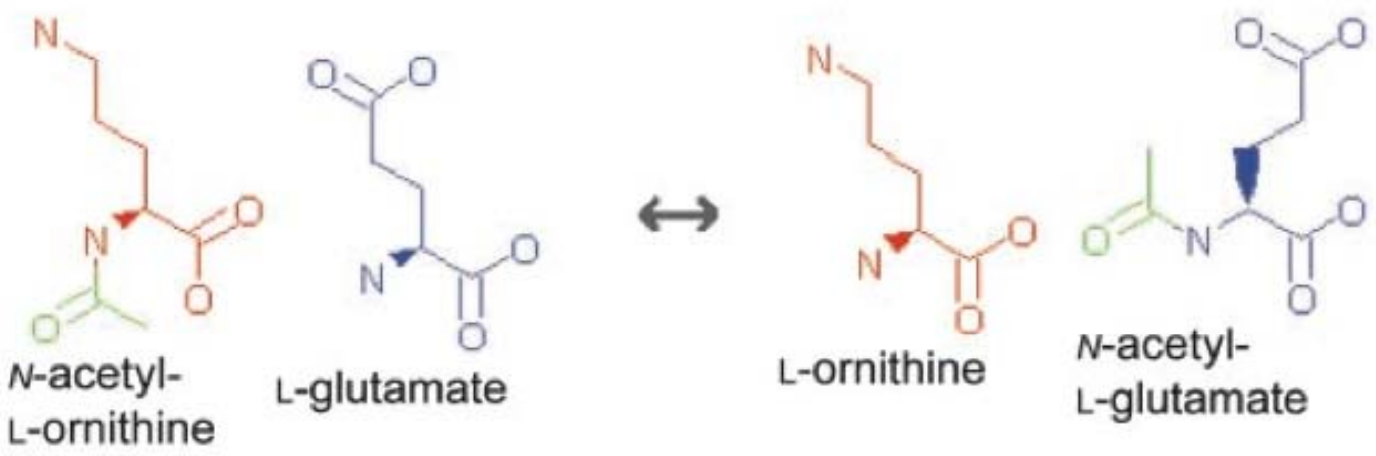
- 酵素による化合物の変換



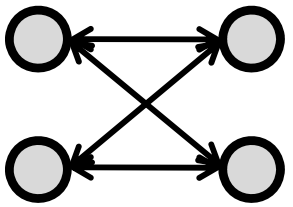
PNAS **101,** 1543 (2004)



代謝反応における表現の注意点

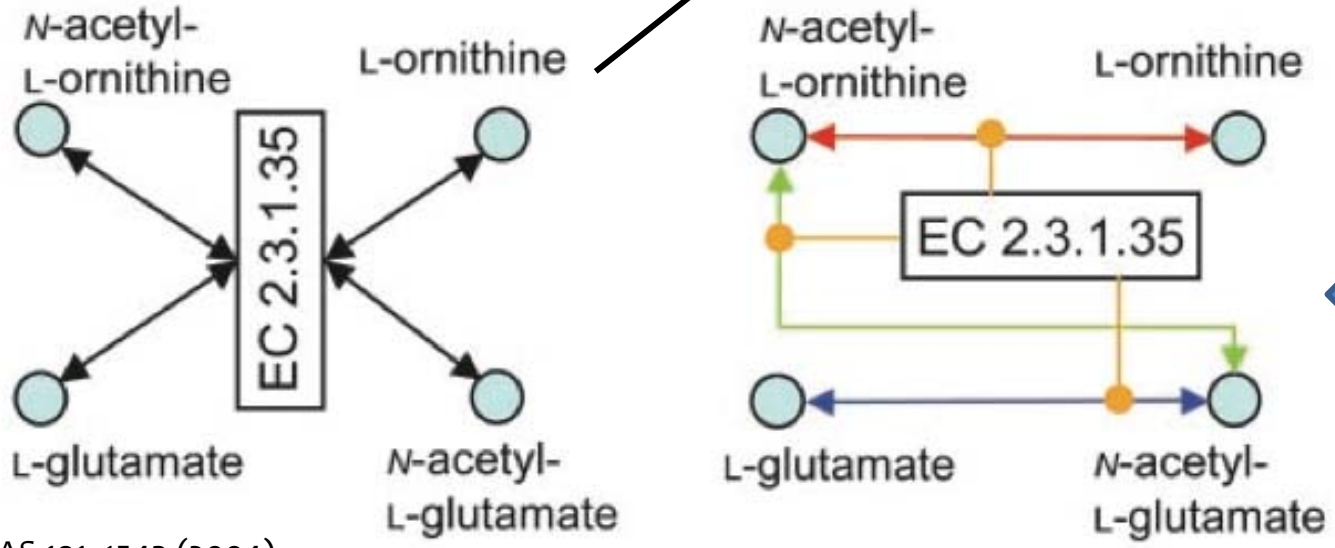


基質と生成物の関係



と書いていいのか？

グルタミン酸とオルニチンはつながっているのか？

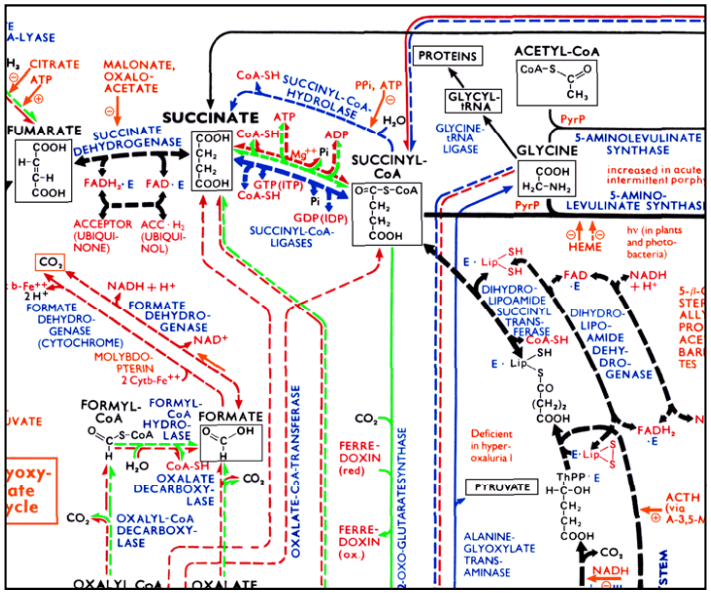


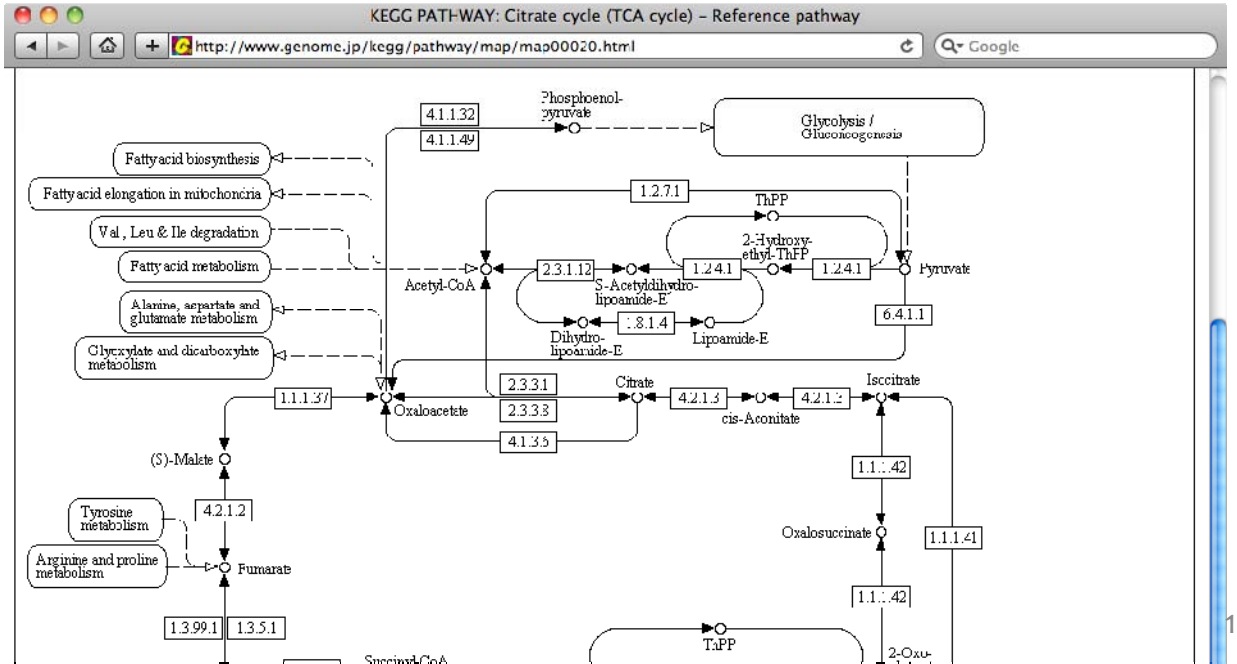
原子の移動に基づくネットワーク

PNAS **101,** 1543 (2004)

代謝経路のデータベース

- KEGG (http://www.genome.jp/kegg/)
- MetaCyc (http://metacyc.org/)
- ExPASy: Biochemical Pathway (http://expasy.org/tools/pathways/)





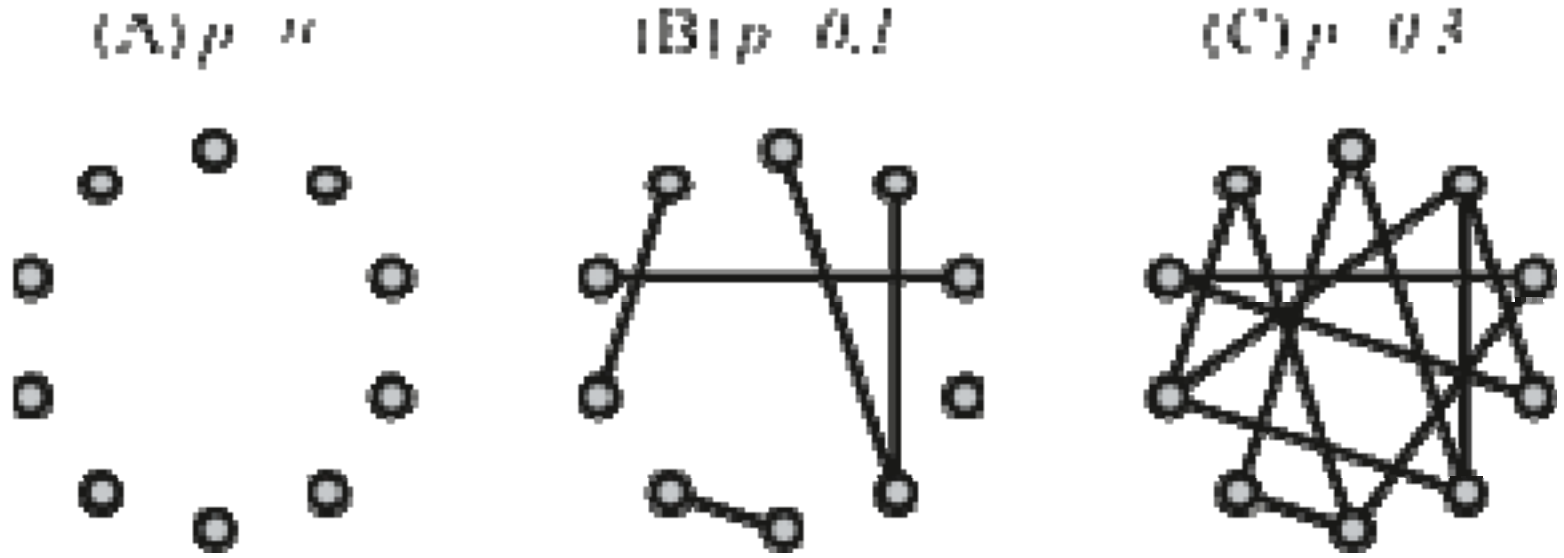
どのようにつながっている？

- ランダムグラフ
- 規則格子モデル

- 昔は、データ量が限られていたので、このようなモデルを使って考察が行われていた。
- また、解析的な取り扱いが容易で、理論の構築にも役立つ。

ランダムグラフ

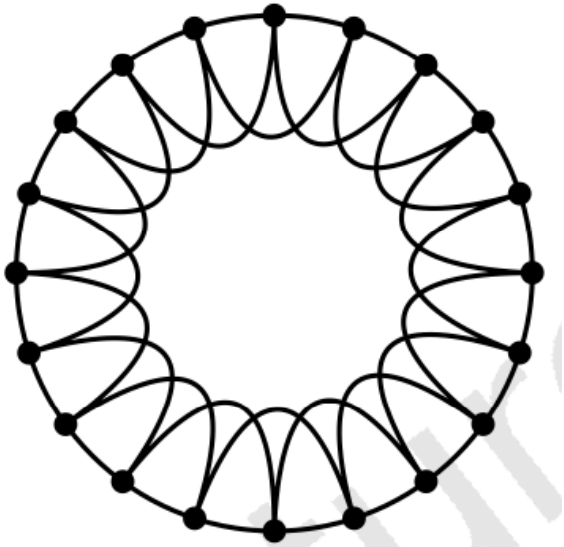
- 任意の２頂点間を確率pでつなぐ。



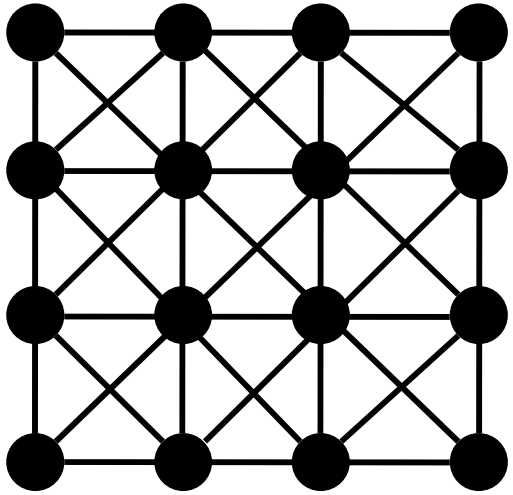
  - 頂点数をN、辺数をEとするとp=2E/[N(N-1)]と見なすことができる。

どのようにつながっている？

- 正則格子モデル
  - 全てのノードの枝数が同じである格子



一次元格子モデルの例



二次元格子モデルの例

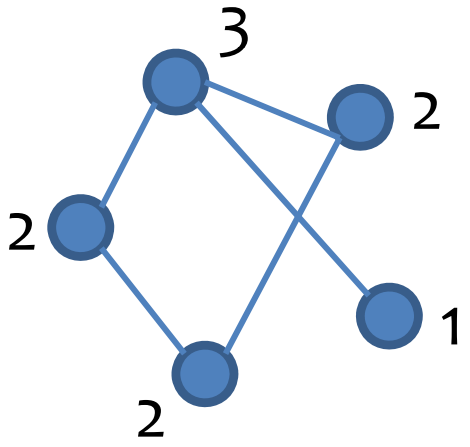
ネットワークの特徴付け

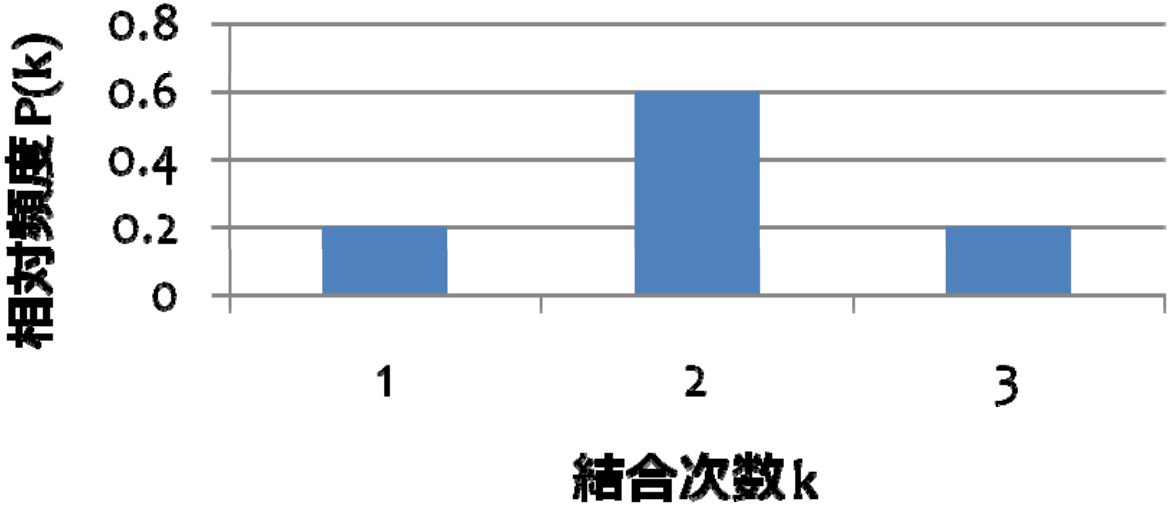
- 次数分布
  - スケールフリー性
  - スケールリッチ性
- クラスタ係数
- 平均最短パス長
  - スモールワールド性

次数分布

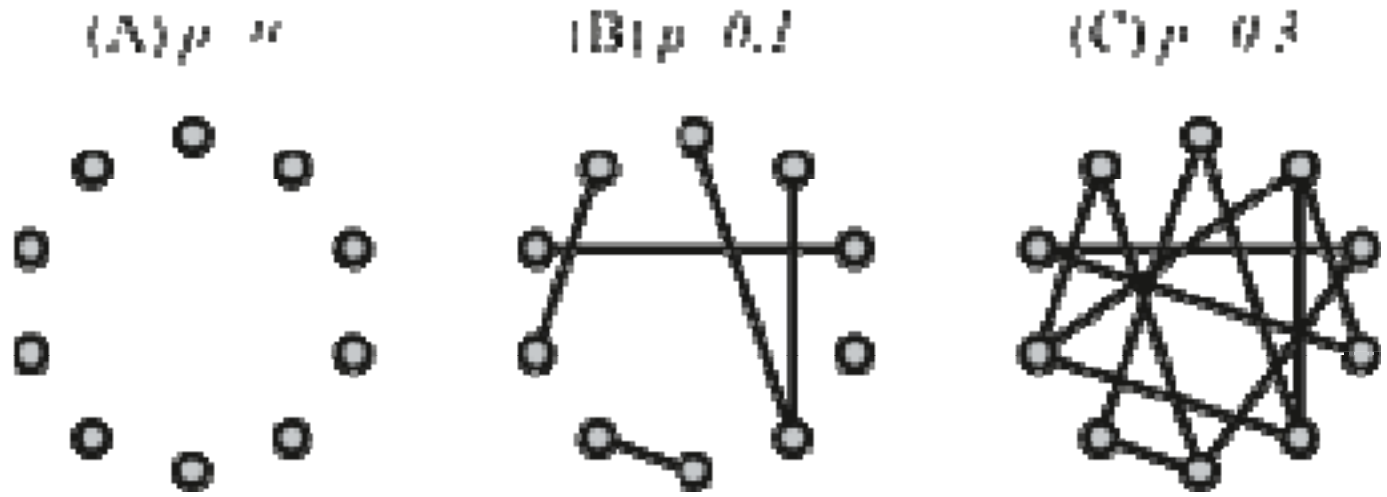
- 次数（結合次数）：ノードが持つ枝数
- 次数分布=

$$\frac{\text{次数kを持つノードの数}}{\text{全体のノード数}}$$



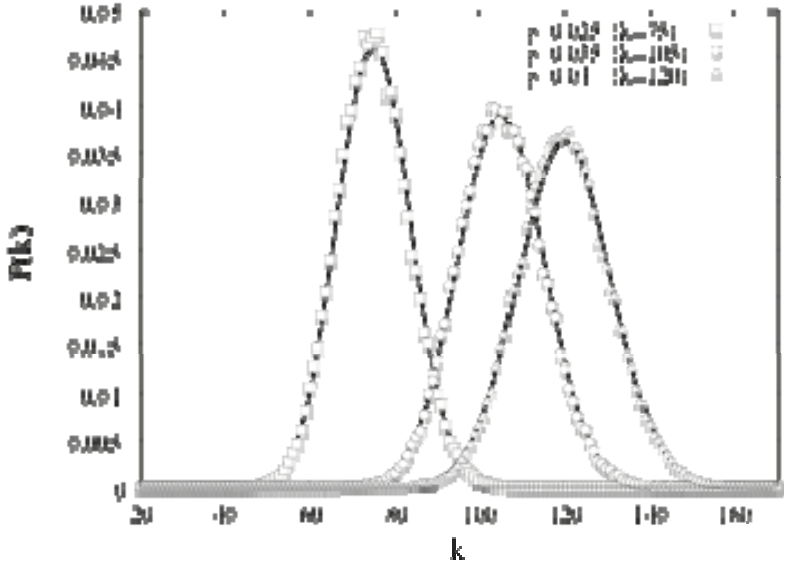


ランダムグラフの次数分布



次数分布は二項分布

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$



各ノードの枝数がmである格子モデルの場合は

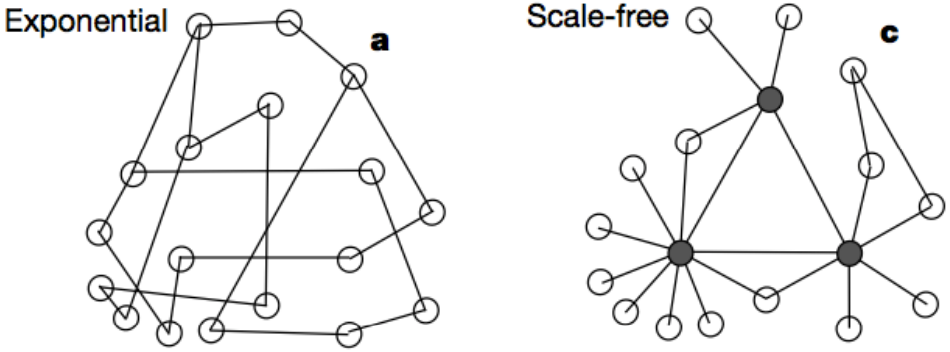
$$P(k) = \delta_{m,k} = \begin{cases} 1\ (m = k) \\ 0\ (m \neq k) \end{cases}$$

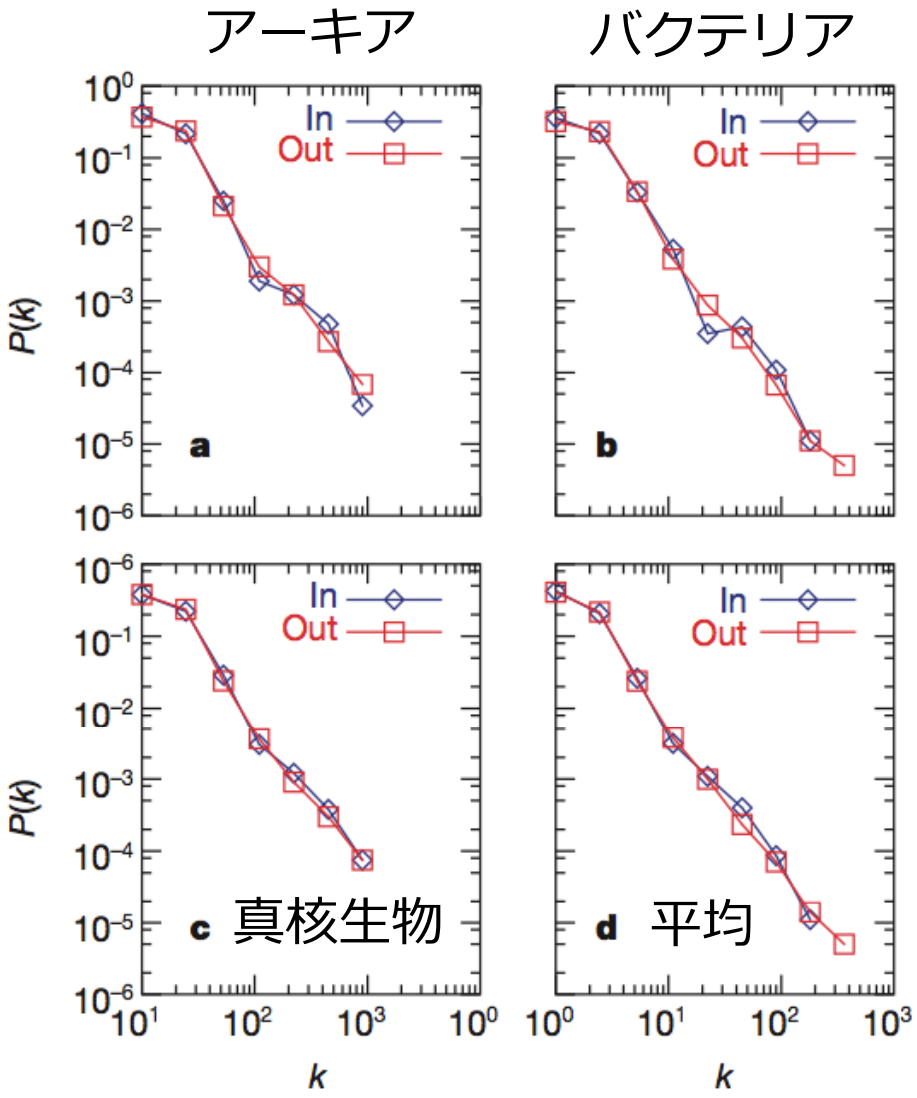
クロネッカーのデルタ関数

現実ネットワークの次数分布
スケールフリー性

$P(k) \sim k^{-\gamma}$

- 次数分布がベキ関数で近似できる。
  - 明確な定義はない
- ランダムグラフとは大きく異なる。





例：代謝ネットワーク

Nature **407**, 651 (2000)

スケールフリー性（２）

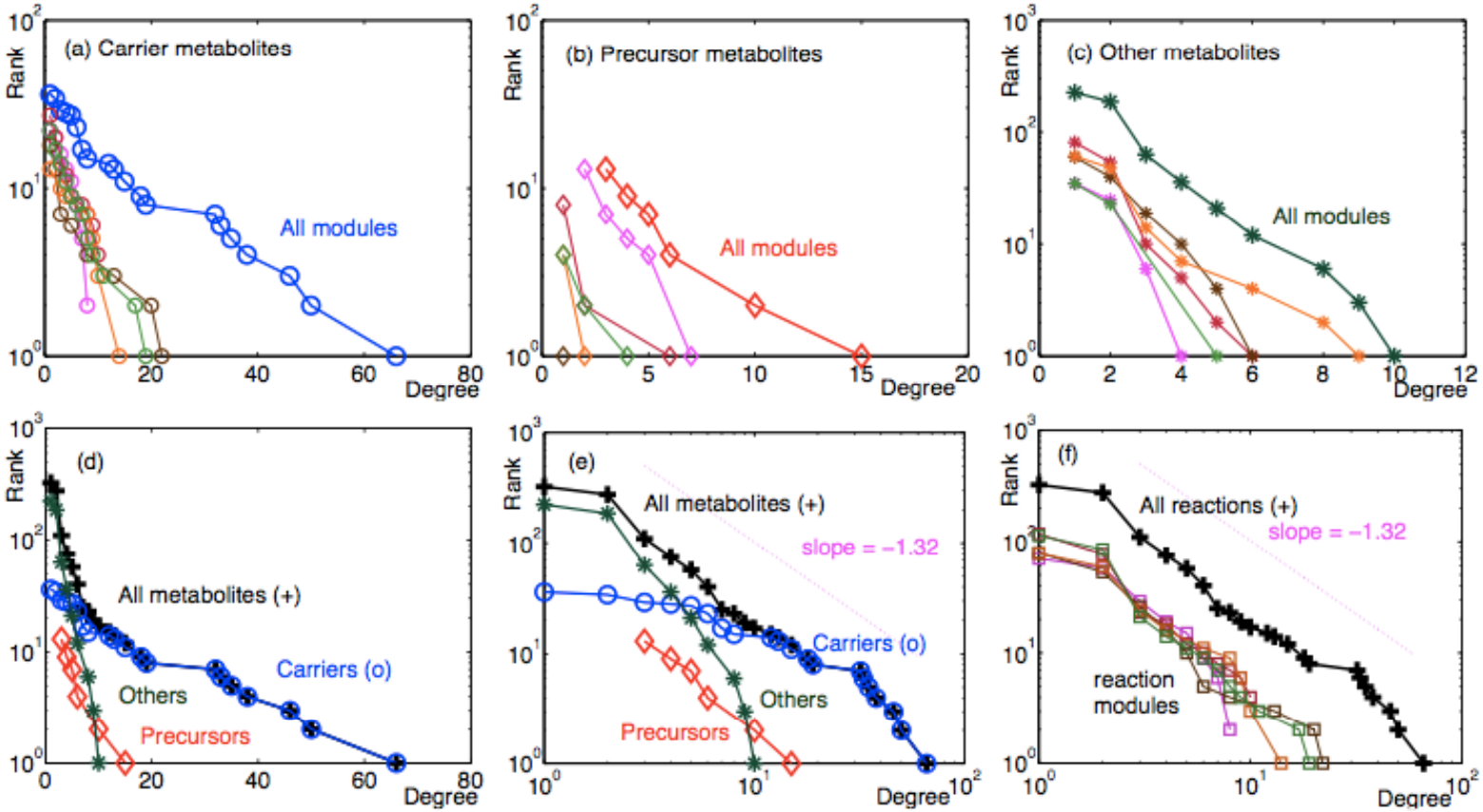
- その他のネットワークでも見られる
  - 転写制御ネットワーク
  - タンパク質間相互作用ネットワーク
  - WWW、人間関係ネットワーク、送電線
- **ベキ分布が最も良いとは限らない**
  - 他の分布での当てはめでも可能
  - ただ、ベキで近似できるというのは重要

スケールフリー：言葉の意味

- サイズが変化しても同様に見られる性質だから「スケールフリー」
  - 代謝ネットワークでは各生物種においてネットワークサイズが異なるが、分布は同じ。
  - フラクタル（自己相似性）との関連
- ベキ関数で表されるので平均（スケール）の概念が適応できないので「スケールフリー」
  - 最近はこちらの意味で使われる。

スケールリッチ性

- 部分に注目すると分布は保存されない。
  そういう意味でスケールフリーではない。

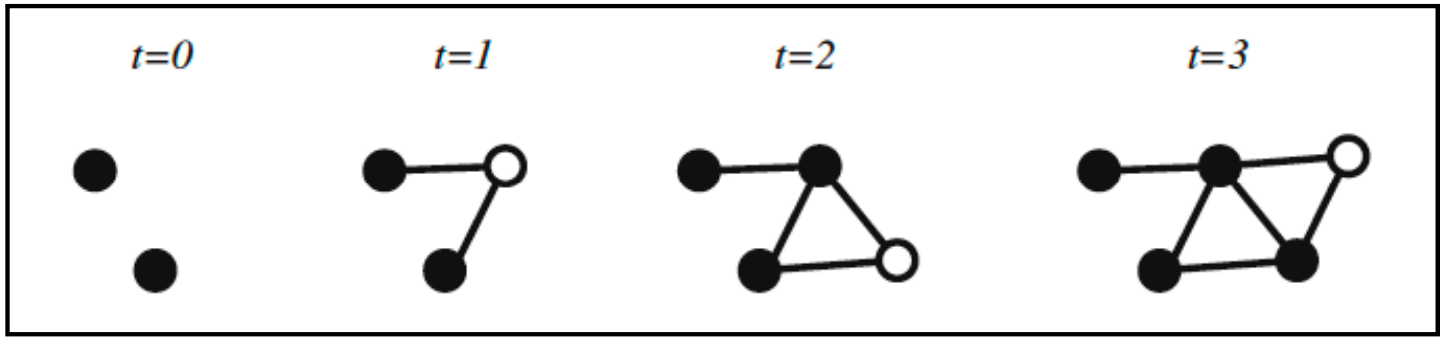


Phys Rev Lett **94,** 168101 (2005)

Barabási-Albertモデル

Physica A **272**, 173 (1999)

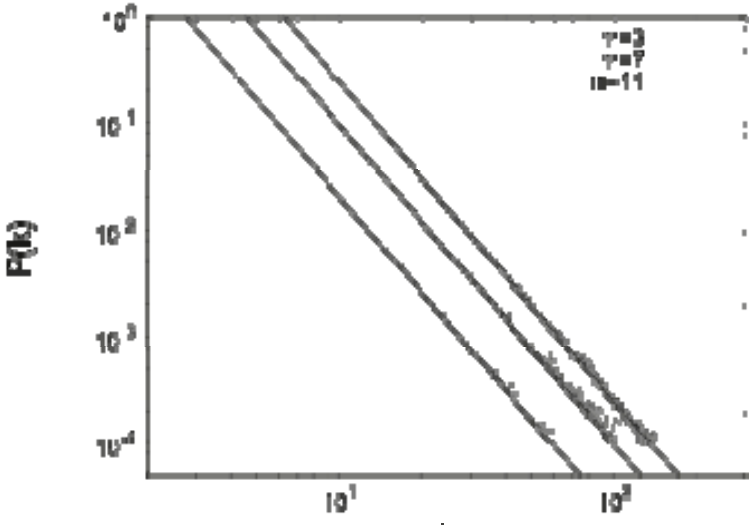
- スケールフリー性のみを説明する。
- 成長性と**優先接続性** $\Pi_i = k_i / \sum_j k_j$



次数分布

$$P(k) \sim k^{-3}$$

拡張モデルで指数は可変

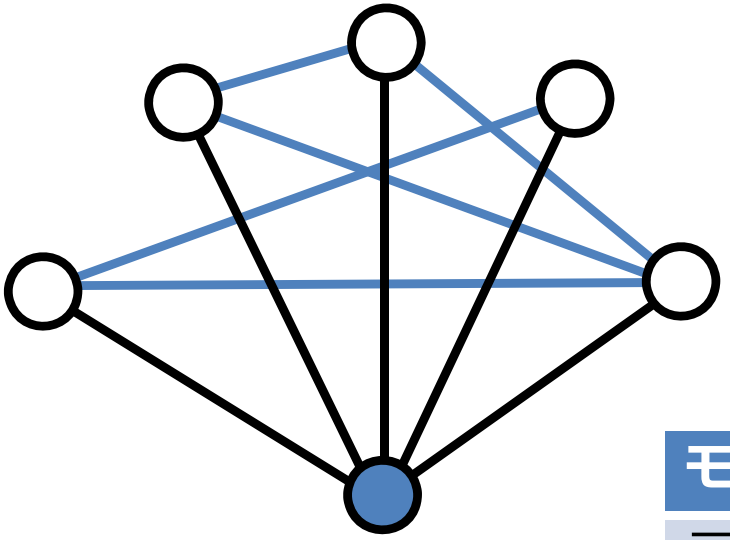


クラスタ係数

- あるノードの近傍間においてエッジが張られる確率

$$C_i = \frac{M_i}{{}_{k_i}C_2} = \frac{2M_i}{k_i(k_i-1)}$$

← 近傍ノード間の辺数



平均クラスタ係数

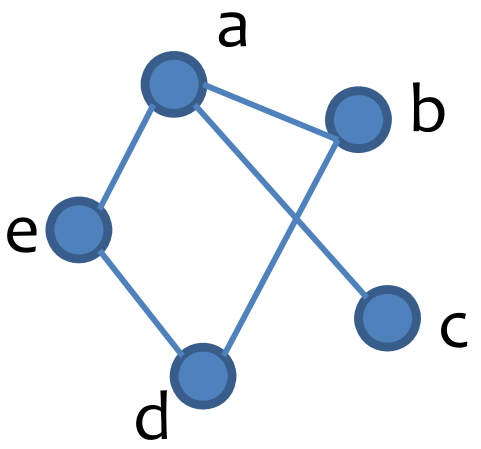
$$C = \frac{1}{N}\sum_{i=1}^{N} C_i$$

モデル	クラスタ係数
ランダムグラフ	p=2E/[N(N-1)]
m-正則一次元格子	(3m-6)/(4m-2)
BAモデル	$\approx (E-N)(\ln N/N)^2/8$

平均最短パス長

$$L = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} d(i,j)$$

距離行列



モデル	平均最短パス長
ランダムグラフ	~ln N/(ln 2E – ln N)
m-正則一次元格子	N/[2m]
BAモデル	$\frac{\ln N - \ln(E/[2N]) - 1.58}{\ln\ln N + \ln(E/[2N])} + \frac{3}{2}$

	a	b	c	d	e
a	-	1	1	2	1
b	1	-	2	1	2
c	1	2	-	3	2
d	2	1	3	-	1
e	1	2	2	1	-

平均次数が
一定とすると

ランダム　~ ln N
格子モデル　~ N
BAモデル　~ ln N / lnln N

スモールワールド性

- 現実のネットワークはクラスタ係数が高く平均最短パス長が小さい。

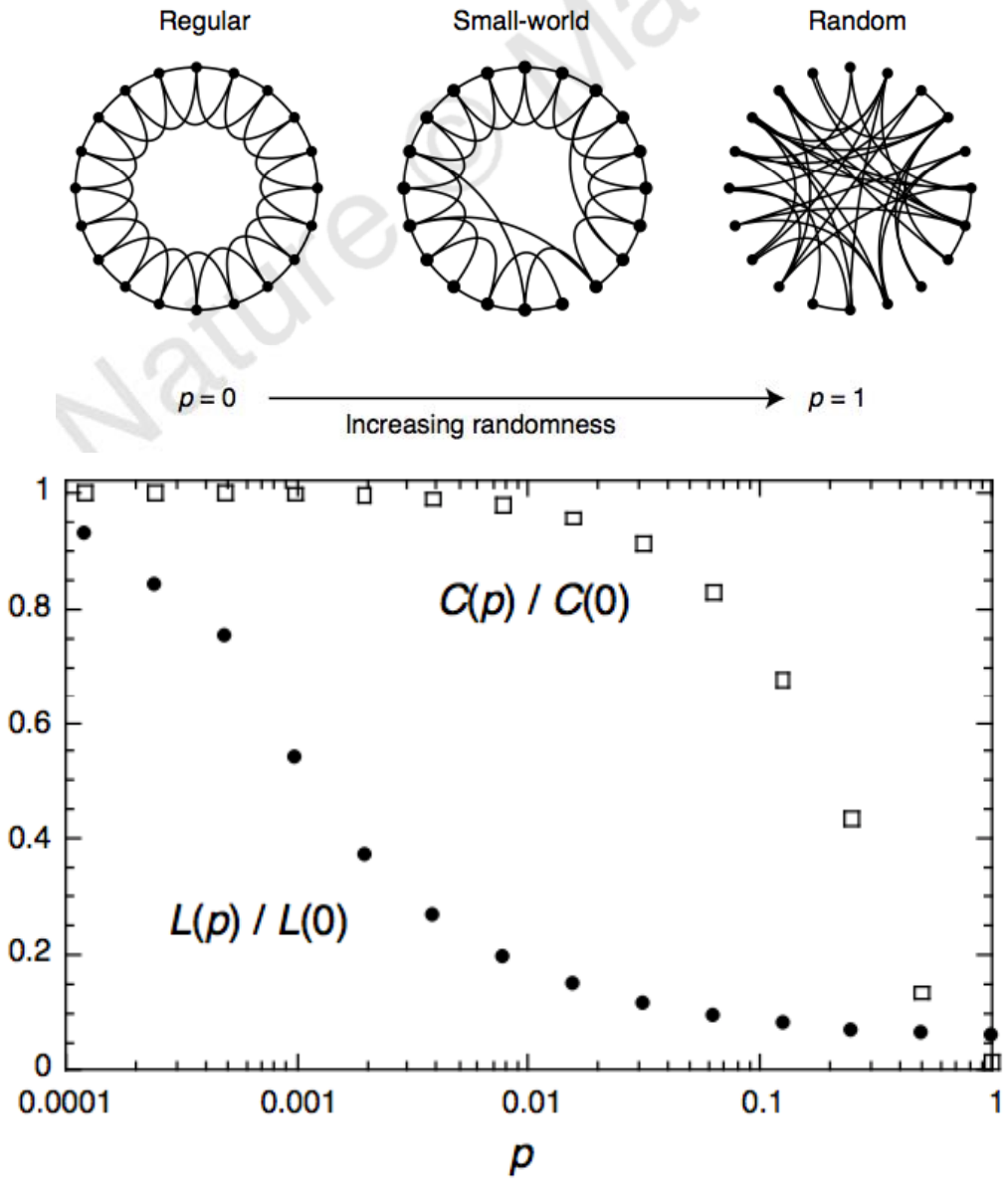
Nature **393**, 441 (1998)

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

- ランダムグラフでは説明できない。
- 格子モデルではクラスタ係数が高くなるが、パス長は大きくなる。

Watts-Strogatzモデル



一次元格子から辺をランダムに張り替えることで、スモールワールド性が出現する。

つまり、秩序構造と無秩序構造の中間

ただし、次数分布は二項分布に近い。

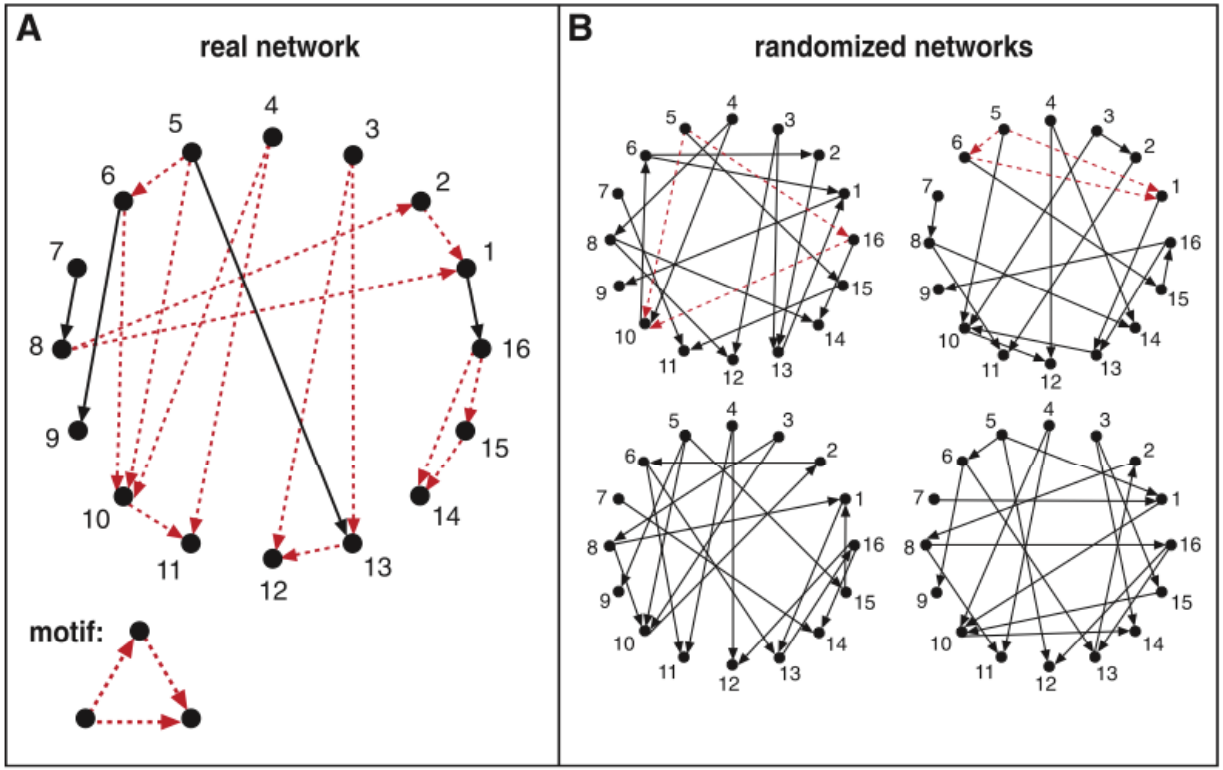
Nature **393,** 441 (1998)

ランダムではない世界

- **現実のネットワークはランダムではない**が、どんな意味があり、何か利点があるのだろうか？

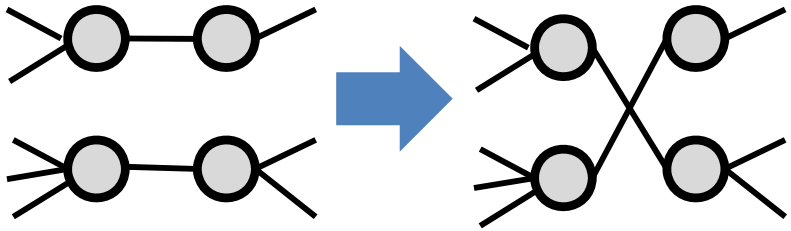
- ネットワークモチーフ
  - 機能モジュール
- ネットワークの頑健性（ロバスト性）
  - 頂点に対する平均最短パス長のロバスト性

ネットワークモチーフ（1）



Science **298**, 824 (2002)

ランダム化：
現実のネットワークから任意に二辺を選び、結合先を交換

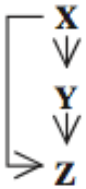
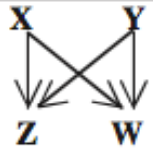
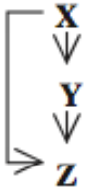
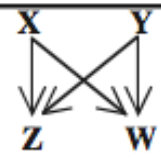
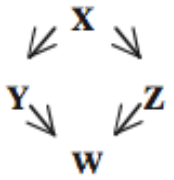


次数分布は変化しない

現実のネットワークにはランダム化されたネットワークに比べて頻出するパタンが存在

ネットワークモチーフ（2）

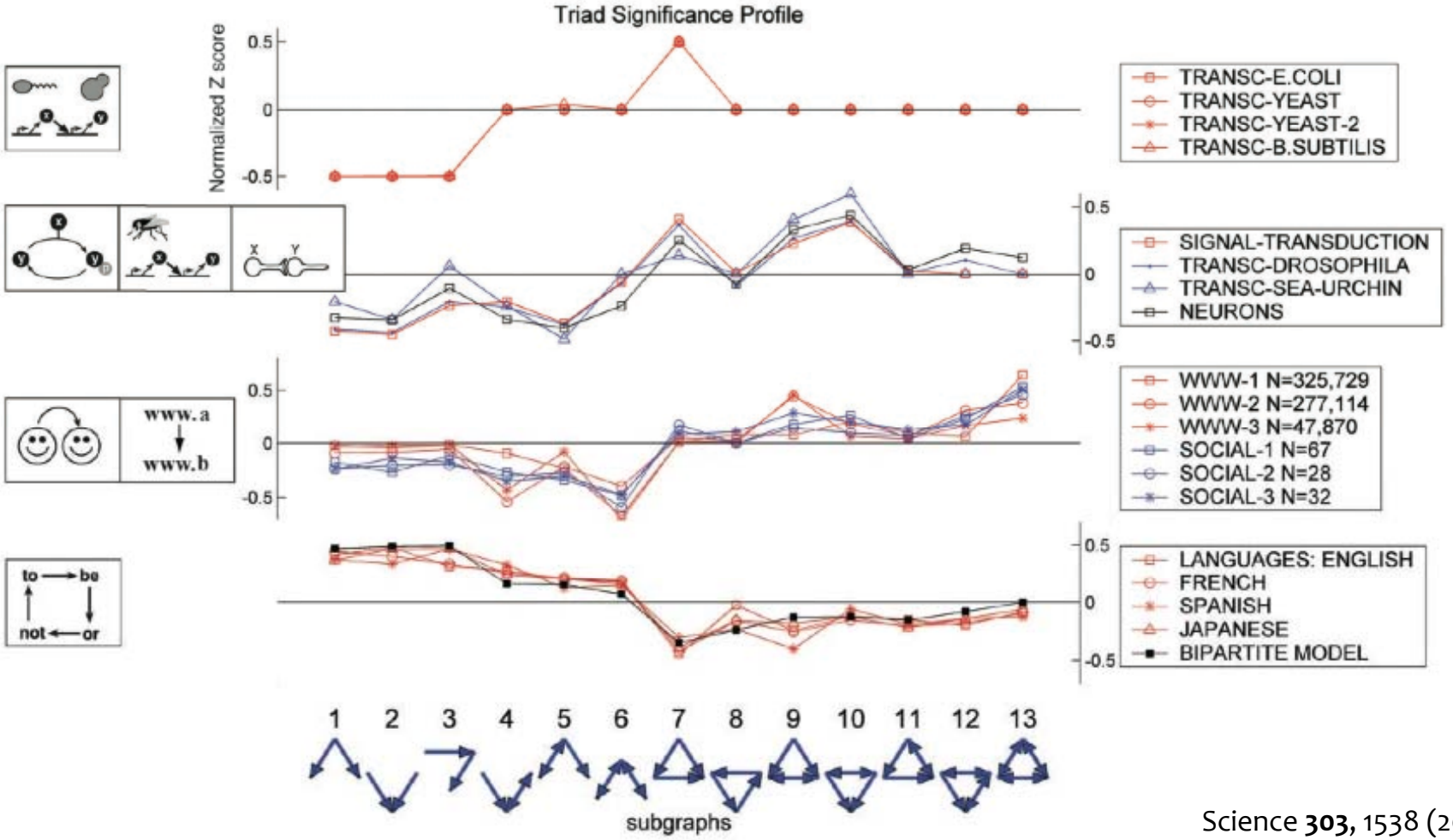
- フィードフォワードループなどが出現
- 制御理論において、このモジュールは重要な役割を果たす。
  - 詳しくは後ほど・・・

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20

Science **298,** 824 (2002)

モチーフのプロファイル

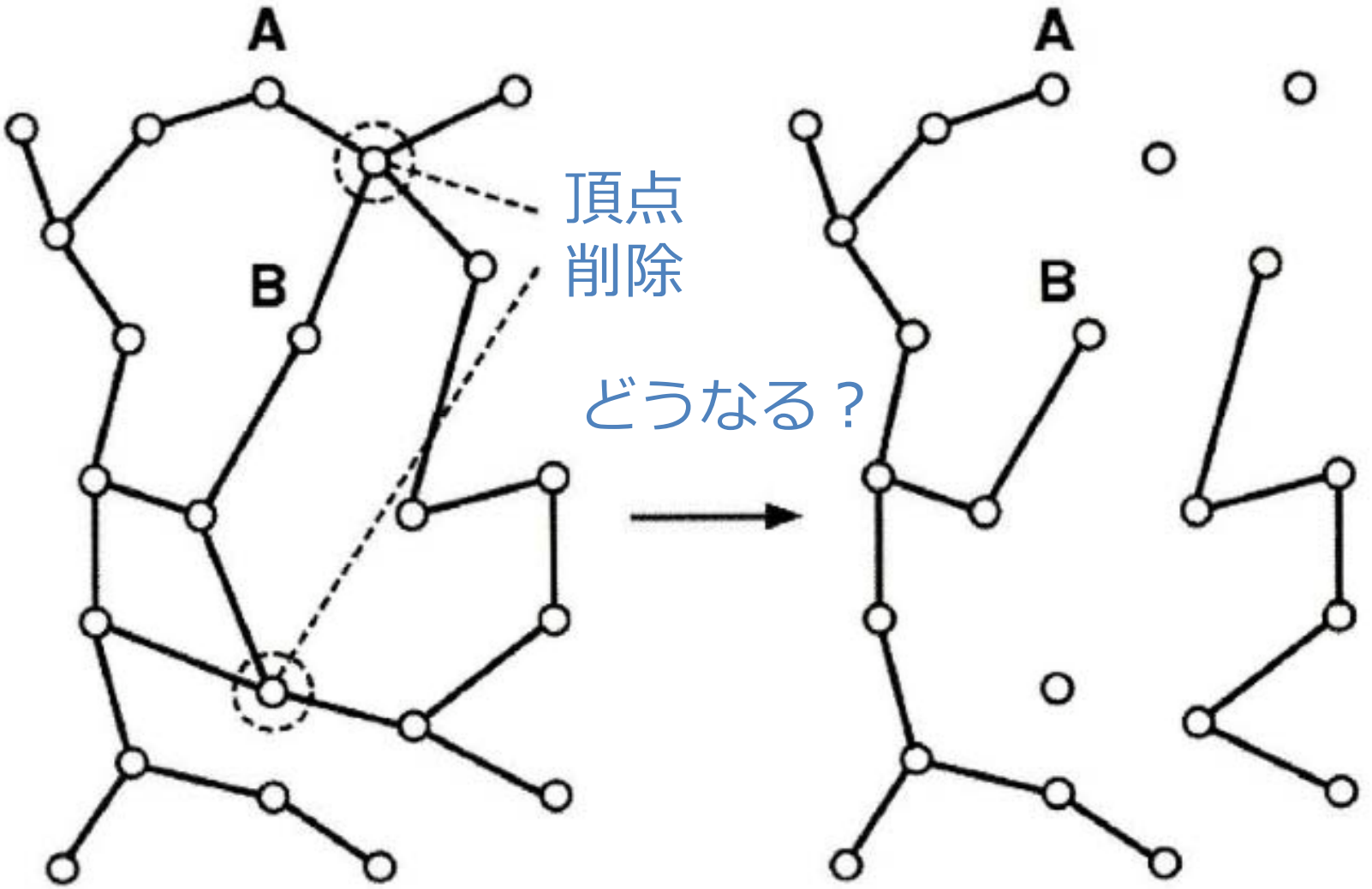
- ネットワークを通して共通性が見出せる。



Science **303,** 1538 (2004)

頂点削除に対する平均最短パス長のロバスト性（１）

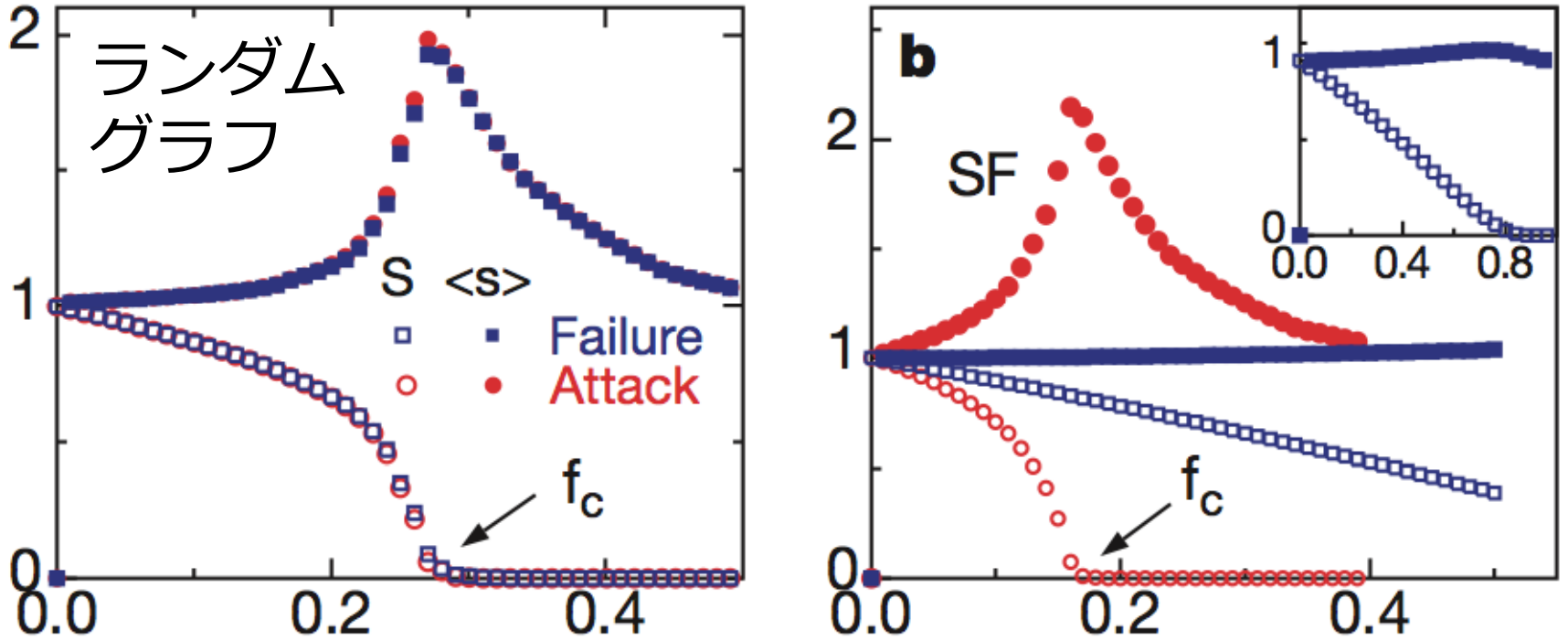
Chapter 3, Handbook of Graphs and Networks: From the Genome to the Internet, Wiley, 2003



例えば、遺伝子のノックアウトに対応

頂点削除に対する平均最短パス長のロバスト性（２）

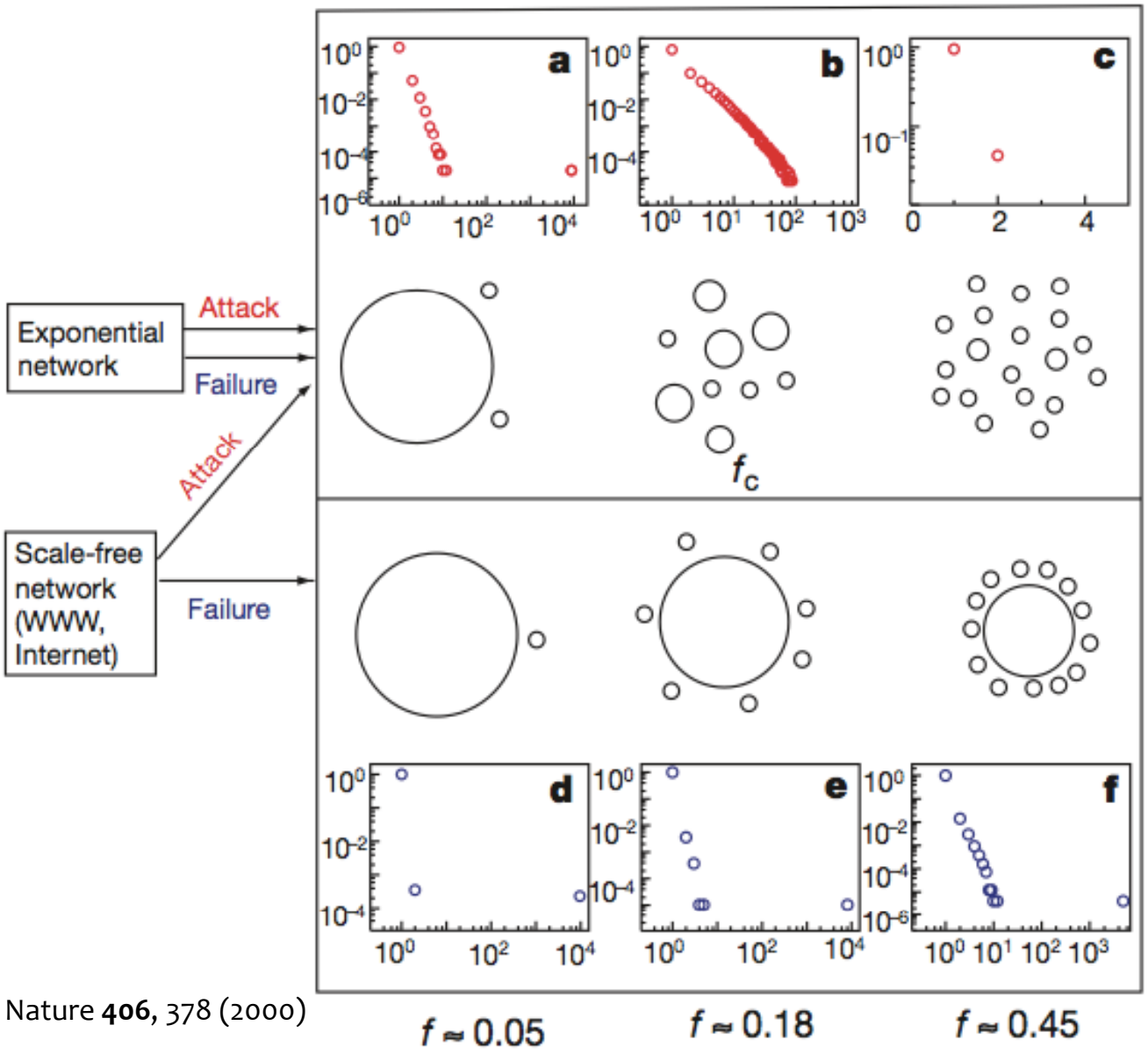
- スケールフリー(SF)ネットワークはランダムな頂点削除に対してロバスト性を示す。
- ただしハブを狙うと弱い（トレード・オフ）。



頂点を削除した割合

Nature **406,** 378 (2000)

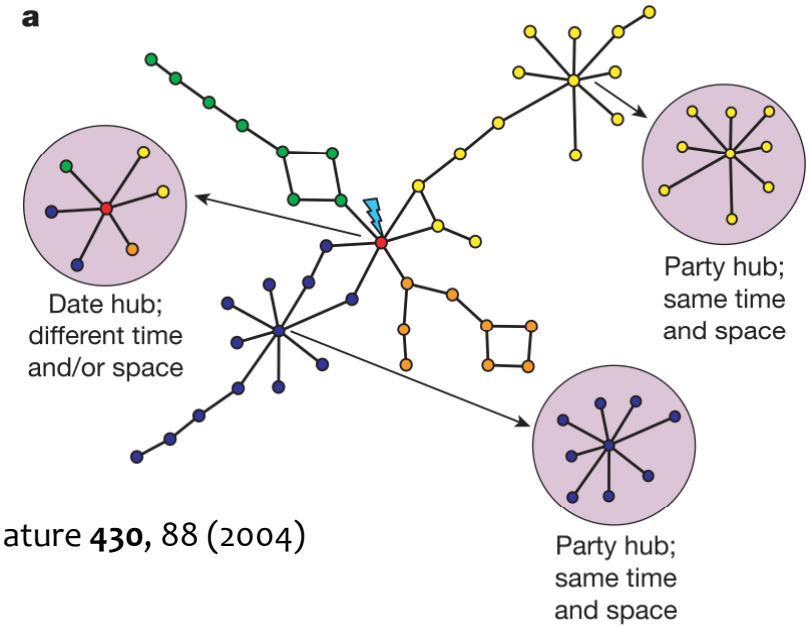
頂点削除に対する平均最短パス長のロバスト性（３）



現実のネットワークは任意のエラーに対してロバストになるように形成された？

Nature **406,** 378 (2000)

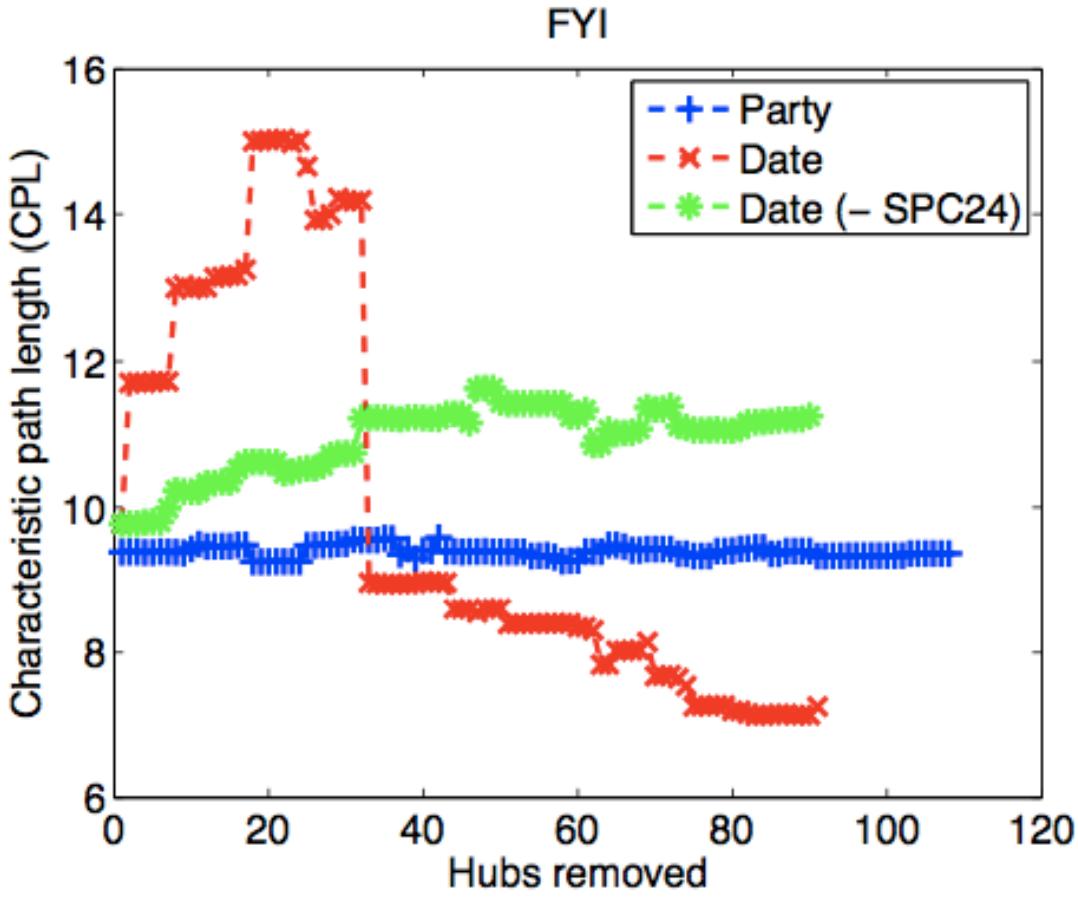
異なるハブが存在する。



Nature **430,** 88 (2004)

異なるモジュールを繋げるハブと、モジュール内のハブが存在する。

単にハブを狙えば良いという訳ではない。



PLoS Comput Biol **6,** e1000817 (2010)

頂点iの重要度を測る

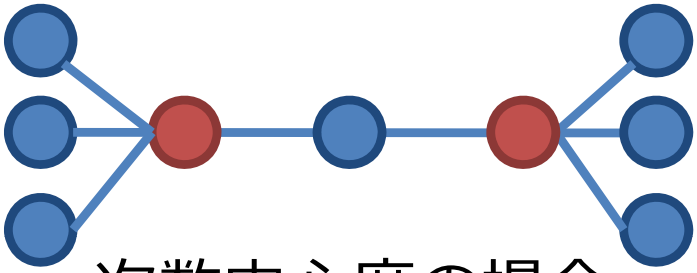
- 次数中心度 $= k_i /(N-1)$
  - 枝数の多い頂点が重要
- 近接中心度

$$= \left[\sum_{\substack{j=1 \\ j \neq i}}^{N} d(i,j) \right]^{-1}$$

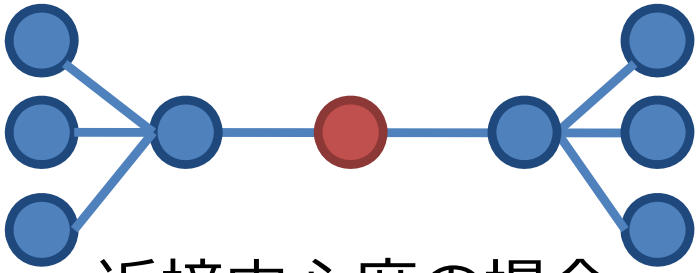
頂点i-j間の最短距離

その他の頂点の短い距離で
つながっている頂点が重要

重要性は尺度によって異なる



次数中心度の場合



近接中心度の場合

その他の尺度

- 媒介中心度
  - 拡張頂点間の最短パスを考える時、よく通過する頂点が重要
- 固有ベクトル中心度
  - 次数中心度の拡張版

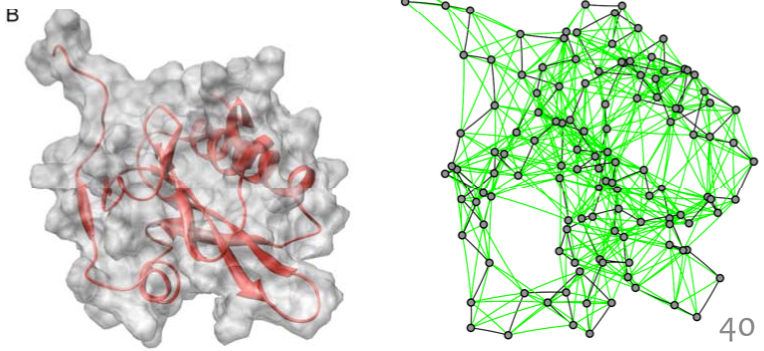
頂点iの重要度 $x_i = \frac{1}{\lambda}\sum_{j=1}^{N} A_{ij} x_j$ ← 隣接行列

⟶ つまり $\mathbf{Ax} = \lambda \mathbf{x}$

  - Page rankとも関係がある。

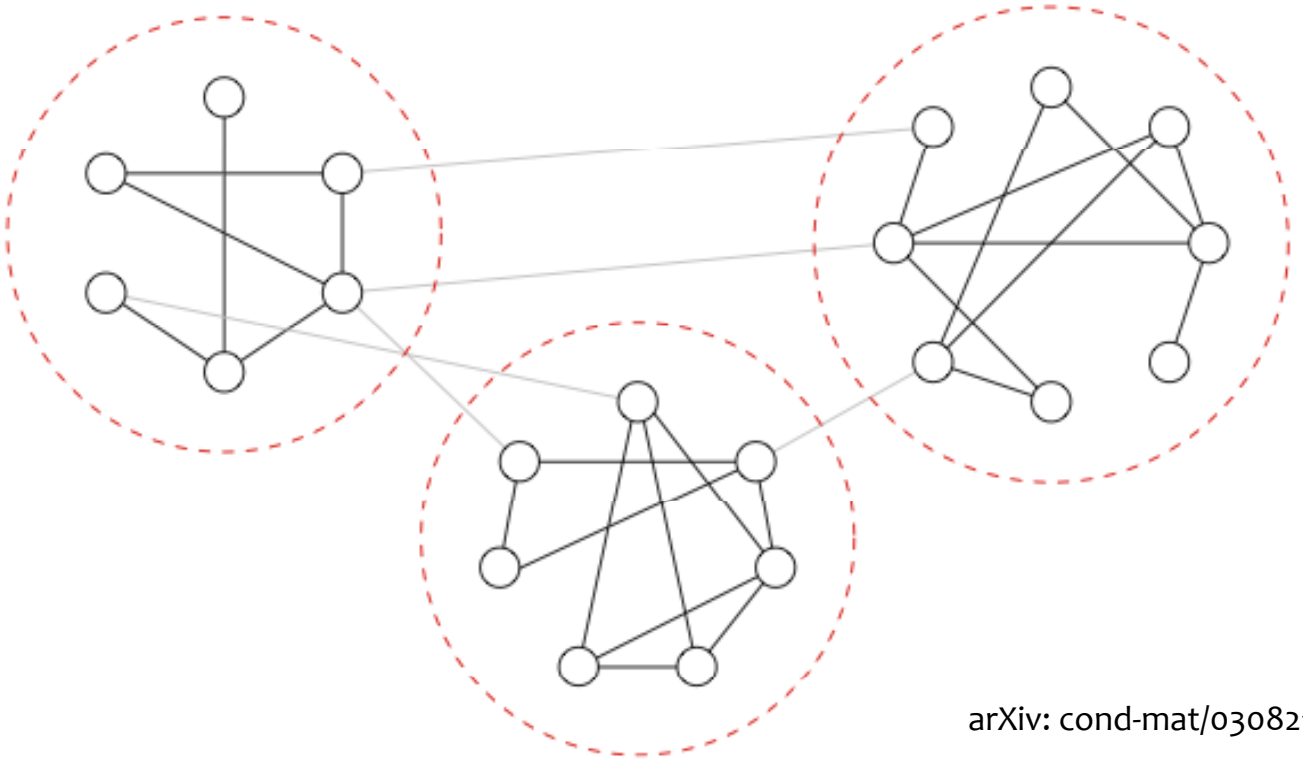
頂点の重要度と生物学的意義

- PPIにおいてタンパク質の次数中心性とその進化速度には負の相関がある。
  - 相互作用の多いタンパク質の進化速度は遅い
  - Science **269**, 751 (2002)
- タンパク質コンタクトネットワークにおいて近接中心性は活性部位を予測できる。
  - Protein Science **15**, 2120 (2006)



コミュニティ構造

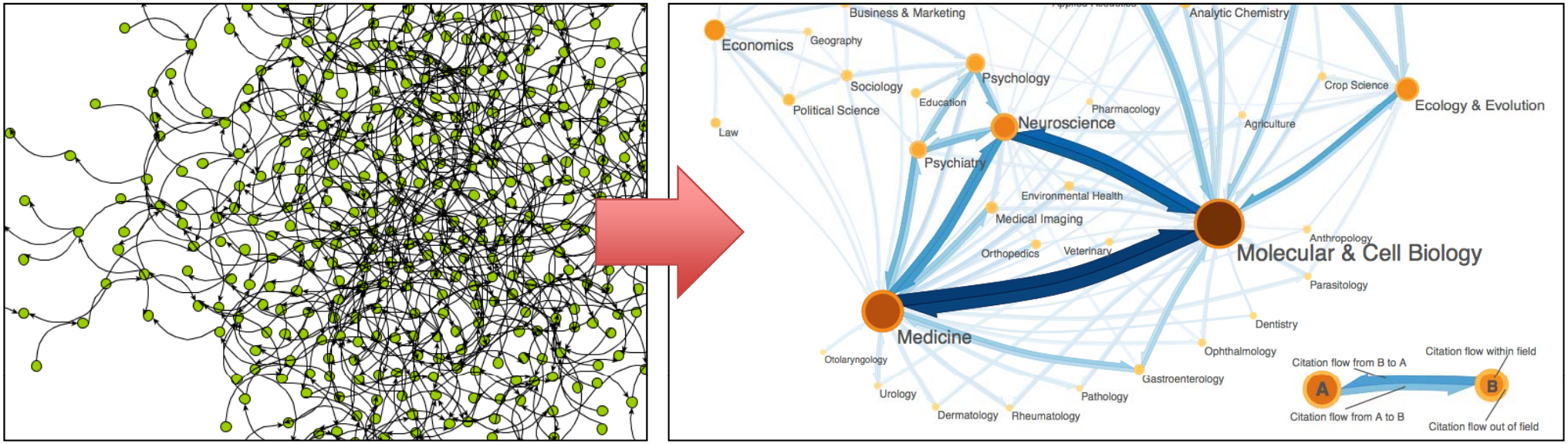
- 比較的密に連結した部分グラフ同士が疎に連結しているような構造
  - グラフクラスタリングに対応



arXiv: cond-mat/0308217

生体ネットワークにおけるコミュニティ構造の重要性

- PPIにおいては機能的に類似なタンパク質が同じコミュニティに属す
  - 機能予測ができる。
- 複雑なネットワークの俯瞰的な理解



PNAS **105,** 1118 (2008)

コミュニティ構造の検出

- 基本的にモジュラリティ最大化
  - コミュニティ内の辺密度が高く、コミュティ間の辺が疎であれば良い分割という前提
  - 定義：

隣接行列

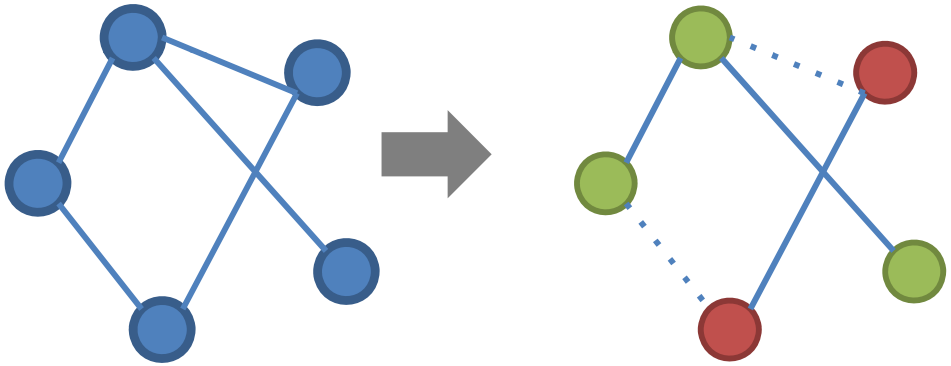
$$Q = \frac{1}{2E}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2E}\right]\delta(c_i, c_j)$$

頂点iとjが同じコミュニティに属している：1
そうではない：0

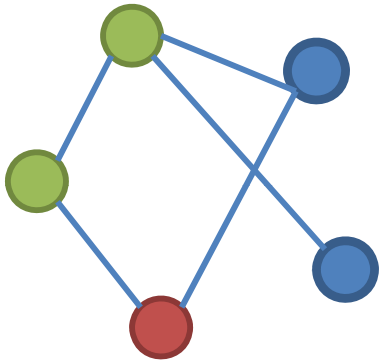
任意の次数列をもつランダムグラフを仮定した場合の結合確率

コミュニティのアサイン

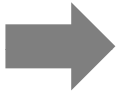
- つまり**c**=(c_1,c_2,… c_N)をどう設定するか
  - なんらかの基準でエッジを削除して部分グラフに分割していく
    - 最小カット
    - 辺媒介度

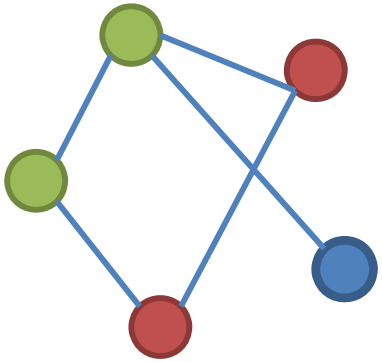


  - 遺伝的アルゴリズムによってQが最大となるアサインを見つける。



アサイン変更





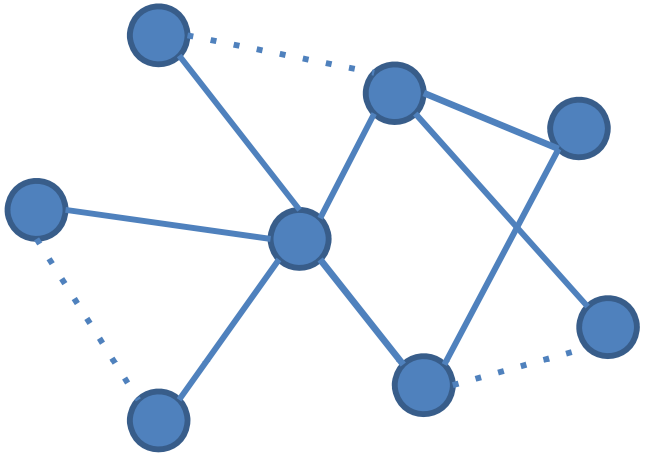
Q_0

Q_1

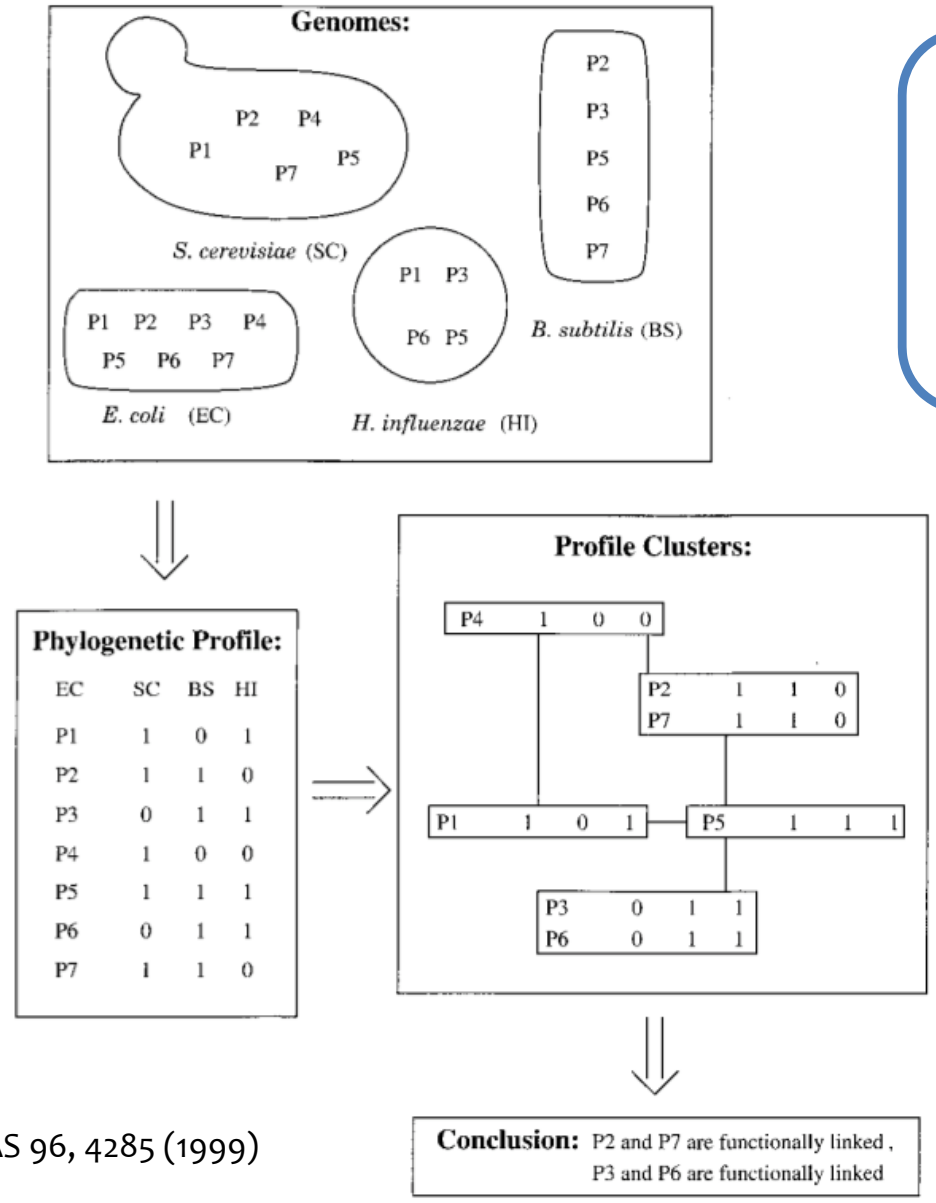
$Q_0<Q_1$: 採択
$Q_0>Q_1$: 棄却

相互作用の予測

- 多くの知られていない相互作用がある
- 相互作用ペアは膨大であるから、ある程度計算機で予測をつけることが重要
  - 要素数をNとすると候補はN^2になる。
  - 酵母の遺伝子は約6000なので、3600万ペアを検証する必要がある。
  - 実験のコストが下がる。



系統プロファイルによる予測



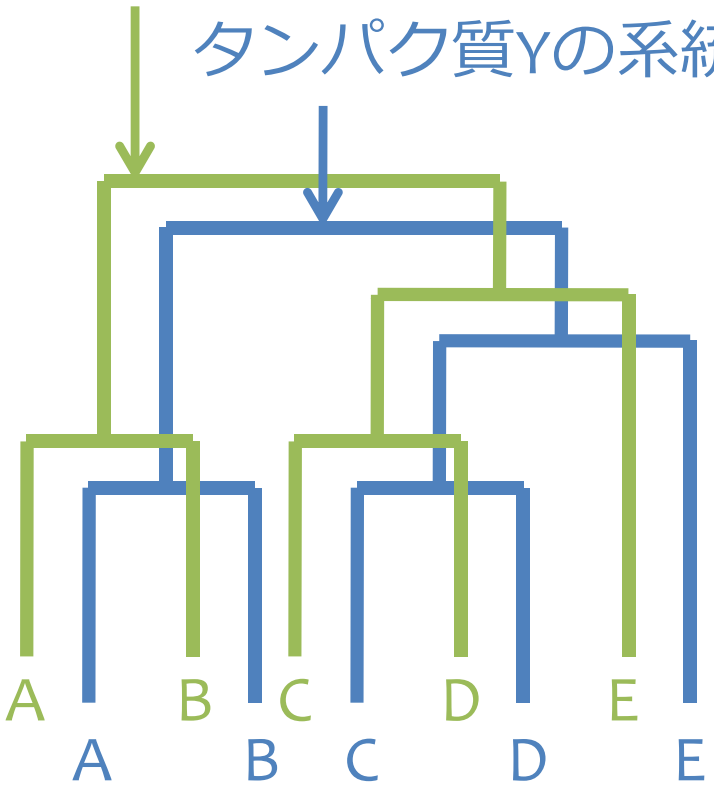
種間で保存されるタンパク質は相互作用すると考えられる。

PNAS 96, 4285 (1999)

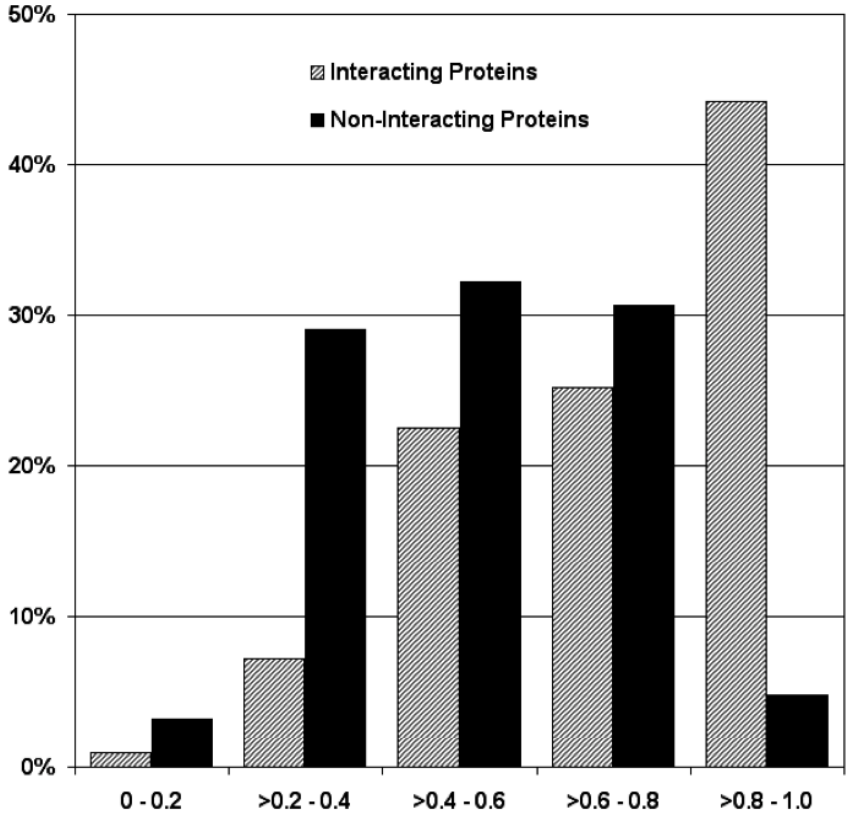
共進化を用いた予測

タンパク質Xの系統樹

タンパク質Yの系統樹



系統距離の関係が似ていれば相互作用する。



Nucleic Acids Res 32, W69 (2004)

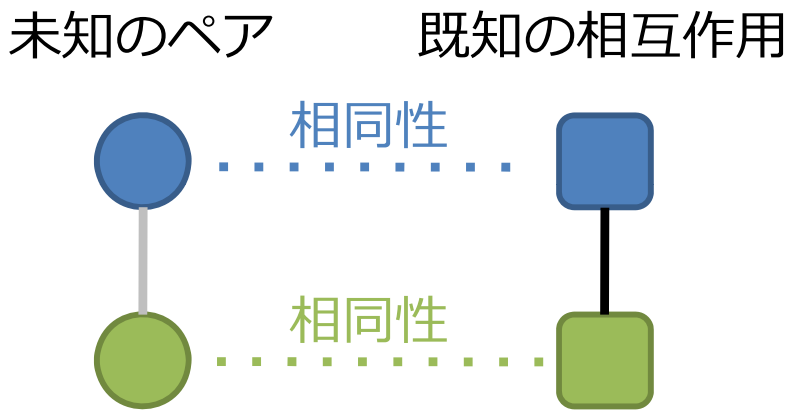
タンパク質Xにおける生物種iとjの間の類似度

$$r = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X_{ij}-\bar{X})(Y_{ij}-\bar{Y})}{\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X_{ij}-\bar{X})^2}\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(Y_{ij}-\bar{Y})^2}},$$

タンパク質Yにおける生物種iとjの間の類似度

事前知識による予測

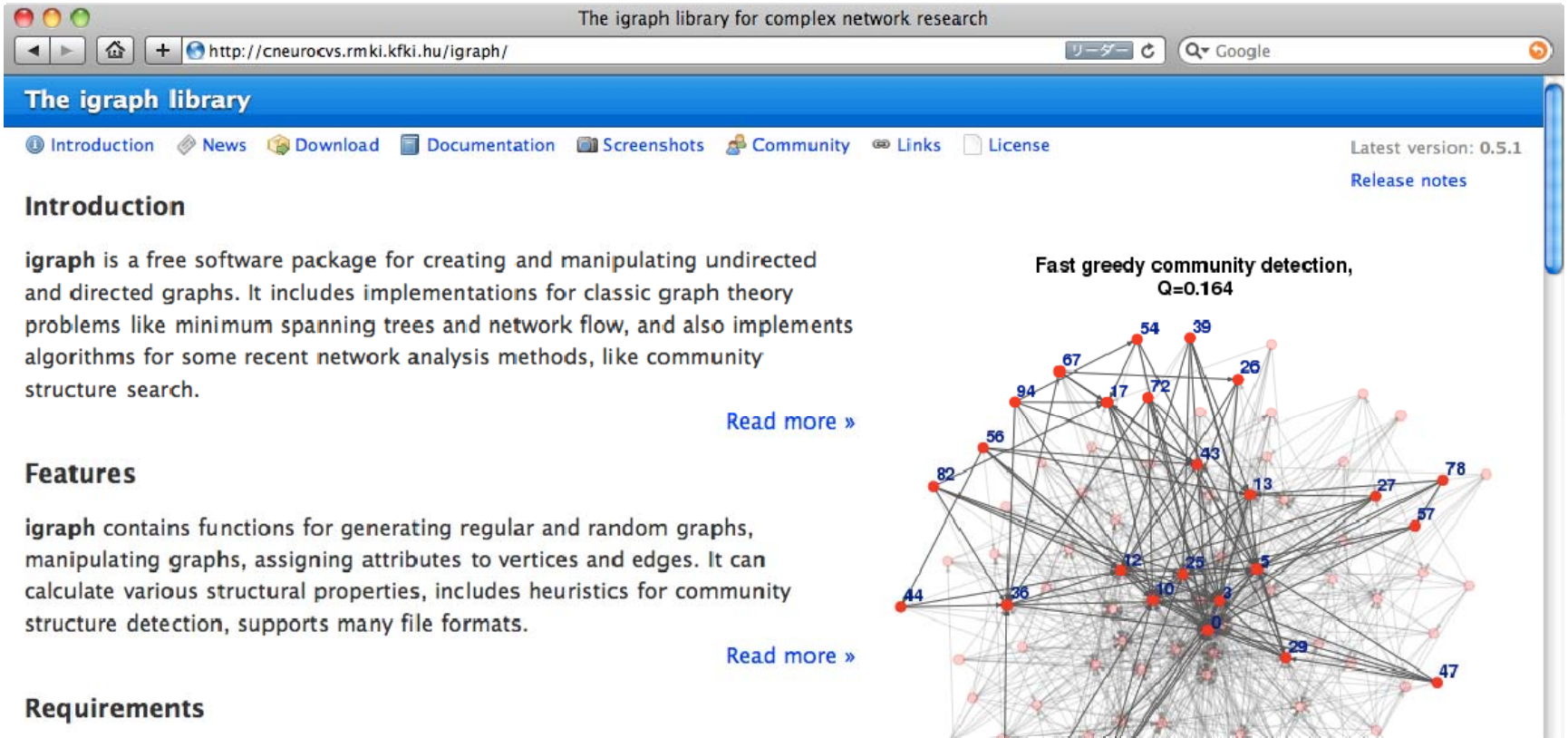
- 既知の相互作用ペアから相互作用の規則を抽出する。
  - アミノ酸配列の相同性
  - タンパク質の構造類似性
  - 遺伝子の共発現パタン



- 高度な統計手法を用いれば精度が向上
  - 機械学習、ベイジアンネットワーク

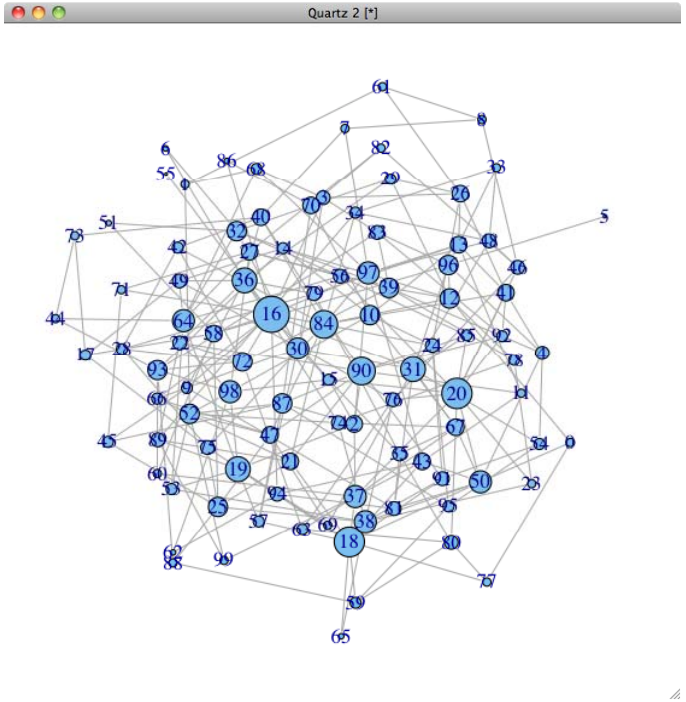
igraph

- Rのパッケージ
  - http://cneurocvs.rmki.kfki.hu/igraph/
- 簡単にネットワーク解析ができる。



簡単な解析

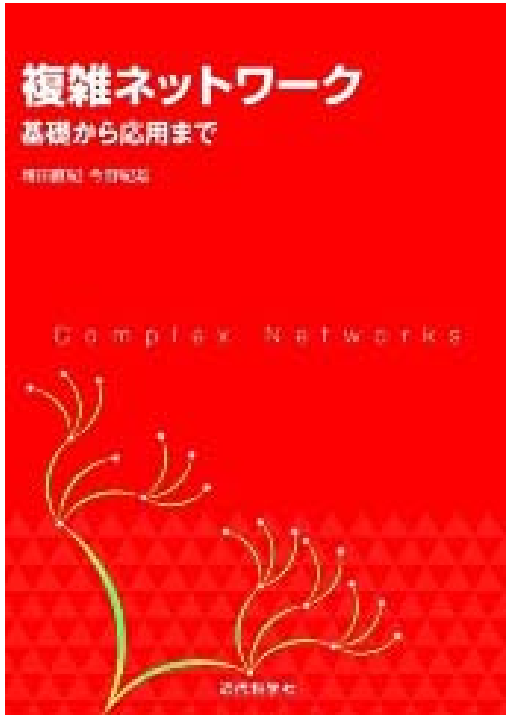
1. N=100, p=0.05のランダムグラフを作成
2. 次数中心度を計算する。
3. 次数中心度を頂点のサイズに反映させてネットワークを表示



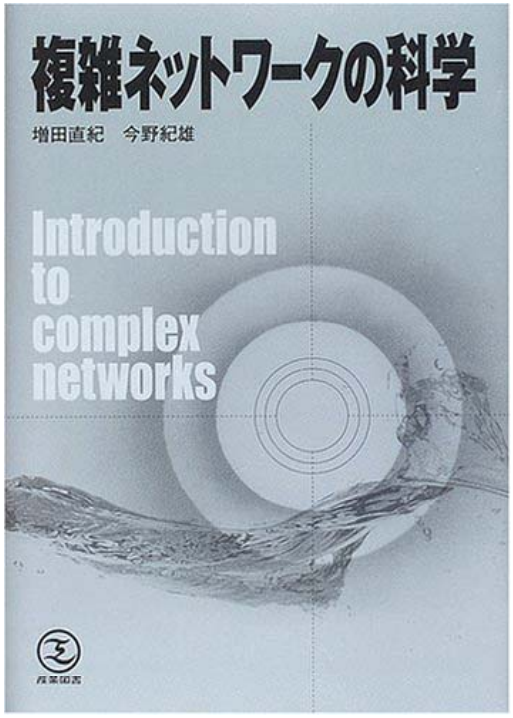

```
> library(igraph)   # ライブラリの読み込み
> g<-erdos.renyi.game(100,0.05)   # 1.に対応
> c<-degree(g)    # 2.に対応
> plot(g,vertex.size=c,layout=layout.kamada.kawai)
                                                  # 3.に対応
```

参考資料

- ネットワークの基礎が学べます。



増田直紀，今野紀雄
近代科学社 (2010)



増田直紀，今野紀雄
産業図書 (2005).