タンパク質相互作用ネットワークからの密結合モジュールの全列挙

津田　宏治
産総研生命情報科学研究センター
2009年9月

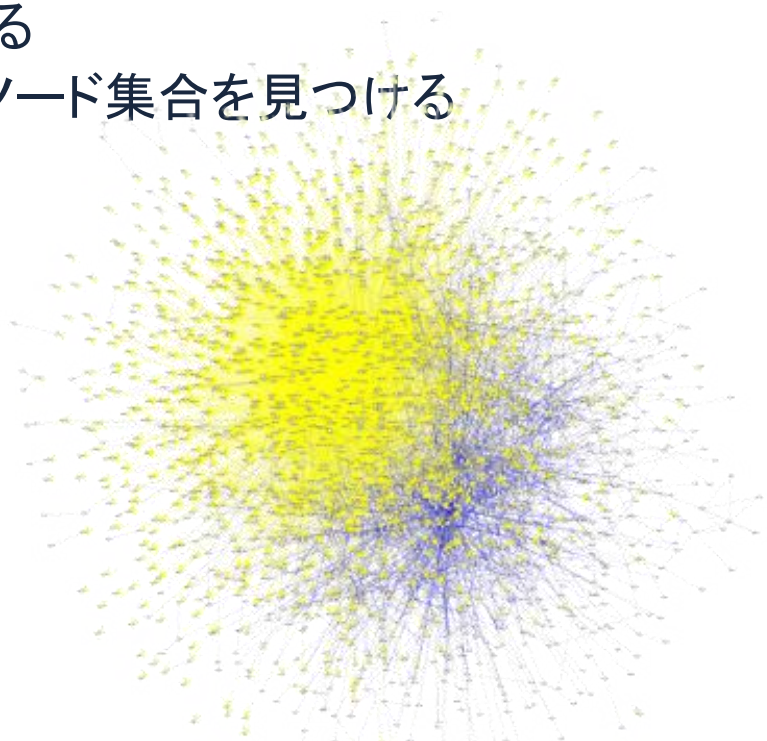
Joint work with
E. Georgii, T. Uno, S. Dietmann, P.Pagel

生物学的な背景

- 細胞の中のプロセスは、おもにタンパク質の複合体（complex)によって行われている
- 近年、実験によって確かめられた相互作用データが手にはいるようになった
- 研究の目的
  - 複合体を、相互作用ネットワークの密結合モジュールとして予測する
  - 付加的情報として、遺伝子発現データを用いて、より精度を高める

タンパク質相互作用ネットワーク

- ノード: タンパク質
- エッジ: 二つのタンパク質の物理的相互作用（結合）
- Challenge 1: 複合体の中のタンパク質の結合は全体全ではない
  - クリークを探してもうまくいかない
- Challenge 2: 誤ったエッジ
  - エッジの確信度スコアを考慮に入れる
- 高確信度のエッジが高密度に存在するノード集合を見つける



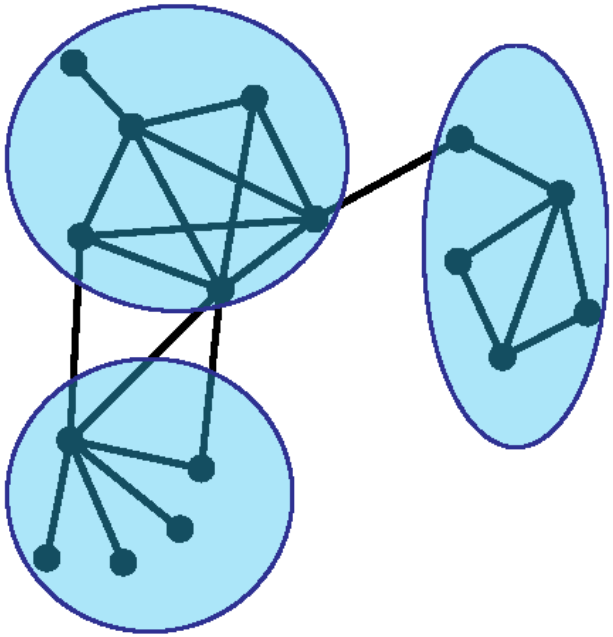
モジュール発見

- 従来法： 完全性に欠ける。重要なクラスタを逃す可能性
  - Clique percolation [Palla et al., 2005]
  - Partitioning
    - Hierarchical clustering [Girvan and Newman, 2001]
    - Flow Simulation [Krogan et al., 2006]
    - Spectral methods [Newman, 2006]
  - Heuristic Local Search [Bader and Hogue, 2003]

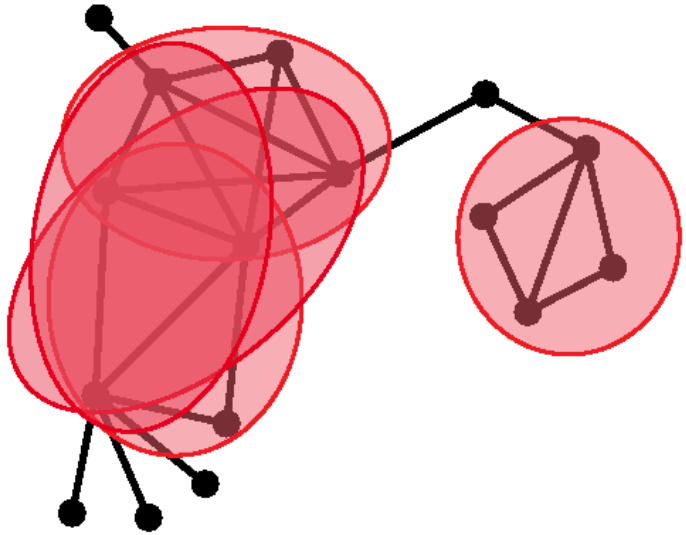
- 我々の方針
  - 密結合モジュールを全列挙する

分割法　対　列挙法

- 列挙法は、密度しきい値を満たすノード集合を列挙
- 1タンパク質が複数の複合体に含まれることが多い
- 分割法ではそのような場合に対応できない



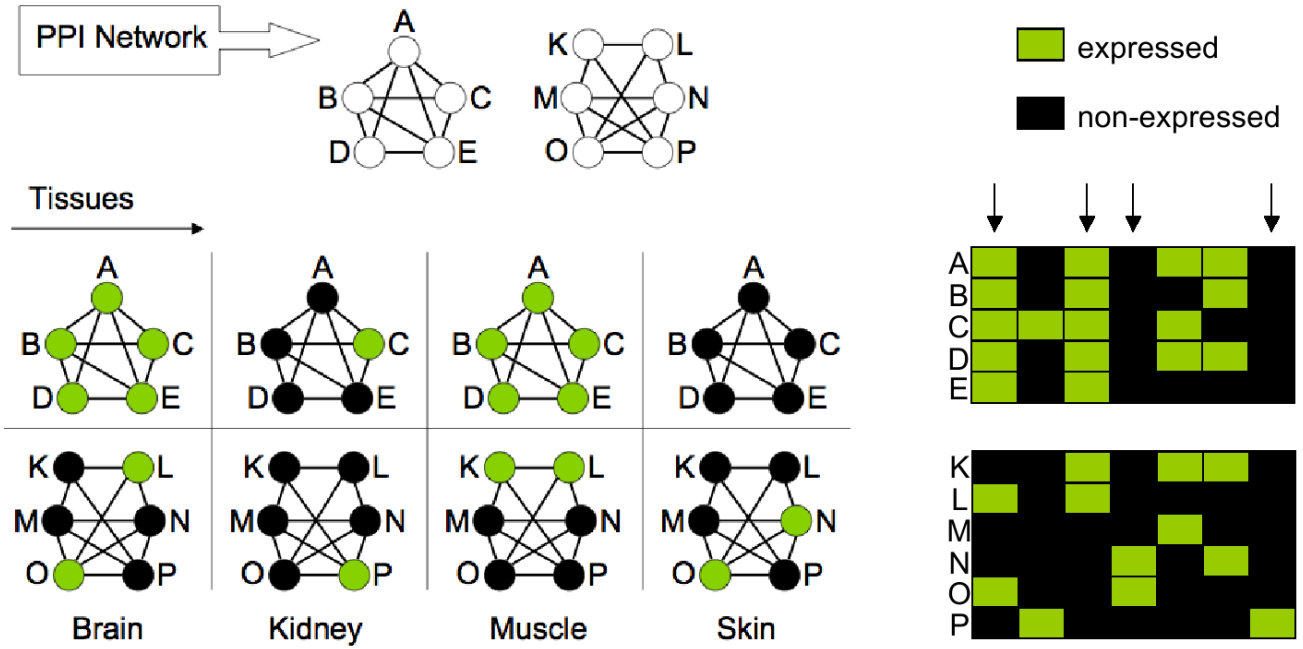
Partitioning



Enumeration

付加的な情報の利用

- 遺伝子発現データの利用
  - 組織に応じた、タンパク質の有無を表す
- モジュールが満たすべき条件
  - e_1：モジュール内の全タンパク質が発現している組織数
  - e_0：モジュール内のタンパク質のどれも発現していない組織数
  - 両方に、しきい値を設ける



問題の定式化

- Interaction network: $G = (V, E(V))$
- Edge weights: $0 \leq w(\{u, v\}) \leq 1$
- Density of $U \subset V$:

$$d(U) = \frac{\sum_{\{u,v\} \in E(U)} w(\{u, v\})}{|U|(|U| - 1)/2}$$

密度

エッジ重みの和

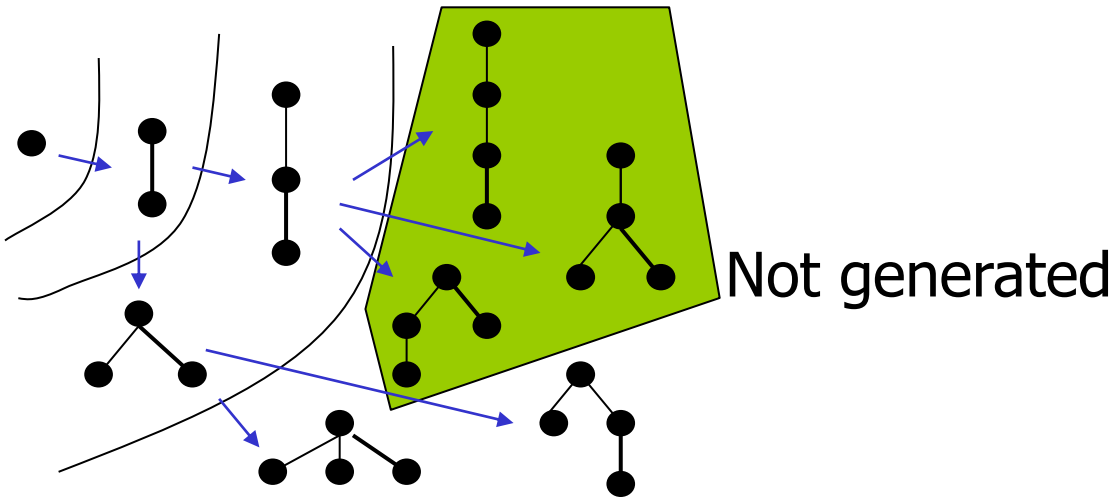
最大のエッジ数

- Find all $U \subset V$ with $d(U) \geq \theta$, $e_1(U) \geq n_1$, and $e_0(U) \geq n_0$

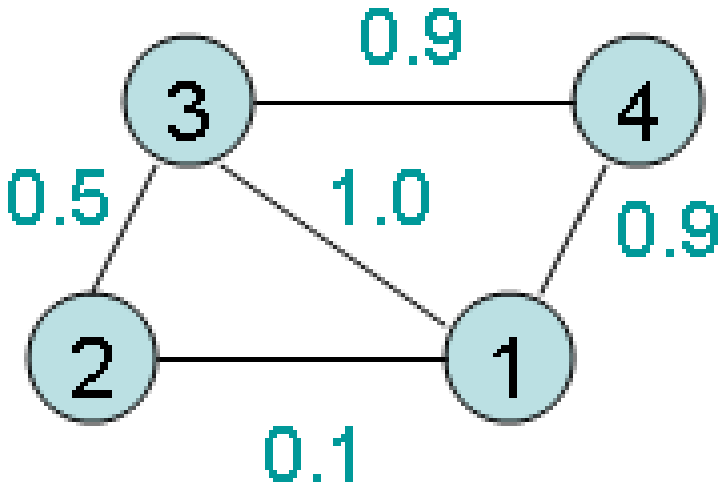
密結合モジュール列挙アルゴリズム

典型的な列挙アルゴリズムの例

- Itemset mining, graph mining etc.
- データベースにおける頻度がm以上の要素を列挙する
- 探索木を設定
- 逆単調性に基づく枝刈り
  - ある要素の頻度は根から葉にかけて単調に減少する



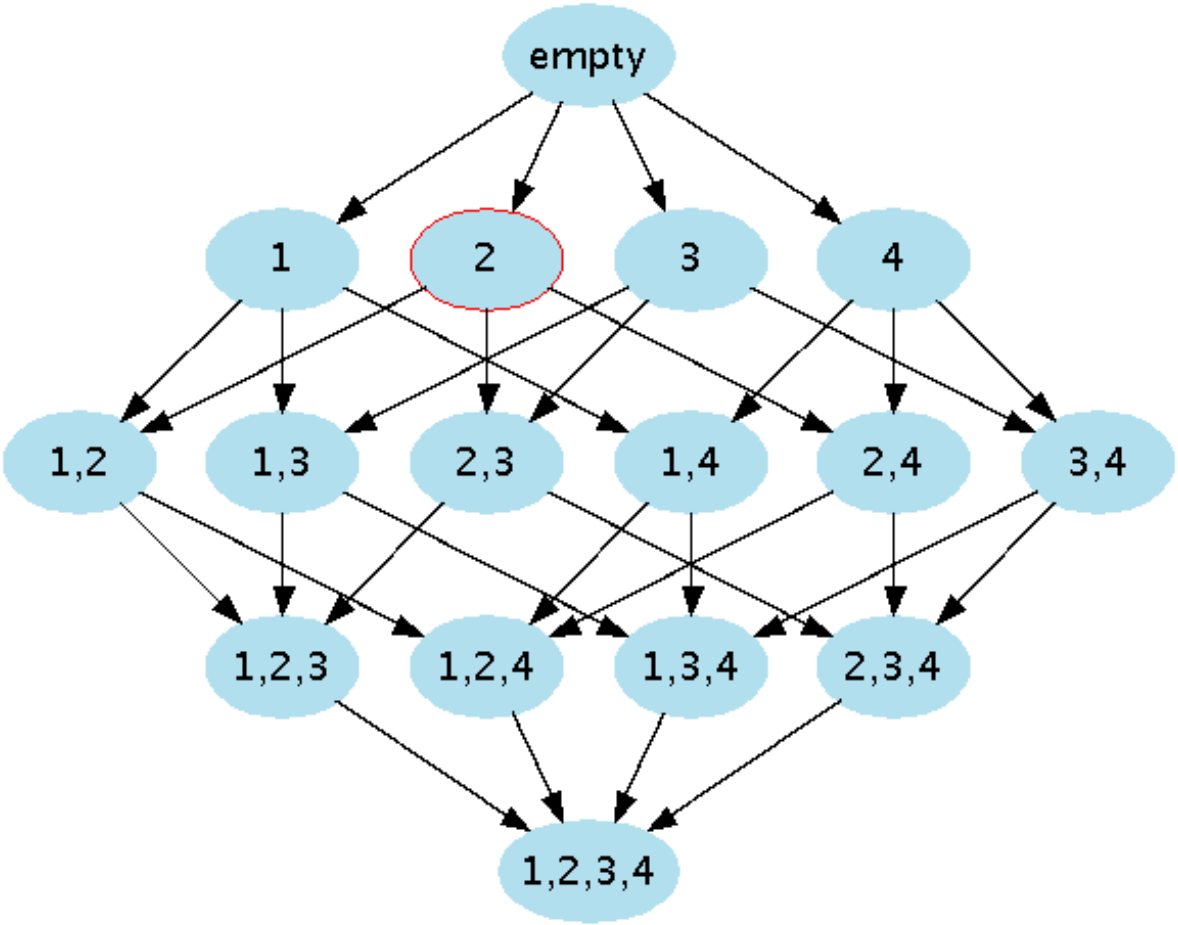
ネットワークの例



Density of Modules

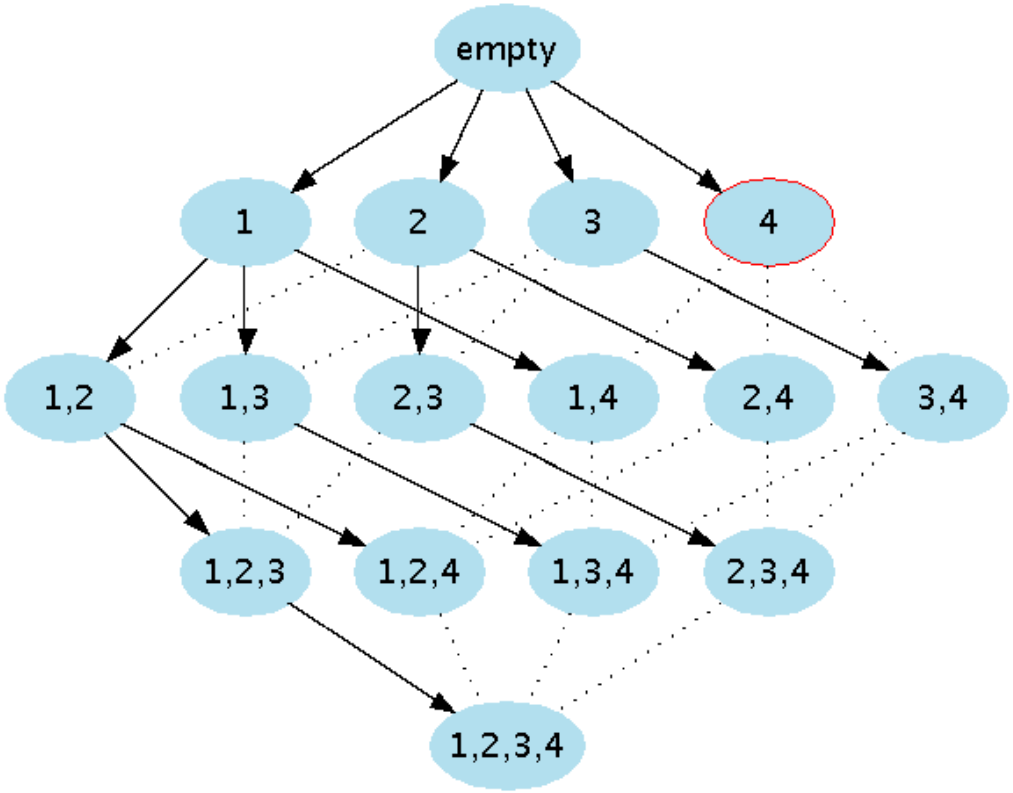
1,2,3	0.5
1,3,4	0.9
2,3,4	0.5
1,2,4	0.3

全モジュールはグラフ状の探索空間をなす



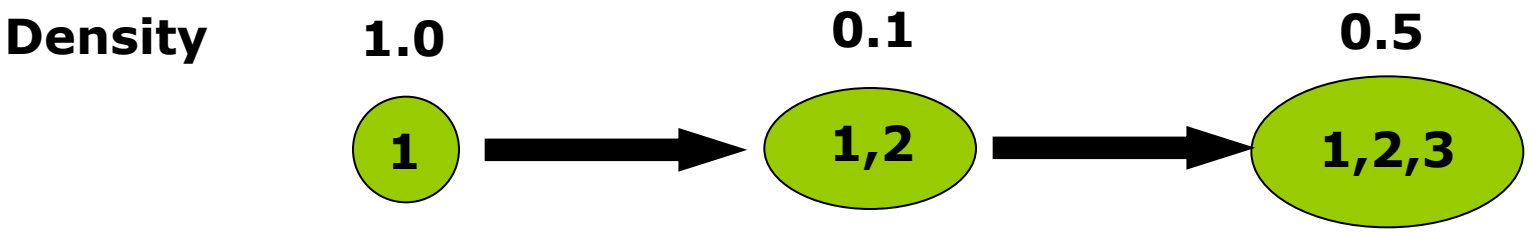
探索木を選択する

- 高速な列挙のためには、探索木が必要
- 探索グラフからは、様々な探索木を選択することができる
- Default: Lexicographical ordering



モジュール密度は単調減少しない

- 密なモジュールの部分集合は、必ずしも密でない
- 密度は、根から葉にかけての経路において、必ずしも単調減少しない
  - 枝刈り不能？



質問

- 密度が単調に減少するような探索木を選択することは常に可能か?

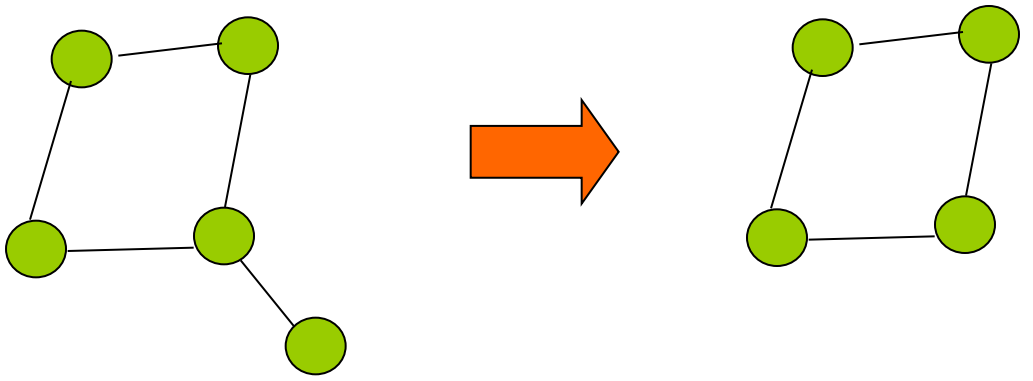
質問

- 密度が単調に減少するような探索木を選択することは常に可能か?

- YES!
  - 逆探索法(Reverse Search, Avis and Fukuda 1993)を用いればよい

逆探索法(Avis and Fukuda, 1993)

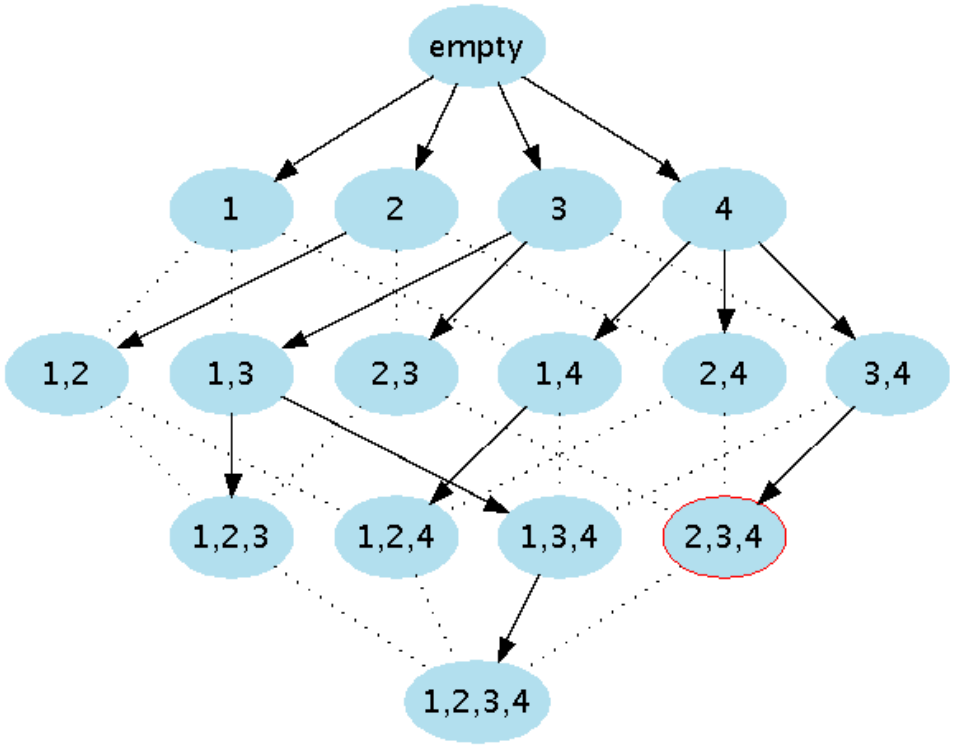
- 探索空間の中で、探索木を指定する方法
- Reduction Mapping
  - 子から親を生成する規則
  - もっとも度数の少ないノードを取り除く
  - 取り除くことによって、密度は必ず増加する



CHILD **PARENT**

Reduction Mappingによって、探索木は一意に指定される

- 必要十分条件: Reduction mappingを有限解施すことによって、すべてのノードが根ノードに収束する



逆探索における列挙アルゴリズム

- Reduction mappingは子から親を作る規則、しかし、列挙では親から子を作る必要

- On-demand construction
  - すべての子候補を用意し、各々にreduction mappingをかける。自分のところに戻ってくる候補を子とする。
- もしも子がなければ枝刈り

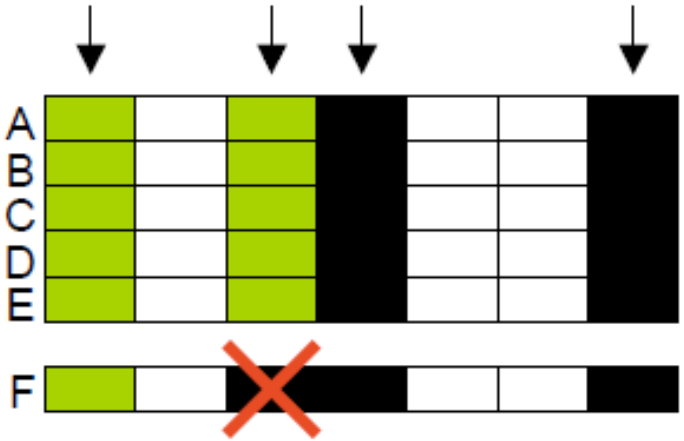
付加的データとの組み合わせ

- 遺伝子発現データからの制約：All-onがn1組織以上、All-offがn0組織以上

$$e_1(U) \geq n_1,\ e_0(U) \geq n_0$$

- 単調性: e0とe1は、根から葉への経路において単調減少

Module	e_1	e_0
ABCDE	2	2
ABCDEF	1	2



- 条件が満たされなかった時点で枝刈りすればいい

モジュールの統計的有意性

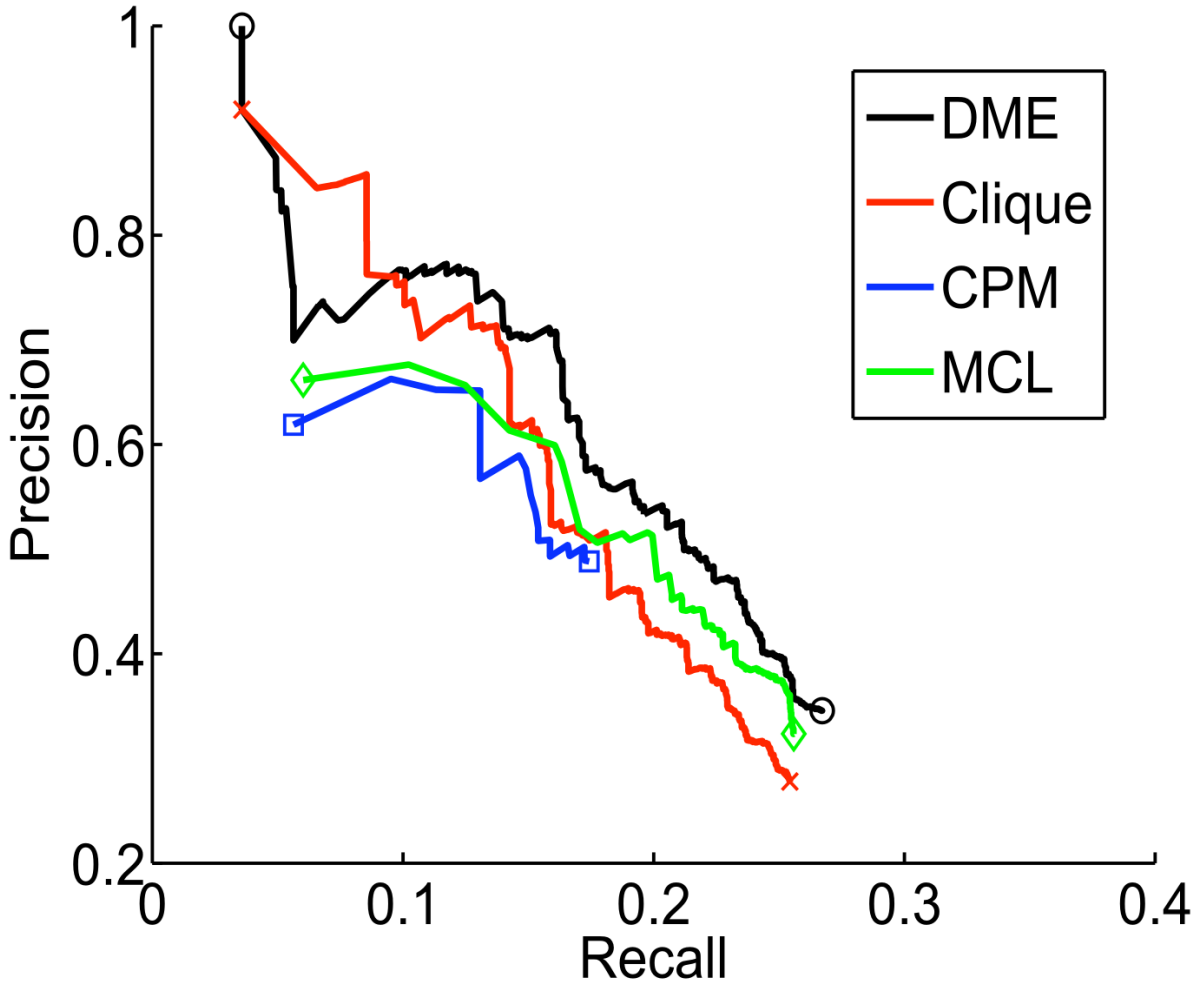
- k：モジュールに含まれるノード数
- ρ ：モジュールの密度
- $m_k(\rho)$: サイズが k で、密度が ρ 以上のモジュールの数
- ランダム選択によって、これよりも密度が高いモジュールがみつかる確率 (p-value)

$$p = m_k(\rho) / \left(\begin{array}{c} n \\ k \end{array} \right)$$

実験結果

イースト菌のネットワークにおける他手法との精度比較

- CYGD-MpactとDIPから相互作用データを取得
- 3559ノード、１４２４２エッジ
- エッジの確信度は (Jansen, 2003)の方法で計算
- 比較手法
  - Clique detection (Clique)
  - Clique Parcolation Method (CPM)
  - Markov Clustering
- Density threshold = 0.95, finished in 2667 secs
- 得たモジュールを MIPS complexesと照合してPrecision-Recallカーブを描く


DME
Clique
CPM
MCL
Precision
Recall
1
0.8
0.6
0.4
0.2
0
0.1
0.2
0.3
0.4

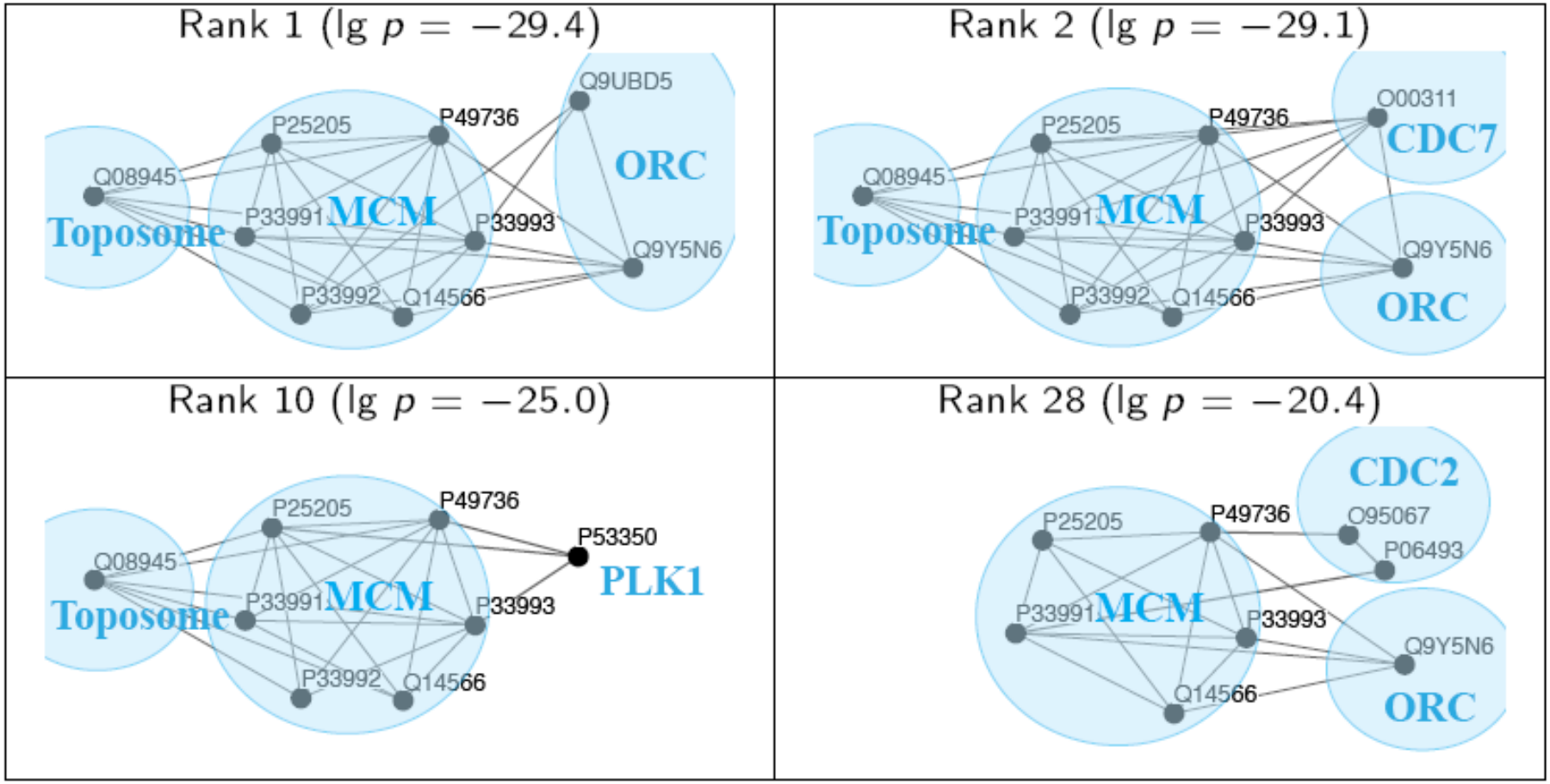
ヒトの相互作用ネットワークの解析

- 組織特異的遺伝子発現プロファイル (Su et al., 2004)
  - 79 組織における発現
- モジュールの全遺伝子が3組織以上で発現、10組織以上で発現なし
- 7763タンパク質, 密度 >= 35%, 5分で終了
- 1021の極大モチーフを発見
- MIPS human complex database (Ruepp et al., to appear)との照合

実験のまとめ

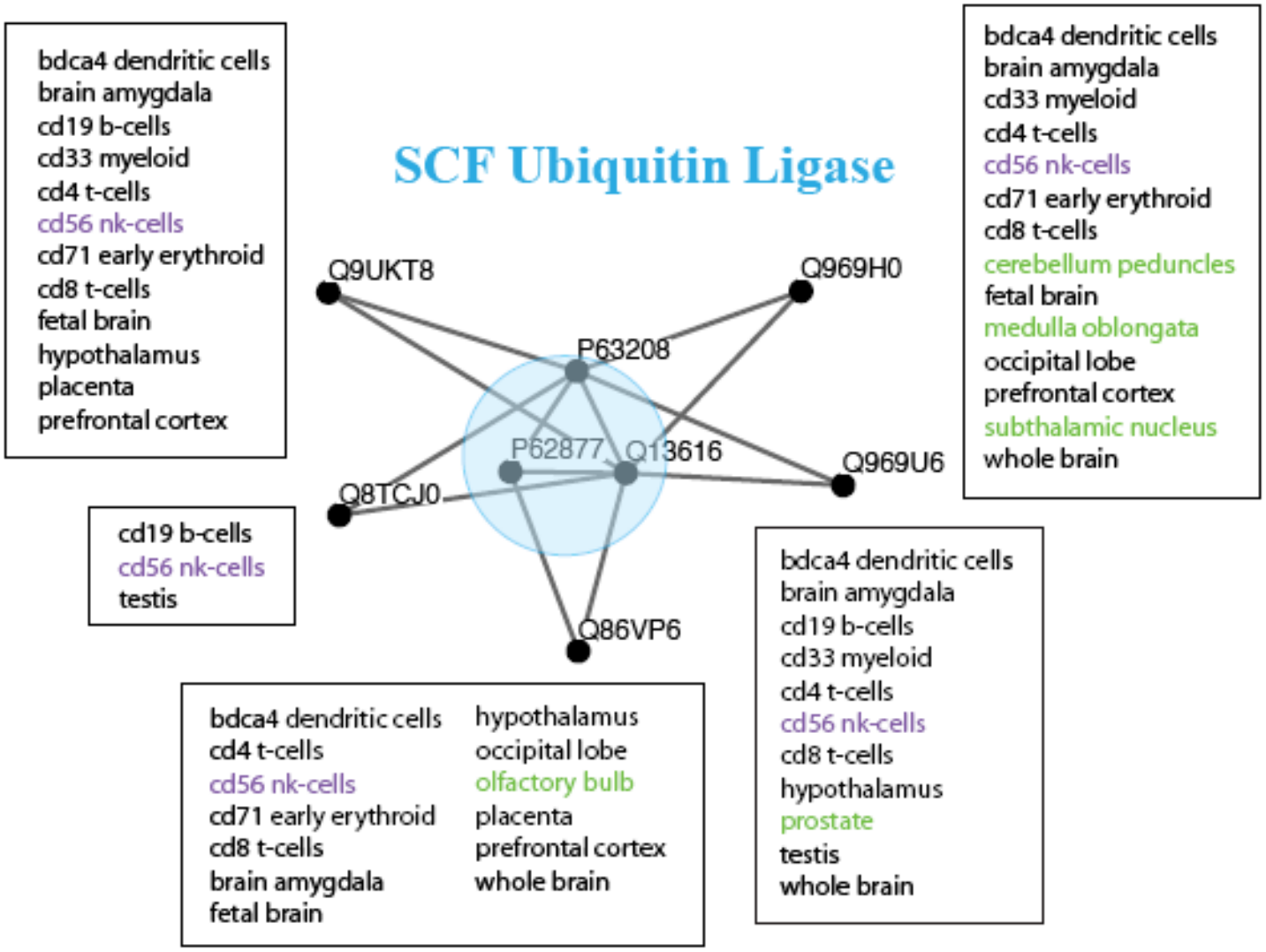
- MCM complexと他の複合体との関係を発見（ORC, CDC7, Toposome, PLK1 protein）
- Uqcrc1, Uqcrc2, Uqcrb, Cyc1からなるモジュールを発見 (lg p = -13)
  - MIPSには登録なし
  - 文献：Ubiquinol-cytochrome c reductase complex
- SCF E3 ubiquitin ligase complex: 他のプロテインにユビキチンを付加し、分解する
  - 組織特異性のことなる5つのバリエーションを発見
  - 周辺のタンパク質: Substrate recognition particles
  - ユビキチンを付加する対象が組織特異的に選ばれていることを示唆
  - Natural Killer cellsは、すべての周辺タンパク質を持つ

MCM complexを含むモジュール



Expressed in bone mallow cells
Not expressed in brain, liver, kidney etc.

SCF ligase complexの組織特異的再構成



まとめ

- 逆探索法に基づく新しいモジュール列挙アルゴリズムを提案
- 補助的な情報ソースとの組み合わせが可能
- 列挙により、各モジュールの統計的優位性を計算可能
- イースト菌とヒトの相互作用ネットワークに適用し、優れた結果を得た