Vol.2009-BIO-18 No.9
2009/9/18

タンパク質機能間関係グラフの構築

寺 田 愛 花[†1]　瀬々 潤[†1]

タンパク質は状況に応じ，異なるタンパク質と相互作用することで機能を発現させている．現在行われているタンパク質相互作用の網羅的観測から構築されるネットワークは複雑であり，どのような状況で相互作用が利用されているか理解することは容易ではない．本研究では相互作用ネットワークから，ネットワークモチーフ— 頻出する部分ネットワーク構造—に着目する事で，タンパク質の機能間に潜む関係を発見し，タンパク質機能間関係を表すグラフを作成する．本手法を酵母の相互作用情報に適用した結果，プリン塩基とピリミジン塩基を有するタンパク質を，エネルギー生成を行う同一の機能として分類できた．また，細胞分裂周期を制御するタンパク質に，相互作用する機能が多いものと少ないものがあることが分かった．

Construction Graph of Proteins Functions

Aika Terada[†1] and Jun Sese[†1]

Function of proteins depends on interaction with other proteins, and the interacting partners are changed in cell conditions. Although observation of comprehensive protein-protein interactions (PPI) tell us the interacting partners, it is difficult to understand the roles and the conditions of the interactions. Here, we develop a novel method to discover large and frequent subgraphs from the PPI, called network motifs, and find a graph about the relations of protein functions. The result apply to yeast dataset shows that we classify proteins associated with purine metabolism and pyrimidine metabolism into the same group to produce enerergies. Furthermore, we discover that the cell cycle proteins are categorized into two different groups; one showing association with various functions and other showing little association.

†1 お茶の水女子大学 大学院人間文化創成科学研究科
Department of Computer Science, Ochanomizu University

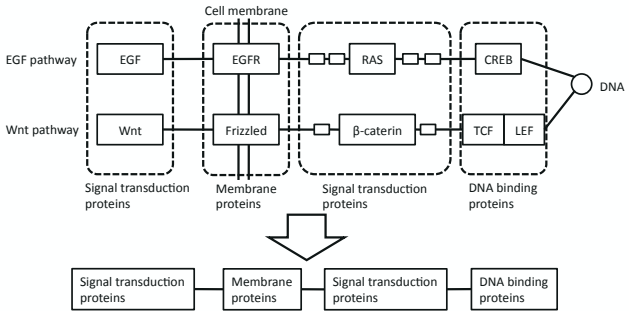

図 1　2 つの pathway に共通するネットワークモチーフの例
Fig. 1　Example of a network motif over different two pathways.

1. はじめに

解析技術の進歩に伴い，タンパク質相互作用をはじめとする生体内ネットワーク情報が採取されるようになってきている．多くのタンパク質は他のタンパク質と相互作用することによって，初めて細胞内で機能を有するため，この情報を基にタンパク質の機能を解析する研究が盛んである[2)–4),7)]．

ネットワークの解析は，大きく三種類に分けられる．まず，ネットワークのスケールフリー性を解析し，ネットワークの構築過程を理解すること[1),2)]，次に，高密度に辺が張られている頂点集合，クリークをネットワークから抽出すること[3)–5)]，最後に，ネットワークに頻繁に現れる構造，ネットワークモチーフを発見することである[6)–8)]．

ネットワークモチーフを発見する手法は，近年盛んに研究されており，生物学的に有益な結果が多数得られている．[7)] スケールフリー性の解析やクリーク抽出とは異なり，頻出するネットワーク構造を抽出できるネットワークモチーフの発見をすることで，ネットワークにおける頂点の役割を知ることができる．例えば，多くのタンパク質が他のタンパク質と相互作用して機能を果たすため，頂点をタンパク質，辺をタンパク質の相互関係で与えるタンパク質相互作用ネットワークのモチーフを発見することで，タンパク質同士の関係性を解析することができ，その機能を理解することができる．

図 1 は，pathway ネットワーク上のネットワークモチーフの例である．この図はタンパク質を四角で表し，関連のあるタンパク質の間に線を引いている．タンパク質 EGF から始まる上のネットワークの流れは，細胞の成長と強い関係がある EGF pathway である．下

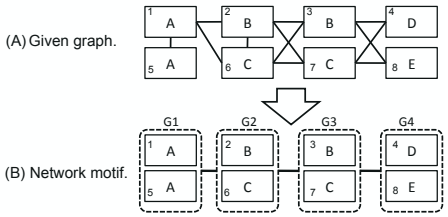

図 2 解析結果の例
Fig. 2 Motivating example.

のタンパク質 Wnt から始まる Wnt pathway は，癌に関連する pathway である．本研究におけるネットワークモチーフとは，ネットワークに頻出する構造を発見することである．図 1 のネットワークで頻出する構造をこの図の最下部に示している．どちらの pathway も，細胞外シグナルで知られるタンパク質から始まり，膜タンパク質に対して辺を張っている．膜タンパク質は，細胞内でもタンパク質変換のシグナルと関連し，最後は DNA 結合タンパクと関連する．このネットワークモチーフ構造から，細胞内のタンパク質の役割を知ることができる．

図 2(A) は本手法の例である．このグラフは頂点をタンパク質で与え，関連しているタンパク質間に辺を張っている．頂点に書かれている番号は ID である．また，タンパク質が有する機能を頂点の属性で与え、それをアルファベットで表している．この例では A から E の 5 種類の属性があり，共通の機能 A を有する頂点が二つある．

図 2(A) のグラフから求めたネットワークモチーフが，図 2(B) である．この図では点線で頂点のグループを表しており，G1，G2，G3，G4 の 4 つのグループがある．頂点 2 と 6 はグループ G2 に属しており，この二つはそれぞれ頂点 1，3，7 と隣接している．G2 に属する頂点は二つとも G1 と G3 に対して辺を張っていることから，求めたネットワークモチーフでも G2 は G1，G3 と隣接している．また，G2 に属する頂点と G4 に属する頂点の間には全く辺がないため，ネットワークモチーフでも G2 と G4 の間には辺がない．同様に，G3 に属する頂点も G2 と G4 に属する頂点と隣接し，G1 に対しては辺がないことから，ネットワークモチーフでも G3 は G2，G4 に対しては辺があるが，G1 との間には辺が無い．

この例では，図 2(B) のような結果を求めることができる．しかし，生物学的なネットワークを表したグラフ構造の多くは，ノイズが含まれており，曖昧であることから，与えられたネットワークに対してモチーフを求めることは難しい．この問題を解決するために，本研究ではネットワークモチーフを計測する新たな指標を定義し，この指標が最大となるネットワークモチーフを発見する手法を提案する．

2. 関 連 研 究

近年，グラフマイニングは盛んに研究されている．頻出するサブグラフを発見する手法[9)–12)]は，グラフデータベースから，共通した構造があり頻出するサブグラフを全て列挙するために用いられる．しかし，生物学的なネットワークにはノイズがあるため，これらのアルゴリズムで頻出する大きなサブグラフを求めることは難しい．本手法の目的は，タンパク質相互作用ネットワーク全体から大きなモチーフのネットワークを抽出することである．

クリークの抽出は，ネットワーク解析の主題の一つである．図 2 では，頂点 2，3，6，7 はクリークである．しかし，2 と 6 はそれぞれ 1 と隣接し，4 と 8 とは隣接していない．一方で，3 と 7 は 4 と 8 に対してそれぞれ辺を張るが，1 に対しては，辺が無い．このように，2 と 6 と，4 と 8 は異なる構造であるため，同じグループに分類するべきではない．しかし，この 4 個の頂点はクリークとして抽出されるため，クリークを抽出する手法ではモチーフを発見することは出来ない．

ネットワークモチーフを抽出する手法は，生物学[7)] とデータマイニング[6)8)] の両方の分野から研究されている．生物学的には，一部の遺伝子やタンパク質だけが研究されている[7)] Chen *et al.*[6)] が研究したネットワークは 25 個以下の遺伝子で構成したものであり，頂点にラベルがないという制限がある．k-SNAP[8)] は膨大なネットワークを解析することができるが，この手法には頂点のラベルの種類が膨大であると，モチーフを発見できないという問題点がある．本手法ではラベルの種類が膨大なネットワークであってもモチーフを抽出することができる．

3. 定　　義

3.1 属性と関係性が最良なグループ集合

この節では，ネットワークモチーフのモデルを定義し，本研究で目標とする最適解を示す．まず，ネットワークモチーフの発見のために，頂点のグループを定義する．

定義 1 （グラフの定義）グラフを，$G = (V(G), E(G), A(G))$ で定義し，重み無しの無向グラフとする．$V(G)$ はグラフの全頂点集合，$E(G)$ は辺の集合である．グラフ G の頂点の属性の集合が $A(G)$ であり，各頂点 v は一つの属性 $a(v) \in A(G)$ を有する．辺には属性がない．本論文では，このようなグラフをラベル有りグラフと定義する．

次に，ラベル有りグラフの頂点のグループを定義する．一つの頂点は必ず一つのグループに属し，オーバーラップはない．

定義 2 （グラフのグループ集合）グラフ G の頂点の分類として，$\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ が以下の条件を満たすとき，$\mathcal{C}$ を G のグループ集合と定義する．

(1) $\forall C_i \in \mathcal{C}, C_i \subseteq V(G)$ かつ $C_i \neq \phi$,

(2) $\bigcup_{C \in \mathcal{C}} C = V(G)$

(3) $\forall C_i, C_j \in \mathcal{C}$, $i \neq j$ について， $C_i \cap C_j = \phi$

次に，頂点の属性で構成するグループ集合を定義する．

定義 3 （頂点の属性が等しいグループ集合）$\mathcal{C}$ は，G のグループ集合とする．$c(v)$ は，頂点 v が属するグループを表し，そのグループは $\mathcal{C}$ に含まれる．本論文では， 以下の条件を満たす $\mathcal{C}$ を，A-compatible と定義する: $\forall u, v \in V(G)$ において， $c(u) = c(v)$ ならば $a(u) = a(v)$ である．

一つのグラフに対して，A-compatible を満たすグループ集合は一意的ではない．そこで，グラフのグループ集合に対する優位関係を定義し，この優位関係を $\preceq$ で表す．A-compatible となるグループ集合の中で優位性が最も高いものが，グループ数が最も少なく，最も良いモチーフ構造を表している．

定義 4 (優位関係) グラフ G のグループ集合 $\mathcal{C}$，$\mathcal{C}'$ を考える．$\mathcal{C}' \preceq \mathcal{C}$ である必要十分条件は，$\forall C' \in \mathcal{C}', \exists C \in \mathcal{C}$ に対し，$C' \subseteq C$ が成り立つことである．

A-compatible である全てのグループ集合には，$\preceq$ で表される優位性が最も高いグループ集合が必ず一つ存在する[8]．このグループ集合は，A-compatible となるグループ集合の中でグループ数が最小の集合である．

A-compatible は，頂点の属性でグループ集合を構成する．しかし，モチーフの発見には，グループ内の頂点の属性だけではなく，他のグループに対する関係性も類似している必要がある．つまり，他のグループに対する辺の張り方も類似する頂点でグループを構成する．そのため，各グループ内の頂点の属性と辺の関係が完全に同一であるグループ集合について定義する．

定義 5 (属性と関係性が最良なグループ集合) $\mathcal{N}(v)$ を $\{c(u) \mid (u, v) \in E(G)\}$ とする．$\mathcal{N}(v)$ は， v が辺を張るグループ集合である．グラフ G のグループ集合を $\mathcal{C}$ とし，次の二つの条件を満たすとき，$\mathcal{C}$ は *(A,R)-compatible* と定義する．

(1) $\mathcal{C}$ は A-compatible である．

(2) $\forall u, v \in V(G)$ について，$c(u) = c(v)$ ならば $\mathcal{N}(v) = \mathcal{N}(u)$ である．

(A,R)-compatible であるグループ集合は，頂点の属性と辺の関係が等しいグループで構成される．しかし，多くの生物学的なネットワークにはノイズが含まれており，まだ解明されていないこともある．*(A,R)*-compatible では各グループに対して類似性の制限が強く，ノイズが含まれていると，目的とは異なるグループとして構成される可能性がある．ノイズを含んだグループは小さいグループを構成する．頻出するネットワーク構造を抽出するために，本手法では属性と関係性の制限を緩和する新たな指標を定義する．またこの指標は正規化項を含み，グループ数を調節することができる．

3.2 ネットワークモチーフの指標

この節では，グループ内の頂点は属性と関係性が同一であるという制限を緩和する．まず，頂点の属性についてエントロピーを用いて緩和する．

定義 6 (頂点属性の類似性)

$$p_A(a) = \frac{|\{v \mid a(v) = a\}|}{|V(G)|} \ , \quad p_A(a \mid C_i) = \frac{|\{v \in C_i \mid a(v) = a\}|}{|C_i|} \text{ として,}$$

グループ C_i に属する頂点の類似性 $Ent(C_i)$ を次式で定義する．

$$Ent(C_i) = \sum_{a \in A(G)} p_A(a \mid C_i) \log_2 \frac{p_A(a \mid C_i)}{p_A(a)}$$

$Ent(C_i)$ は，グループ C_i の頂点の属性が同一であるほど値は大きく，異なるほど小さい．グラフのグループ集合が A-compatible であるかは $\sum_{C_i \in \mathcal{C}} Ent(C_i)$ で計測でき，一定のグループ数の下では，A-compatible な分割のとき最大となる．

次に，関係性に関する制限を緩和する．グループ C_i の頂点で，C_j と隣接している頂点のグループを $V_R(C_i, C_j)$ で表し，グループ間の隣接性を $r(C_i, C_j)$ で定義する．

定義 7 $V_R(C_i, C_j)$ を $\{v \in C_i \mid (v, v') \in E(G), v' \in C_j\}$ とする．$V_R(C_i, C_j)$ は， 非対称型である．グループ C_i と C_j の隣接性 $r(C_i, C_j)$ を次式で定義する．

$$r(C_i, C_j) = \frac{|V_R(C_i, C_j)| + |V_R(C_j, C_i)|}{|C_i| + |C_j|}$$

$r(C_i, C_j)$ が大きいほど，グループ C_i と C_j の頂点がお互いに隣接していることを表している．C_i と C_j の全頂点が互いに隣接しているとき $r(C_i, C_j)$ の値は 1 であり，1 つも隣接していないとき 0 である．

隣接性から，二つのグループの関係性を表す式を定義する．$r(C_i, C_j)$ が 0 か 1 に近いほど，C_i と C_j は関係性をよく表していることを示している．逆に 0.5 に近いほど，グルー

プ間の辺はランダムに張られており，グループの関係性は曖昧である．二つのグループ間の関係性を表す指標を定義する．

定義 8 グループ C_i と C_j の関係性を， $c(C_i, C_j)$ で計測する．

$$c(C_i, C_j) = \begin{cases} |V_R(C_i, C_j)| & \text{if } r(C_i, C_j) > 0.5 \\ |C_i| - |V_R(i, j)| & \text{otherwise} \end{cases}$$

グループ集合が *(A,R)*-compatible のとき，全てのグループ C_i，C_j について，$c(C_i, C_j) = |C_i|$ である．また，$c(C_i, C_j)$ は非対称型である．

以上の値を用い，グループ集合を評価する指標を定義する．この指標は，グループの頂点の属性の類似性と，グループ間の辺の関係性の両方が高いときに最大値をとる．

定義 9 (属性と関係性の良さを表す指標) $\mathcal{C}$ はグループ集合である．$\mathcal{C}$ の良さを計測する指標 $w(\mathcal{C})$ を次式で定義する．

$$w(\mathcal{C}) = \sum_{C_i \in \mathcal{C}} \left(Ent(C_i) \times \sum_{C_j \in \mathcal{C} \text{ and } C_j \neq C_i} c(C_i, C_j) \right) - \lambda |\mathcal{C}|^2$$

$w(\mathcal{C})$ を *attribute relationship index* と呼び，本論文では AR index と表記する．

AR index では，グループ数に対してペナルティを与えている．このペナルティの項にはパラメータ λ があり，この値でグループの個数を調節することができる．λ が小さいほど，グループの個数は多くなる．

AR index を用いることで，この指標が最大となるグループ集合を求める最適化問題を定義する．本論文では，この問題を *Attribute Relationship Index ANAlysis* (*ARIANA*) と呼ぶ．この問題の解から関係性の高いグループ間を求め，ネットワークからモチーフを発見できる．

しかし，AR index が最大となるグループ集合を求めることは難しい．そのため，次の章で指標の値が最大とは限らないが，大きい値となるグループ集合を求める，シンプルな手法を導入する．

4. 提 案 手 法

この章では，ARIANA 問題を解くためのアルゴリズムを導入する．$w(\mathcal{C})$ が最大となるグループ集合を求めることは難しいため，本研究では二段階で構成される新たなアルゴリズムを提案する．また，この新しいアルゴリズムで生物学的なネットワークを解析する．

Algorithm 1 : ARIANA(J, K, λ)

Require: グラフ $G = (V(G), E(G), A(G))$ である．
1: A-compatible で優位性が最も高いグループ集合 $\mathcal{C}$ を構成する．
2: // 併合段階
3: **for** while $|\mathcal{C}| \geq J$ **do**
4: 全組み合わせ $C_i, C_j \in \mathcal{C}$ について，$cos(C_i, C_j)$ を算出する．
5: $cos(C_i, C_j)$ が最小のグループを求める．
6: from $\mathcal{C}$ から C_i，C_j を除き，新たなグループ $C_i \cup C_j$ to $\mathcal{C}$ を追加する．
7: **end for**
8: // 分割段階
9: **for** while $|\mathcal{C}| \leq K$ **do**
10: $max\_w \leftarrow 0$.
11: **for** 各 $C_i, C_j \in \mathcal{C}$ **do**
12: $C' \leftarrow \{v \in C_i \mid (v, v') \in E(G) \text{ for } v' \in C_j\}$.
13: 新たなグループ $\mathcal{C}' \leftarrow \mathcal{C} - \{C_i\} + \{C'\} + \{C_i - C'\}$ を作成する．
14: $max\_w \leftarrow w(\mathcal{C}')$ if $w(\mathcal{C}') > max\_w$.
15: **end for**
16: $\mathcal{C} \leftarrow \mathcal{C}'$ $\mathcal{C}'$ は max_w に値を代入したグループ．
17: **end for**
18: return $\mathcal{C}$.

本手法は，グループの併合と分割の二段階で構成される．まず，A-compatible であるグループ集合の中で，グループ間の優位性（定義 4）が最も高いものを求める．次に，グループの併合を行うが，この段階ではコサイン類似度を用いた以下の式でグループ C_i，C_j の類似度を定義する．

$$cos(C_i, C_j) = \frac{\sum_{C_k \in \mathcal{C}'} |V_R(C_i, C_k)||V_R(C_j, C_k)|}{\sqrt{\sum_{C_k \in \mathcal{C}'} |V_R(C_i, C_k)|^2} \sqrt{\sum_{C_k \in \mathcal{C}'} |V_R(C_j, C_k)|^2}}$$

ここで，$\mathcal{C}' = \mathcal{C} - \{C_i, C_j\}$ である．この定義によって，C_i と C_j の他のグループに対する関係性の類似度を定義できる．本手法では，グループ数が J 個になるまでコサイン類似度が最大のグループから順に併合する。この手順により，異なる属性でも，関係性が類似している頂点が同一グループに属するように，グループ集合を構成できる．

次に，グループを分割して AR index を最大化する．分割するグループは，属性は類似しているが，他のグループに対する関係性が異なる頂点が多く属するグループである．この分割では，一つのグループを二つに分割するため，A-compatible なグラフに対しては，分割の前後ではグループ内の属性の類似性は同一である．そのため，グループを分割する場合に

は頂点の属性は考慮せず，AR index が上がるようにグループを二つに分類する．

グループ C_i を，あるグループに対して辺を張るグループ C' と辺を張らないグループ $C_i - C'$ の二つに分割する．この分割によって，AR index が上がる分割を行う．本手法では，C_i とは異なるグループ C_j を用いて C_i を二つに分割し，新たなグループ C' を以下のように構成する．

$$C' = \{v \in C_i \mid (v, v') \in E(G) \text{ for } v' \in C_j\}$$

グループ $C_i, C_j \in \mathcal{C}$ を全組み合わせでグループを分割し，AR index を計算する．その中で，値が最大のグループ集合を求める．

アルゴリズム 1 は，ARIANA の疑似コードである．1 行目で，A-compatible のグループ集合の中で，優位性が最も高いものを構成する．構成されるグループ数 $|\mathcal{C}|$ は，属性の種類数である．2 行目から 7 行目では，グループ数が J 個となるまでグループの併合を行い，8 行目から 17 行目では，グループ数が K 個になるまで，AR index が最大となるようにグループの分割をする．

ARIANA には，グループ数 J と K をパラメータとして入力する．この二つのパラメータの値を変更して生物学的なネットワークを解析し，AR index が最大となるグループ集合を求める．

5. 解析結果

5.1 疑似データの解析結果

この節では，疑似データの解析をする．疑似データは，頂点が 400 個，辺の本数が 1200 本，属性が 20 種類のグラフで，以下の方法で作成する．

(1) 20 種類の属性，$a_1, \ldots, a_{20}$ を生成する．
(2) 属性 a_i を有する頂点を 20 個作成する．頂点の総数は 400 個である．
(3) 10 個のグループ，$C_1, \ldots, C_{10}$ を構成する．$C_i (i = 1, \ldots, 10)$ の頂点は，$i \bmod 2 = 1$ のとき半数の 10 個が属性 a_i，残りの 10 個が a_{i+1} であり，$i \bmod 2 = 0$ のとき 10 個が a_{i-1}，残りの 10 個が a_i である．
(4) 10 個のグループ，$C_{11}, \ldots, C_{20}$ を構成する．グループ $C_i (i = 11, \ldots, 20)$ に属する 20 個の頂点全ての属性は a_i である．
(5) グループ間に 100 本の仮の辺をランダムに張る．各グループは，他のグループ 5 個に対して辺を張る．
(6) 仮の辺が張られているグループ C_i と C_j について，C_i と C_j の間に各頂点が 1 本ずつ辺を張るよう，20 本の辺を生成する．
(7) 各グループ $C_i (i = 1, \ldots, 20)$ の頂点間に辺を張る．C_i からランダムに選択した 2 つの頂点の間に，10 本の辺を張る．各頂点がグループ内に張る辺は 1 本である．

この手法で，グループ数が 20 個，グループ間の関係性が最良であるネットワークを生成する．構成したグループの 10 個は各グループ内の点の属性は等しく，残り 10 個は 2 種類の異なる属性で構成される．

次に，作成した疑似グラフにノイズを加える．

(1) ノイズの割合を r とする．
(2) $400 \times r$ 個の頂点をランダムに選択し，属性をランダムに変更する．
(3) $1200 \times (r/5)$ 本の辺をランダムに選択し，グラフから取り除く．
(4) $1200 \times (r/5)$ 本の辺をランダムに張る．

このノイズを加えたグラフを，ARIANA と k-SNAP[8] の結果で比較する．k-SNAP は，最初に A-compatible で優位性が最高のグループ集合を求めた後，グループ数が k 個になるまでグループの分割をする手法である．ARIANA のグループ数も k-SNAP のグループ数も 20 個の結果で比較する．グループ数が等しい結果を比較しているため，$\lambda = 0$ で AR index を算出する．

図 3 はノイズの割合による AR index の変化を表しており，ARIANA と k-SNAP は各結果の AR index の値である．Initial grouping は，ARIANA や k-SNAP で解析する前の，

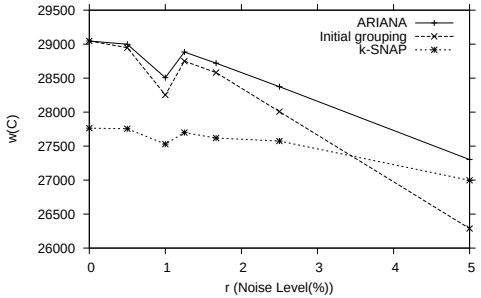

図 3 ノイズの増加に対する AR index の変化
Fig. 3 Changes of AR index with respect to adding noise.

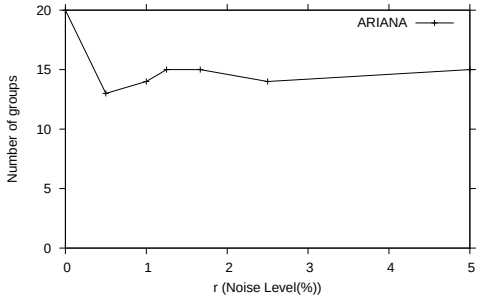

図 4 AR index が最大となる結果の併合段階終了時のグループ数の変化
Fig. 4 Changes of number of aggregated clusters when the maximum AR index is computed.

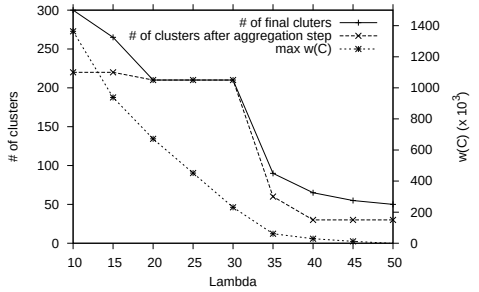

図 5 λ によるグループ数の変化
Fig. 5 Changes of number of groups and $w(\mathcal{C})$ according to λ.

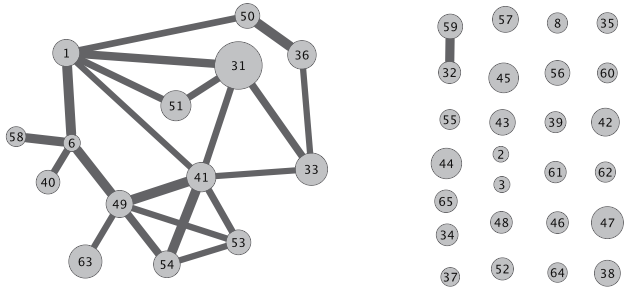

図 6 ARIANA で求めたネットワークモチーフ（$\lambda = 30$）
Fig. 6 Network motifs discovered by ARIANA ($\lambda = 30$)

ノイズを加えたグラフのグループ集合の AR index である．

ノイズが増加するほど，どの結果の AR index も減少している．しかし，どのノイズの割合でも，ARIANA の解析結果の方が k-SNAP よりも大きく，ARIANA の方が良いグループ集合を求められることを示している．ノイズの割合が低い場合，k-SNAP の AR index は Initial grouping よりも小さい．これは，グループの関係性を抽出するためには，k-SNAP のグループ内の属性に対する制限が厳しいことを表している．一方で，ARIANA と Initial grouping の AR index を比較すると，ARIANA の方が大きく，良い結果を求められていることがわかる．ノイズの割合が上がると，k-SNAP の分類が Initial grouping よりも良くなっているが，同じレベルの ARIANA の値の方がより大きい．これらの結果から，ARIANA はデータに含まれる属性と関係性のノイズを考慮して，グループ集合を構成できることがわかる．

図 4 は，ARIANA の解析で AR index が最大となる結果の，併合段階終了時のグループ数である．例えば，$r = 1$ の値が 14 であることから，ARIANA の結果で AR index が最大であるものは，グループ数が 14 個になるまで併合し，その後 20 個まで分割した結果であると分かる．ノイズを含まない $r = 0$ を除いて，ARIANA はグループをある程度併合してから分割を行っている．このことから，各グループの属性と関係性が類似しているグループ集合を構成するために，ARIANA の併合段階も分割段階も効果的であることがわかる．

5.2 タンパク質相互作用ネットワークの解析

次に，生物学的なネットワークを本手法で解析する．解析したネットワークは，複数のデータベースのデータを複合している iRefIndex[13] のタンパク質相互作用ネットワークである．頂点はタンパク質，辺はタンパク質の相互作用を表し，頂点の属性をタンパク質が属する pathway で与える．このネットワークの頂点数は 1,184 個，辺は 9,807 本、属性は 221 種類である．

ARIANA を様々なパラメータで解析し，3 章で定義した $w(\mathcal{C})$ が最大のグループ集合を求める．ARIANA には，併合段階終了時のグループ数 J，結果のグループ数 K，正規化項の λ の三個のパラメータがあるが，ここでは λ の値のみ指定して，J と K を 10 から 300 の間の様々な値で解析し，その結果の中で $w(\mathcal{C})$ が最大のものを求める．

図 5 は，λ の値による，結果のグループ数と $w(\mathcal{C})$ の変化を示している．この図では，λ が小さいほど，グループ数は多くなっており，λ の値でグループ数を調節できることがわかる．k-SNAP ではグループ内の属性の制限が厳しく，属性の種類数より少ないグループ数で集合を構成することはできない．そのため，k-SNAP では，ARIANA が抽出した $\lambda > 20$ の結果を求めることはできない．

図 6 は，$\lambda = 40$ で抽出したネットワークモチーフである．この λ の値では，併合段階終了時のグループ数が 30 個，最終的なグループ数は 65 個である．この図の頂点は一つのグループで与え，辺は隣接性が高いグループ間に与えている．属するタンパク質が 5 個以上のグループのみ頂点として描画しており，$r(C_i, C_j) \geq 0.80$ のグループ間に辺を張っている．辺の太さで $r(C_i, C_j)$ の値を表しており，太い辺ほど値は大きい．結果のグラフでは，辺は 22 本である．グループの大きさは属するタンパク質の個数を表し，大きい頂点ほど多くのタンパク質が属している．頂点に書かれている番号は，そのグループの ID である．

この図には，連結した頂点で構成される大きな要素がある．この集合は，生物学的に基礎的な過程を表している．例えば，cell cycles（細分分裂周期の制御），ribosomes（タンパク質の生成），purine と pyrimidine（エネルギーの生成）である．また，孤立頂点で表され

るグループもあり，これは特定の環境で単独で機能するタンパク質のグループを表している．例えば，DNA replication（DNA 損傷の修復），グリカンの生成である．この結果から，生物の基礎的な過程は強い相互関係があることがわかる．

左上にある頂点 1 は，多くの頂点に対して辺を張るハブである．このグループに属する頂点の多くが，属性 cell cycle を有している．この機能は細胞が機能し始める時間をコントロールすることで知られており，多くの機能と相互作用する．このことから，本手法の結果が，生物学的な知識と一致していることがわかる．頂点 1 と隣接する頂点 51 も，同じく属性が cell cycle のタンパク質が多く属する．しかし，この頂点は頂点 1 と 31 にのみ辺を張っている．この結果から，cell cycle を有するタンパク質でも，多くの機能と相互作用をするグループもあれば，少ないグループもあり，果たす役割の異なる二個以上のグループに分類できることが分かる．

頂点 54 は，purine metabolism と pyrimidine metabolism の両方に関連した頂点の集合である．このどちらも，細胞内の要求に対してエネルギーを生成する，類似した機能である．patyway の分類が細かいため，この二つは解析前には異なるグループに属していたが，本手法では類似した機能として同じグループに分類できる．

これらのことから，本手法の解析で，生物学的に有益なネットワークモチーフを抽出できることがわかる．

6. まとめと今後の課題

タンパク質相互作用ネットワークからネットワークモチーフを抽出することで，タンパク質の機能間の関係を表すグラフを構築した．既存の手法では，ノイズと曖昧さを多く含む生物学的なネットワークの解析をすることは困難であるが，本手法では，ネットワークにノイズと曖昧さがある場合でも，大きなネットワークモチーフを抽出できた．タンパク質相互作用ネットワークを本手法で解析した結果，細胞分裂の周期を調節する機能を持つタンパク質でも，既知の知識に当てはまる多くの機能と相互作用するグループと，少数の機能とだけ相互作用するグループがあることを抽出した．また，異なる pathway に関連していても，エネルギー生成という類似した機能を持つタンパク質を同じグループに分類することができ，果たす役割でタンパク質を分類することができた．

本手法は生物学的なネットワークの解析で有益な結果を得られた．しかし用いた指標は，統計や情報理論の裏付けが無いため，AR index に理論に基づいて改良することが今後の課題である．また，ARIANA では指標が大きい結果を求めることはできるが，最大の結果になるとは限らない．そのため，最も良い分類を求めるようアルゴリズムの改良が必要である．

ネットワークモチーフを発見する問題は，膨大で複雑なネットワークの可視化に応用できる．本論文では，タンパク質相互作用ネットワークを解析したが，遺伝子ネットワークなどの他の生物学的なネットワークにも適応することができ，これらのネットワークを簡潔な形で可視化することで，機能や関係性について新たな見解を得ることに役立つだろう．

参 考 文 献

1) Barabasi, A.L. and Albert, R.: Emergence of scaling in random networks, *Science*, Vol.286, No.5439, pp.509–512 (1999).
2) Barabási, A.L. and Oltvai, Z.N.: Network biology: understanding the cell's functional organization., *Nature Reviews Genetics*, Vol.5, No.2, pp.101–113 (2004).
3) Xiong, H., He, X., Ding, C., Zhang, Y., Kumar, V. and Holbrook, S.R.: Identification of functional modules in protein complexes via hyperclique pattern discovery, *PSB '05*, pp.221–232 (2005).
4) Spirin, V. and Mirny, L.A.: Protein complexes and functional modules in molecular networks, *Proc Natl Acad Sci*, Vol.100, No.21, pp.12123–8 (2003).
5) Bachman, P. and Liu, Y.: Structure discovery in PPI networks using pattern-based network decomposition, *Bioinformatics*, Vol.25, No.14, pp.1814–21 (2009).
6) Chen, J., Hsu, W., Lee, M. and Ng, S.-K.: NeMoFinder: dissecting genome-wide protein-protein interactions with meso-scale network motifs, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006).
7) Alon, U.: Network motifs: theory and experimental approaches, *Nature Reviews Genetics* (2007).
8) Tian, Y., Hankins, R.A. and Patel, J.M.: Efficient aggregation for graph summarization, *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, New York, NY, USA, ACM, pp.567–580 (2008).
9) Huan, J., Wang, W., Prins, J. and Yang, J.: SPIN: mining maximal frequent subgraphs from graph databases, *KDD '04*, pp.581–586 (2004).
10) Inokuchi, A., Washio, T. and Motoda, H.: An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, *PKDD '00* (2000).
11) Kuramochi, M. and Karypis, G.: Frequent Subgraph Discovery, *ICDM*, pp.313–320 (2001).
12) Yan, X. and Han, J.: gSpan: Graph-Based Substructure Pattern Mining, *ICDM '02*, p.721 (2002).
13) Razick, S., Magklaras, G. and Donaldson, I.M.: iRefIndex: a consolidated protein interaction database with provenance, *BMC Bioinformatics*, Vol.9, p.405 (2008).