ROBIN: アイテム集合を付与した大規模ネットワーク解析による副作用の発見

福崎 睦美　(指導教員：瀬々 潤)

1　まえがき

本研究では薬の副作用を解析する手法の一つとしてRelatiOn Between Items and Network(ROBIN) を提案する．信頼性のある副作用候補を発見するためには，遺伝子ネットワークと遺伝子発現情報を統合した解析が重要である．そこでROBINでは，遺伝子ネットワークをあらわすグラフの頂点 (遺伝子) に遺伝子発現量から得た活性条件のアイテム集合を付与することで，2つの情報を統合する．そして，このグラフから共通するアイテム集合を持つ部分グラフ集合の列挙によって，副作用の出る可能性のあるネットワーク群とその条件を同時に発見する．ROBINは以下のような手順で構成される．(1) 共通アイテム集合を有する部分グラフの列挙 (2)(1) で列挙した部分グラフを組み合せ，部分グラフ集合を包含関係が無いように列挙する．本研究では，(1) ではCOPINE[3] を使用し，(2) における効率よい共通アイテム集合を持つ部分グラフ集合の列挙を提案する．

図1(a) は無向グラフであり，(b) はその各頂点のアイテム集合を表す．ROBINでは結果のネットワークの信頼性を高めるため，共通アイテム数が閾値 θ_I，大きさが閾値 θ_S を満たす部分グラフのみ列挙する．図1を，$\theta_I = 2$，$\theta_S = 2$ で解析した一例が図2である．図2において太線で示した二つの部分グラフは $\{i_2, i_3\}$ という条件で副作用を起こす可能性のあるネットワークを表すと言える．

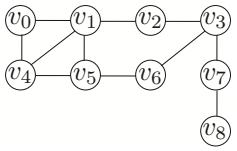
関連研究としては，グラフデータベース解析手法である頻出サブグラフ列挙手法 [1] やデータベース上の遺伝子群の持つアイテム集合からの頻出アイテム集合マイニング [2] も存在するが，遺伝子ネットワークとアイテムの両方を考慮することはできない．COPINE[3] は，二つのデータベースを統合した解析を可能にするが，連結グラフのみを考慮している．

2　問題設定

本研究で使用する定義を以下に示す．G は非連結でラベルや重みのない無向グラフであり，その各頂点はアイテム集合を持つ．このようなグラフ G を itemset-associated graph(IA graph) と定義する．$V(G), E(G), \mathcal{I}(G)$ および $I(v)$ は，それぞれ G の頂点集合，G の辺集合, G に含まれるアイテムの種類，$v \in V(G)$ のアイテム集合を表す．$|E(G)|$ は, G の大きさを表す．また，G の部分グラフの中でも重要なのはより大きなグラフであると言えるので，次のItemset-Sharing Graph (ISS) を，G の興味深い部分グラフであるとし，その共通アイテムとともに定義を示す．

定義 1：(ISS)　G' を IA graph G の部分グラフとする．$I(G')$ を $I(G') = \bigcap_{v \in V(G')} I(v)$ と定義したとき，$I(G')$ をグラフ G' の共通アイテムとし，$I(G') \neq \phi$ かつ，G' に隣接する全てのノード v' について $I(v') \not\supseteq I(G')$ のとき，G' を $I(G')$ についてのISSと定義する．■

このISSを用い，本研究は次の定義の部分グラフを求める．



(a) 遺伝子ネットワークを表すグラフ

v_0	$\{i_1, i_2, i_3\}$
v_1	$\{i_1, i_2\}$
v_2	$\{i_1, i_2\}$
v_3	$\{i_2, i_3, i_4\}$
v_4	$\{i_1, i_2, i_3\}$
v_5	$\{i_1, i_2, i_3, i_4\}$
v_6	$\{i_2, i_4\}$
v_7	$\{i_1, i_2, i_3\}$
v_8	$\{i_2, i_3\}$

(b) アイテム集合

図 1: IA graph の例

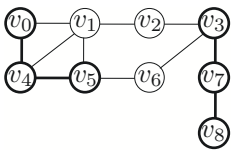


図 2: ISS 集合の例:太枠の部分グラフは共通アイテム $\{i_2, i_3\}$ を持つ

定義 2：(ISS 集合列挙問題) $\mathcal{G} = \{G_1, \ldots, G_n\}$ を ISS の集合とする．ここで G_i は ISS である．θ_S をユーザ定義の値とすると，(1) $V(G_i) \cap V(G_j) = \phi$ (ただし i と j は異なる), (2) $\mathcal{G}$ に隣接する頂点 v に関し $I(v) \not\supseteq I(\mathcal{G})$, (3) $|G_i| \geq \theta_S$，(4) $\mathcal{G}$ 内の ISS 以外に θ_S 以上の大きさの $I(\mathcal{G})$ に関する ISS は存在しない，を全て満たすグラフを ISS 集合と呼ぶ．本研究ではこの中から，θ_F 個以上の部分グラフを持ち，θ_I 以上の大きさのアイテム集合に関連づけられた ISS 集合を列挙する．この問題を，ISS 集合列挙問題と呼ぶ．■

3　提案手法

ROBIN では ISS 列挙問題を解くため，閾値を満たす ISS を COPINE によって列挙し，その中から共通するアイテムをもつ ISS 集合を発見する．このとき ISS の全通りの組み合せを調べる必要がある．その方法の一つとして ISS の深さ優先による組み合せがあるが，ISS 数が多い場合，組み合せを探索する木が深くなる．そこで本手法では，ISS を関連するアイテム集合でグループ化することで，集合の組み合わせを可能にし，高速化する．

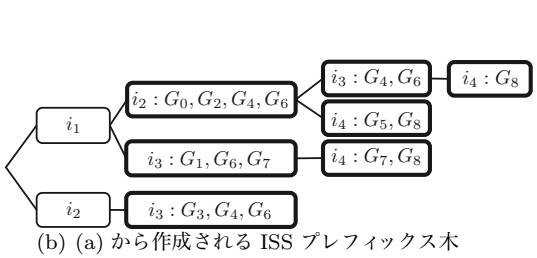
定義 3：(陽な ISS 集合と陰な ISS 集合)　ISS に関連したアイテム集合を $\mathcal{I}$ とすると ISS 集合 $\mathcal{G}$ に関し，$I(\mathcal{G}) \in \mathcal{I}$ の時，$\mathcal{G}$ を陽な ISS 集合，それ以外を陰な ISS 集合と呼ぶ．■

ROBIN では列挙した ISS をアイテム集合に着目してグループ化することで陽な ISS 集合を求める．次に，陽な ISS 集合を組み合わせることで陰な ISS 集合を求める．ISS のグループ化にはアイテム集合のプレフィックス木を利用する．また，アイテムは添字通りの順序を持つ．

定義 4：(ISS プレフィックス木)　T_G を木，n をそのノードとする．この木は次の性質を持つ．n はアイテム i_n と ISS 集合 $\mathcal{G}(n)$ を持つ．$n_i, n_j \in T_G$ のノードと

G_0	$\{i_1, i_2\}$
G_1	$\{i_1, i_3\}$
G_2	$\{i_1, i_2\}$
G_3	$\{i_2, i_3\}$
G_4	$\{i_1, i_2, i_3\}$
G_5	$\{i_1, i_2, i_4\}$
G_6	$\{i_1, i_2, i_3\}$
G_7	$\{i_1, i_3, i_4\}$
G_8	$\{i_1, i_2, i_3, i_4\}$

(a) ISS 集合



(b) (a) から作成される ISS プレフィックス木

図 3: ISS プレフィックス木:太枠のノードが陽な ISS 集合を表す

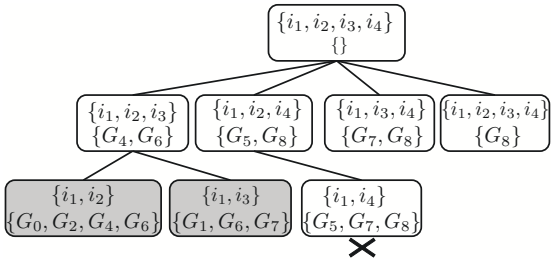


図 4: 図 3 から作成される ISS 木

し，n_j が n_i の子ノードの時，$i_{n_j} < i_{n_i}$ を満たす．ISS 上の根からあるノードまでのパスは，パス上のノードに関連したアイテムの集合からなるアイテム集合を表す．ISS の共有するアイテム集合は全てこの木上の根からのパスで表される．この木を ISS プレフィックス木と呼ぶ．■

これにより，完全一致またはサブセットになるようなアイテム集合に関連した陽な ISS 集合へ高速にアクセスできるようにする．ISS プレフィックス木の例を図 3 に示す．図 3(a) は，ISS とそれに関連するアイテム集合を表す．ただし，$V(G_4) \subset V(G_1)$, $V(G_5) \subset V(G_0)$, $V(G_8) \subset V(G_6)$ という包含関係があるとする．

全 ISS 集合に関連するアイテム集合は，ISS 木によって列挙される．

定義 5：(ISS 木)　T_I を木とし，T_I の各ノード n は，アイテム集合 $I(n)$ と ISS 集合 $\mathcal{G}(n)$ を持つ．$n_i, n_j \in T_I$ について n_j が n_i の子ノードの時，$I(n_j) \subset I(n_i)$ を満たす．この性質を満たす木を ISS 木と呼ぶ．■

定義より，アイテム集合の大きさは単調減少であり枝刈りが可能である．また，ROBIN では新たにノード n' が作成されるとき，$I(n')$ に関連する全ての ISS を $\mathcal{G}(n')$ に加える．図 3 をもとに ISS 木を作成した例を図 4 に示す．各ノードには，アイテム集合と ISS 集合を示した．$\theta_I = 2$ の時，灰色のノードは既出のアイテム集合であり，$\times$ の子ノードはアイテム集合のサイズが閾値 θ_I を満たさないため，枝刈りされる．

4　実行結果

4.1　疑似データによる実験

ROBIN の性能の評価のため，疑似データを作成し実験を行った．特に指定しない場合，頂点数 3000，平均次数 10，アイテムの種類 100，平均アイテム集合サイズ 10，解のアイテム集合サイズ $|I|$=10，解の CCPG サイズ S=7，解の CCPG 集合サイズ 5，解のパターン数 10，閾値 $\theta_I = |I| - 1, \theta_S = S - 1$ とする．実行結果は図 5 に示した．図 5 は，疑似データ上で頂点数を変化させた場合の実行時間である．比較のため，ISS の単純な深さ優先探索による組み合せの実行時間を計測した．頂点数が増加するほど列挙される ISS 数も増加するため，提案手法が有利になることがわかる．

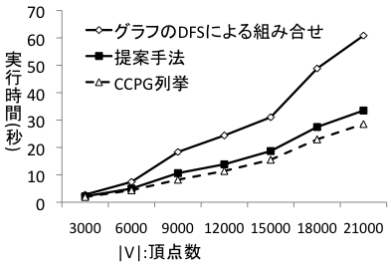
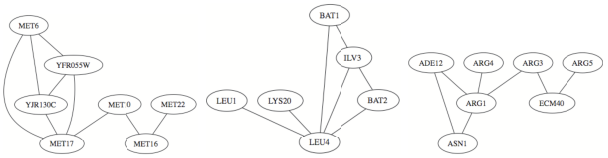


図 5: 疑似データによる実行結果



(a) 硫黄代謝　(b) バリン，ロイシ (c) アラニン，アスパン，イソロイシン合成 ラギン酸代謝

図 6: 酵母の実データ解析結果

4.2　実データによる実験

図 6 は，実データによる実験結果を表す．使用したデータは遺伝子ネットワークのエッジ数が 3,324，各遺伝子の持つ活性化条件 (アイテム) 数の平均が 5.7 となっているデータであり，閾値を $\theta_I = 3, \theta_S = 7$ とした．列挙された 3 つのネットワークは共通に {erg2，yhl029c，ERG11(tet promoter)} の条件において活性化している．また，これらのネットワークの機能を KEGG を用いて調べると，(a)，(b)，(c) の何れもアミノ酸代謝に関連する既知のパスウェイと高い相関があり，生物学的な知識と一致する結果が得られた．

5　あとがき

本論文では，遺伝子ネットワークと遺伝子発現量を統合して解析する手法として，無向グラフの頂点にアイテム集合を付与したグラフから複数の部分グラフ間に共通するアイテム集合の列挙アルゴリズム ROBIN を提案した．今後の課題としては，ISS の列挙時にアイテムの欠損値を許した枝刈りをする方法や，実データの実行結果の生物学的な解釈を進めていきたい．

参考文献

[1] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In ICDM ’02, page 721, 2002.

[2] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. Data Mining and Knowledge Discovery, 14(1), 2007.

[3] M. Seki and J. Sese. Identification of active biological networks and common expression conditions. In BIBE ’08, 2008.