公募研究：2008 ～ 2009 年度

活性化情報に基づくパスウエイ横断ネットワークの推定

●瀬々 潤
お茶の水女子大学大学院人間文化創成科学研究科

＜研究の目的と進め方＞

パスウエイ情報の蓄積が進んでいるが、現在データベースに登録されているネットワークは生体全パスウエイに比べ極一部であると考えられる。パスウエイの理解を深めることは、生体のシステム的挙動の理解及び創薬ターゲットの発見において重要な役割を果たす。この目的のため新規パスウエイの予測が行われてきたが、生体から観測されたデータは揺らぎやノイズが大きいため、成功を見てはいない。

本研究では、既知のパスウエイ間には協調して働く関係にあるパスウエイがあるであろうという既知のパスウエイから見られる仮説から、新規ネットワークの予測を行う。特に (1) 既知のパスウエイやたんぱく質相互作用のネットワークとマイクロアレイ等で得られた遺伝子発現量の情報を照らし合わせることで、どのネットワークがいつ利用されているかのネットワーク活性化情報を抽出し、(2) 得られた活性化情報を基に、接続されるべきだが、現在はネットワークが描かれていないパスウエイを求める。

本研究により、パスウエイや相互作用ネットワークのどの部分がいつ働いているかの注釈付けが行え、現在のパスウエイ情報を超えて、副作用が起きているネットワークの推定が期待できる。

＜研究開始時の研究計画＞

2008 年度の計画は主に、既知のパスウエイやたんぱく質相互作用ネットワーク情報と、遺伝子発現量の情報を照らし合わせることで、どのネットワーク（サブネットワーク）が、どのような環境下で活性化しているかの注釈付けを行う。特に、次の 3 点を行う (1) ネットワークの注釈付けに必要なアルゴリズムの構築、(2) 発現量データの収集と選別、(3) 遺伝子発現やネットワークは、ほ乳類などの多細胞で高等な生物において、複雑な調整が成されていると推定できるため、本研究では酵母などの単細胞生物から研究をスタートし、より高等な生物への適用，さらに高等生物へ応用する際の問題点を発見するである．

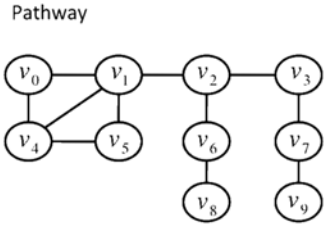
2009 年度は，2008 年度に単細胞生物に適用する事で明らかになる問題点を解決し，更に酵母に比べて 5 倍程度の遺伝子数を持ち，たんぱく質相互作用情報の全てを実験的に網羅できていない高等生物のデータベースに対しても本アルゴリズムを適用する事で，高等生物特有の問題点を解決する．

＜研究期間の成果＞

2008 年度はパスウエイ横断ネットワークを見つける要素技術となる活性化パスウエイ発見技術の開発を行い，発表した (Seki and Sese, 2008)．本手法の概要を図 1 に示す．本研究ではパスウエイもしくはたんぱく質相互作用 (PPI) と複数の環境下で観測された遺伝子発現情報を統合する．図 1(A) に本研究で利用するデータ例を示す．Pathway はグラフ構造で表され，ノードは遺伝子もしくはタンパク質を，辺はタンパク質間の相互作用を表している (ここでは簡単のため無向グラフとしてある)．図では v0 から v9 まで 10 個の遺伝子とその相互作用を示している．たとえば，v0 は v1 と v4 の 2 つのタンパク質と相互作用することを示している．遺伝子発現情報は，高発現環境の集合で表す．図では i1 から i4 まで 4 種類の環境における遺伝子発現を示しており，v0 は環境 i1，i3 で高発現することを示している．

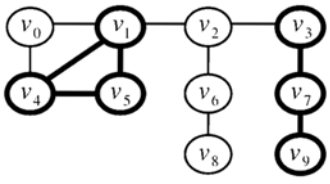
本研究で導入した活性化パスウエイ抽出法を図 1(A) に適用したものを，図 1(B) で示す．活性化パスウエイは相互作用グラフの上で連結しており，かつ，複数の共通の条件下で高発現を示している部分グラフの事である．図 1(A) において v1,v4,v5 から成る部分グラフに着目すると，この 3 つの遺伝子は全て i1 及び i2 で活性化をしている．本解析より，v1,v4,v5 の部分グラフは条件 i1 及び i2 において活性化するネットワークである可能性が高い事が示唆される．なぜなら，発現が変化した条件が遺伝子毎にランダムに現れると仮定した場合、ネットワークで連結している 2 遺伝子で発現変化条件を共有している確率は低い。更に、より大きなネットワーク中全ての遺伝子で条件が共有されている確率は低くなる。よって、一定以上の大きく、条件を共有するネットワークを抽出することで、生物学的に活性化している可能性の高いネットワークの抽出が可能となる。更に，本解析により i1 及び i2 の条件は v1,v4,v5 の働きが共通であることから，近しい条件であることが示唆される．本手法は COmmon Patteern Itemset NEtowork (COPINE) と名付けた．

本部分グラフ列挙は計算機的に容易ではない．部分グラフの数はグラフのサイズが大きくなるに従って，あるいは，条件の数が増えるに従って，指数級数的に考えるべき組み合わせが大きくなる．この増加を防ぎ, 現実的な時間で最適な回答を得るために, (1) 条件集合の単調減少性を利用した枝刈り (2) 同一条件を持つ重複



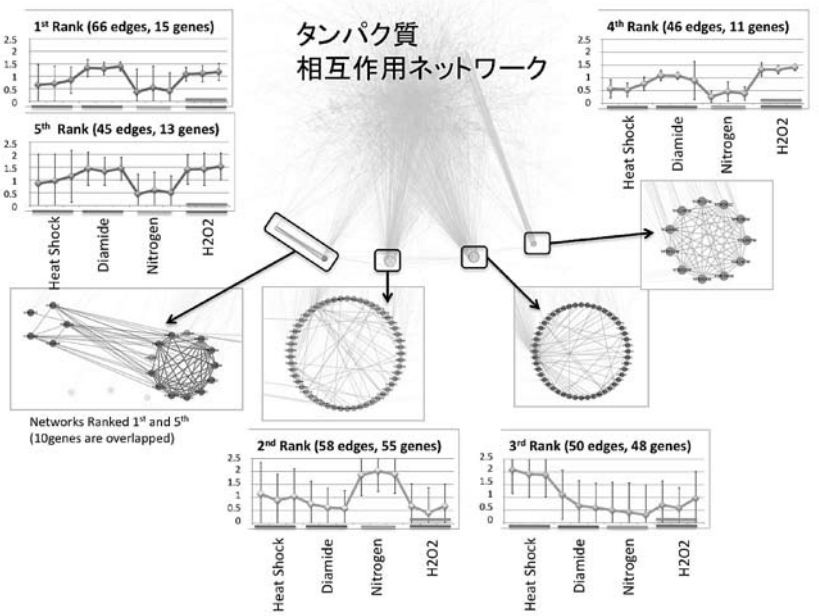
Gene	Expressed conditions
v_0	$\{i_1,i_3\}$
v_1	$\{i_1,i_2,i_4\}$
v_2	$\{i_2,i_3,i_4\}$
v_3	$\{i_1,i_2\}$
v_4	$\{i_1,i_2,i_4\}$

Gene	Expressed Conditions
v_5	$\{i_1,i_2,i_3\}$
v_6	$\{i_2,i_4\}$
v_7	$\{i_1,i_2,i_3\}$
v_8	$\{i_1,i_2\}$
v_9	$\{i_1,i_2,i_4\}$



(A) Example of IA Graph (Pathway and Conditions of Each Gene)　(B) Cooperative Pathways with $\{i_1,i_2\}$

図 1．活性化したネットワーク発見の模式図．及び，協調パスウエイの模式図．



した部分ネットワーク探索を避ける枝刈り，を導入し全探索に比べ 100 倍以上の高速化を行った．

本手法の妥当性を調べる為，酵母の PPI ネットワーク (Ito et al. 2000, Uetz et al. 2000, Nevan et al. 2006) 及び 173 の様々なストレス環境下における遺伝子発現量情報 (Gasch et al, 2000) を組み合わせることで，ストレス環境における酵母の活性化パスウエイを抽出した．遺伝子発現量は，野生酵母の 1.5 倍の遺伝子発現を示すものを高発現であると見なした．この結果を図 2 に示す。図 2 には、大きなネットワーク、つまり、より同時に活性化している可能性の高いネットワーク上位 5 つを示した。特筆すべき点は 3 点有る。1 点目は環境特異的に活性化しているネットワークが自動的に発見できている点である。図の中で (A) はジアミド条件下で、(B), (E) は過酸化水素過剰な条件下で、(C) は窒素枯渇の条件下で、(D) はヒートショック条件下で活性化しているネットワークであり、膨大な相互作用ネットワーク、多様な条件から、条件特異的に働いているネットワークとその条件が自動的に抽出できることが示唆される。第 2 点目は既知の情報と強く関連したネットワークが抽出されていることである。(A), (B), (E) はプロテアソーム複合体の一部であり、プロテアソーム複合体は 20S と 19S の 2 つの大きなサブユニットを持つが、(A), (B) は 20S の一部分、(E) は 19S の一部分である。(A) と (B) は共通する遺伝子が多いが、(A) に含まれ (B) に含まれない遺伝子には、環境が変化すると局在箇所が変化するタンパク質が含まれており、本手法により環境によるタンパク質の特性の変化を調査できる可能性が示唆されている。(C), (D) はそれぞれオートファジー関連、サイトスケルトンの立体構造変化に関連した遺伝子群が多く含まれている。第 3 点目の特筆すべき点は、疎な結合のネットワークの機能を明らかに出来る可能性である。今までグラフを用いた注釈には主に密な結合を持つ部分に特化しているものが多かった。一方、パスウエイは結合が疎であっても重要な情報伝達を担う遺伝子が存在しており、このようなネットワークは困難であった。我々のアルゴリズムで抽出されているネットワークには、(A), (B), (E) の様に密なネットワークも含まれるが、(D) の様にハブを有しているネットワークも抽出可能となっている。更に、一般にハブのネットワークは機能的に重要であることが叫ばれつつ、どのような環境で、どのような機能を果たしているか細かい機能が議論されることが少なかった。本実験結果により、ハブの遺伝子がどのような環境で、どの遺伝子群と協調しているかが明らかになり、遺伝子機能の理解が進む物と考えられる。

更に本手法をヒト等高等生物に適用する場合には，アルゴリズムのスケーラビリティーが問題となる．特に、ネットワークを探索する際、ネットワークの大きさが大きくなるに従い爆発的に探索すべきネットワーク数が大きくなる。近年では、KEGG や Reactome.org で公開されているパスウエイ、相互作用情報は 50 万以上の相互作用を有しており、これらの解析には上記で示した条件共有ネットワークを効率よく、高速に探索できる必要がある。我々の提案手法は、このようなネットワークの探索を高速に行う手法を含んでおり、100 万以上の相互作用、100 以上の実験条件がある環境下でも、解を得られるアルゴリズムである事を確認した．

次に，発見した活性化パスウエイを組み合わせる事でパスウエイ横断ネットワークの発見を行うアルゴリズムを開発した．図 1 の例に戻ると，図 1(B) の部分グラフ v1, v4, v5 及び v3,v7,v9 は当初提案しているパスウエイ横断ネットワークを示している．なぜなら，2 つのネットワークは同一環境下で働き，この間に未知のつながりが予想されるからである．このように（複数の）同一環境下で活性化する部分グラフが複数ある場合，協調パスウエイと名付けた．

この協調パスウエイを効率よく列挙するため我々は新たなアルゴリズム CoopeRativE Pathway Enumerator (CREPE) を開発した (Fukuzaki et al. 2009)．図 1(A) から CREPE で列挙できるグラフは図 1(B) の太線で示された 2 つの部分グラフの集合とその部分グラフ集合が関連した条件集合 i1, i3 となり，条件 i1,i3 においてこのグラフ上は離れた 2 つのネットワークは協調して働く可能性があることを示唆できる．また，協調パスウエイの列挙は，活性化パスウエイの組み合わせをすべて調べる組み合わせ問題を解く必要がある．この組み合わせ列挙を高速化に行うため，我々はグラフの組み合わせを求める代わりに，存在するアイテム（条件）集合の接頭辞木を作成した．

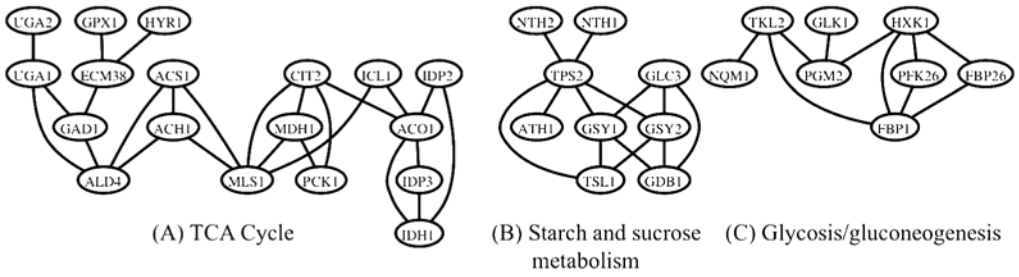
表 1 に図 1 に示したデータと同一のデータを用いて協調パスウエイを計算した結果を示す．YO1 から YO6 が高発現環境から，YS1 と YS2 が低発現環境から抽出した協調パスウエイである．更に図 3 は YO1 についてグラフを図示した物，図 4 は YO1 内の遺伝子について発現量を図示した物である．これらの図より CREPE により協調パスウエイが抽出できることがわかる．

表 1 の GO の列は各協調パスウエイ内の遺伝子が，どの遺伝子オントロジーで定義される機能に関連しているかを示したものである．各協調パスウエイに示した結果から本手法が既存の遺伝子機能間での関連が深いにもかかわらず，パスウエイ上は異なった位置に属している隠れた遺伝子間の関係を，正しく発見できることを示唆している．

更に，我々はマウスの ES 細胞とその分化後の遺伝子発現に関しても協調パスウエイ発見を行った．その結果を表 2 に示す．酵母の例に比べると，各遺伝子と既存の知識との対応が弱いが，これは遺伝子に未知の機能が多数含まれるためと思われる．この内容より，我々の手法が高等生物に於いても，協調パスウエイを発見する事が可能であり，今後隠れたパスウエイ間の接続を見つけ出せることを示唆している．

表 2. 酵母ストレス環境下からの協調パスウエイ発見

ID	# of subgraphs	# of edges	# of genes	common pattern conditions	GO (Biological process)	p-value
YO1	3	48	35	Stationary phase 6h (25 degC), 12h (25 degC), 2d (25 degC)	Disaccharide metabolic process	3.2e-10
YO2	3	44	31	Stationary phase 10h (30 degC), 12h (30 degC), 3d (30 degC)	Cellular carbohydrate biosynthetic process	6.8e-10
YO3	3	43	32	Heat shock 17 to 37 degC, Stationary phase 10h (30 degC), 12h (30 degC)	Acetyl-CoA metabolic process	6.7e-16
YO4	3	42	31	Stationary phase 5d (30 degC), 5d (25 degC), 7d (25 degC)	Amino catabolic process	1.3e-9
YO5	2	55	39	Heat shock 17 to 37 degC, 21 to 37 degC, 25 to 37 degC	Disaccharide metabolic process	1.1e-15
YO6	2	48	26	Nitrogen depletion 1h, 2h, 4h	Glutamine family amino acid metabolic process	1.1e-15
YS1	3	88	57	Stationary phase 5d (30 degC), 13d (25 degC), 22d (25 degC)	Glycolysis	6.1e-18
YS2	2	21	18	Stationary phase 6h (30 degC), 2d (30 degC), 3d (30 degC)	Ribonucleotide metabolic process	2.0e-12



(A) TCA Cycle (B) Starch and sucrose metabolism (C) Glycosis/gluconeogenesis

図 3. 表 1 中の協調パスウエイ Y01 を図示したもの

図 4. 表 1・ 協調パスウエイ Y01 内の遺伝子の発現量を示した図

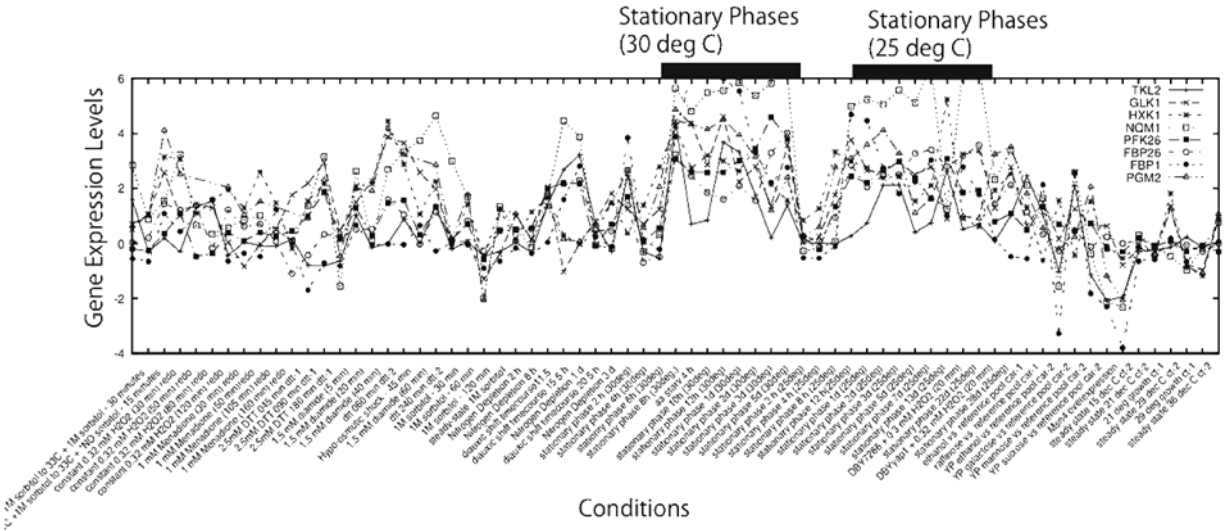


表 2. マウス ES 細胞から取得した発現量より求められる協調パスウエイ

ID	# of subgraphs	# of edges	# of genes	common pattern conditions	GO (Biological process)	p-value
MO1	8	106	71	mp53/Ras Rep5, 6, 7, 8, 9	Regulation of insulin receptor signaling pathway	3.5e-4
MO2	7	118	82	mp53/Ras Rep1, 5, 6, 7, 10	Regulation of insulin receptor signaling pathway	5.3e-4
MO3	7	104	71	mp53/Ras Rep1, 3, 5, 7, 8	Negative regulation of epithelial cell differentiation	1.4e-5
MS1	6	232	115	mp53/Ras Rep1, 2, 6, 9, 10	Positive regulation of cyclase activity	4.3e-4
MS2	5	195	92	mp53/Ras Rep1, 4, 5, 7, 10	Neuron fate commitment	2.9e-6
MS3	2	146	67	mp53/Ras Rep1, 4, 5, 8, 10	Extracellular matrix organization	2.7e-12

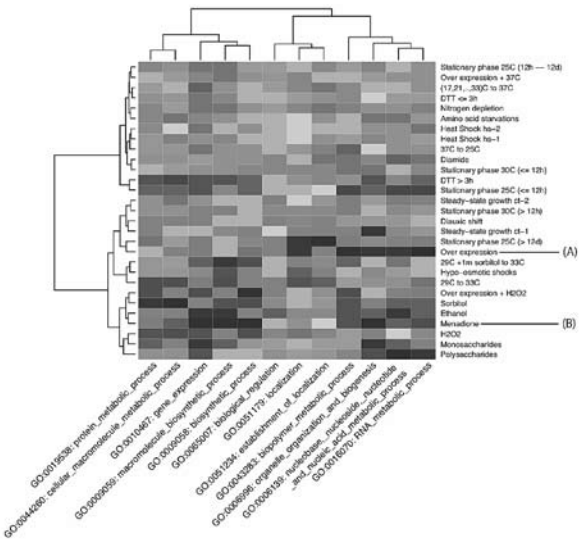


図 5. ストレスに関連して変化する発現とその機能の関係

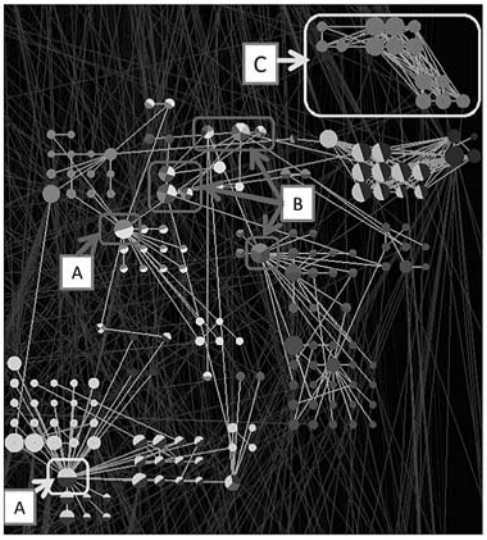


図 6. 図 2 の結果の可視化の別例

また、上記アルゴリズムは既知のネットワークと実験条件の組み合わせを詳細に把握する事に適しているが、パスウエイを横断するネットワークの検出には、既知のネットワークでは捕らえられていない、より大域的なネットワークの検出も必要である。しかし、遺伝子発現量を含む生物学データは観測ノイズを多く含むだけでなく、全遺伝子が一斉に変化するのではなく、一部の遺伝子の状態が変化することで適用する事が多い。よって、弱い相関を発見する、逆に言うと無相関であることを発見する技術が必要であった。この問題に対し、密度比を利用することで、事前に分布を与える必要が無く、モデル選択可能な相互情報量推定手法を開発した。また、本手法を上記同様ストレス環境下における酵母発現量情報に適用することで、ストレス刺激に対しての酵母の大域的な応答を把握する事が可能となったことを確認した。図 2 は、その相関を表した物で、DNA や RNA にダメージを起こすストレスは、細胞活動全体に刺激を与え、ビタミン K による刺激はタンパク質の局在にのみ影響を与えることが確認できた。

最後に，本研究で発見できる活性化パスウエイ，及び，協調パスウエイには，既存のネットワーク解析研究で仮定されている密なグラフの発見を仮定していない．このため，ネットワークの図示に既存手法を適用すると，見づらい図ができあがってしまうため，可視化の手法も検討を行った (Itoh et al. 2009)．その例を図 6 に示す．本手法の特徴である疎なネットワーク，ハブ，密なグラフの混在をそれぞれ正しく可視化することに成功した．

＜国内外での成果の位置づけ＞

本研究はデータマイニング分野において研究されている頻出グラフ列挙、グラフクラスタリング、制約付きクラスタリングの分野を跨ぐ研究成果である。頻出部分グラフでは、多数のグラフを含むデータベースから共通のグラフを発見するが、遺伝子の発現条件を扱うことができない。グラフクラスタリングでは密なネットワークを見つけることが出来るが、パスウエイの様に疎な結合でも重要なネットワークを把握する事ができない。制約付きクラスタリングでも、密なネットワークを要求するアルゴリズムが多いこと、また、一般に全ての実験条件で協調していないとクラスタが構成できない。よって、我々の手法は、他の手法では発見できず、かつ、生物学的な解釈が行いやすい結果を得られる点で優れている。

また，本研究の論文は採択率 18.9% の難関を突破して口頭発表となる程，注目を浴びている．

＜達成できなかったこと、予想外の困難、その理由＞

本結果に至るまで，Gene Expression Omnibus において公開されている様々なデータに対し，本手法を適用してきた．しかしながら，手法の制約のみだけでなく，遺伝子発現データによっては実験が失敗しているのかノイズの大きなデータも散見され，解析に苦労した．また，GEO に掲載されている注釈情報が正しくなかったり，情報が欠けていたりするために，解析結果の解釈に困難を要する例もあり，公共データベースの整備の重要性を改めて感じた．

＜今後の課題、展望＞

現在までの解析は，計算機上の解析にとどまっているため，本結果が実験上正しい物であるかを検証していく必要があり，今後の研究に於いて計算機手法の改良だけでなく，実際の実験へのフィードバックを行えるようにしていきたい．

＜研究期間の全成果公表リスト＞

1. 0901120751
Mio Seki and Jun Sese. Identification of active biological networks and common expression conditions. 8th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2008), Athens, Greece, Oct. 8-10, 2008

2. 0901120757
Taiji Suzuki, Masashi Sugiyama, Takafumi Kanamori, and Jun Sese.. Mutual information estimation reveals global associations between stimuli and biological process. BMC Bioinformatics 2009, 10 (Suppl 1):S52.

3. 0912032245
Takayuki Itoh, Chris Muelder, Kwan-Liu Ma, and Jun Sese. A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs. *IEEE Pacific Visualization Symposium 2009, pp.121–128*, Beijing, China, Apr. 20-23, 2009

4. 0912032241
Mutsumi Fukuzaki, Mio Seki, Hisashi Kashima, and Jun Sese. Side effect prediction using cooperative pathways. *IEEE International Conference on Bioinformatics and Biomedicine 2009 (IEEE BIBM 2009)*, pp.142–147, Washington D.C., USA, Nov. 1-4, 2009.