Vol.2014-MPS-98 No.4
Vol.2014-BIO-38 No.4
2014/6/25

無限次数多重検定法へのグラフ制約の導入と ゲノムワイド関連解析への応用

齊藤有紀[†1]　寺田愛花[†2 †3]　瀬々潤[†2]

概要：ゲノムワイドなデータが容易に観測できるようになり，遺伝子毎，あるいは，1塩基多型（SNPs）毎に検定を行う解析も頻繁に行われている．この解析の中で問題になるのが，検定結果の偽陽性であり，これを抑えるためにBonferroni補正をはじめとした多重検定補正法が利用されている．ところが，単一の遺伝子やSNPsならまだしも，複数の組合せを考えると近似が甘くなり，補正後に有意な結果が現れなくなる事が問題となっていた．この問題に対処する方法として，寺田らは無限次数多重検定法（LAMP）を導入した．

ところが，LAMPには，特にゲノムワイド関連解析（GWAS）の様な超大規模なデータへの適用を行おうとすると，実行が終わらない問題点が存在し，応用の幅に限りがあった．そこで本研究では，現実的な時間で超大規模データに対してもLAMPが実行できるよう，取れる組合せに制約をいれる手法の提案を行う．特に遺伝子やタンパク質の間の相互作用情報を元にしたグラフ構造の制約をもうけ，関連性が予想されるSNPの組み合わせのみに注目して解析を行う．これにより探索空間の削減と，偽陽性の発生を抑制することが可能となり，シロイヌナズナのGWASデータの解析が可能となった．

1. はじめに

解析機器の性能向上により，一塩基多型（SNPs）をゲノムワイドに観測したデータが容易に得られるようになり，着目した表現型に影響を及ぼすSNPを発見するゲノムワイド関連解析（GWAS）が盛んに行われている[1]．このとき，遺伝子毎，あるいはSNP毎に検定を行うことで，表現型に有意に関わるSNPの網羅的な発見が期待できるが，SNPがM個ある場合，それぞれに対して検定を行うため，M回の検定が必要である．この複数の検定によって問題となるのが検定結果の偽陽性である．例えば，有意水準0.05の検定を100回行う場合，結果の中に1個以上偽陽性が生じる確率は$1-(1-0.05)^{100}\approx 0.994$である．つまり，99.4%以上の確率で，結果の中に偽陽性が生じる．このように，検定回数が増加するほど偽陽性が起こる可能性が高まり，ゲノムワイドなデータのような大規模データでは偽陽性の発生が避けられない．これを抑えるために，このM回の検定で偽陽性が1回以上生じる確率（Family-Wise Error Rate; FWER）が頻繁に制御され，Bonferroni補正をはじめとした，多重検定補正が利用される[2]．

遺伝子型と表現型の関係は，より複雑であり，単体のSNPでは影響は無くとも，複数のSNPsがあることで初めて表現型に影響を及ぼすものが知られている[3]．このような組み合わせの影響を発見しようとすると，検定数はSNPの数Mに対して2^M-1回であり，指数関数的に増加する．この結果，少数のSNPの調査であっても，組合せの効果を網羅的に調査すると，膨大な数の検定が必要となる．この組合せ爆発により，全ての組合せを網羅する事は難しい．また，例え網羅的に調査できたとしても，よく使われているBonferroni補正などの多重検定補正法では，偽陽性の生起確率の近似が緩くなってしまい，補正後に有意な結果が現れないという問題があった．

この問題に対し，近年，寺田らはBonferroni補正の過剰な偽陽性の見積もりをより厳密に補正し，効率的な枝刈り法を導入することで網羅的な組合せの調査ができる，無限次数多重検定法（Limitless-Arity Multiple testing Procedure; LAMP）を提案した[4]．しかし，LAMPが解析できるのは高々数百個程度の因子がもたらす組合せの効果であり，数万〜数百万のSNPsを調査するGWASの様な超大規模なデータに対しては，このLAMPは現実的な時間で解を返すことができない，あるいは，解答できても補正項が非常に大きくなり，有意なSNPsの組み合わせの検出が期待できないという問題があった．特に後者に関しては，LAMPの内部でBonferroni補正と同様，全てのSNPsが独立で起こることを仮定して補正していることが大きい．ゲノム上の近くのSNPsは，互いに相関が高いなど，SNPs間は必ずしも独立ではない．

そこで，本研究では互いに従属関係がある遺伝子同士を組み合わせの考慮に入れるという制約条件を考慮した上でLAMPを行う枠組みを作成する．特に，グラフ構造で与えられる，生物学的に既知な相互作用の情報を利用し，考慮する組み合わせに制約を入れる．その例を図1に示す．LAMPでは，図1(A)のように，すべての組み合わせを考慮していたが，細胞内で相互作用のないSNPsの組み合わせは表現型に影響を与えているとは考えにくい．そのため，本研究では，図1(B)のように，取れるSNPsの組み合わせにグラフ構造の制約を入れ，部分グラフで表せる組み合わせのみを考慮する．

また，この様なSNPsの組合せを効率的に列挙するため，グラフ構造も考慮して頻出パターン列挙をするアルゴリズ

†1 東京工業大学大学院情報理工学研究科
Department of Computer Science, Tokyo Institute of Technology
†2 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
†3 日本学術振興会特別研究員
Research Fellow of the Japan Society for the Promotion of Science

Vol.2014-MPS-98 No.4
Vol.2014-BIO-38 No.4
2014/6/25

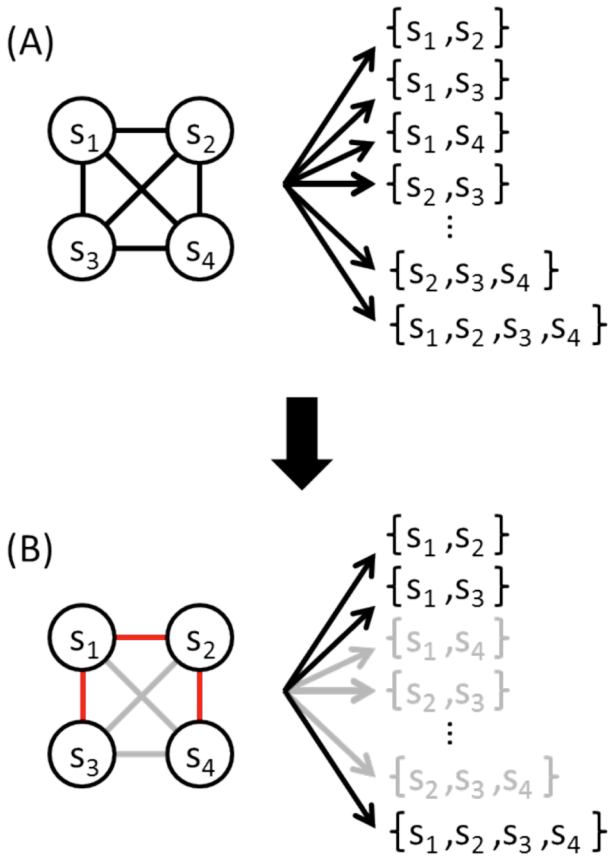

図 1　生物学的に既知な相互作用を考慮した組み合わせの例. (A) LAMP で考える組合せの効果. 各頂点は SNP を表す. LAMP では全ての組み合わせを考えるため，完全グラフの全部分グラフを考えることに相当する．(B) 本研究で扱う組合せの効果．各頂点は SNP を，赤い辺は，生物学的に既知な相互作用を表している．提案法では，与えられたグラフの部分グラフのみを考える．

ム [5]を利用することで，表現型に有意に関わる SNP の組合せを列挙する手法を提案する．さらに，シロイヌナズナのデータを使用し，遺伝子やタンパク質の間での相互作用情報を用いてグラフ構造の制約を設けることで，関連性が予測される SNP の組み合わせのみに注目して解析を行う．

2. 関連研究

関連研究としては，Gungor *et al.* [6),7)]によって行われた，GWAS 情報とタンパク質間相互作用（PPI）ネットワークから，疾病等に関連がありそうなパスウェイを探すツールが発表されている．しかし，このツールは機能既知の SNPs を用いてパスウェイの機能を予測するものであり，SNPs の機能予測を行うことや，組み合わせによる表現型への影響に関する予測などは行うことができない．

表 1　分割表

		表現型 1	表現型 0	
SNPs 集合	有	a	$\lambda - a$	λ
	無	$n - a$	$N + a - n - \lambda$	$N - \lambda$
		n	$N - n$	N

また，Fang *et al.*[8)]は，SNP 単体についてスクリーニングを行った後，その SNP に他の SNP を組み合わせたものの関連を調べる手法を提案した．しかしこの手法では，単体では発現しない SNP を最初のスクリーニングの段階で落としてしまう可能性がある．

さらに，Zhang *et al.*[9)]によって，Westfall-Young 法を高速化して 2 個の SNPs の相乗効果を網羅的に調べる手法が提案されたり，Croiseau *et al.*[10)]によって，回帰を用いた手法が提案されたりしたが，いずれも少数の組み合わせに対してのみ適用できるものであった．

3. 準備

3.1 カイ二乗検定

この章では， SNP の組合せが与えられたとき，表現型との関係を表す統計的な計量方法を定義する．

本研究では，SNP 集合Sが与えられたとき，N個の個体を表現型と SNP 集合の有無で分類する．これを，表 1 に示す．SNP 集合有りとは，S に含まれる全ての SNPs を持つ個体である．a は，Sを持つ個体集合のうち，表現型が 1 の個体数を示している．

このとき，この SNP 集合Sの有無と表現型の間に相関があるかどうかを検定する．ここでは，GWAS でよく利用されるカイ二乗検定を利用するが，本稿で提案する手法は，Fisher の正確確率検定なども利用可能である． 表 1 の分割表に対して，カイ二乗検定で用いるカイ二乗値は次式で計算する．

$$\chi^2(\lambda, a) = \frac{N\big(a(N + a - n - \lambda) - (\lambda - a)(n - a)\big)^2}{n(N - n)\lambda(N - \lambda)}$$

こうして得たカイ二乗値が，有意水準を下回るかどうかで関連性の有無を計るのがカイ二乗検定である．この時のカイ二乗値の有意水準は，3.841 なので，$\chi^2(\lambda, a) > 3.841$であれば有意とみなす．カイ二乗検定の P 値は，カイ二乗値を用いて，次式で表される．

$$P(\lambda, a) = \int_{\chi^2(\lambda,a)}^{\infty} \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}\, dx$$

この P 値は，カイ二乗分布の分布関数を，現在のカイ二乗値から正の無限大まで積分した形で与えられるため，カイ二乗値と P 値は逆相関にある．

3.2 LAMP

この節では，FWER の上限を α に抑えつつ，有意な組み合わせを列挙する無限次数多重検定法（LAMP）について述べる．

転写因子や SNP などの因子の組み合わせの検定を考えた場合，2 つの問題が存在する．第一の問題は膨大な計算時間である．100 個の因子が構成する全通りの組み合わせを網羅すると，10^{30}通りを考えねばならず，因子数に対して検定数が指数関数的に爆発する．第二の問題は，過剰な多重検定補正である．Bonferroni 補正を考えた場合，検定数をm，有意水準をαとすると補正後の有意水準δとしてα/mを利用する．この補正は，全検定が均等にδの偽陽性を産むと考えた場合を想定しており，現実には過剰に偽陽性を見積もっている場合が多い．

LAMP では以上の問題点を解消するために，前者の問題に対しては頻出パターン列挙の手法 [11]を，後者の問題に対しては，偽陽性を生まない検定を Bonferroni 補正の補正項から除くことで偽陽性の過剰な見積もりを正した Tarone 法 [12]を利用している．偽陽性を生まない検定とは，ある SNPs 集合の検定で，周辺分布が与えられた時，考えられる P 値の中で最も小さな値が補正後の有意水準δよりも必ず大きい検定である．このような検定は，有意な SNPs 集合として検出されないため，偽陽性も起こらない．よって，補正の際の項目から除去でき，Tarone 法では，偽陽性を生まない検定の数がm_λのとき，補正後の有意水準をα/m_λとする．カイ二乗検定の場合，カイ二乗値を利用するので，カイ二乗値の上限を見積もればよい．

ある SNP 集合Sを有する個体群Iに着目する．Iの個体数をλ，全個体の中で表現型が 1 である個体数をnとする．この時，I中で表現型が 1 である個体数に寄らず，Sのカイ二乗値は, 以下の上限が計算できる. 最も大きくなるのは, Iが全て表現型 1 を持つときであり，そのとき，$a=\lambda$となるので，カイ二乗値の上限は次式で定義できる．

$$\chi^2_{upper}(\lambda)=\frac{N\lambda^2(N-n)^2}{n(N-n)\lambda(N-\lambda)}=\frac{N\lambda(N-n)}{n(N-\lambda)}$$

このカイ二乗値の上限は，λの値が増加するにつれて，増加する．

$$\begin{aligned}&\chi^2_{upper}(\lambda+1)-\chi^2_{upper}(\lambda)\\&=\frac{N(\lambda+1)(N-n)^2}{n(N-n)\big(N-(\lambda+1)\big)}-\frac{N\lambda(N-n)^2}{n(N-n)(N-\lambda)}\\&=\frac{N(\lambda+1)(N-n)(N-\lambda)-N\lambda(N-n)\big(N-(\lambda+1)\big)}{n\big(N-(\lambda+1)\big)(N-\lambda)}\\&=\frac{N(N-\lambda)\{(\lambda+1)(N-\lambda)-\lambda\big(N-(\lambda+1)\big)\}}{n\big(N-(\lambda+1)\big)(N-\lambda)}\\&=\frac{N^2(N-n)}{n\big(N-(\lambda+1)\big)(N-\lambda)}\geq 0\end{aligned}$$

先述のように，カイ二乗値と P 値は逆相関にあるため，P 値の下限は，カイ二乗値の上限を用いて，以下のように表される．

$$f(\lambda)=\int_{\chi^2_{upper}(\lambda)}^{\infty}\frac{1}{\sqrt{2\pi}}x^{-1/2}e^{-x/2}\,dx$$

LAMP は，λと P 値の下限が逆相関の関係にある統計量であれば利用することができる．上記の証明より，λの増加に従ってカイ二乗値の上限が大きくなり，P 値の下限は小さくなるので，カイ二乗値も LAMP の枠組みで利用可能であることがわかる．

事前に補正後の有意水準δが分かっていれば，上記の計算が可能であるが，実際にはδは偽陽性を生まない検定の数に依存するので，これらの平衡点を見つけなければならない．LAMP では各検定での適切な補正後の有意水準を，出現頻度の閾値λを変えて探索する．

出現頻度の閾値をλ，出現頻度がλ以上の組み合わせの数をm_λ，出現頻度がλの時の P 値の下限を$f(\lambda)$とする．

- λを上限に設定する
- 頻出パターン列挙を利用しm_λを計算する
- $f(\lambda)$を計算する
  - $m_\lambda f(\lambda)\leq\alpha$ならば$\lambda=\lambda-1$として再計算
  - $m_\lambda f(\lambda)>\alpha$ならば$\lambda=\lambda+1$として終了

各検定の補正後の有意水準は$\delta=\alpha/m_\lambda$とし，P 値≦ δ 以下の組み合わせを，列挙したm_λ個の組合せから求める．これにより，Bonferroni 補正の過剰な補正を抑え，検出力を向上した．

しかし，ゲノムワイドなデータのように非常に大規模なデータの場合，LAMP では，頻出パターン列挙の際に，λが小さいと組み合わせの数が非常に大きくなってしまうため，現実的な実行時間では実行が終了しない．そこで本研究では，グラフ構造を用いて取れるアイテム間に制約を入れることで，この問題点を解消する．グラフ構造として，タンパク質間相互作用（Protein-Protein Interaction: PPI）情報を用いる．

3.3 アイテムセット付き部分グラフの列挙

LAMP における SNPs 集合の列挙は，各 SNP を頂点とする完全グラフから，部分グラフを全列挙することに対応する．本稿では，SNPs 間の関係を表すグラフとして，完全グラフ以外のグラフ構造も許し，その部分グラフで表される全 SNPs 集合を考えることで，SNPs 間の関連性に制約を入れる．

１つの案として，与えられたグラフの部分グラフを全て列挙し，その上で LAMP 同様の多重検定補正を行う方策がある．しかし，この方法は冗長である．その理由は，あるグラフGに対し，その部分グラフG'を考えた時，GとG'が同一の SNPs 集合に関連付いている可能性があり，その場合はGとG'が従属関係にあるので，Gのみを補正項に考慮すればよいためである．

部分グラフを全列挙の後，部分グラフ間の関係を調べることも可能ではあるが，部分グラフの判定は必ずしも容易ではなく，時間を要する．代わりに，部分グラフを列挙する際に，部分グラフの判定及び各 SNP を保有する個体の集合（アイテム集合）の調査も行い，検定に必要な部分グラフを抽出する策を考える．

このような方法として COIN[5]が提案されている．COIN では，アイテム集合共有グラフを導入し，その列挙を行っている．アイテム集合共有グラフとは，グラフにおいて頂点にアイテム集合が関連付けられているグラフである．また，その部分グラフは，グラフ上の部分グラフであると同時に，その部分グラフに属する全頂点に共通するアイテム集合（共有アイテム集合）にも着目する．その状況下で，部分グラフのサイズ，共有アイテム集合のサイズ共に，予め決められた閾値以上の値となり，かつ，部分グラフの共有アイテム集合が同一にならない部分グラフを全て抽出するアルゴリズムが COIN である．

COIN では共有アイテム集合の最小サイズを予め決める必要があるが，提案の条件では予め決めることができない．そこで，提案法では探索中にそれらの閾値を変化させて部分グラフを列挙し，理論上計算しうる FWER と比較をすることで，偽陽性の生起確率が有意水準以下に必ずなるようにする．

4. 提案手法

本章では，3 章で導入した LAMP 及び COIN の手法を組み合わせることで，超大規模データに対しても，表現型に有意に関わる SNPs 集合を列挙できる手法を提案する．

本提案手法の背景思想としては，検定対象に関する予備知識を用いてグラフに制約を加えるというものである．今回は， PPI 情報を用いてグラフ制約を行う．これは，タンパク質同士が相互作用するかどうかが，そのタンパク質をコードする遺伝子上に存在する SNP 同士が関連するかどうかの指標として利用できるという考えに基づくものである．今回は PPI 情報で相互作用するとされるタンパク質をコードする遺伝子上に存在する SNP に関して総当たりで組み合わせを考えることにする．これにより SNP 同士の関係性を，各 SNP を頂点で与えたグラフ構造で表現できる（図 1(B)）．また，グラフの頂点にあたる SNP は，その SNP を持つ個体をアイテムセットとして保持する．

本提案手法は，GWAS データ(各個体の表現型及び，保有する SNPs の情報)と PPI 情報，有意水準αを入力として，表現型の違いに有意に関わる SNP の組み合わせを出力する．表現型データは，各個体の表現型を 0,1 の 2 値で表したものとする．SNP の有無の情報は，参照ゲノムと同一か否かの二値で考える．グラフの頂点には各 SNP を保有する個体の集合をアイテムセットとして与え，辺は翻訳後のタンパク質に PPI があることを表す．λ以上の個体を共有する部分グラフを効率的に列挙し，その頂点が表す SNPs の集合についてカイ二乗検定をする．求めた P 値が，補正後の有意水準以下であれば有意な SNPs の組み合わせとして列

表 2 個体と SNPs と表現型の一覧表

個体	SNPs	表現型
a	s_1, s_2, s_3, s_7	1
b	s_1, s_2, s_3, s_6, s_7	1
c	s_1, s_2, s_3, s_4, s_7	1
d	s_1, s_2, s_3, s_7	1
e	s_1, s_2, s_6	0
f	s_2, s_3, s_4	0
g	s_5, s_6	0
h	s_7	0

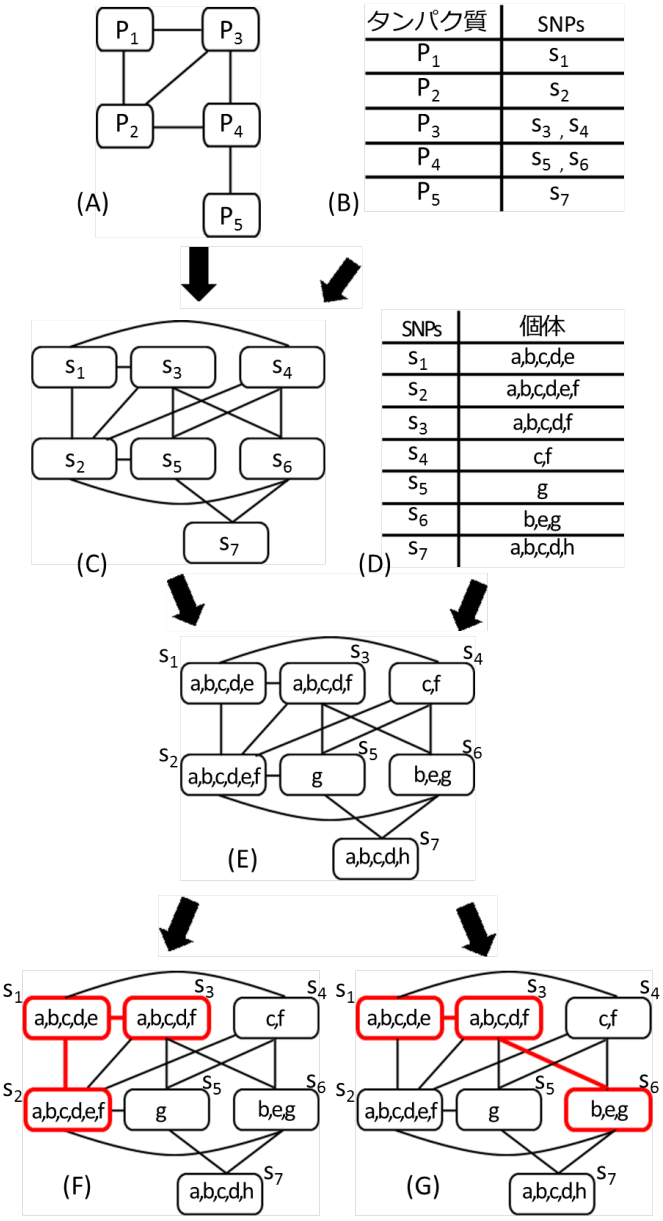

図 2 提案手法の流れ

挙する．

出現頻度λが小さいほど，COIN は部分グラフの列挙に時間を要するため，本研究では，λを大きい値から徐々に減らすアルゴリズムを提案する．λの初期値は，表現型が 1 である個体数に設定し，FWER の上限がαに達するまでλを小さくする．

λを減少させると，検定対象となる SNPs 集合の数m_λが増加し，P 値の下限$f(\lambda)$は逆に大きくなる．このとき，補正後の有意水準 δ は α/m_λ であり，FWER の上限は$m_\lambda f(\lambda)$である．$m_\lambda f(\lambda)$がα以下であれば，まだ偽陽性を生じ得ない検定が存在するため，$\lambda = \lambda - 1$とする．それ以外の場合には， FWER がαを超えてしまったため，$\lambda = \lambda + 1$として探索を終了する．最後に，列挙されているm_λ個の SNPs 集合に関して検定し，P 値が最適な補正後の有意水準 δ 以下の SNPs 集合を表現型に有意に関与する SNPs の組み合わせとして出力する．

4.1 組み合わせの列挙

本研究では，λ 個以上のアイテム集合が共有する部分グラフを列挙するため，COIN を利用し，提案手法を実装した．COIN が求める部分グラフの例を図 2 に示す．入力は，個体と SNPs と表現型の関係（表 2），PPI 情報（図 2(A)），タンパク質をコードする遺伝子領域に含まれる SNPs の情報（図 2(B)）である．これらの情報から，SNPs の間の関連の有無を表すグラフを構成する（図 2(C)）．このグラフの頂点はその SNPs を持つ個体集合をアイテム集合として保有する（図 2(D)）．これらの情報を合わせ，図 2(E)に示したグラフを構築する．

COIN を利用し，図 2(E)のようなグラフから，部分グラフを列挙することで SNPs 集合を列挙し，そのアイテム集合（今回はその SNPs を持つ個体）について共通集合を取る．COIN では，閾値 λ を指定し，λ 個以上の個体が共有する部分グラフを列挙する．COIN で列挙される部分グラフの関係を，図 2(F)と(G)に示した．

図 2 の例では，図 2(F)の赤で示した部分グラフの頂点が SNPs 集合$\{s_1, s_2, s_3\}$の組み合わせに相当し，個体集合$\{a, b, c, d\}$がこの SNPs 集合を共有する． 図 2(G)の部分グラフは，SNPs 集合$\{s_1, s_2, s_6\}$に相当し， SNPs に共通の個体集合は$\{b\}$である．個体数に対する閾値$\lambda = 3$のとき，SNPs 集合$\{s_1, s_2, s_3\}$は個体数が 4 なので部分グラフとして列挙されるが，SNPs 集合$\{s_1, s_3, s_6\}$は，個体数が 1 なので列挙されない．

λが小さくなると，列挙される部分グラフの数 m_λ は大きくなる．このときの偽陽性の生起確率の上限は$m_\lambda f(\lambda)$である．$m_\lambda f(\lambda) \leq \alpha$であれば，$\lambda = \lambda - 1$として探索を続ける．それ以外の場合は，偽陽性の生起確率が上限αを越えたため，$\lambda = \lambda + 1$として探索を終了する．このように，補正項で考慮する組み合わせの数を徐々に増やし，適切な補正後の有意水準を探索する．

5. 実験

5.1 入力データの準備

本研究の有用性を示すため，シロイヌナズナの GWAS データ及び PPI 情報を用いて実験を行った．GWAS データは Atwell *et al.*によって観測されたデータ [13]を使用した．107 の表現型に関して，計 216,130 件の SNPs データが含まれている．表現型のデータは 2 値のものはそのまま 0,1 に変換した．連続値のものは，数値の大きいものから 25%を 1，残りを 0 として扱った．また，表現型のデータでは，表現型毎に観測された個体数や SNPs が異なり，欠損値があるため，表現型毎にデータ数が異なる．個体数の平均は，約 134 個体，最大は 194 個体で，最小は 76 個体である．PPI 情報は，ATPIN の AllPPI 情報 [14]を使用した．96,827 件の PPI 情報が含まれている．また，このデータはタンパク質間相互作用を列挙している．SNPs の情報と PPI 情報を結びつけるためには，各 SNPs がどのタンパク質，あるいは遺伝子上に存在するかを整理する必要がある．そのため，今回は AtSNPtile1 のアノテーションデータ [15]のうち，TAIR9 のデータを使用して関連付けを行った．

これらのデータを元に，相互作用情報が明らかになっているタンパク質をコードする領域に存在する SNPs に関してのみデータを抽出し，グラフデータを作成した（図 2(A)及び(B)から(C)を作成）．辺の数（PPI 情報から予測した SNP 同士の関連の数）は 13,241 本，頂点数は SNPs 数と同じであるため 216,130 個である．

5.2 実験環境

実験は，CPU は Intel(R) Xeon(R) CPU E7- 4870 (2.40GHz) 40 cores，メモリは 512GB，OS は Ubuntu 12.10 である．

また，言語は Java を用いて実装し，シングルスレッドで実行，1GB のメモリを使用した．

5.3 実験結果

実験は 107 の表現型毎に行った．これらの結果を表 3 にまとめる．探索・検定を行った SNPs 集合の数は，平均で約 180,998 個，最大では 230,150 個となった．これは，データセットに含まれる SNPs 数 216,130 とほぼ同程度であり，2 個の SNPs がもたらす組み合わせの影響を網羅的に検定する場合と比較しても，提案法によって探索空間が大きく削減されていることが分かる．

検出した有意な SNPs 集合の数は，平均で 189 個，最大で 2,064 個であった．また，実行時間も平均で約 315 秒，最大でも約 826 秒であり，現実的な実行時間である．

さらに，各表現型で検出した SNPs 集合の最大の大きさを調べたところ，107 表現型の平均では 2.52 個，最大では 23 個の組み合わせまで発見することができた．これは，従来手法では探索できない規模の組み合わせである．

表 3　実験結果

	最大	平均
探索・検定を行った数	230,150	180,998
有意な組み合わせの数	2,064	189
実行時間	826 秒	315 秒
組み合わせの最大サイズ	23	2.52

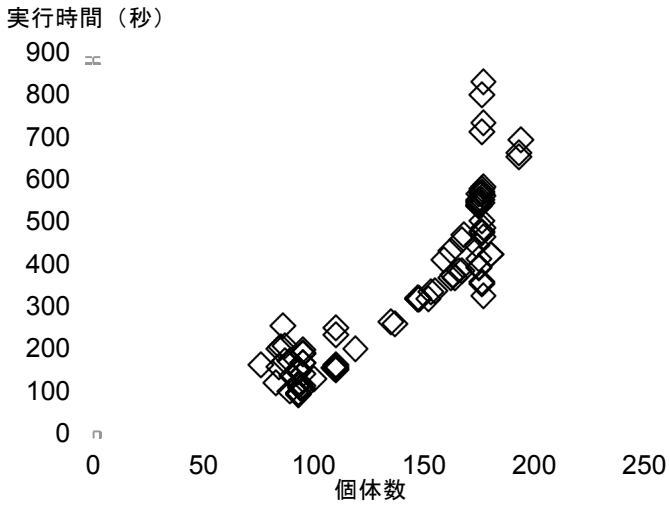

図 3　個体数と実行時間の関係

また，107 の表現型について，その個体数と実行時間の関係を表したグラフを図 3 に示す．この結果から，個体数の増加に伴って実行時間が増加することがわかった．また，個体数 170 から 180 あたりで実行時間のばらつきが大きくなっていることが観測されたため，この要因に関して検討してみたところ，個体数が同程度であるが実行時間が大きくなっているものは，表現型が 1 である個体の割合が高いためにλの初期値が大きく，COIN の実行回数が多いことがわかった．例えば，個体数が 170 から 180 の時，最も実行時間を要したデータセットでは，COIN の実行回数は 79 回であり，一方で，実行時間が最小のデータセットでは 25 回であった．表現型の割合によって実行時間に差はあるが，どのデータセットでも 15 分以内で全通りの SNPs の組み合わせの効果の検定ができ，これまでは発見できなかった 23 個の SNPs の組み合わせも検出できた．本提案手法を用いることで，GWAS のような超大規模データに対しても，現実的な実行時間内で，これまでは発見できなかった有意な組み合わせの列挙が可能になったことが示された．

6. 結論及び今後の課題

本提案手法により，ゲノムワイドなデータに対しても現実的な実行時間で多重検定を実行可能になった．また，発見された組み合わせのサイズも従来手法では探索できないサイズであり，データサイズ及び，発見できた組み合わせの数の面から本提案手法の有用性が示された．

また，今回は検定手法としてカイ二乗検定を用いている．カイ二乗検定は高速に実行できるが，要素数が少ない検定では誤差が大きくなってしまう可能性がある．そのため，今後 Fisher の正確確率検定などを用いてその性能の検証を行っていくことを考えている．

参考文献

1) Visscher PM, *et al.*, “Five years of GWAS discovery”, Am. J. Hum. Genet 90(1), 7-24, (2012)
2) Bonferroni CE “Teoria statistica delle classi e calcolo delle probabilità.”, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8, 3-62, (1936).
3) Sladek R, *et al.*, “A genome-wide association study identifies novel risk loci for type 2 diabetes.”, Nature 445, 881-885, (2007).
4) Terada A, *et al.*, “Statistical significance of combinatorial regulations” Proc. Natl. Acad. Sci. USA 110 32, 12996-13001, (2013).
5) Sese J, Seki M, Fukuzaki M, “Mining networks with shared items.” Proceedings of the 19th ACM international conference on Information and knowledge management, 1681-1684, (2010).
6) Gungor BB, Egemen E, Sezerman OU, “PANOGA: a web server for identification of SNP-targeted pathways from genome-wide association study data.”, Bioinformatics 30,1287–1289, (2014).
7) Gungor BB, Sezerman OU, “Identification of SNP Targeted Pathways From Genome-wide Association Study (GWAS) Data”, Protocol Exchange, (2012).
8) Fang G, *et al.,* "High-order SNP combinations associated with complex diseases: efficient discovery, statistical power and functional interactions.", PLoS One 7, e33531, (2012).
9) Zhang X, Zou F, Wang W, "FastANOVA: an Efficient Algorithm for Genome-Wide Association Study.", Proceedings of the 14th ACM international conference on Information and knowledge management, 821‑829, (2008).
10) Croiseau P, Cordell H, "Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach.", BMC Proc. 3, S61, (2009).
11) Agrawal R, Srikant R, “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th VLDB Conference, 487-499, (1994).
12) Tarone RE, “A Modified Bonferroni Method for Discrete Data”, Biometrics 42, 515-522, (1990).
13) Atwell S, *et al.*, “Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines”, Nature 465, 627-631, (2010).
14) Brandao MM, Dantas LL, Silva-Filho MC, “AtPIN: Arabidopsis thaliana Protein Interaction Network”, Bioinformatics 10, 454-461, (2009).
15) Array platform AtTILE1 and AtSNPtile1 http://aquilegia.uchicago.edu/naturalvariation/cisTrans/ArrayAnnotation.html