Vol.2014-MPS-98 No.3
Vol.2014-BIO-38 No.3
2014/6/25

Westfall-Young法を用いたエンリッチメント解析の感度改善と高速化

金 韓永[1,a)]　寺田 愛花[2,3,b)]　瀬々 潤[2,c)]

概要：ゲノム網羅的な遺伝子機能の解析が頻繁に行われるようになり，着目した遺伝子群に統計的に有意に関わる細胞機能を明らかにするエンリッチメント解析が重要になっている．ところが，この解析の結果，その遺伝子群に頻繁に見られる機能であるにも関わらず，有意に関連した機能として検出できないという直感に反した結果が得られることがしばしば起こり，問題となっている．本稿では，この原因の 1 つとして，多重検定補正で用いられる Bonferroni 補正が，機能間の独立性を仮定して偽陽性の生起確率を計算するため，機能解析では非常に保守的に補正する可能性を示した上で，Westfall-Young 法で多重検定補正を行う手法を提案する．Westfall-Young 法は，並べ替え検定を用いて確率分布を推定することで機能間の独立性を仮定せずに偽陽性の生起確率を計算でき，Bonferroni 補正に比べて検出力が高い．一方で，分布の推定に膨大な計算を要するため，時間がかかることが問題となる．この問題を解決するため，本研究では，各検定が p 値の下限を有することを示し，下限を利用した探索空間の枝刈りを導入することで，高速な Westfall-Young 法を提案する．マイクロアレイで観測した多様なヒト組織の発現プロファイルをクラスタリングした結果に対し，各クラスタに有意に関連する遺伝子オントロジーを求めたところ，有意水準に変更が無いにも関わらず，提案法の補正後の有意水準は Bonferroni 補正に比べ，5 倍以上大きくなった．また，枝刈り導入前の Westfall-Young 法に比べ 1,000 倍以上の高速化を達成した．

1. はじめに

マイクロアレイや新型シーケンサなど，近年，遺伝子網羅的な実験が可能となっている．大規模なデータが取得できる一方で，その実験から得られた結果が，どのような生体機能に関連しているかを見出すのは容易ではない．頻繁に用いられる方法として，様々な環境下で遺伝子発現量を観測，そのデータをクラスタリングし，その後，各クラスタに有意に関連する遺伝子機能を検出するエンリッチメント解析が行われている．機能の一覧としては，遺伝子オントロジー [1] の項目 (GO ターム) が頻繁に利用されている．

あるクラスタに関連する GO タームを求めるエンリッチメント解析は，以下の手順で行われる．クラスタに含まれる遺伝子群と，各遺伝子がどの GO タームに関連しているかの一覧を用意する．その遺伝子群に各 GO タームが高頻度で関連づいているか否かを判定するために，超幾何分布を用いて p 値を計算する．全 GO タームについて p 値を計算した後，多重検定補正によって，p 値を補，その後，有意水準 (0.05 など) と比較して，有意水準より小さな補正後の p 値を持つタームを有意に関連しているとみなす．

この手順で用いられる多重検定補正は，複数の GO タームに対して検定を行った場合に発生する高い偽陽性を避け，偽陽性の生起確率を有意水準以下に抑えるために行われる．例えば，有意水準 $\alpha = 0.05$ で 100 個の GO タームを検定すると，一個以上の偽陽性が起きる確率 (Familywise error rate，FWER) は 0.994 となり，99.4%以上の確率で偽陽性が生じる．そのため，偽陽性が一定以下になるよう有意水準を調整する [2]．広く使われている Bonferroni 補正 [3] は，全ての GO タームが独立に偽陽性を生じると見なして FWER の上限を算出し補正する．このため，上限値を過剰に見積もる傾向があり，一つも有意な GO タームが現れない事も多い．本研究では並べ替え検定を用いて帰無分布を推定する Westfall-Young 法 [4] を利用し，より厳密に偽陽性の上限を見積もることで関連する機能を十分に検出可能にする．また，Westfall-Young 法は計算時間が長いという欠点が存在するため，高速化を行う．

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
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
[a)] hanyoung@ss.cs.titech.ac.jp
[b)] terada.aika@ocha.ac.jp
[c)] sesejun@is.ocha.ac.jp

2. 関連研究

エンリッチメント解析とは，遺伝子の発現量とアノテーションデータを基づき，遺伝子群に網羅的に関わる機能を見つける方法である．

エンリッチメント解析の手法として，超幾何分布を用いた検定以外には，Gene Set Enrichment Analysis(GSEA)[5] が行われている．この方法は遺伝子発現量データから特定の遺伝子群 (クラスタ) を求め，その中にどんな GO タームが有意であるかを解析する GO 解析とは異なっていて，特定の遺伝子群を遺伝子セットとして予め準備しておいて，発現量データから遺伝子がどんな遺伝子セットに含まれるかを計算する方法である．

一方，多重検定補正に対して様々な方法が提案されている．例えば，よく使われ，すべての検定を独立に考える Bonferroni 補正の低い検出力を改善し， FWER を抑える Holm 法 [6] や Westfall-Young 法 [4] などが存在する．Holm 法は Bonferroni 補正から有意とみなした検定以外に残った検定の数で，改めて p 値の閾値を決める方法であり，Westfall-Young 法は，検定の従属関係を考慮して FWER を計算するため，並べ替え検定を利用して帰無分布を推定し，有意水準を補正する方法である．その他にも，偽陽性の割合 (False discovery rate, FDR) を抑える Benjamini-Hochberg 法 [7] や Storey 法 [8] など存在し，p 値から新たな基準値を生成し，閾値と比較する方法である．だが，Westfall-Young 法以外の各方法は，実際には親子関係のある GO タームが全て独立であることを仮定しているため，条件によって Westfall-Young 法より検出力が低いことが多い．

今までの研究ではエンリッチメント解析において Bonferroni 補正や Benjamini-Hochberg 法のよう p 値から容易に閾値を計算することは可能であるが，検出力が低い方法しか行われていない．GO 解析では GO の特徴から下位と上位のタームが強い関連性をもっている．そのため，GO 解析では， GO タームの関連性も考慮できる多重検定補正が有効になり，そこで本研究では，FWER を抑える検定の中，Bonferroni 補正などの p 値のみ使用する検定に比べて，GO 解析に対して，より厳密に FWER を抑えられ，p 値の分布を使用する Westfall-Young 法を使用して GO 解析を行う [9]．また，GO 解析に対して検出力が高くなることで，以前の手法から求めた結果とは関連性が低かった GO タームを見つけることが可能である．

3. 手法

3.1 遺伝子オントロジー

遺伝子オントロジー (Gene Ontology，GO[1]) とは生物学的概念を記述するために作られているテータベースである．機能を表す各項目を GO タームと呼ぶ．生物学的プロセス，細胞の構成要素，分子機能の三つの項目を根とした 3 つの非循環有向グラフ (Directed acyclic graph，DAG) で構成されている．GO の最上位階層 3 つは独立して，各遺伝子がどの GO タームに関連しているかについてもデータベースがまとめられている．

クラスタリングなどにより求まった着目する遺伝子群を $\mathcal{C}$ とすると，ある GO ターム G に関連しているとは，$\mathcal{C}$ の多くが機能 G を持っていて，他の遺伝子は G を持っていないとき，その GO ターム G は $\mathcal{C}$ に関連しているとみなせる．この検定に非復元抽出を表す超幾何分布が利用できる．全体の遺伝子が N 個，機能 G を持った遺伝子が M 個，$n = |\mathcal{C}|$ の場合，$\mathcal{C}$ の中に G を持った遺伝子数が m 個になる確率は，超幾何分布を用いて

$$P_{H(N,M,n)}(m) = \frac{{}_MC_m \cdot {}_{(N-M)}C_{(n-m)}}{{}_NC_n} \tag{1}$$

と表される．超幾何分布は，2 つの分類が存在する集合からランダムに非復元抽出を行った時，その状態が現れる確率である．求めたい確率 p 値は $\mathcal{C}$ の中で G を有する遺伝子数が m 個以上の場合なので，p 値 $P(N, M, n, m)$ は以下の片側検定として表せる．

$$P(N, M, n, m) = \sum_{x=m}^{n} P_{H(N,M,n)}(x) \tag{2}$$

GO を用いたエンリッチメント解析ではすべての GO タームに対して $P(N, M, n, m)$ を求めた上で，多重検定補正を行い，補正後の有意水準以下の GO タームを $\mathcal{C}$ に有意に関わるものとして列挙する．

3.2 Bonferroni 補正

Bonferroni 補正は，有意水準 α，検定の集合を $\mathcal{T}$ とした時，$\alpha/|\mathcal{T}|$ を補正後の有意水準として利用する．これは，補正後の有意水準を δ とすると，以下の式で FWER の上限が求められることに由来する．

$$\begin{aligned} \text{FWER} &= 1 - P\left(\cap_{i\in\mathcal{T}'}\{p_i > \delta\}\right) \\ &= P\left(\cup_{i\in\mathcal{T}'}\{p_i \le \delta\}\right) \le P\left(\cup_{i\in\mathcal{T}}\{p_i \le \delta\}\right) \\ &\le \sum_{i=1}^{|\mathcal{T}|} P(p_i \le \delta) \le |\mathcal{T}|\delta \end{aligned} \tag{3}$$

このとき，p_i は検定 i の p 値，$\mathcal{T}'$ は帰無仮説に従う検定の集合である．式の 2 行目の変形には，$\mathcal{T}' \subseteq \mathcal{T}$ であることを利用している．

式 (3) が α 以下になるようにすると，閾値 δ の最大値は，$\alpha/|\mathcal{T}|$ となる．よって，$\delta = \alpha/|\mathcal{T}|$ とすることで，FWER を α 以下に抑えることができる．この補正は検定数のみを用いて計算でき，計算速度が速いことから，多くの解析に使われている．

一方で，Bonferroni 補正は検出力が低い事が知られてい

る．これは，式 (3) の 2 行目から 3 行目の不等式で，最悪の場合として，検定間が全て独立と仮定しているためであり，検定間に従属性がある場合は，補正後の有意水準が本来の値より緩くなる．今回対象としている GO では，ターム間に親子関係が存在し，非独立であるため，Bonferroni 補正では非常に厳しい補正が行われている可能性が高い．

3.3 Westfall-Young 法

Westfall-Young 法 (WY 法) は，並べ替え検定を利用して帰無分布を求め，その分布を基に補正後の有意水準を計算する方法である．検定間の独立性を用いずに FWER の上限を計算するため，一般に Bonferroni 補正よりも厳密に FWER の計算ができる．

WY 法では，帰無仮説が真の検定集合 $\mathcal{T}'$ が既知であると仮定した場合，FWER は $\mathcal{T}'$ の最も小さい p 値が有意水準を下回る確率であることを利用する．その確率は次式で表せる．

$$\begin{aligned}\mathrm{FWER} &= P\left(\cup_{i\in\mathcal{T}'}\{p_i \le \delta\}\right) \\ &= P\left(\min_{i\in\mathcal{T}'}\{p_i\} \le \delta\right) \le P\left(\min_{i\in\mathcal{T}}\{p_i\} \le \delta\right) \quad (4)\end{aligned}$$

FWER は一つでも有意ではない検定を有意とみなす確率なので，$\mathcal{T}'$ の中で p 値が最小である検定が有意とみなされる確率と同様である．しかし，帰無分布に従う検定がどれであるかは予め知ることは出来ないため，$\mathcal{T}$ 中で最も小さな p 値を用いて FWER を抑える．この特徴を使うと，p 値の最小値の分布が分かれば， FWER の上限が計算できる．しかし，p 値の分布はわからないため，WY 法では並べ替え検定を用いて確率分布の計算を行う．

本研究では，遺伝子群 $\mathcal{C}$ の中に GO ターム G を持っている遺伝子の数から G が有意か否かの検定を行う．並べ替え検定は，遺伝子とクラスタの対応を入れ替え，その上で帰無分布を計算し，その分布を用いて検定を行う方法である．WY 法に従うと， 並べ替えた各状態から，最小の p 値を計算し分布を求めることで帰無分布を推定することができる．推定した分布を下から積分し，α になるところが補正後の有意水準 δ に対応する．並べ替えの回数を増やすことで，より正確な確率分布が得られる事が理論的に保証されている [4].

この方法は Bonferroni 補正に比べて，GO ターム同士の独立性を仮定せずに，確率分布を推定しているため，Bonferroni 補正より p 値の閾値が大きくなり，検出力が高くなることが期待できる．特に，GO タームは親子関係を持つため，その効果は顕著であると予想される．だが，確率分布の推定に繰り返し並べ替え検定を行うので，1 回の並べ替えに対する p 値の計算時間を t_b，並べ替え検定の回数を K とした場合，WY 法の実行時間は約 $K \cdot t_b$ の計算時間を要してしまう．

4. Westfall-Young 法の高速化

WY 法における計算時間の問題を解消するため，本研究では，p 値の下限を用いた枝刈り [10] と，キャッシュを用いた p 値の重複した計算の除去を導入することで，WY 法を高速化する．

4.1 p 値の下限を用いた枝刈り

WY 法では，個々の並べ替えたデータセットに対しては，式 (4) より，最小値のみ計算できればよい．本研究に関して並べ替えたデータセットはクラスタ $\mathcal{C}$ である．ここでは，その最小値を高速に計算する手法を導入する．

p 値の計算は，式 (2) より，N, M, n, m の 4 つの変数が存在する．しかし，GO ターム G に着目した時，並べ替えによって変化するのは m のみであり， N, M, n に変更は無い．そのため，G に対して，$\mathcal{C}$ が変わると m を改めて数える必要がある．G を持つ遺伝子数が $\mathcal{C}$ の中に多いほど p 値は小さくなるので，p 値は m の値が増加することで単調的に小さくなる．だが，m の値はクラスタサイズ n または全遺伝子の中，G に関連する遺伝子数 M より小さいため，上限が存在し，そこから p 値の下限が存在する．その p 値の下限は $m = \min(M, n)$ ときの p 値になる．

WY 法で各々の並べ替えたデータセットから求めるのは，並べ替えた $\mathcal{C}$ における p 値をすべての GO タームに対して計算し，その中で最小の p 値である．そのため，p 値の下限が分かれば，その値が既に検定した最小の p 値より大きい場合は，その GO タームの p 値は検定中の最小の p 値にはならないことが保証できるので，その GO タームに対して p 値を計算せずに枝刈りができる．本章ではその p 値の下限を計算し，下限から各検定に対して枝刈りを行うことで WY 法の高速化を行う．

p 値の下限 $P_{\mathrm{LB}}(N, M, n)$ は N, M, m から次式で計算することができる．

$$P_{\mathrm{LB}}(N, M, n) = \sum_{x=\min(M,n)}^{n} P_{H(N,M,n)}(x) \quad (5)$$

式 (1) と式 (2) の定義より，式 (2) の m は $m \le M$ かつ $m \le n$ であり，m が M と n の内，小さい値になるときの p 値がその検定に対する p 値の下限になる．

本研究ではその WY 法の計算過程を p 値の下限を用いて枝刈りを行う．枝刈りの方法は一回の検定の中で最小の p 値と各並べ替えから求めた最小の p 値の分布，二つ存在する．

第一の枝刈りでは，並べ替え毎に行うものである．並べ替えを行った後，全ての GO に対し p 値を計算していくが，計算途中までに求めた p 値の内，最も小さい値を枝刈

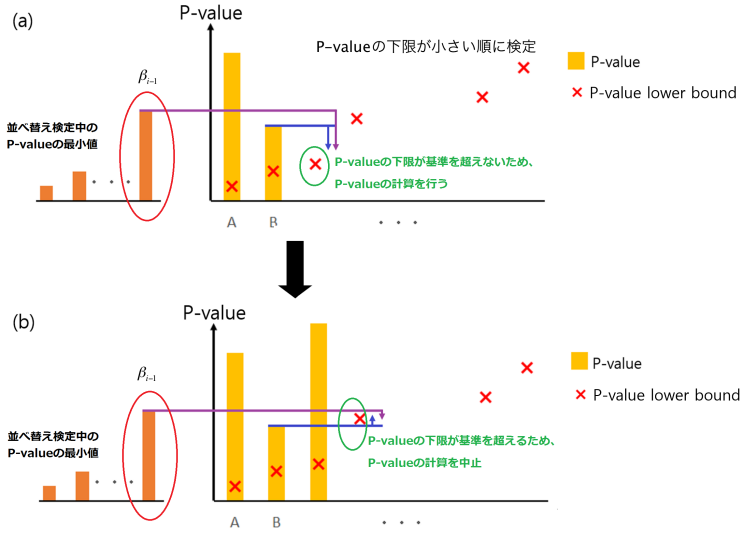

図 1 i 回目の並べ替え検定中の枝刈りによる高速化アルゴリズム，GO タームの p 値の計算は式 (5) の値が小さい順で計算する．(a)GO タームの p 値の下限が，p 値の最小値と β_{i-1} より小さいため，p 値の計算を続ける (b) 小さくない場合，その計算を中止する．

りに用いる．WY 法の一回の並べ替えから求めるべき値は全 GO タームの中で最も小さい p 値であるため，p 値がその値を下回らない事が言えれば，そのタームに関しては p 値を計算すること無く，最小の p 値を求めることができる．

第二の枝刈りとして，補正後の有意水準を求めるためには，p 値の最小値の分布の中，下から α%の p 値のみしか必要が無いことを利用する．i 回目の並べ替え検定中，並べ替え総回数を K，P_j を最小の p 値の分布から j 番目の p 値 (ただし，$1 \leq j \leq i-1$，$i = 1$ の時は存在しない，$j < k$ なら $P_j \leq P_k$ にソート)，i 回目の並べ替え検定を終えた時の $P_{K\alpha}$ を β_i とおくと，i 回目の検定で β_{i-1} より p 値の下限が大きい検定は枝刈りができる．

一回の並べ替え検定が終わったとき，求めた p 値の最小値と P_j(ただし，$j \leq K\alpha$) と比較し，改めて小さい順でソートすることで，P_j を更新し，β_i を計算する．

以上２つの枝刈りを導入した，高速版 WY 法の計算アルゴリズムは以下の様になる．

Step1. すべての GO タームに対して，式 (5) から p 値の下限を計算し，GO タームの検定を p 値の下限が小さい順番で行う．**図 1** では p 値の下限，赤の $\times$ が大きくなる順で検定を並べて行う．

Step2-A. i 回目の並べ替え検定中で，求めた p 値の最小値に対して，p 値の下限が大きい検定は計算を除く．図 1(a) で着目した GO タームの p 値の下限が検定中の p 値の最小値 (青線) 以下になるため，検定を続ける．図 1(b) は p 値の下限が検定中の p 値の最小値以上になるため，計算を中止できる．そのとき，検定を p 値の下限が小さい順番で行うため，一回計算を除くことは次にあるすべての GO タームに対して除くことになる．

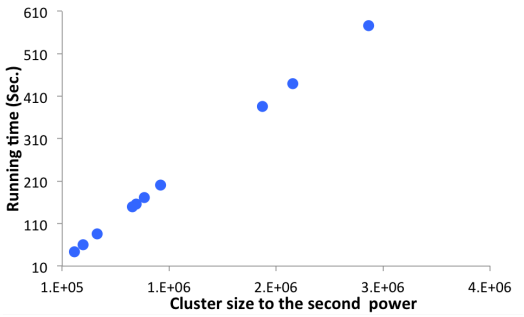

図 2 クラスタサイズの 2 乗に対する p 値の計算時間

Step2-B. i 回目の並べ替え検定中，β_{i-1} より p 値の下限が大きい検定は計算を除く．図 1 では着目した GO タームの p 値の下限が β_{i-1}(紫線) より小さいため，計算を中止しない．

Step3. step2-A または step2-B から i 回目の並べ替え検定を中止した時の p 値の最小値を P_j(ただし，$j \leq K\alpha$) と比較を行い，P_j(ただし，$j \leq K\alpha$) と p 値の最小値を合わせて順番でソートし，P_j(ただし，$j \leq K\alpha$) の更新と β_i の計算を行う．$P_{K\alpha} \leq p$ 値の最小値の場合，更新されない．

Step4. Step1 から Step3 からの作業を繰り返し，i 回目の並べ替えを K まで行う．β_K を計算し，その値が WY 法の閾値 δ になる．

4.2 p 値計算経過のキャッシュ

超幾何分布を用いた一個の p 値の計算量はクラスタのサイズを n とした場合，$O(n^2)$ であり，全体の GO タームに対して p 値の計算時間 t_b を**図 2** で表す．図 2 で計算時間は n の 2 乗に比例して，t_b が大きくなることが確認できる．

表 1 クラスタサイズ，WY 法の閾値 (Bonferroni 補正の閾値 =3.09E-6)，検定方法による有意に関連しているターム数

クラスタ名	クラスタサイズ	δ_{WY}	N_B	N_{WY}
A	1722	2.32E-5	32	45
B	1501	2.05E-5	356	421
C	1405	2.30E-5	91	113
D	1010	1.98E-5	148	193
E	932	2.02E-5	131	176
F	889	2.30E-5	163	210
G	869	2.29E-5	229	293
H	652	1.95E-5	25	35
I	543	2.28E-5	26	32
J	462	2.16E-5	223	282

一方，p 値の計算の際，p 値の 4 つの変数 N, M, n, m の中，GO タームとクラスタに対して，N と n は変わらず，M と m の値のみに変化し，p 値は 2 つの変数のみに依存する．本研究では計算した M, m の値と p 値の関係を保存しておくことで，p 値の計算時間を減らし，そのときのメモリの量は $M \times m$ に関わる．

5. 実験

本研究では提案手法が既存の WY 法に比べ高速に動作すること，また，同時に，WY 法が Bonferroni 補正に比べて，FWER は同じ水準に制御している場合でも，より多くの GO タームを関連していると判定する事を示す．

本研究で使用したデータは BioGPS に登録されているマイクロアレイで観測したヒトの 84 組織における 13,145 個の遺伝子発現データ [11] である．発現量は底が 2 の対数を取った後，その値が少なくともひとつの組織で 4 以上かつ GO タームがアノテーションされている 9,985 個の遺伝子を用いた．発現量は，MultiExperimentViewer[12] を用いてクラスタリングを行い，各クラスタに有意に関連する GO タームを求めた．クラスタリングの際，メソッドは k-means を用い，10 個のクラスタを生成，距離はピアソンの相関係数を用いた．GO タームは 16,177 個である．実行環境は OS：Linux，CPU：Intel Xeon2.60GHz，プログラミング言語は C，分布の推定に用いた並べ替えたデータセットの数は 10,000 回，有意水準は 0.05 である．

まず，Bonferroni 補正と WY 法の結果を比較する．各クラスタで補正後の有意水準を計算し，有意に関連する GO タームを求め，Bonferroni 補正で有意とみなされる GO タームの数と，WY 法で有意とみなされる数を比較した．

表 1 にクラスタのサイズ，Bonferroni 補正と WY 法の閾値とその閾値を持って有意とみなす GO ターム数を示す．本研究で使用した WY 法の閾値は Bonferroni 補正に比べて 5 倍以上大きくなり，その結果で Bonferroni 補正では発見できなかった有意な GO タームが多くあることが確認できる．

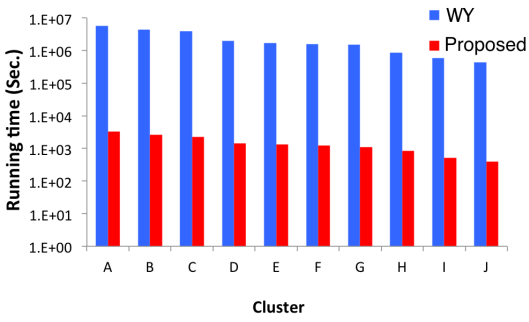

図 3 並べ替え回数 K=10,000，WY 法の計算時間の比較，青：WY 法，赤：高速化した WY 法

図 3 では Bonferroni 補正の計算時間，10,000 回の並べ替え検定に対する WY 法の計算時間と本研究で行った高速化後の計算時間を棒グラフで表す．

WY 法の計算時間は p 値の計算時間 t_b から並べ替え回数 K をかけた $t_b \times K$ であり，図 3 の青で示したように WY 法は長い計算時間を要する．クラスタサイズによって 60 日以上かかることもあり，現実的ではない．本研究の高速化後の時間は図 3 の赤で表し，その結果は一般の WY 法より 1,000 倍以上の速い．

枝刈りによって計算を省略した GO タームの数は平均 44%(7,118 個)，分散 36,7261 であり，なおかつ p 値のキャッシュにより，より速い計算時間の結果であった．本研究の高速化方法は並べ替えの数が多くなることで，枝刈りの効果と p 値のキャッシュの効果が大きくなる．

また，p 値を保存しておくため，大容量のメモリが使用される恐れが存在するため，WY 法の使用メモリと提案法で使用したメモリの実使用メモリ量を比較した．クラスタサイズに対して使用メモリが大きくなるため，クラスタサイズが最も大きい 1,722 個の場合を比較した．この時，WY 法のメモリは 0.3GB 使用したのに対し，提案法は 0.5GB 使用と，微増に留まったため，現在の計算機上で実行する上ではメモリ上の問題は起きないと考えられる．

検出された GO ターム数について考察すると，クラスタ A に対して，Bonferroni 補正で検出した GO ターム数は 32 個であり，本研究で用いた WY 法を用いる場合，p 値の閾値が約 6 倍になったことから有意とみなす GO タームも 44%増加した 46 個の GO タームが有意とみなされるようになる．

検出された GO タームの増加によって，生物学的発見を誘導できるか検証するために，**図 4** はクラスタ A で有意とみなした GO タームの一部を表す．赤のタームは Bonferroni 補正と WY 法の両方で有意とみなした GO タームで，緑のタームは WY 法のみで有意と見なした GO タームである．また，下位の GO タームは上位の GO タームに対して”下位 is a 上位”の関係を持っている．図 4 の最上位は GO:0003674，*molecular_function* であり，その下位に GO

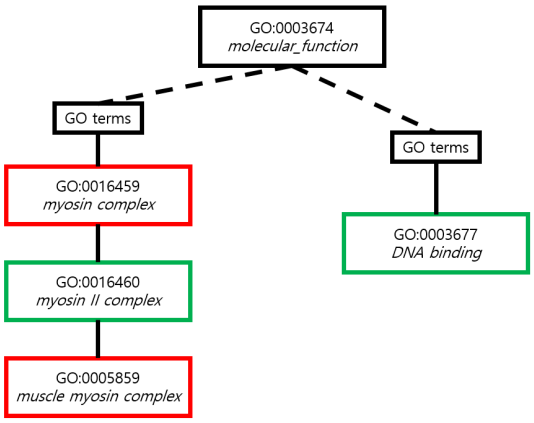

図 4 遺伝子オントロジーによる関係．赤：Bonferroni 補正，WY 法両方で有意とみなした GO ターム，緑：WY 法のみで有意とみなした GO ターム，黒：どちらの方法も有意とみなしてない GO ターム．左の赤と右の緑の最短距離は 5 で，関連性が低い．

タームが存在し，その GO タームらを黒四角で GO terms と表示している．その GO タームらを下位に各色をつけている Bonferroni 補正でも有意とみなした赤に囲まれた GO タームと WY 法のみで有意とみなした緑に囲まれた GO タームがある．各 GO terms は最小 2 個以上の深さである．

左の WY 法のみで有意とみなした GO:0016460, *myosin II complex* は Bonferroni 補正でも有意とみなした GO:0016459, *myosin complex* と GO:0005859, *muscle myosin complex* の間に存在する．そのため，Bonferroni 補正で有意とみなすことはできないが，Bonferroni 補正の結果と GO の DAG の特徴からある程度予想することが可能である．

右の GO:0003677, *DNA binding* について考える．*DNA binding* と Bonferroni 補正で有意とみなした GO タームは最短距離でも 4 階以上の差を持ち，*DNA binding* は Bonferroni 補正で有意と見なした GO タームと関連が少ない．そのため，WY 法を用いると Bonferroni 補正では有意とみなせず，また Bonferroni 補正の結果と関連性が少ない GO タームを有意とみなすことでできた．

また，今回の WY 法の結果が偽陽性ではないことを確認するため，簡単な方法として発現量が高い組織と GO タームの関連性を見る．クラスタ A に対する組織別すべての遺伝子の平均発現量が一番高い組織は *Skeletal Muscle* であり，また，クラスタ A の WY 法で有意とみなした最高の p 値 (一番偽陽性起きる確率が高い) を持つ GO タームは GO:0090257, *regulation of muscle system process* である．発現量が高い組織と GO タームは同様な筋肉に関しており，実際 *skeletal muscle* に存在する遺伝子 CASQ1 は GO ターム *regulation of muscle system process* を持っている．またその上位の GO タームは Bonferroni 補正で有意とみなした関連性が高い GO タームであるため，本研究で検出した GO タームは偽陽性ではないと判断できる．

6. おわりに

本研究ではクラスタに属する遺伝子に有意にアノテーションされている GO タームを求める際に，従来の Bonferroni 補正では過剰に補正するため，改善法として Westfall-Young 法を利用することを提案し，補正後の有意水準が 5 倍以上になることを示した．

また，Westfall-Young 法の欠点である計算時間の遅さを改善するため，枝刈りと p 値の計算とキャッシュを行い，平均 1,000 倍以上の高速化を達成した．

参考文献

[1] The Gene Ontology Consortium, Gene Ontology: toll for the unification of biology, Nature genetics, Vol. 25, No. 1, pp. 25-9 (2000)
[2] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, Multiple Hypothesis Testing in Microarray Experiments, Vol. 18, No. 1, pp. 71-103 (2003)
[3] C. E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, Vol. 8, pp. 3-62 (1936)
[4] P. H. Westfall, S. S. Young, Resampling-based multiple testing : Examples and methods for p-value adjustment, Wiley (1993)
[5] A. Subramanian *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci, Vol. 102, No. 43, pp. 15545-15550 (2005)
[6] S. Holm, A simple sequentially rejective multiple test procedure, Scand J Stat, Vol. 6, No. 2, pp. 65-70 (1979)
[7] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Series B, Vol. 57, No. 1, pp. 289-300 (1995)
[8] J. D. Storey, R. Tibshirani, Statistical significance for genome-wide studies. Proc Natl Acad Sci, Vol. 100, No. 16, pp. 9440-9445 (2003)
[9] 金 韓永，寺田 愛花，瀬々 潤, Westfall-Young 法を用いた遺伝子機能解析の感度改善，第 76 回 (平成 26 年) 全国大会 講演論文集 情報処理学会, Vol. 4, pp. 583-584 (2013)
[10] A. Terada, K. Tsuda, and J. Sese. Fast Westfall-Young permutation procedure for combinatorial regulation discovery. In Proceedings of IEEE Bioinformatics and Biomedicine 2013 (BIBM 2013), pp. 153-158 (2013)
[11] A. I. Su *et al.*, A gene atlas of the mouse and human protein-encoding transcriptomes, Proc Natl Acad Sci U S A, Vol. 101, No. 16, pp. 6062-6067 (2004)
[12] E. Howe *et al.*, MeV: MultiExperiment Viewer, Biomedical Informatics for Cancer Research, pp. 267-277 (2010)