2ZD-6

Westfall-Young 法を用いた遺伝子機能解析の感度改善

金 韓永[†1]　寺田 愛花[†2]　瀬々 潤[†3]
東京工業大学 大学院情報理工学研究科 計算工学専攻

1. はじめに

マイクロアレイや新型シーケンサなど，昨今，一度に大量の遺伝子を対象にした実験が可能となっている．このデータを解析する上で，得られた遺伝子群が関わる機能の調査が重要であり，遺伝子群に有意に関連する遺伝子オントロジー[1]の項目(GO ターム)を統計的検定で検出する解析がよく行われている．GO タームは全体で約４万個存在し，それぞれに対して検定を行うと，検出した GO タームに高確率で偽陽性が含まれる多重検定の問題が生じるため，偽陽性が一定以下になるよう有意水準を調整する多重検定補正が必要である．よく使われる Bonferroni 補正[2]は，偽陽性が生じる確率の上限を理論的に算出し補正するが，上限値を過剰に見積もる傾向があるため，一つも有意な GO タームが現れない事も多い．本研究ではランダムパーミュテーションを用いて帰無分布を推定する Westfall-Young 法[3]を利用し，より厳密な多重検定補正を行うことで，偽陽性を抑えつつ，関連する機能を十分に検出可能にする．

2. 手法

2.1 遺伝子オントロジー(Gene Ontology:GO)

GO 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[1]．各項目を GO タームと呼び，生物学的プロセス，細胞の構成要素，分子機能の三つの大分類で構成されている．この大分類を最上位とし，各 GO タームは非循環有向グラフ(DAG)で作られている．各遺伝子の機能は，GO タームの集合で表されている．

2.2 多重検定補正

統計的検定の際に本来は無関係の項目を有意と判定してしまうことを偽陽性と呼ぶ．一回の検定では有意水準 α 以下の P 値を有意とみなすことで，偽陽性が生じる確率を α 以下に抑えることができる．ところが，検定を繰り返すと，一回でも偽陽性が現れる確率(Familywise error rate，FWER)が大きくなる．FWER を一定以下に抑えるため，多重検定補正が必要である．

Bonferroni 補正

理論的に FWER が一定以下になるように補正する方法である．FWER の上限は，検定 i の P 値が pi であり，M 個の検定がある時，以下で計算できる

$$FWER = 1 - \Pr\left(\bigcap_{i=1}^{M}\{p_i > \delta\}\right) = \Pr\left(\bigcup_{i=1}^{M}\{p_i \le \delta\}\right) \le \sum_{i=1}^{M}\Pr(p_i \le \delta) \le M\delta \quad (1)$$

δは補正後の有意水準である．よって，$\delta=\alpha/M$でおくことで，FWER をα未満にできる［2］．

Westfall-Young 法（WY 法）

実際に帰無仮説に基づく分布が分かるとしたら，補正後の有意水準δは下から α 番目の値にすればよい．しかし，どのような分布か分からないため，下式より，ランダムに n 回サンプリングした場合にサンプリングし，P 値の最小値の分布を調べ，α 番目の値をδとすれば良いことが分かる［3］.

$$\mathrm{FWER} = \mathrm{P}\left(\bigcup_{i=1}^{M'}\{p_i \le \delta\}\right) = \mathrm{P}\left(\min_{i\in\{1,\ldots,M'\}}\{p_i \le \delta\}\right) \quad (2)$$

GO の有意性の判定

ある GO タームが有意となるには，着目する遺伝子群の多くがそのタームの機能を持っていて，他の遺伝子はそのタームの機能を持っていないことである．このような状況を検定する方法として，超幾何分布を用いた方法があり，GO 解析でも頻繁に利用される．式(4)で P 値が計算できる．

$$P_{H(n,M,N)}(x) = \frac{{}_{M}C_{x} \times {}_{N-M}C_{n-x}}{{}_{N}C_{x}} \quad (3)$$

$$p = \mathrm{P}(x \ge m) = \sum_{x=m}^{n} P_{H(n,M,N)}(x) \quad (4)$$

式(3)と式(4)の N は全遺伝子の数，M は GO タームに関連する遺伝子の数，n は着目している遺伝子の数，m は着目している遺伝子の中で，その GO タームの機能を有している遺伝子の数，p は P 値である．
有意に関連する GO タームを求めるため，すべての GO タームの P 値を計算する．

3. 実験結果

Bonferroni 法と WY 法の結果を比較するため，ヒトの線維芽細胞の発現量データ[4]から GO タームがアノテーションされている遺伝子を検出，更に発現量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 タームを求める．有意水準 α は 0.05 とした．

実行環境は OS として Linux，CPU は，Intel Xeon2.60GHz，C 言語使用．WY 法の実行時間は 1,000 回繰り返した時，平均 9,389 秒，Bonferroni 法の実行時間は 9.389 秒であった．

クラスタリングは gepas.org[5]を利用し，メソッドは非加重結合法，距離は Euclid 距離で行った．結果，378 個の遺伝子，2,857 個の GO タームで，6 個のクラスタに分けられた．各クラスタを A から F まで番号

Detection power improvement of gene function analysis using Westfall-Young method
†1 Hanyoung Kim, Tokyo Institute of Technology
†2 Aika Terada, Tokyo Institute of Technology
†3 Jun Sese, Tokyo Institute of Technology

付け，表 1 で各クラスタの大きさを示す．

表 1　クラスタの大きさ

クラスタ名	クラスタの大きさ
A	14
B	206
C	32
D	41
E	82
F	3

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する．

表 2 はその結果を表す．N_B と N_{WY} は，Bonferroni 補正と WY 法で有意と見なしたターム数，δ_{WY} は WY 法の有意水準である．Bonferroni 補正の補正後の有意水準は，どのクラスタでも δ_B=1.75E-5 である．

表 2　クラスタと有意な GO ターム数

Cluster	N_B	N_{WY}	δ_{WY}
A	0	0	2.78E-4
B	0	0	3.83E-4
C	0	0	2.02E-4
D	12	18	3.39E-4
E	0	2	3.36E-4
F	0	1	4.54E-4

δ_{WY} は，すべてのクラスタにおいて δ_B の 10 倍以上である．ところが，クラスタ A，B，C のでは，WY 法でも一つも有意とならない．

図 1 は，クラスタ D で有意とみなした GO タームの一部を表す．緑は WY 法でのみ有意とみなした GO タームで，赤は WY 法と Bonferroni 補正の両方で有意とみなした GO タームであり，下位 GO タームが“is a”上位 GO タームの関係である．a)の場合は WY 法でのみ有意とみなした GO タームが Bonferroni 補正でも有意とみなせた GO タームの上位に位置する．本来は小胞体（organella）全体に関連するのに，Bonferroni 補正では，この関係が見いだせない．b)の場合は a)と逆の関係であり，WY 法でのみ有意とみなした GO タームが下位のに位置する．Bonferroni 補正を使用する場合，上位の細胞内小胞体(intracellular organelle part)からは“is a”関係で細胞骨格(cytoskeletal part)を求めることができない．また，c）の場合は WY 法でのみ有意とみなしたタームは Bonferroni 補正で有意とみなしたタームとの関係がないため，b)，c)から Bonferroni 補正では推測できないタームも WY 法では有意とみなすことができることが分かる．

表 3 は，クラスタ D で WY 法でのみ有意とみなせる GO タームの数を示す．各記号の意味は M，m は式(3)と式(4)と同様である．クラスタの大きさは全遺伝子の 1/10 程度であることに比べ，有意になるタームを持つ遺伝子がクラスタに属する割合は 1/4 から 1/2 程度であり，クラスタに属する遺伝子でその GO タームを有意と判定することは適切であると言える．

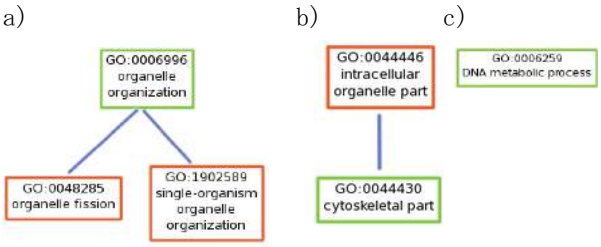

図 1　GO タームの関係

表 3　WY 法でのみ有意とみなした GO ターム

GO ターム(ID，名称)	M	m
GO:0006996，*organelle organization*	41	13
GO:0005874，*microtubule*	14	7
GO:0016043，*cellular component organization*	63	16
GO:0006259，*DNA metabolic process*	28	10
GO:0071840，*cellular component organization or biogenesis*	64	16
GO:0044430，*cytoskeletal part*	29	10

4．まとめと今後の課題

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 WY 法を利用することを提案し，補正後の有意水準が 10 倍以上になることを示した．一方で，実行に時間がかかるので寺田ら[6]の方法などを応用した高速化が必要である．

参 考 文 献

[1] The Gene Ontology Consortium, Gene Ontology: tool for the unification of biology, Nature genetics, Vol. 25, may (2000)

[2] 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.

[3] P. H. Westfall and S. S. Young, Resampling-based multiple testing : Examples and methods for p-value adjustment. Wiley, 1993.

[4] Vishwanath R. Iyer et al., The Transcriptional Program in the Response of Human Fibroblasts to Serum, Science 283, 83 (1999)

[5] Javier Herrero et al., GEPAS: a web-based resource for microarray gene expression data analysis, Nucleic Acids Research, Vol. 31, No. 13 3461–3467 (2003)

[6] Aika Terada, Koji Tsuda, and Jun Sese. Fast Westfall-Young permutation procedure for combinatorial regulation discovery." IEEE Bioinformatics and Biomedicine (BIBM 2013). pp. 153-158. Shanghai, China. December 18-21, 2013.