

GOMA:遺伝子発現量の簡便機能解析 Web アプリケーション

水谷 枝理子 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

概要: 近年、DNA マイクロアレイは遺伝子の網羅的解析に欠く事の出来ないツールとして普及している。膨大な遺伝子を扱うには統計的知識が必要であり、実験者が容易に解析できない場合も多い。また解析の結果得られた何らかの特徴ある遺伝子群に、どのような機能があるかを調べる事もまた容易ではない。そこで本研究では、DNA マイクロアレイなどで得られた遺伝子発現量データの入力を受けて、特徴ある発現を示す遺伝子群と関連した遺伝子オントロジータームを抽出し表示する Web アプリケーション GOMA を開発した。GOMA を用いる事で、実験でどのような機能の遺伝子群に変化が見られたかを容易に調べる事が可能である。GOMA は <http://goma.sel.is.oscha.ac.jp/> から利用できる。

GOMA: Web Utility for Direct Finding of Enriched Gene Ontology Terms from Gene Expression Profile

Eriko Mizutani and Jun Sese

Graduate School of Humanities and Sciences, Ochanomizu University

Abstract: Microarrays have become popular devices used to elucidate changes of cell status. Currently, we validate analyses of several or hundreds of samples using clustering and visualization. However, comparison is not an easy task because knowledge of data analysis is required and hundreds of clusters might be produced. This study presents GOMA, a web-based expression enrichment checking server using Gene Ontology. By putting sets of gene names and expressions into the server, which GO categories' genes are enriched can be verified immediately, and changes of cell status can be detected. GOMA is available online at <http://goma.sel.is.ocha.ac.jp/>

1 序論

近年、マイクロアレイを利用した遺伝子発現量の網羅的採取が進んでいる。マイクロアレイによる発現量の採取後一番始めに行われる解析は、そのアレイのデータが正しく採取されたものか否かのチェックであり、その後どのような遺伝子群に特徴的な変化が起きたかをクラスタリングによる解析で調べている。ところがこれらの作業には、複数枚のマイクロアレイが必要な上、専用のソフトウェアもしくは計算機言語に習熟する必要がある、実験者が手軽に行えるものではない。そこで本研究では実験でどの遺伝子機能に変動が見られたかを容易に確かめられるよう、マイクロアレイデータの入力から発現量に有意な変化があった遺伝子群の機能を、遺伝子オントロジー (GO) を用いて自動的に調べ表示する Web アプリケーション GOMA を開発した。GOMA ではマイクロアレイなどで得られた遺伝子発現量 1 サンプルを入力とする。データの入りは Web ブラウザから可能である。GOMA の内部では遺伝子オントロジーコンソーシアム [1] で配布しているデータベースを保持しており入力遺伝子から特徴ある発現の見られる GO Term を抽出する。ここで問題になるのが膨大にある GO Term からどのように Term を抽出するか、またどのように GO のグラフを表示させるかという事である。前者については一般的に用いられている統計指標である Mann-Whitney 検定の U 値を高速に計算して GO Term のランク付けを行い、上位 N 個の Term を注目すべき Term (以下では注目 Term と呼ぶ) として抽出している。後者については、膨大な GO の DAG (非循環有向グラフ) の中から注目 Term を含む部分に範囲を絞り表示させ、注目 Term を見やすい位置に配置させる事でユーザが変化のあった機能を把握しやすくし

ている。本稿は 2 章で関連研究を示し、3 章では Term の抽出方法と GO グラフの表示方法について述べる。また 4 章では実際のデータを用いた実行例を示し、5 章でまとめを行う。

2 関連研究

これまでに様々な GO viewer が提案されている。その中でも DAVID [2] や GO::TermFinder [3] は入力した遺伝子からそれらの機能を有意に含む GO Term を抽出し表示することができるアプリケーションである。これら 2 つは予め複数枚のアレイから採取した発現量を用いてクラスタリングを行い、遺伝子セットを取り出す必要があった。1 つのマイクロアレイサンプルから解析を行う手法も提案されているが [4]、この手法では p -value の計算に非常に時間がかかるため Web 上で扱うアプリケーションとしては有効ではない。そこで GOMA では、1 枚のアレイデータから特徴的な動きを見せる遺伝子群を高速に抽出させる手法を提案する。また表示の際 DAVID では Tree 形式を採用しているが、GO は本来 DAG 構造をしているため親を二つ持つ Term は重複して表示されてしまう。GO Term::Finder においては DAG 構造をそのまま表示させる事でその問題を回避している。しかし一般的なグラフの描画方法であるため DAG 構造の GO を描くには不向きであると言える。なぜならば、抽出された注目 Term 同士がグラフの構造上ではさほど離れていないにもかかわらず、実際に描画されたグラフでは非常に離れた位置に配置されてしまう場合があるからである。GOMA では注目 Term を中央に配置することで注目 Term 同士を近づけ、Term 間の関係が分かり易くなっている。

3 手法

3.1 Term の抽出

ここではユーザが入力した遺伝子発現量データをもとに、注目すべき GO Term を統計的手法を用いてランク付けする手順について述べる。図 1 (A) において四角いノードは GO の Term を、エッジは Term 同士の親子関係を示している。それぞれ Term II には遺伝子 D,F,G が、Term III には遺伝子 B,E,F が関連づけられている。また、ある Term T に関連づけられている遺伝子は T の祖先の Term にも同様に関連づけられるので、Term I は Term II と Term III の遺伝子の和となる B,D,E,F,G が関連づけられていることになる。図 1(B) では各 Term での遺伝子の発現量について示している。丸で示しているのが遺伝子で、発現量の高い順に右から配置されている。灰色で示している遺伝子がそれぞれの Term で関連づけられている遺伝子である。Term II と Term III を比べて、Term II の方が関連づけられている遺伝子の発現量が高い位置にまよっており、注目すべき Term であることが直感的には理解できる。しかしこれを計算によって求める必要がある。そこで我々は Term の重要度の指標に統計学的指標として知られる Mann-Whitney 検定の U 値を用いた。この指標は値の順位にのみ注目しているため、実験で得た発現量の大小関係さえ分かればどのようなデータでも用いる事が出来る。Mann-Whitney 検定の U 値を求める手順は以下の通りである。

1) ユーザが入力した全ての遺伝子を発現量によって順位付けをする。

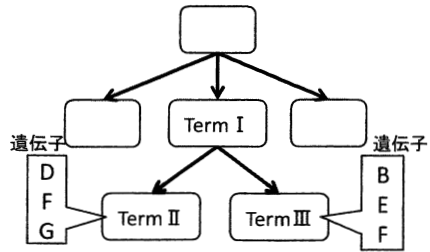
2) 各 Term ごとに、関連づけられている遺伝子群の順位を足し合わせ、その合計を r とする。またその Term の遺伝子群を G とする。

3) $U = r - |G| \times (|G| + 1) / 2$ を求める。

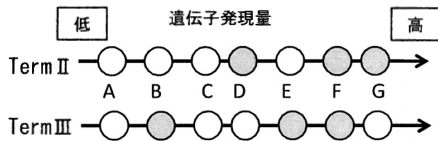
そして最終的に求められた U から p -value を算出し最も値の低い N 個の Term を注目 Term として抽出する。

例えば図 1(B) ではユーザが入力した遺伝子が発現量の順に並べられ、最も発現量の低い遺伝子が“A”，高い遺伝子が“G”となっている。つまり遺伝子の順位は“A”が1番で“G”が7番目である。この順番は Term II ,Term III の間で変わらないので発現量によるソートははじめに1度行えばよい。Term II に関連づけられている遺伝子の順位は4,6,7番目なので $U = 4 + 6 + 7 - 3 \times 4 / 2 = 11$ で p -value は0.114である。一方 Term III は $U = 7$ で p -value は0.857となり Term II の方がより重要な Term であることが示される。

また本来 U 値はその Term に関連づいた遺伝子一つ一つについて、その Term 以外の遺伝子における順位を求めてそれらを足し合わせる事で求まるが、約 25,000 もある Term で毎回そのような順位付けを行うと計算に非常に時間がかかってしまう。そのため本手法でははじめに入力遺伝子全てで順位付けを行い、各 Term では計算のみを行っている。これにより注目 Term を高速に抽出する事が可能となる。



(A) Gene Ontology の DAG 構造



(B) GO Term と発現量の関連

図 1 GO Term と遺伝子発現量

3.2 Term の配置

次に抽出した N 個の注目 Term の表示について述べる。GO を表示するには Tree 形式や DAG 構造をそのままグラフとして描く形式が一般的である。しかし DAG の特性である一つの Term に対して 2 つの親が存在する場合を Tree 形式で表示するとユーザに Term 同士の関係を誤って理解させてしまう恐れがある。また DAG のまま描く場合も、広く使われているグラフ描画ソフトである Graphviz[5] を用いるとエッジの交差を少なく配置できるが、抽出した Term が互いに離れ全体として非常に横に長い図になってしまう事がある。そこで我々は抽出した N 個の Term 同士を近くに配置し、さらにできるだけエッジの交差をなくす手法を提案する。

(a) 表示 Term の選択と階層

表示させる GO グラフの範囲は注目 Term を全て含む最小の範囲とする。そのためまず N 個の注目 Term の共通の祖先となる Term a を見つける。そして a と各注目 Term を結ぶパス上にある Term を、 a からの距離が 1 のものを 1 階層目、距離が 2 のものを 2 階層目... のように振り分ける。

(b) Term の入れ替え

階層ごとに並べて表示するだけでは注目 Term の位置が離れていたりエッジの交差が多くあるために Term 同士の関係が見えにくい。ここでは注目 Term 同士が近くなりかつエッジの交差が出来るだけ少なくなるよう、各階層の中で Term の入れ替えを行う。まず各階層ごとに Term 間の幅が均等になるよう図 2(A) のように配置させる。図 2 で四角で示されているのが Term で特に灰色の Term は注目 Term である。ここで t を Term とし、その直接の親である Term の集合を $parents(t)$ 、直接の子である Term の集合を $children(t)$ とする。例えば図 2 (A) では $children(t_3) = \{t_6, t_7\}$ となる。また、各階層の Term の位置を左から 0, 1, 2... として t の位置を $x(t)$ で表すとする。

例えば $x(t_2) = 0, x(t_6) = 2$ である。ここで

$$avg(T) = \frac{1}{|T|} \sum_{t \in T} x(t)$$

と定義する。 T は Term の集合で $avg(T)$ は T に含まれる要素の位置の平均である。また d 階層目に含まれる Term の集合を L_d , N 個の注目 Term の集合を T_N とする。注目 Term が各階層で中央に位置するように入れ替える手順は以下になる。

- (1) $t \in T_N \cap L_d$ なる t を選択する。
- (2) $t' \in T_d$ なる t' をランダムに選択する。
- (3) もし $|x(t') - avg(L_d)| < |x(t) - avg(L_d)|$ ならば t と t' を入れ替える。

次にエッジの交差を減らす入れ替えを行う。Term の入れ替えは、注目 Term の共通の祖先となる Term a の階層から一番下の階層まで走査する事を 1 回とし全部で 4 回行う。1,3 度目の走査では各階層で次の操作を 100 回行う。

- (1) 一度目は注目 Term から、3 度目は注目 Term 以外から t_1, t_2 (t_1 は t_2 よりも左にある Term とする) の 2 つの Term をランダムに取り出す。
- (2) t_1, t_2 のいずれかに子が存在しなければ再び t_1, t_2 を選び直す。
- (3) もし $avg(children(t_1)) > avg(children(t_2))$ なら t_1 と t_2 を入れ替える。

例えば図 2(B) において t_2 と t_3 を入れ替える事を考える。このとき $children(t_2) = \{t_4, t_5, t_7\}, children(t_3) = \{t_6, t_7\}$ で、 $avg(children(t_2)) = 2, avg(children(t_3)) = 1.5$ となり $avg(children(t_2)) > avg(children(t_3))$ であることがわかる。さらに $x(t_2) < x(t_3)$ なので t_2, t_3 が入れ替わることになる。次に 2、4 度目の走査では、各階層で Term の親の位置に着目して入れ替えを行う。

- (1) 2 度目は注目 Term から、4 度目の時は注目 Term 以外から t_1, t_2 (t_1 は t_2 よりも左にある Term とする) の 2 つの Term をランダムに取り出す。
- (2) もし $avg(parents(t_1)) > avg(parents(t_2))$ なら t_1 と t_2 を入れ替える。

例えば図 2 で、 t_5 と t_7 が選択されたとすると $parents(t_5) = \{t_2\}, parents(t_7) = \{t_2, t_3\}$ である。 $avg(parents(t_5)) = 1, avg(parents(t_7)) = 0.5$ となるので t_5, t_7 で入れ替えがおこりエッジの交差が解消される。

以上の操作によって注目 Term を中央に配置しながら、エッジの交差を減らす入れ替えを行う事が出来る。

4 実行例

この章では実際の GOMA の利用を紹介し、DNA マイクロアレイなどの実験結果の確認として有用である事を示す。ここでは入力データとして、酵母遺伝子 6,152 個に対して行った実験で観測された発現量を用いている [6]。

GOMA では遺伝子とその発現量の入力と、調べたい遺伝子の種の選択のみでそのデータの簡単な解析を行う事が出来る。図 3 では酵母に sorbitol を与えて 120 分後に観察された発現量データを入力した結果である。左から Molecular Function, Biological Process, Cellular Component の

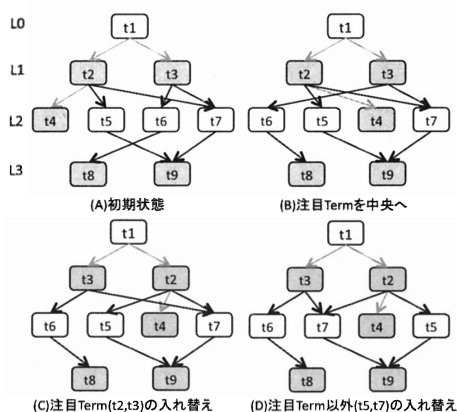


図 2 Term の入れ替え

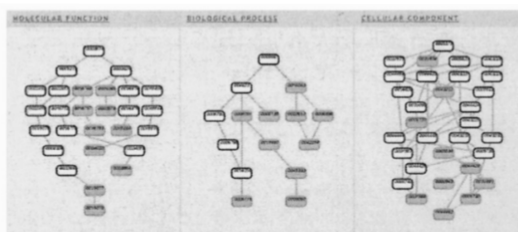


図 3 酵母に sorbitol を投与して 120 分後のデータでの実行結果

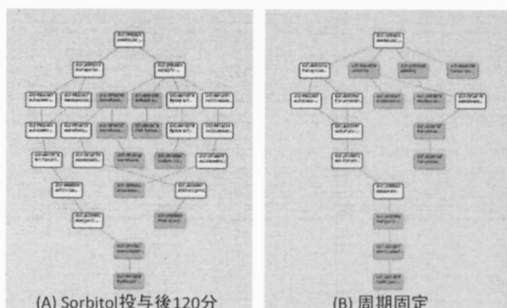


図 4 異なる実験間での抽出された上位 Term (Molecular Function) の比較

それぞれの GO カテゴリに関する DAG である。四角で示されるのが Term で特に着色されている Term は注目 Term である。一番左のグラフをクリックすると拡大された図 4(A) のようなグラフとその注目 Term についての詳細が書かれた表 1 が現れる。図 4(A) を見ると、各階層では Term が予め決められた幅の中で均等に配置され、必要以上に図が横に長くなる事が防がれている。また図の左側で親 Term を 2 つ持つ Term があるが、Tree 形式での表示のような重複は起きず正確に Term 同士の親子関係が示されている。さらに Term の入れ替えにより注目 Term は中央に配置されかつエッジの交差も少なく描かれている。Term の詳細が書かれた表 1 ではその Term の順位、GOID、Term

表 1 sorbitol 投与後 120 分 (Molecular Function)

Rank	GO ID	Term Name	p-value	Num Of Genes
1	GO:0004386	helicase activity	9.56e-07	69
2	GO:0016740	transferase activity	2.92e-06	520
3	GO:0016757	transferase activity, transferring glycosyl groups	1.24e-04	81
4	GO:0004033	aldo-keto reductase activity	1.74e-04	8
5	GO:0003964	RNA-directed DNA polymerase activity	2.10e-04	21
6	GO:0015077	monovalent inorganic cation transmembrane transporter activity	5.59e-04	48
7	GO:0003678	DNA helicase activity	5.75e-04	25
8	GO:0016758	transferase activity, transferring hexosyl groups	8.09e-04	64
9	GO:0015078	hydrogen ion transmembrane transporter activity	1.32e-03	43
10	GO:0016884	carbon-nitrogen ligase activity, with glutamine as amido-N-donor	1.54e-03	8

表 2 周期固定 (Molecular Function)

Rank	GO ID	Term Name	p-value	Num Of Genes
1	GO:0003735	structural constituent of ribosome	3.22e-54	187
2	GO:0005198	structural molecule activity	4.95e-47	288
3	GO:0016491	oxidoreductase activity	3.81e-22	215
4	GO:0005488	binding	1.96e-15	842
5	GO:0003676	nucleic acid binding	9.20e-14	407
6	GO:0008135	translation factor activity, nucleic acid binding	2.48e-13	42
7	GO:0003743	translation initiation factor activity	4.09e-12	28
8	GO:0045182	translation regulator activity	3.43e-11	50
9	GO:0034062	RNA polymerase activity	5.34e-11	32
10	GO:0003899	DNA-directed RNA polymerase activity	5.34e-11	32

名、p-value、Term に関連づけられている入力遺伝子の数が書かれている。入力で与えた酵母遺伝子とその発現量は sorbitol を与えた後に観察されている。酵母は糖分をエネルギーに変換するとともにエタノールを作り出し、その過程の中ではアルド-ケト還元酵素が必須である事が知られている。表 1 の中で aldo-keto reductase activity が 4 位である事からこの実験が酵母に正しく作用している事が確認できる。また再現性を示すため、同じ条件下で sorbitol を与えて 45 分後に観察されたデータを入力に用いた。このときも、aldo-keto reductase activity は p-value が 2.55e-06 で 6 位にランクされ、120 分後の場合と同様に特定の遺伝子群の発現量の変化を確認する事ができた。

さらに上記とは異なる実験 (8 時間育成した時点で周期を固定) で観測されたデータを入力した場合、表 2 で示すように Molecular Function での上位 10 個の Term の中には表 1 で示されたものは 1 つも含まれていない。つまり本手法で上位に選ばれる Term は、細胞の状態変化を敏感に捉えていると言える。またこのような上位 Term の変化は、表を比較するよりも、描かれた DAG の形状の変化を見ることでより簡単に発見することが出来る (図 4)。

以上のことから GOMA を用いて、実験による細胞内の変化の検出が可能であると言える。

5 まとめと今後の課題

本稿ではマイクロアレイで採取した遺伝子の発現量を入力し、重要な変化の見られる GO Term を自動で手軽に計算し見やすく表示する Web アプリケーション GOMA を提案した。GOMA を利用することで、実験で遺伝子にどのような変化が生じたかを容易に調べる事ができ、本格的な解析前に正しく実験が行われたか否かの確認が可能となる。今後の課題として、より注目 Term の算出の精度を上げる

事、複数毎のアレイでの解析、ユーザの利便性を考えたリンク情報の充実などが挙げられる。

参考文献

- [1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler et al., "Gene ontology: tool for the unification of biology. the gene ontology consortium.", *Nat Genet*, vol. 25, no. 1, pp. 25-29, May 2000.
- [2] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "David: Database for annotation, visualization, and integrated discovery", *Genome Biology*, vol. 4, p. R60, 2003.
- [3] E. Boyle, S. Weng, J. Gollub et al., "GO: Termfinder-open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes", *Bioinformatics*, vol. 20, pp. 3710-3715, 2004.
- [4] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles", *Proc. Natl. Acad. Sci.*, vol. 102, pp. 15 145-15 550, 2005.
- [5] "Graphviz: <http://www.graphviz.org/>".
- [6] A. P. Gaschet et al., "Genomic expression programs in the response of yeast cells to environmental changes", *Mol. Biol. Cell*, vol. 11, no. 12, pp. 4241-4257, 2000.