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　(指導教員：瀬々 潤)

1　はじめに

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り、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)[1] 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。

2　関連研究

DAVID[2] や GO::TermFinder[3] は、予めユーザがアレイなどのデータからクラスタリング等を用いて取り出した遺伝子セットを入力し、それらの機能を有意に含む GO Term を表示できるアプリケーションである。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ので、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表示する手法を提案する。

3　研究内容

3.1　Term の抽出

本章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 の Term を抽出する手法について述べる。図 1 の (A) において、それぞれ TermII 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。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とまっており、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

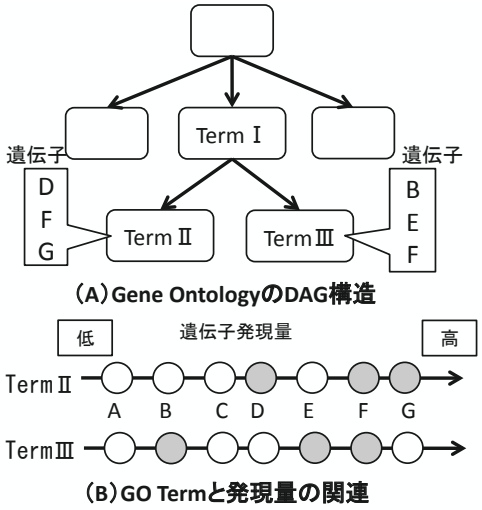

図 1: GO Term と遺伝子発現量

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。

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。

3.2　Term の配置

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 (図 2(A))。

(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。入れ替えの手順は次のようになる。

(1) 各階層で注目 Term を中央に配置させる (図 2(B))。
(2) 注目 Term 同士を結ぶエッジの交差を解消する (図 2(C))。
(3) 注目 Term は固定し、それ以外の Term を結ぶエッジの交差を解消する (図 2(D))。

以上の手順により注目 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可能となる。

4 実行結果

図 3(A) 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。着色された四角が抽出された注目 Term である。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ことが防がれている。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。表 1 は Molecular Function において抽出された注目 Term 上位 5 個である。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、その過程の中でアルド-ケト還元酵素が必須である事が知られている。表中で aldo-keto reductase activity が 4 位で抽出されている事から、実験が酵母に正しく作用している事が確認できたと言える。

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) で観測されたられたデータを入力した場合、Molecular Function での上位 10 個の Term の中には図 3(A) で示された注目 Term は 1 つも含まれていなかった。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 の変化は描かれた DAG の形状の変化を見ることで簡単に発見することが出来る (図 3)。

5 まとめ

本論文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である事が示された。また、結果の表示では抽出された GO Term を DAG の構造そのままに見やすく配置することができた。

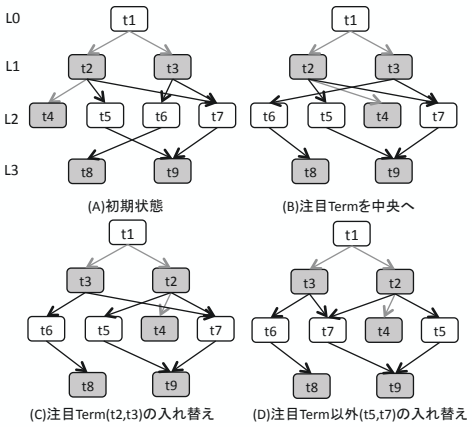

図 2: Term の入れ替え

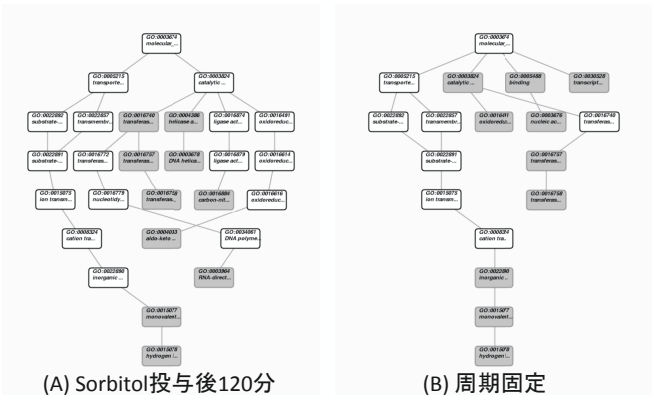

図 3: molecular function での Top10 の Term を描いたグラフ

表 1: 抽出された GO Term 上位 5 個抜粋 (Molecular Function)

	Term Name	p-value
1	helicase activity	9.56e-07
2	transferase activity	2.92e-06
3	transferase activity, transferring glycosyl groups	1.24e-04
4	aldo-keto reductase activity	1.74e-04
5	RNA-directed DNA polymerase activity	2.10e-04

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 sortware for accessing gene ontology information and finding significantly enriched gene ontology terms associated with alist of genes”,*Bioinformatics*,vol.20,pp.3710-3715,2004.

[4] A. Subramanian, P. Tamayo*et al.*,“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/”.