遺伝子群の共通機能に着目した遺伝子オントロジー表示

水谷 枝理子　(指導教員：瀬々 潤)

1　はじめに

ポストゲノム時代を迎えた今日，遺伝子に関する多種多様なデータが大量に得られるようになった．これらの実験から得られた遺伝子に関する知識は，遺伝子オントロジー（GO)[1] により統合され始めている．GO は生物学的機能を記述する Term（語彙，GO Term）を，非循環有向グラフ (DAG) によって階層的に表現し，各 Term にはその機能を持つ事が知られる遺伝子が関連付けられている．図 1 に GO の例を示した．図 1 において，転写活性という生物学的機能を表す Term に着目すると，この Term の子 Term（詳細な分類）として転写因子活性があり，親 Term（おおまかな分類）として機能が存在していることが分かる．更に，転写活性という Term には CAF20,NGR1, CBS1 が関連付けられており，これら 3 つの遺伝子が転写活性を持つ遺伝子として知られていることが分かる．更に，CBS1 は転写因子活性の機能なども知られていることが，この図から分かる．

現在のゲノム科学研究の現場では，既知の遺伝子機能が集約された GO を利用して，実験で採取できた遺伝子群がどのような機能に関連しているかの検索が行われている．しかし，GO では現在 Term が 21,000 個，関連付けられた遺伝子が 185,000 個と膨大であり，検索が困難なだけでなく，結果も膨大となる傾向に有る．

そこで本研究では，着目した遺伝子群に関連する Term 群を二項検定で抽出した後，その Term 群を DAG 構造は崩さないまま，見やすいように配置する手法を提案する．

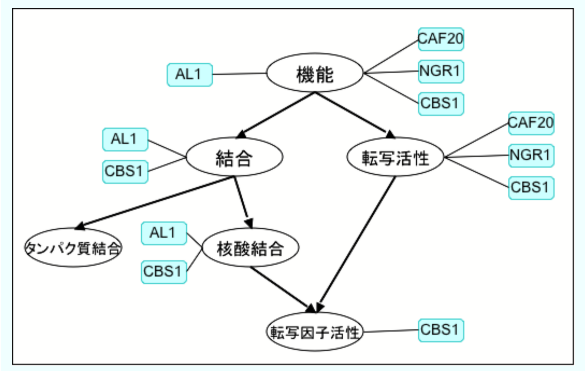

図 1: GO の一部の模式図

2　本研究の動機付けとなる例

図 2 は GO Term を Tree 形式で表示したものである [1] ．GO は DAG 構造であるため，Term に親が二つ以上存在する場合，その子供が重複して表示されてしまい，着目している Term の階層がわかりにくい．一方 DAG 構造を一般的なグラフ表示ソフトで表示するとエッジに交差が起きたり，注目している Term が離れて表示されたりし，決して見やすくない．(図 3 に注目している Term の関係が分かりにくい表示例を示した．数字の書かれた丸印が Term を示し，注目 Term

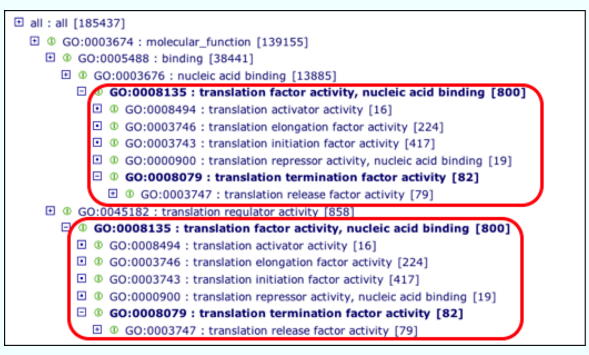

図 2: Tree 形式による表示の例 (囲みが重複部分)

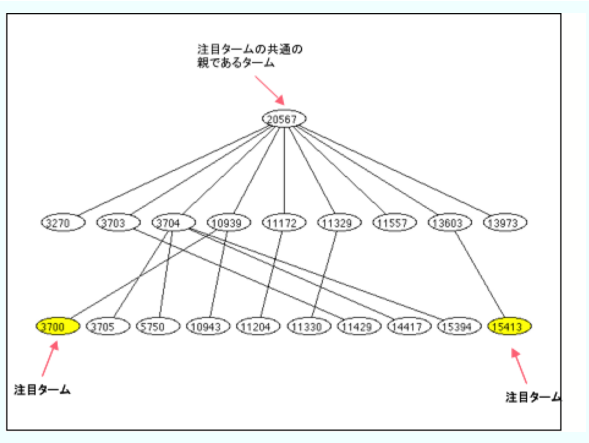

図 3: DAG 構造の見にくい表示例

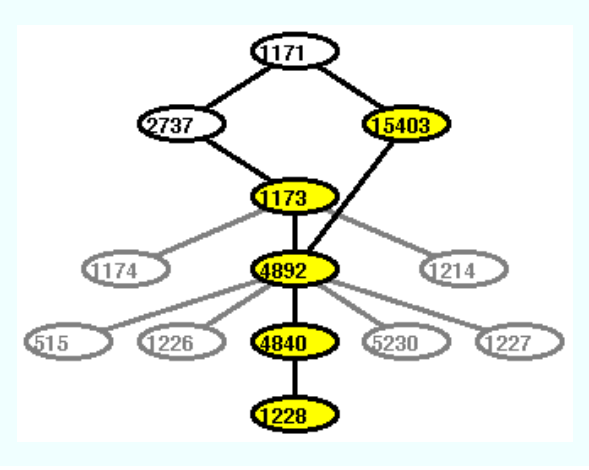

図 4: 提案手法による表示の例

を矢印で示した）．そこで我々は図 4 のように DAG 構造は保ったまま，エッジの交差を防ぎ，注目している Term を固めて表示することで視覚的に理解が容易な GO 表示を提案する．

3 関連研究

GO Term は Tree 形式で表示する手法 [1, 2] が一般である．また，グラフによる可視化も行われている [3] が，図 3 の様に着目する Term 群が非常に離れた位置に配置されることがあるため，Term 間の関係を理解することが容易ではない．一般に交叉の少ないグラフを描く手法も研究されているが，着目した Term が近くに来る制約は考慮されていない [4]．本研究の提案手法ではエッジの交叉を少なくし，かつ着目する Term 群を近くに配置する事が可能である．

4 手法

提案手法は，入力された遺伝子群から注目している Term（以下注目 Term と呼ぶ）を抽出する部分と，それらの Term の表示を行う部分に別れる．我々は，入力された遺伝子群が有意に現れる注目 Term として二項検定で P 値が最も低い 5 個の Term を選択した．グラフの表示には，2 個の問題点がある．一つ目は GO Term は全体で約 21,000 個存在し，多くは有意な Term(以下注目 Term と呼ぶ) に関連が無いため，全てを表示するのは冗長であること．もう一つは，Term の配置を工夫しないと注目 Term が離れたり，エッジの交叉が多くなったりし，見にくい図になる可能性があることである．

4.1 表示する Term の選択

表示する Term は，注目 Term の共通の親から注目 Term までの階層にあり，かつ以下の条件のいずれかに当てはまるものに限定する．

1. 注目 Term（以下, この Term 群を T_s とする）
2. 子孫に T_s 内の Term を全て含む Term の内，GO DAG 上で最も深さの深いもの（以下この Term を r とする）
3. r と $t \in T_s$ を結ぶ全てのパス上にある Term (T_p とする)
4. $t \in T_s$ が親である Term（T_c とする．ただし，冗長な表示を防ぐため兄弟が 10 個より多い場合を除く）

$T = T_p \cup T_c$ とする．

4.2 Term の配置順序を決定

Term r から子を辿り Term $t \in T$ を結ぶパスの内，最も長いものの長さが L の時，t は第 L 階層に属すると定義し，第 L 階層の Term 群を $T(L)$ と表す．最も深い階層を H とする．提案手法では，次の手順で階層毎 Term の順序を決める

(1) 初期配置: $d = 0, \cdots, H$ について,(1-1) を行う

(1-1) 要素数 $|\ T(d)\ |$ の配列を用意する．配列は中央を高い優先度とする．例えば 5 個の配列の場合，[3,1,0,2,4] と優先順を付ける．$x \in T_s$ の Term を $y \notin T_s$ より優先度の高い枠に配置する．

(2) $d = 0, \cdots, H$ について (2-1)(2-2) を行う

(2-1) untieChildrenEdges($T_s \cap T(d)$)

(2-2) untieParentsEdges($T_s \cap T(d)$)

(3) $d = 0, \cdots, H$ について (3-1)(3-2) を行う

(3-1) untieChildrenEdges($T \cap T(d)$)

(3-2) untieParentsEdges($T_c \cap T(d)$)

- underlineuntieChilrenEdges (T)
  (1) 子のエッジに交差が無くなるまで (2) を行う．
  (2) $x, y \in T$ を T からランダムに選び，x, y から子のエッジに交差があれば二つのノードの位置を入れ替える．
- underlineuntieParentsEdges(T)
  (1) 親のエッジに交差が無くなるまで (2) を行う．
  (2) $x, y \in T$ を T からランダムに選び，x, y から子のエッジに交差があれば二つのノードの位置を入れ替える．

untieChildrenEdges, untieParentsEdges は，エッジの交差を減らす手順である．

4.3 描画位置と彩色

Term は描画ウィンドウのサイズに対し，縦は階層の数，横は各階層のノードの数で均等に分割し配置した．また T_c 内の Term 及びそれらに接続するエッジは薄い色描画し，T_s 内の Term は色をつけて示した．図 3 に示したものが，実行結果である

5 まとめと今後の課題

本研究ではユーザの注目した遺伝子群が有意に現れる GO Term をグラフ上で近くに配置し, また不要な部分を省略する事で，注目すべき Term をわかりやすく表示する手法を提案した．今後の課題として，注目 Term 以外で表示する Term の選択方法を工夫する事や，表示の際にノードの Term 名などの必要な注釈を付加する事，注目 Term を P 値によって色分けする事などが上げられ，ユーザビリティーを高める工夫をしていきたい．

参考文献

[1] The Gene Ontology Consortium,
Gene Ontology: tool for the unification biology. Nat. Genet. 2000. Vol. 25, 25-29.
http://www.geneontology.org/

[2] J. Ye, et al. WEGO: a web tool for plotting GO annotations.
Nucleic Acids Res.,2006.34.W293-297.

[3] J.M.Cherry, et al. SGD: Saccharomyces Genome Database. Nucleic Acids Res.
1998. 26(1).73-80.
http://www.yeastgenome.org/

[4] E.R.Gansner, et al. A Technique for Drawing Directed Graphs. IEEE-TSE, 1993. 19:3.