Gene Ontology解析を理解する

ADVAITA
Bioinformatics for the Extraordinary

パスウェイ解析や、遺伝子変異解析に関するソリューションをラインナップしています。

Gene Ontology（GO）とは

Gene OntologyとGene Ontology analysisに関する混乱は、その用語そのものから始まることがあります。一般にGene Ontologyまたは「GO」と呼ばれるものには、実は2つの異なる実体があります：

1. **オントロジーそのもの**は、正確な定義と定義された用語のセットであり、それらの間の関係も定義されています。

2. **遺伝子産物とGO termの関連性**を示すもので、各遺伝子が何をすることが知られているかという既存の知識を把握するために使用されます。

しかし、 Gene Ontology （GO）という言葉は、一般的に両方を指す言葉として使われており、時に混乱を招く可能性があります。これを避けるため、ここでは用語のセットとその階層構造を表すものは「Gene Ontology」という用語を使用し、遺伝子とGO termの関連付けのセットを表すため「GO annotations」という用語を使用することにします。

Gene Ontologyには、3種類の用語（ドメイン）があります：

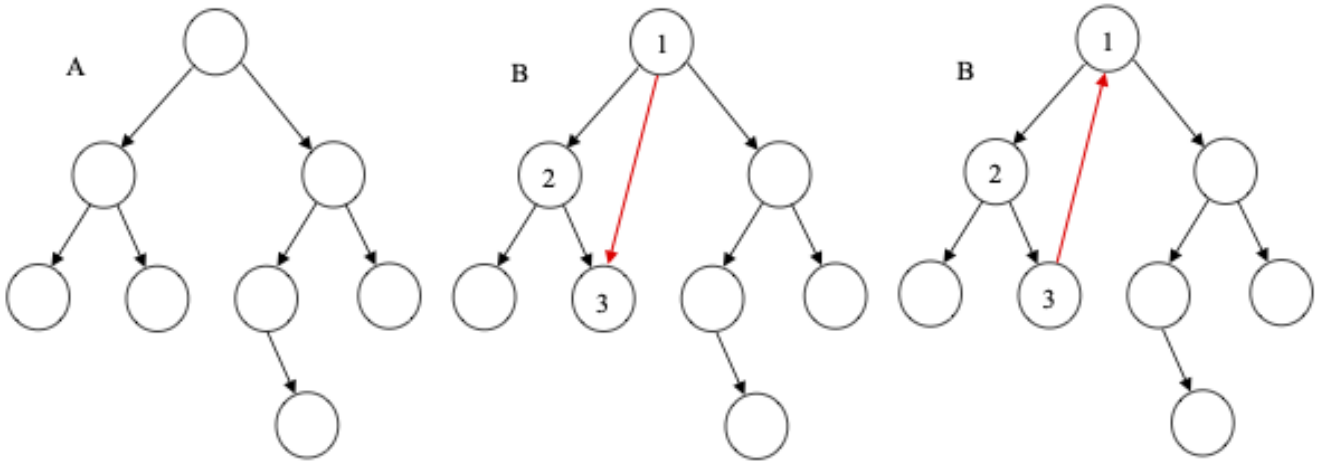
- Biological Processes (BP)
- Molecular Functions (MF)
- Cellular Components (CC)

GOの構造およびデータ表現

一般に、 Gene Ontologyのようなオントロジーは、生物学的な対象や事象の名前である多数の明示的に定義された用語で構成されています。これらの用語は、グラフのノードとして描かれ、ノード間の関係を表します。例えば、「cytoplasm (細胞質)」はノードであり、その親である「intracellular part (細胞内部)」とエッジで結ばれています。このエッジのタイプは「is-a」であり、この構造は単に細胞質は細胞内部分であることを意味します。

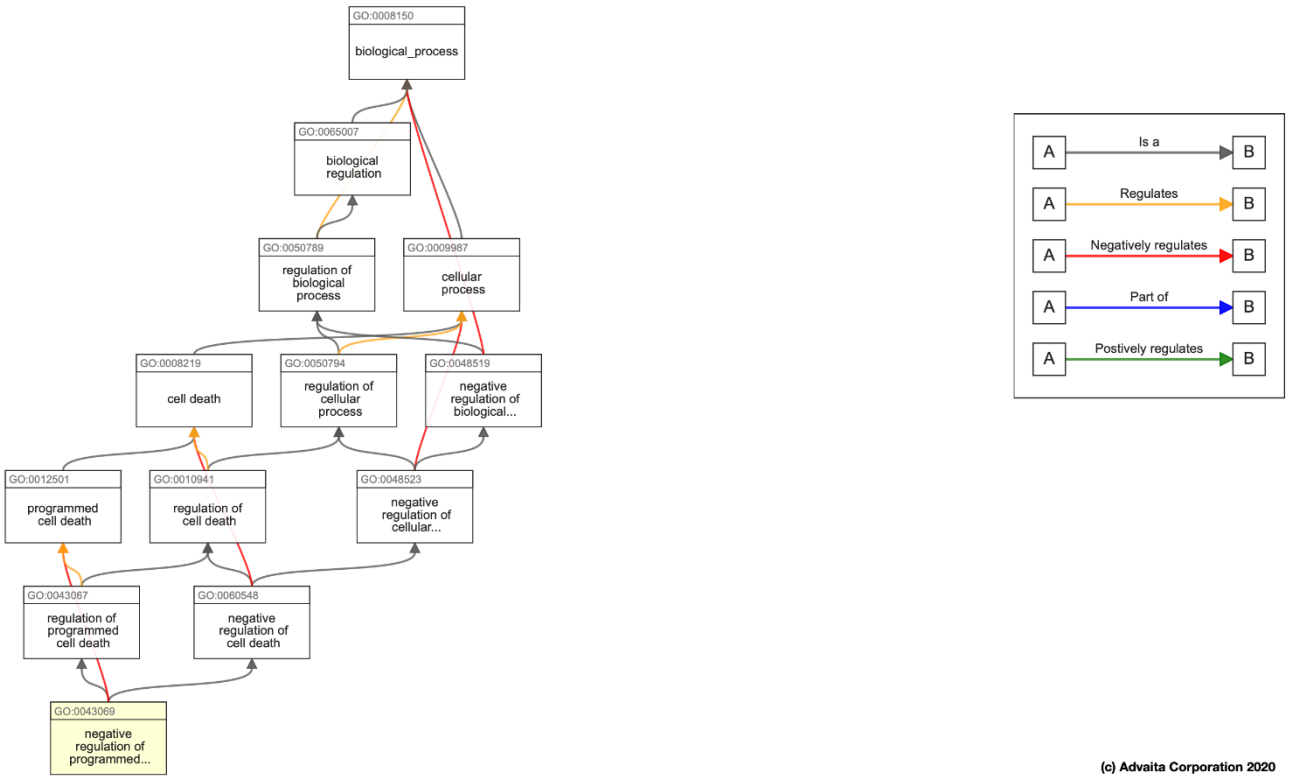
このノードとエッジで形成されるグラフは、ただのグラフではありません。いわゆる「有向非巡回グラフ (Directed Acyclic Graph: DAG)」、と呼ばれるものです。DAGにはいくつかの重要な特徴があります。まず、エッジは方向付けされます。つまり、各エッジにはソースとデスティネーション(行先)があります。 Gene Ontologyでは、ソースが親ターム、デスティネーションが子タームと呼ばれます。これは、細胞内部分が細胞質であるのではなく、細胞質が細胞内部分であることを示しています。

第二に、一般的なグラフと異なり、DAGはサイクルを持ちません。つまり、有向辺を辿ってループを完成させることはできません。特に、この制限は、2つのタームが互いの親と子の両方になることはできず（そうでなければ、互いの間でループを形成してしまうため）、また子のないノードが少なくとも 1 つ存在する必要があることを意味します。DAGはツリーと似ていますが、ツリーでは各ノードが1つの親しか持てないのに対し、DAGではノードが複数の親を持つことができる、という違いがあります。以下の図では、A はツリー (各ノードには 1 つの親のみがあります)、B は DAG (ノード 3 には 2 つの親があります)、C は一般的なグラフ (ノード 1、2、および 3 がループを形成します) を示しています。



ツリー、有向非巡回グラフ、一般グラフの違い

ここでは、「negative regulation of programmed cell death」という生物学的プロセスと、そのすべての先祖とのさまざまな関係を示す例を紹介します。



Gene Ontologyにおける階層構造と祖先の関係。
下位の用語はより具体的であり、上位の用語はより一般的である。

Gene Ontology analysisとは何ですか？

基本的に、Gene Ontology解析は、次のような非常に単純な質問に答えることを目的としています。

> *「表現型（例：疾患）とコントロール（例：健康）で発現差があることが分かった遺伝子のリストがある場合、この表現型に関与する生物学的プロセス、細胞の構成要素、分子機能は何ですか？」*

一言で言えば、ここでの前提は、ある生物学的プロセスに関連する遺伝子の多くが、ある疾患において差次的に発現していれば、その生物学的プロセスはその疾患に関与しているということです。Gene Ontology解析では、研究対象の疾患において影響を受ける生物学的プロセス、細胞の構成要素、分子機能を特定することを目的としています。

しかし...ここで問題になるのは、あるGene Ontology termが重要かどうか、どのように判断するかということです。結局のところ、どのような生物学的用語であっても、偶然に、あるいは、その遺伝子が研究対象である疾患にとってより重要な他の生物学的プロセスに関連しているために、差次的に発現する遺伝子があることになるのです。

そこで、この問題を解決するための主なアプローチを簡単に紹介します。良くなるものもあれば、悪くなるものもあります。実際、完全に間違っているものもあります。しかし、この問題に対する考え方がどのように変化してきたかを理解し、自分が使いたいアプローチを賢明に選択することができるように、ここでそれらをレビューしていきます。

以下は、数学的な詳細を伴わない簡単な概要であることに留意してください。本書の第16章でこれらのアプローチについてより詳しく説明しています。

最もシンプルなGene Ontology解析： Over-representation analysis (ORA) またはエンリッチメント解析

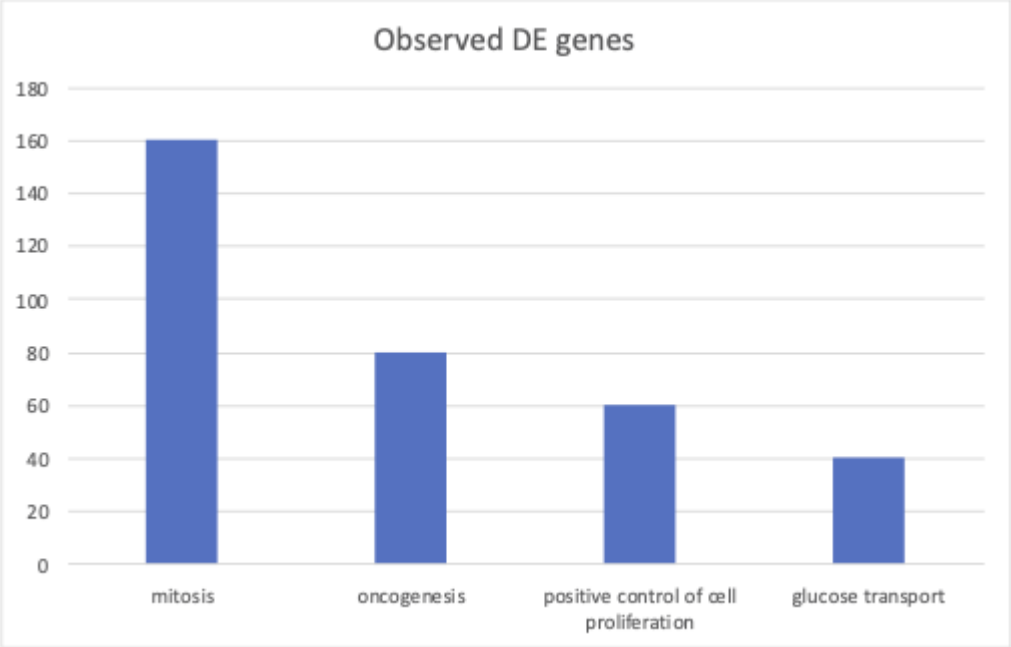
もし、differentially expressed （DE）遺伝子リストの処理を手作業で行うとしたら、DE遺伝子に対応するアクセッション番号を一つ一つ取り、様々な公開データベースを検索して、例えばその遺伝子が関与する生物学的プロセスをリストにまとめることになります。生化学的機能、細胞の役割など、他の機能カテゴリに対しても同じ種類の分析を実行できます。この作業を遺伝子ごとに繰り返し行うことで、少なくとも1つの遺伝子が関与するすべての生物学的プロセスのマスターリストを作成することができます。このリストをさらに処理すると、いくつかのDE遺伝子の間で共通するそれらの生物学的プロセスのリストが得られます。このリストの中で、より頻繁に出現する生物学的プロセスは、研究された疾患との関連性が高いと直感的に予想されるのです。 では、もし200個の遺伝子の発現に差があり、そのうちの160個が例えば有糸分裂に関わることが分かっていれば、有糸分裂は与えられた条件下で重要な生物学的プロセスであると直感的に結論づけられるでしょうか？

それは間違っています！

次の例でわかるように、この直感的な推論は間違っており、本当に関連する生物学的プロセスを特定するためには、より慎重な分析が必要です。

ある物質Xを摂取した場合の影響を調べるために、2,000個の遺伝子を含むパネルを使用する場合を考えてみましょう。例えば、200個のdifferentially expressed geneがあるとします。

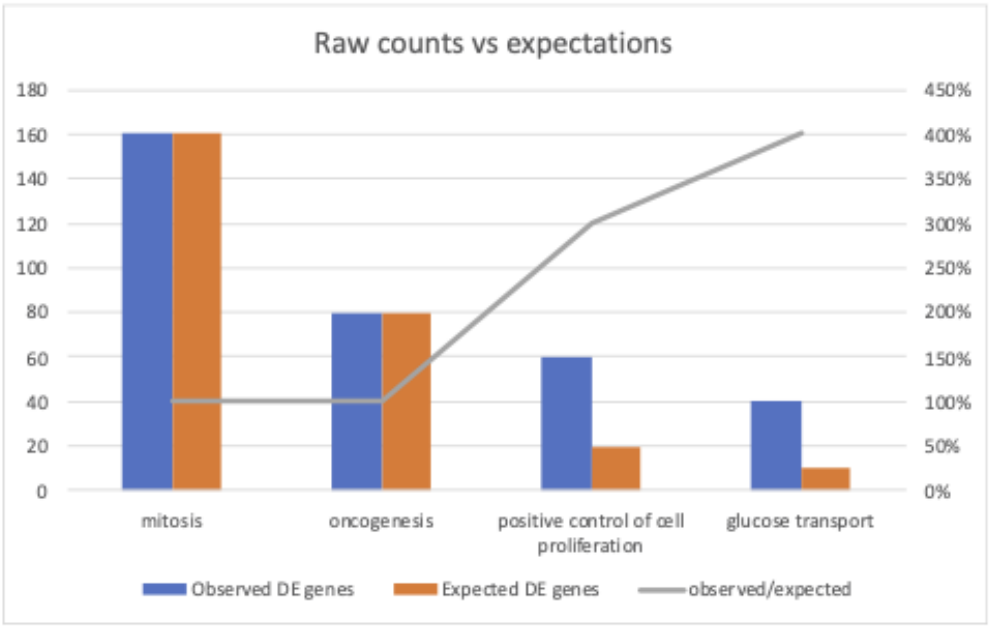
生物学的プロセスに着目し、200個の差次的に制御された遺伝子の結果が以下の通りであったとします：
200個の遺伝子のうち160個が有糸分裂に、80個が腫瘍形成に、60個が細胞増殖のポジティブコントロールに、40個がグルコース輸送に関与している。



Gene Ontology解析でよくある間違いは、生の数を使うことです。
ここでは、有糸分裂がこの実験で最も重要な生物学的プロセスであるように見えます。

ここで、上記の機能プロファイルを見ると、有糸分裂、腫瘍形成、細胞増殖がすべてその文脈で意味をなすので、物質Xはがんに関係しているかもしれないと結論づけることができるかもしれません。しかし、使用したパネルのすべての遺伝子が有糸分裂の経路の一部であった場合、何が起こるのでしょうか？有糸分裂は引き続き重要な意味を持つのでしょうか？明らかに、答えはノーです。したがって、正しい結論を導くためには、個々のカテゴリーについて、**常に実際の発生回数と予想発生回数を比較する必要があります。**

この比較は、下の図に、観察されたものと予想されたものの比率を示す線で示されています (パーセンテージは右側の縦軸に示されています)。この観点から見ると、同じデータでもまったく異なる同じデータでも全く違う話になってきます。確かに160個の有糸分裂遺伝子がありますが、これは最大数であるにもかかわらず、実際には160個のこのような遺伝子が観察されると予想していたので、これは単なる偶然よりも優れたものではありません。腫瘍形成についても同様です。細胞増殖のポジティブコントロールは、20と予想していたのに60も観察されたので、これは興味深いといえます。これは予想の3倍です。しかし、最も興味深いのはグルコース輸送です。このような遺伝子は10個しか観測されないと予想していましたが、40個も観測され、予想の4倍も多いことがわかりました。予想される遺伝子数を考慮することで、データの解釈は根本的に変わりました。これらのデータを考慮すると、X とがんではなく糖尿病との相関関係を検討したくなるかもしれません。



Gene Ontology解析：生カウントと期待値を比較する必要がある。有糸分裂が最も多くdifferentially expressed遺伝子を有していたにもかかわらず、これは偶然に予想されたもの以上ではなかった。一方、グルコース輸送は、絶対数が少ないにもかかわらず、偶然に予想される数の4倍と最も有意にエンリッチメントされていた。

この例は、制御されていることが判明した遺伝子の中で、特定の機能カテゴリーの出現頻度が単純に高いだけでは誤解を招く可能性があることを示しています。正しい結論を出すためには、**観測された頻度を予想される頻度を照らし合わせて分析する必要があります**。

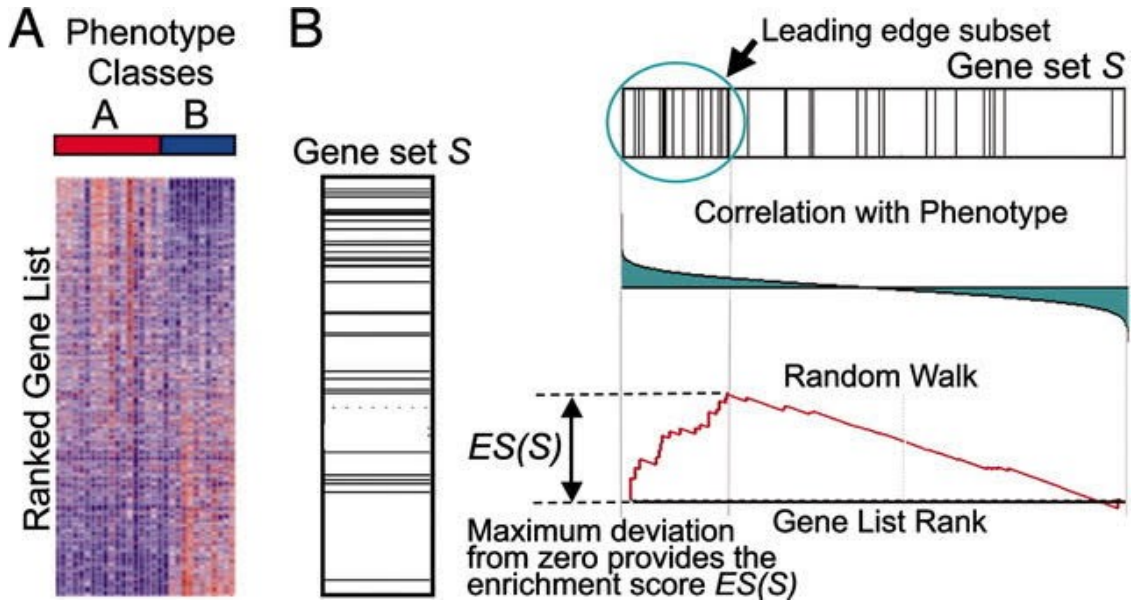
問題は、10個の遺伝子を期待していたのに、40個の遺伝子を観察するような事象が、偶然に起こりうるということです。可能性は低いですが、起こりうることです。肝心なのは、偶然に観測された値が現れる確率で、これらのカテゴリの有意性を評価する必要があるということです。これは、超幾何分布、フィッシャーの正確確率検定、カイ二乗検定などのさまざまな統計モデルを使用して実行できます。

詳細に興味がある方は、別の場所（本書の23章）で計算式を説明しましたが、ここでは、結論として、上記のようなカウントグラフから結論を出そうとせず、各項のp値を計算するソフトを使い、多重比較の補正を忘れないようにしましょうということです。

Gene Ontology解析の一歩先を行く： Functional Class Scoring (FCS)

健全な科学的結果をもたらす最も単純なアプローチは、上で説明したover-representationアプローチです。しかし、時代とともに、より洗練された手法が開発されてきました。重要なカテゴリーとして、 Functional Class Scoringメソッドを挙げることができます。このカテゴリーで最もよく知られている方法は、Gene Set Enrichment Analysis（GSEA）です。

最初のステップとして、GSEAは各遺伝子と表現型との関連性に基づいて遺伝子をランク付けします。この関連性は、t 検定などの任意の検定を使用して確立されます。ランク付けされた遺伝子リストLが作成されると、遺伝子セットリストの各セットについてエンリッチメントスコア（ES）が計算されます。リスト L は上から下にたどられ、そのセットに属する遺伝子が見つかるたびに統計量が増加し、そうでない場合は減少します。増加 (または減少) の値は、遺伝子のランキングに依存します。リストの上位にあるすべての遺伝子が、ある生物学的プロセスに関連している状況を想像すると、そのプロセスのスコアは、遺伝子が増えるごとに増えていくことになります。最終的なエンリッチメントスコアは、Walk中に遭遇したゼロからの最大距離です。



Gene Set Enrichment Analysis(GSEA) によって計算された統計の例。
Image from Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov PNAS October 25, 2005 102 (43) 15545-15550; first published September 30, 2005; https://doi.org/10.1073/pnas.0506580102

この図では、上のグラフが遺伝子リストをウォークスルーする際のエンリッチメント値を示しています。縦線は、セットSに属する遺伝子がランク付けされたリストに現れる位置を表しています。下段のグラフは、各遺伝子が表現型とどの程度相関しているかを示しています。

原則として、グラフがゼロから大きく離れるほど、高いエンリッチメントスコアが得られます。しかし、エンリッチメントスコア単体で有意性を評価することは、遺伝子の生カウントをそのように使うことと同様にできません。理由は同じです。原理的には、任意のスコアがゼロ以外の確率で出現する可能性があります。偶然に予想よりも頻繁に出現するものだけに焦点を当てなければなりません。これをブートストラップ方式で行っています。要するに、ブートストラップ法は、ラベルをランダムに並べ替えることによって、偶然に何かが現れる頻度を評価するものです。並べ替えの基準は、表現型サンプルの並べ替えと遺伝子ラベルの並べ替えの2つが考えられます。一般に、遺伝子間の相関関係が保存されるため、 label permutation methodが好まれます。このアルゴリズムでは、empirical p valueの計算を可能にするnull hypothesisが生成されます。 empirical p valueは、正しいラベルで観察されたエンリッチメントスコアと同等かそれ以上をもたらしたランダムブートストラップ実行の数として計算されます。

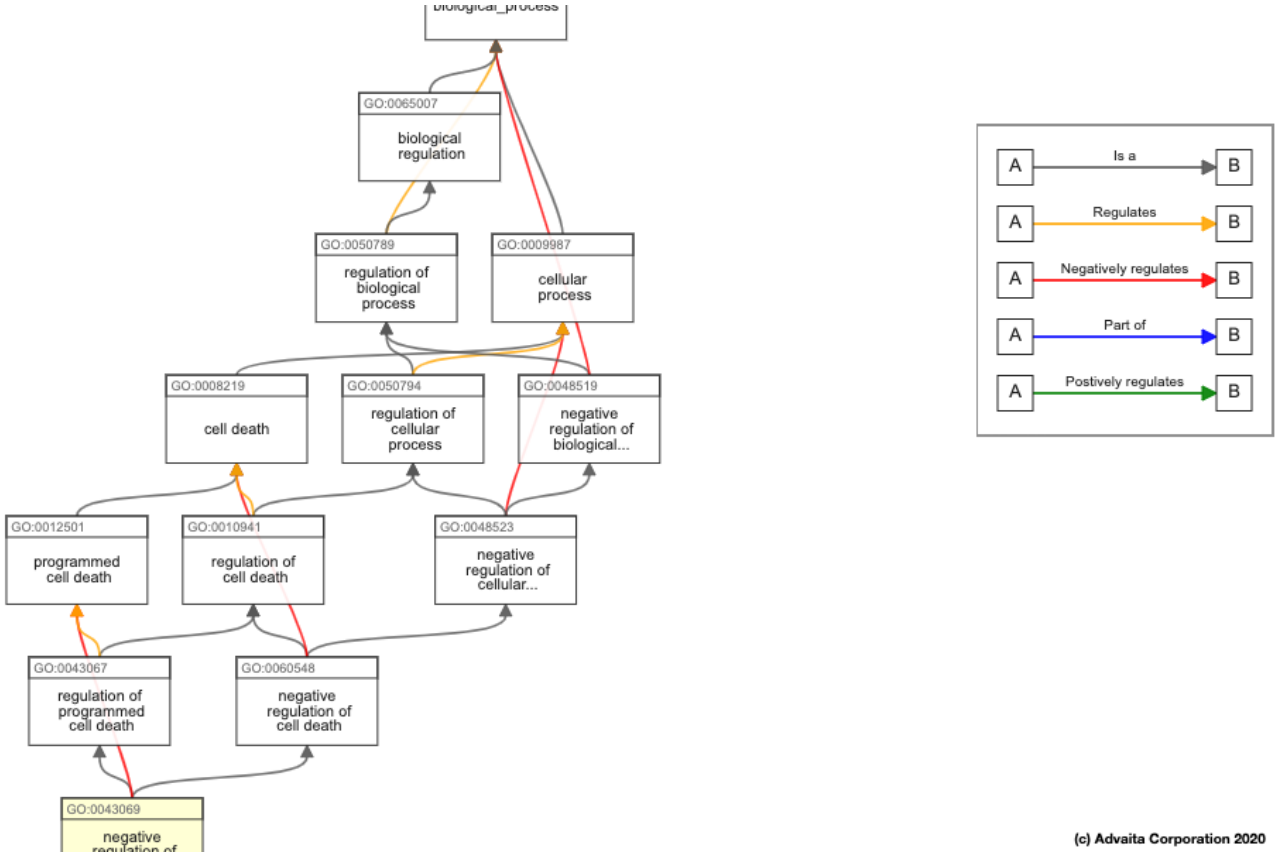
次の最後のステップでは、複数の仮説検定のために有意水準が調整されます。各エンリッチメントスコアは、セットのサイズで正規化され、Normalized Enrichment Score（NES）となり、各NESのfalse discovery rate（FDR）が計算されます。

より洗練されたGene Ontology手法: elim とweight

上記のアプローチは、Gene Ontology termに関連する発現量の異なる遺伝子の数を正確に解釈する問題に焦点を当てたものです。しかし、これらのアプローチは、Gene Ontologyの構造と様々な用語間の関係を無視しています。より洗練されたGO解析手法を理解するためには、Gene Ontologyについてさらにいくつかのことを学ぶ必要があります。

GOは、上記の関係のタイプを使用する階層構造で構成されています： “is a”、“part of”、“regulates ”の3つの関係で構成されています。例えば、「細胞外シグナルによるアポトーシスの誘導」は「アポトーシスの誘導」の一種であり、それが「アポトーシスの正の制御」となり、それが「アポトーシスの制御」の一種である、など。一般に、上記のような“is a”の関係に従ってDAGを横断することは、抽象度を超えて移動していると見なすことができます。ルートであるBP（または“All” ）は、最も高い抽象度、または最も低い詳細度に対応することになります。一方、「ホルモンによるアポトーシスの誘導」のようなリーフノードは、最も抽象度が低く、最も詳細な情報に対応することになります。同様に、「細胞外膜」は「細胞外皮」の一部です。 “part of”関係を横断することは、スケールの変化と解釈できます。一般に、根に近い用語（BP、CC、MF）はより一般的であり、葉に近い用語はより具体的です。

GO Consortium’sの用語で言えば、親より子の方が専門性が高いです。遺伝子にGO termsを付与する際には、可能な限り詳細な情報を付与するように努めています。一般的な言い方をすれば、最も抽象度の低いレベルに相当します。例えば、ある遺伝子がホルモンに応答してアポトーシスを誘導することが知られている場合、その遺伝子は「ホルモンによるアポトーシス誘導」という用語でアノテーションされ、単に「アポトーシス誘導」や「アポトーシス」といった上位の用語の1つではありません。



ある遺伝子にある用語がアノテーションされている場合、GOの構造から推測できるすべての推論も成立しなければなりません。言い換えれば、子用語が遺伝子産物を記述する場合、その親用語もすべてその遺伝子産物に適用されなければなりません。先ほどのontologyをもう一度確認してみましょう（ここでは便宜上、図を繰り返しています）。例えば、遺伝子が「プログラムされた細胞死の調節」の役割を持つとアノテーション付けされている場合、それは必然的に「細胞プロセス」に関与しているはずです。なぜなら、「プログラムされた細胞死の調節」は「プログラムされた細胞死」を調節しており、それは「細胞死」の一種であり、「細胞プロセス」の一種だからです。この性質は、**"True Path Rule "**と呼ばれています。しかし、これにより、GO解析が独立して行われる場合、各用語ごとに、differentially expressed geneは複数回数、その遺伝子がアノテーション付けられている最も低い用語からルートまでのすべての用語について１回ずつ数えられることを意味します。このことは、2つの結果をもたらします。 第一に、このプロセスには非常に多くの冗長性があり、第二に、表現型研究に関してあまり有益でない一般用語が大量に報告される傾向があります。

これに対応するために、Alexa et al (2006)によって2つの方法が提案されています。**Elim**は、GOの最も低いレベル、最も具体的な用語から開始し、そのエンリッチメントのp値を計算します。それが有意でない場合、アプローチは階層を上げ、通常通りp値を計算します。「細胞外シグナルによるアポトーシスの誘導」が有意でない場合、その親である「アポトーシスの誘導」についても、2つの用語が独立しているものとして、それに関連するすべてのDE遺伝子がカウントされることになります。しかし、より具体的な用語が本当に有意である場合、elimはそれに関連するすべての遺伝子をすべての祖先から排除するため、冗長性を排除し、より具体的な用語が有意であると報告される機会が与えられます。

Alexa が提案する 2 番目の方法は**weight**と呼ばれます。
このアプローチの背景にある考え方は、多くの非常に特殊な用語が有意であり、それらの用語をすべて包含するようなやや一般的な用語がある場合、このより一般的な用語を特定することが有益である可能性があるというものです。

Gene Ontology解析の落とし穴

Gene Ontology解析は強力なツールですが、他の強力なツールと同様に、誤用や誤解が生じる可能性があります。

Gene Ontology解析で最も多い間違いは、誤った背景を選択すること（あるいは明確な背景を選択しないこと）です。

エンリッチメント解析に戻りましょう。1,000個の遺伝子をパネルで測定し、200個の遺伝子が差次的に発現していることを発見したとします。これは20%の割合になります。では100個の遺伝子に関連する生物学的プロセスを考え、この100個の遺伝子のうち30個が差次的に発現していることを発見したとしましょう。この場合、もしこのプロセスが表現型と関係がないのであれば、その遺伝子の約20%が偶然に、つまり20個ほど差次的に発現していると予想できます。20 ではなく 27 が見つかりました。これが偶然に起こる確率は、約 0.076 または 7.6% であると計算できます。これは、通常受け入れられている1%、あるいは5%という有意水準に満たないので、このプロセスが表現型に関与していることを示す十分な証拠はなく、このプロセスの研究にあまり時間を費やすべきではないでしょう。

しかし...ほとんどの人は、この解析を上記のような手動ではなく、何らかのソフトウェアで行っています。また、そのようなソフトウェアでは、ユーザーが差次的に発現した遺伝子のリストをアップロードするだけでよいこともあります。これが200個の遺伝子のリストとなるのです。もしこのようなソフトウェアを使用していて、1,000遺伝子のパネルだけを使用したことを指定していない場合、すなわち、この分析の統計的背景から、ソフトウェアは、あなたがゲノム全体のRNA-Seq実験からこれらの200遺伝子を選択したと考えるかもしれません。その場合、数値はかなり異なるものになります。現在、DE遺伝子の割合は200/30,000、つまり2/300（または0.0066）であり、100個の遺伝子を持つ生物学的プロセスは、最大で1個の遺伝子を偶然に得るだけだと思われます（実際には0.666・・）。偶然に25個の遺伝子が存在する確率、つまり報告されているp値は、実質的にゼロです。

このような非常に有意なp値の場合、この生物学的プロセスの実験への関与を理解または検証しようとして数週間または数か月を費やす可能性があります。しかし、実際には、この非常に有意なp値は背景セットの誤った選択によるものであるため、注意が必要です。

この話の教訓は、エンリッチメント解析は常に測定する遺伝子のセットを明示的に指定して行わなければならないということです。最も安全な方法は、測定した遺伝子のリスト全体をアップロードし、発現差があるとみなしたい遺伝子を指定できるソフトウェア プラットフォームを使用することです。それができない場合は、参照セットを別途アップロードするか、別の方法で指定する必要があります。

この分野で2番目に多い間違いは、**多重比較の補正**を失敗していることです。この補正の必要性については、本書第16章で解説しています。 Ontologyは階層的な関係を持つため、適切な補正係数を適用して誤差を最小限に抑えることが重要です。例えば、FDR（False Discovery Rate）やFamily-wise Error Rateの補正係数を使用することは、GO解析には適していない可能性があります。

3番目に多い間違いは、様々な生命現象間の**クロストークを検出できない**ことです。Gene Ontologyの階層構造からくる冗長性に加え、多くの遺伝子が複数のプロセスに関与しています。時には、同じ群のDE遺伝子が複数のプロセスを有意に見せることがあります。このオーバーラップとクロストークが検出されて除去されない限り、多くの時間が無駄になる可能性があります。ここでは、クロストークを検出し、除去するための詳細な数学的アプローチを提案しています。より実践的でユーザーフレンドリーなクロストークの識別方法も、iPathwayGuideにに含まれています。

その他、Gene Ontology解析にまつわるわずかな間違いもあります：

- アノテーションを誤って解釈してしまう
    - ND = no biological data available.解析の結果、NDエビデンスコードが見つかった場合、追加の文献検索に時間を割く必要はありません。
    - NOT = it can be confusing to interpret the negative annotation NOT. NOTは完全なリストではありません。予期せぬ場所でNOTが現れたリストです。
- GOの出力であるdirected acyclic graph（DAG）の矢印の向きを誤認する。ここでは、明確なラベリングが必須です。

これらおよびその他の落とし穴についてのより詳細な説明は、この文書に記載されています。

Rhee, S., Wood, V., Dolinski, K., Draghici, S. (2008). Use and misuse of the gene ontology annotations Nature Reviews Genetics 9(7), 509-515. https://dx.doi.org/10.1038/nrg2363.

そして最も重要なことは、遺伝子発現実験の最終的な目的は、数字ではなく生物学的な知識を得ることであることを心に留めておくことです。

iPathwayGuide を使用したGene Ontology解析

iPathwayGuideは次世代シークエンサーやマイクロアレイで得られた遺伝子発現データを使用してパスウェイやGene Ontology情報を特定するソフトウェアです。他のパスウェイ分析ソフトウェアとは異なり、遺伝子相互作用を加味した高度なアプローチを採用しています。

遺伝子発現解析について何ができる？
パスウェイ解析、 Gene Ontology解析、薬剤・疾患・上流調整因子などの予想機能、バイオマーカーの探索、ネットワーク図作成　など

⇒製品詳細情報

【お問い合わせ先】
フィルジェン株式会社　バイオインフォマティクス部
TEL：　052-624-4388 (9:00～17：00)
FAX：　052-624-4389
E-mail：　biosupport@filgen.jp