様式 C-19

# 科学研究費補助金研究成果報告書

平成２２年　６月２１日現在

研究種目：若手研究（B）
研究期間：　2008　～　2009
課題番号：20700268
研究課題名（和文）
部分順序情報を用いた遺伝子発現量検索エンジンの構築
研究課題名（英文）
Constructing gene expression search engine with partial rank information
研究代表者
瀬々　潤　（ SESE JUN ）
お茶の水女子大学・大学院人間文化創成科学研究科・准教授
研究者番号：　40361539

研究成果の概要（和文）：
近年マイクロアレイをはじめとする網羅的遺伝子発現量の観測が盛んになり，データベースには 40 万件を超える量の遺伝子発現量が蓄積されている.その一方でその検索インタフェースは実験者の付けた注釈情報からの検索が主であり，注釈が適切な単語でない限り検索難しい．本研究では，遺伝子発現量を基に１．特徴ある機能情報の検索，２．類似した遺伝子発現量の検索を実装する事で，異なる視点からの検索手法を提案した.

研究成果の概要（英文）：
Recent progress for observation of comprehensive gene expression such as microarray produces many experimental results. The number of records in microarray database is now more than 400,000 experiments. The database provides the search interface by using keywords, but it is difficult to select proper keywords because of the difference of keywords between experimental scientists and users. To overcome these problems, we studied (1) a search method to characterize gene expression with their functions from user-given gene expressions (2) an implementation of the Web site to search the nearest gene expressions.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２００８年度 | 1,900,000 | 570,000 | 2,470,000 |
| ２００９年度 | 1,400,000 | 420,000 | 1,820,000 |
| 総　計 | 3,300,000 | 990,000 | 4,290,000 |

研究分野：情報学
科研費の分科・細目：生体生命情報学
キーワード：遺伝子発現・情報推薦・遺伝子オントロジー

## １．研究開始当初の背景

近年マイクロアレイ及び次世代（高速）シークエンサの登場により，遺伝子発現量の観測が盛んに行われている．2010 年 5 月末次点で

NCBI Gene Expression Omnibus（以下 GEO）には 44 万件を超す網羅的発現量が採取されている．また，EBI Array Express にも同様に発現量データが蓄積されている．このように大規模なデータが蓄積されている一方で，これらの Web サイトで提供されている検索機能は実験 ID によるもの，及び，情報提供者が書き込んだアノテーションによるものである．この検索機能は十分とは言えず，ID による検索は目的の ID が分からないと検索する事が出来ない．情報提供者のアノテーションによる検索では，実験者が意図する用語とデータベースに保持されている単語が一致するケースでは，検索が可能であるが，実験結果が必ずしも意図しない反応が含まれていたり，分野が異なる事による用語の違いがあったりするために，適切な結果が発見出来ないケースが想定される．

大規模にマイクロアレイ実験を行える大規模プロジェクトであれば，実験結果のクラスタリングや実験環境に関連した情報を用いてクラス分類手法を用いる事が可能である．しかし，たとえば患者から採取した遺伝子発現量を用いて診断を行うと考えた場合は，サンプルは１つであり上記の様な手法は用いることができず，また，小規模の研究室では数サンプルの実験を行うことが限界である．本研究で提案する検索エンジンを利用する事で，現在データベースに存在するマイクロアレイの発現量も利用して研究室で得たマイクロアレイの結果を解釈できる可能性が有る．

## ２．研究の目的

本研究では，現在 GEO や Array Express に蓄積された遺伝子発現量に対し，既存の文字列検索によるデータ検索だけでなく，新たな検索手法として，発現量をもとにした検索手法を提供する．特に，遺伝子発現量を基にしてその実験で特徴的な遺伝子機能を発見するために，遺伝子オントロジーの検索，及び，類似した遺伝子発現量を観測する手法として，遺伝子発現量自身の検索という２種類について検索エンジンを構築する．

## ３．研究の方法

本研究では，以下の３つの順に研究を行った．

A. 遺伝子オントロジーを検索する Web サイトの実装と評価
B. 遺伝子発現量を検索する Web サイトの実装
C. 検索の高速化に向けた次元圧縮及びインデックス手法の検討

遺伝子発現量を観測した際に問題となることの一つは，その実験により細胞内のどの機能が変化したかを検出する事である．既存のサイトでは着目する遺伝子セットをユーザが決定して入力する必要があったが，本研究では，遺伝子とその発現量をペアで入力する事で，ユーザが遺伝子を選択する必要を無くし，発現に変化の起こった機能を自動で抽出できるようにした．どの機能に変異が起こっているかは遺伝子オントロジーで提供されている機能と遺伝子の対応を用いた．計算に際し４点の工夫を行った，(1)計算は遺伝子発現量を順位に直し，順位統計量を用いた．(2) 1 万個近いタームに対する順位統計量を高速に計算するため，重複した計算を省く計算手法を導入した．(3)結果の提示に際しては，全ての関連あるオントロジータームを表示するのではなく，遺伝子オントロジーのタームは DAG 構造をしていることに着目し，関連の深い部分 DAG 構造のみを抽出した．(4)関連する機能が多い場合，DAG 構造を単純に

表示すると非常にタームとタームを結ぶ線に交差の多い図が現れる．また，既存のグラフ描画手法では，DAG の階層構造が崩れて分からない事が多い．そこで，本研究では，階層構造を保ったままターム間の関係を見やすくする描画手法を提案した．図 1 に本システムの概要を示す．また，図 2 に本システムを用いて描画した DAG 構造のサムネールを示す．システムでは，このサムネールをクリックすることで，詳細を記した大きな遺伝子オントロジーの階層構造と関連した遺伝子の詳細を表示することが可能である．

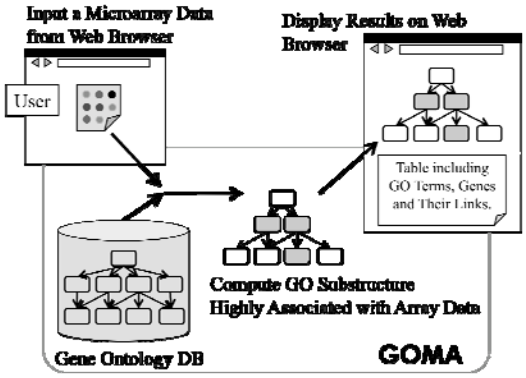



図 1. 遺伝子オントロジーの検索と表示システムの概要

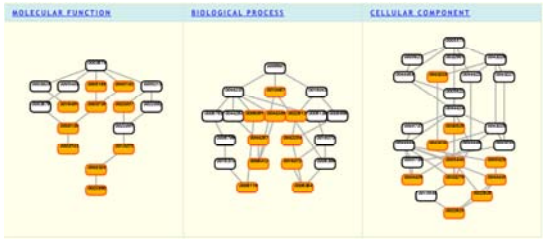



図 2. 遺伝子オントロジー表示の例

２．遺伝子発現量を入力として，類似の遺伝子発現量を持つサンプルを検索するサイトの構築を行った．図 3 はその検索結果ページを記しており，最も入力した遺伝子-発現量に類似した実験を表示している．各行が実験を記しており，列は左から順位，実験名称，類似度，類似度を記した棒グラフである．

事前の酵母データを用いた少量の実験では順位相関係数の方がピアソンの相関係数に比べて精度が高かったが，酵母及びヒト 8700 サンプルに関して調査したところ，ピアソンの相関係数の方が少し良い精度の高い傾向を示した為，ピアソンの相関係数を用いて構築した．

本サイトの構築により数点の問題点が明らかになった．まず，速度面である．実装上の工夫を含め，順位相関係数及びピアソンの相関係数を用いて計測をしたが，数分の待ち時間を要する状態であった．検索エンジンとしては長くても数秒単位での待ち時間に押さえたく，速度改善の必要があった．次に，精度である．非常に高い相関を持つサンプルに関しては精度（F 値を用いて計算した）が高いものの，相関が少し悪くなると突然精度が悪くなる傾向にあった．これら 2 点に関し，前者の問題は低次元への射影を用いた改善を実験し，前者と後者の問題に同時に対処するため多次元インデックスと k 最近傍グラフを組み合わせた検索手法を行った．

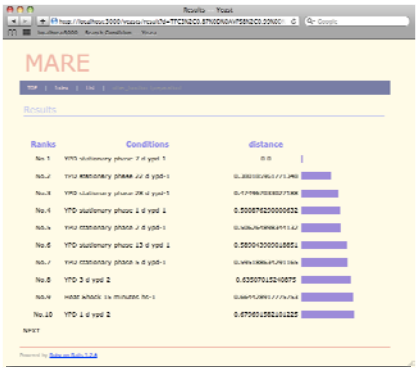



図 3. 遺伝子発現量を検索するサイト

3.

高速化の手法として D. Achliptas(J. Comp. and Sys. Sci, 2003)のランダムプロジェクションを用いた次元削減手法の実装を行った．これは，一定の法則に従った射影行列を作成して射影した次元における距離が，確率的に一定値以下と成ることが理論的に保証される手法である．まず第一に，この射影行

列を用いて射影した空間上での距離を持ちてピアソンの相関係数を計算した．次元数を 1 から 50 次元まで変更し F 値を計算したが 0.6 以下と低い値であり，射影が有効に効いていない様子がうかがえた．遺伝子発現量データが Achliptas の仮定するデータの分布から外れている事が原因である事が調べた結果分かった．

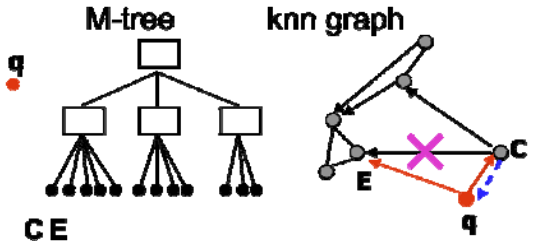
次に，非常に近傍であれば検索精度が高い点を用い，k-近傍グラフを用いた検索を行った．K-近傍グラフを用いた検索とは，与えられたクエリに対し最も近い点を求める．次に，k-近傍グラフ上を順に辿る形で近傍の点を求める事である．グラフ上を順に辿る事でピアソンの相関係数とは異なり空間内の距離が局所的に異なる様なデータに対しても対応可能である．予め全ての実験に対して k-近傍グラフを作成しておくことで，最も近い点間を高速に知る事ができる．

しかし，この k-NN グラフ手法では検索は速くならない．最も問題となるのは，最近点となる点を探索する事である．そのため，本研究では，多次元インデックスである M-tree を用い，最近傍の探索を高速化した．これにより，検索時間を 1 秒未満で終了する事が可能となった．

更に，検索精度をピアソンの相関係数を用いた全探索と比較したところ，非常に類似した実験の近さは同一であるが，類似したサンプルが 10 番目以降となると，k-NN グラフを用いた方がピアソンの相関係数より F 値が良い結果となった．これは，遺伝子発現量の空間上での点の密度が実験によって異なるため，近傍グラフを辿る事でその変化を検出でき，よりよい結果が得られたと考えられる．

## ４．研究成果

本研究では，遺伝子発現量を入力として，その実験に特徴的な機能及び類似した実験を検索するサイトの構築，また，類似した実験




に関しては高速化を試みた．

## ５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

①Eriko Muzutani and Jun Sese. GOMA: Web Utility for Direct Finding of Enriched Gene Ontology Terms from Gene Expression Profile. *8th IEEE International Conference on BioInformatics and BioEngineering (BIBE2008)*, Athens, Greece, Oct. 8-10, 2008. 査読有

〔学会発表〕（計 7 件）

全て査読無

①距離木と k 近傍グラフを用いた超高次元データの近傍検索．ユスフ ムカルラマー，渡辺 知恵美，瀬々 潤 (DEIM, 淡路島, 2010/3)
②NearestNeighbors for a High-dimensional Data ユスフ ムカルラマー，渡辺 知恵美，瀬々 潤 (SIGBIO, 東京, 2009/12)
③梅澤香矢乃，瀬々潤: MARE：遺伝子発現量検索エンジンの構築に関する一考察 (DEIM, 静岡, 2009/3)
④水谷枝理子，瀬々 潤: GOMA:遺伝子発現量の簡便機能解析 Web アプリケーション (分子生物学会, 神戸, 2008/12)
⑤水谷枝理子，瀬々 潤: GOMA:遺伝子発現量の簡便機能解析 Web アプリケーション (SIGBIO, 札幌, 2008/8)
⑥水谷枝理子，瀬々 潤: GOMA: 複雑な遺伝子オントロジーの分かりやすい表示(第 7 回オープンバイオ研究会)
⑦水谷枝理子，瀬々 潤: 遺伝子群の共通機能に着目した GO 表示の研究（第 69 回情報処理学会全国大会）

〔その他〕
ホームページ等
http://goma.sel.is.ocha.ac.jp/

６．研究組織
(1)研究代表者
瀬々　潤（ SESE JUN ）
お茶の水女子大学・大学院人間文化創成科学研究科・准教授
研究者番号：40361539