様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２２　日現在

機関番号：　１２６０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０３４２

研究課題名（和文）異質データの相関解析による潜在的概念モジュールの同定

研究課題名（英文）Identification of hidden concept modules using correlation analysis of heterogenious data

研究代表者

村上　勝彦（Murakami, Katsuhiko）

東京大学・医科学研究所・特任研究員

研究者番号：３０３４４０５５

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：ヒト遺伝子（タンパク質）に関するデータには様々なタイプがある。この複数データを統合したとき情報(term)は相互に関連している。本研究では、termの相関を解析しデータの潜在的因子を新たに定義することが目的である。遺伝子機能のtermについて相関検出を行うと、タンパク質の細胞局在情報が他の機能の情報と相関が高いことがわかった。さらに大規模化の方法を検討し、厳密な相関計算をせずに行列因子分解を適用する方法により大規模化が可能であることがわかった。同時に複数termで表現される複合概念を自動抽出できた。さらに遺伝子発現や変異等のオミックスデータに適用し、新規バイオマーカーの候補を得た。

研究成果の概要（英文）：It is not explicit what relationship exists between the various information (term) in the database. In this research, the purpose is to find a new relationship between information by integrated analysis. First, correlation detection of term information to describe gene function was performed, and it was found that the cell localization information of protein has high correlation with information such as other functions. As a result of investigating the method for enlarging the scale, we applied matrix factorization approach to make it large scale, by which we can skip for computation rigorous correlation factors. At the same time, complex concepts expressed by plural terms can be extracted automatically. When the technique was applied to omics data such as gene expression and mutation, we could find new candidates of cancer biomarkers.

研究分野：　生命情報

キーワード：　遺伝子機能　遺伝子変異　遺伝子発現　オミックス解析　バイオマーカー　データ統合　細胞内局在

１．研究開始当初の背景

ヒト遺伝子（タンパク質）に関するデータベースには遺伝子発現、機能、蛋白質間相互作用、立体構造、蛋白機能ドメイン、細胞内局在など様々なタイプがある。データベースとしても GEO（遺伝子発現データ）、総合的なものでは Uniprot、RefSeq、WikiGene など多岐にわたる。各データベースでは、アノテーションデータをグループ化して表示しているが、”Zinc finger motif”といった個別アノテーション情報（以下、term）を列挙していることが多く、各 term や記号を知っていることが前提となっている。システム生物学や発現解析では、遺伝子全体を対象とするため、未知の term に出会うことも多い。古くからある上位の概念は理解できるが、あらたな発見もあるので term を把握することは困難である。一方で、蓄積したこれらのデータを統合して可能となるような高度な解析手法が待たれている。そのためには情報の冗長性を排除し、意味的な関連性を整理することが必要である。

我々はヒト遺伝子データを高度に利用するために、遺伝子発現解析などで与えられた遺伝子セットを解析するシステムを公開していた(Takeda, 2013)。これは入力の遺伝子セットには、どんな共通の機能・モチーフを持ったものが多いか、という疑問に答えるもので、有意に多いアノテーションを提示するシステムである。このような解析は Gene Set Enrichment Analysis (GSEA)と呼ばれ、多くのバリエーションがある。これらは遺伝子セットに特徴的なアノテーションデータを列挙するに留まっている。そのため、研究者がそれらの関係性を含めて正しく解釈するのは、全遺伝子や現象についての知識がなければ困難である。

２．研究の目的

本来生命現象は化学反応が個別に起こるものではなく、関連して起こる一連のイベントの総体である。よって、多くの term 間には関連があることが多い。これらを機械的に大量に関連付けて整理することができれば GSEA にも応用でき、妥当な部分や意外な部分や、本質的な部分を考察し判断することが容易となる。

本研究では、term の相関を解析すること、およびデータの潜在的因子を新たに定義することが目的である。

３．研究の方法

ヒト遺伝子（タンパク質）に関する各種データベースである UniProt、RefSeq、H-InvDB 等からヒトに関する疾患、蛋白質間相互作用、立体構造、蛋白機能ドメイン、細胞内局在等のアノテーション情報（term）を収集した。当初は term 同士でフィッシャー検定による相関検出を行い、関連する term を調べた。一定数のデータでは成功したが、この方法では計算時間がかかり、大規模化は困難であった。そのため複数の方法を検討して、疎行列に対する行列分解を用いる方法を採用することとした。

以下では、統合されたデータベースの遺伝子 n 個と、それらに付与されていた term m 個から、n 行 m 列の行列 V を作成した。この行列要素は、遺伝子 i に term j が付与されていれば V(i, j)が 1、付与されていないならば 0 となる行列である（図 1 左）。この行列に対して解析することで、遺伝子と term からなる関係性を見いだそうということである。

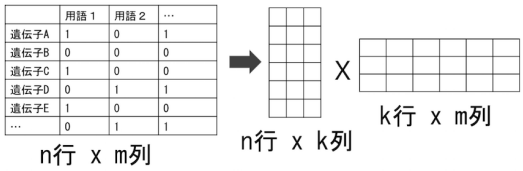

図１　gene-term 行列の構成と行列分解(左から行列 V, W, H)

NMF は解析対象のデータ行列 V を、図のように基底行列 W と係数行列 H の積の形に近似的に分解するものである（図１）。 ここで、V, W, H についてすべての行列要素が非負値である。行列の次元は、V が n×m、W が n×k、H が k×m である。また、k は基底ベクトルの数で、解析時に与える。行列分解後、基底ごとに各行列の要素値の大きさから成分分析を行う。すなわちランクごとに、縦軸（遺伝子等の行ラベル）と横軸（細胞等の列ラベル）の要素値の大きいカテゴリ群をクラスターとして取り出す（ソフトクラスタリング）。同じクラスターに複数の gene, term が属するような結果が得られる。

また、がんゲノムパネルデータに対して本手法を利用すれば、あらたな関係性を見出すことができ、バイオマーカーを探索することが出来ると考え研究を行った。そこでは、The Cancer Cell Line Encyclopedia (CCLE)から 1,055 個の細胞株について、薬剤感受性、変異情報、コピー数多型, mRNA の発現量、および腫瘍タイプ（細胞株の系統）のデータを取得した。このデータから解析対象となる行列（分解前）を作成するが、上記の項目のうち、行列の行に細胞株名が並び、のこりのカテゴリは各列に並べた。要素値の数値は 0,1 のバイナリか、0 から 1 までの連続値に変換（規格化）した。その後 NMF の解析を行った。このデータでは欠損値が多いため、マスク行列を作成する対策をほどこした。

４．研究成果

生命系データベースのアノテーション項目について、当初は term 同士でフィッシャ

ー検定による相関検出を行い、関連する term を調べた。これから「DNA 修復」と「ヌクレオソーム構造」などの関連する情報を取り出せることがわかった。さらに、大きなカテゴリ間で関連性をみると、たんぱく質の細胞局在情報が他の機能や立体構造などの情報と相関が高いことがわかった。

次に大規模化のための方法として、非負値行列因子分解（NMF）を適用する方法により大規模化が可能となった。同時に複数 term で表現される複合概念を自動抽出できることがわかった。例えば「核内で DNA に結合することで転写制御を行う」という複合概念が自動抽出できた。この結果は、term 間に隠れていた潜在的概念、高度な抽象概念の自動獲得に成功したということになる。

さらに新規バイオマーカーの候補を得ることを目的に、遺伝子発現や変異等のオミックスデータに本手法を適用した。その結果、BRAF 阻害剤感受性と，悪性黒色腫（メラノーマ），BRAF 変異，MITF 活性度が同じクラスターとして関連が指摘された。Ingenuity Pathway Analysis (IPA)解析を行うと、新規の MITF の活性度がバイオマーカー候補であることの状況証拠を複数得ることが出来、本手法のマルチオミックスデータに対する有効性を確認することができた。

今後は、得られた gene,term の関連データを GSEA の発展や予測ツールに応用していくことが考えられる。また、関連付けの手法をさまざまなデータに応用して新たな関係性の発見や予測に適用することが期待される。

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　5　件）

(1) Fujita, N., Mizuarai, S., Murakami, K. & Nakai, K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses, Scientific Reports, 査読有, 2018, in press.

(2) Murakami K, and Sugita M. “Evaluation of Database Annotation to Determine Human Mitochondrial Proteins” International Journal of Bioscience, Biochemistry and Bioinformatics, 査読有 2018, in press, http://www.ijbbb.org/

(3) Takamatsu K, Bannaka K, Kirimura T, Noda I, Murakami K, Mitsunari K, et al. Tag-based knowledge network models. Bull Kobe Tokiwa Univ. 2017;10:51–60. 査読有

(4) Takamatsu K, Murakami K, Lim R-JW, Nakata Y. Novel visualization for curriculum in silico using syllabus by a combination of cosine similarity, multidimensional scaling methods, and scatter plot: Dynamic curriculum mapping (DCM) for syllabus. Bull Kobe Tokiwa Univ. 2017;10:99–106. 査読有

(5) K. Murakami, "Extraction of latent concepts from an integrated human gene database: Non-negative matrix factorization for identification of hidden data structure," 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2015, pp. 346-350. 査読有, DOI: 10.1109/SOCPAR.2015.7492771

〔学会発表〕（計　13　件）

(1) Takamatsu K, Murakami K, Kirimura T, Bannaka K, Noda I, Omori M, et al. A new way of visualizing curricula using competencies: Cosine similarity, multidimensional scaling methods, and scatter plotting. 6th Int Conf Data Sci Institutional Res (DSIR 2017). 2017]

(2) Murakami K (2017) Prediction of human mitochondrial proteins from various resources（Intelligent Systems for Molecular Biology (ISMB), 2017）

(3) 村上勝彦、「PubMed Central 論文ネットワークの解析」第 38 回日本分子生物学会年会、2016

(4) 村上勝彦「A tool and analysis of citation network」生命医薬情報学連合大会、2016

(5) Katsuhiko Murakami “Clustering of gene ontology annotation by matrix factrization”, Intelligent Systems for Molecular Biology, 2016

(6) 村上勝彦「遺伝子オントロジーの階層的クラスタリング」人工知能学会全国大会 2016 年

(7) 村上勝彦（2015） 遺伝子データベースアノテーションのクラスタリング（第 38 回日本分子生物学会年会、第 88 回日本生化学会大会 合同大会）

(8) 村上勝彦 Identification of latent factors in gene databases using non-negative matrix factorization（2015）生命医薬情報学連合大会

(9) 村上勝彦 （2015）行列因子分解による遺伝子データからの潜在的因子の抽出（2015 年度人工知能学会全国大会（第 29 回））

(10) 村上勝彦、間宮健太郎、世良実穂、今西規 (2014)，ヒト遺伝子データベース H-InvDB を中心とした RDF データ統合，第 37 回日本分子生物学会

(11) Murakami K, Imanishi T. (2014), Construction of human gene and protein

ontology which connects related concepts from different data sources, ISMB2014. (Boston, USA)
(12) 村上勝彦, 今西規 (2014) 高度な推論にむけたヒト遺伝子データベース H-InvDB の RDF 化, 生命医薬情報学連合大会
(13) 村上勝彦, 今西規 (2014), 遺伝子データからの相関する概念抽出と関係づけオントロジーの作成, 人工知能学会

〔図書〕(計 2 件)

(1) 村上勝彦、他 (図書)「医薬品・医療機器・再生医療開発におけるオープンイノベーションの取り組み事例集」2018 年 共著
(2) 村上勝彦、他 (図書)「in silico 創薬におけるスクリーニングの高速化・高精度化技術」2017 年 共著 ISBN：978-4-86104-688-9

〔産業財産権〕
○出願状況 (計 0 件)
○取得状況 (計 0 件)

6．研究組織
(1)研究代表者
村上 勝彦 (KATSUHIKO MURAKAMI)
東京大学・医科学研究所・特任研究員
研究者番号：30344055