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　２９　年　　６　月　１８　日現在

機関番号：　３２６２０
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６８７０１７５

研究課題名（和文）部分最小二乗回帰を用いたシステム創薬戦略データベースの構築

研究課題名（英文）System-based drug discovery database using partial least square regression

研究代表者
茂櫛　薫（MOGUSHI, Kaoru）

順天堂大学・医学（系）研究科（研究院）・助教

研究者番号：６０５６９２９２

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：肝細胞癌は国内および世界的な統計でも、癌死の原因として上位に位置している。本研究では、網羅的遺伝子発現データに加え、臨床病理学的因子をはじめとするさまざまなデータを融合し、システム創薬の実現に向けたデータベースを構築することを目指した。多変量データ同士の回帰を行うことができる部分最小二乗法(partial least square regression: PLSR)を用い、臨床病理学的因子などの付随情報を目的変数群、遺伝子発現データを説明変数に取り、これらの相関関係をデータベースとして活用することを目的とした。

研究成果の概要（英文）：Hepatocellular carcinoma (HCC) is one of the most common cause of cancer-related death in the world. There is an urgent need for novel drug other than sorafenib for treatment of HCC. In this study, we aimed to develop multi-dimensional database for system-based drug discovery in HCC using various clinicopathological characteristics and gene expression profiles obtained from cancer tissues. Partial least square regression (PLSR) was used to analyze complex correlation between gene expression patterns and clinicopathological features.

研究分野：　クリニカルバイオインフォマティクス

キーワード：　肝細胞癌　遺伝子発現解析　バイオインフォマティクス　部分最小二乗回帰　多変量解析　シグナル伝達

１．研究開始当初の背景

近年の遺伝子解析技術の進展により、細胞内に発現している転写産物の包括的解析(トランスクリプトーム解析)が広く行われるようになった。2000 年代には、DNA マイクロアレイを用いたトランスクリプトーム解析が盛んに行われてきた。最近では次世代シーケンシングによる解析の低価格化により、RNA の塩基配列を直接シーケンシングしてリード数をもとに定量する RNA-seq を用いた解析が広まっている。また、疾患と遺伝子発現の関連性の解析手法やモデル化、分子ネットワーク解析、そして創薬への応用方法については、「システム創薬」として国内外問わず広く研究されている。がん研究の分野でも活発に行われており、システム創薬に基づく新しい新規薬剤の開発が期待される。

さて、肝細胞癌(hepatocellular carcinoma: HCC)は国内および世界的な統計でも、癌死の原因として上位に位置している。外科的切除術が有効であるものの、感染した B 型および C 型肝炎ウイルス(HBV、HCV)や肝内転移のため再発が多く、予後が不良であることが知られている。またウイルス性肝炎だけでなく、最近ではメタボリック・シンドローム関連因子(高血圧、高血糖、高脂血症、肥満など)に起因する非アルコール性脂肪性肝炎が肝細胞癌の原因として増えており、新たな問題となりつつある。

肝細胞癌の治療薬は現段階では非常に限られており、近年では血管内皮細胞増殖因子レセプターVEGFR2/3 や血小板由来成長因子レセプターPDGFR などを同時にターゲットとするマルチキナーゼインヒビターであるソラフェニブが HCC で初めて分子標的薬として承認された。他にも分裂期キナーゼの１つである Aurora kinase B の阻害剤など、いくつか画期的な HCC の治療薬の開発が進められているものの、治療戦略の選択肢を増やす、あるいは患者の予後をさらに改善する上で、継続的な新薬開発は急務である。

一方、肝細胞癌をはじめとした患者検体を対象とし、遺伝子発現情報などの分子情報や臨床病理学的因子等の患者情報を収集した Integrated Clinical Omics Database (iCOD) などのデータベースが存在している。近年ではがんゲノムまで含めたデータベース化も進んでおり、The Cancer Genome Atlas (TCGA) だけでも 300 症例以上の肝細胞癌に対する体細胞変異、コピー数異常、RNA-seq による発現情報などがデータセットとして公開されている。このような状況を鑑み、既存の公開データを活用して、新たな医学的知見を得るような枠組みの開発が必要とされている。

２．研究の目的

(1) データベースと解析システム構築

本研究では網羅的遺伝子発現データをもとに、臨床病理学的因子、シグナル伝達系や代謝系などのパスウェイ情報をはじめとするさまざまなデータを融合し、システム創薬の実現に向けたデータベース・アプリケーションを構築することを目指した。これにより、各種の臨床病理学的因子に相関する遺伝子群の解析・同定が可能になり、疾患機序の解明や、新たな創薬ターゲットが得られることが期待される。さらに、肝細胞癌をモデルケースとして確立したのち、将来的には他の癌や神経疾患等の難病などの疾患への応用も想定することとした。

(2) 本システムの応用のシーズ探索

構築したデータベースを活用し、医師や基礎系研究者とのディスカッションや共同研究を通じて、どのような解析デザインや機能が望まれているかの調査を行い、本システムの機能に反映させることを目的とした。

さらに、肝細胞癌の基礎および臨床研究に携わる医師・研究者が研究のアイディアを得ることができるよう、実践的に使えるデータベースを作成することを念頭におきながら、適宜ディスカッションを行いながら開発を進めることとした。

３．研究の方法

(1) 部分最小二乗回帰

本研究の核となるアルゴリズムには、多変量データ同士の回帰を行うことができる部分最小二乗回帰(partial least squares regression: PLSR)を用いた。PLSR は、同様な多変量データ同士の回帰を行う正準相関分析に対し、変数の数がサンプル数を上回った場合(遺伝子発現データは一般にこの状態である)においても、逆問題の正則化などの検討を行うことを必要とせず、より直接的な解析が可能であるという利点がある。

PLSR は多様な分野でのデータ解析に用いられている。例えば、代謝物の網羅的解析であるメタボローム解析、分析化学分野における赤外吸収スペクトルからの化合物解析、既知の化合物の構造や物性から新規化合物の生理活性などを予測する定量的構造活性相関(QSAR: quantitative structure-activity relationship)、MRI などの脳画像と神経症状を組み合わせた定量解析、顔認識や物体追跡などの画像認識など、さまざまな用途での応用が進んでいる。

また、一般的な多変量回帰において説明変数間の相関が強い場合に、係数の推定が不安定になる「多重共線性」がしばしば問題になる。PLSR の特徴として、多重共線性を持ったデータの回帰に優れていることが挙げられる。特に遺伝子発現情報を解析する際には、発現パターンが連動する共発現遺伝子群が存在するため、非常に有効である。例えば肝細胞癌の網羅的な遺伝子発現情報を解析すると、肝代謝に関わる酵素(シトクローム P450、アルデヒド脱水酵素、脂質代謝酵素など)や、細胞周期関連遺伝子群(サイクリン・ファミリー、DNA 複製因子、有糸分裂の調節因子等)、

浸潤したリンパ球に由来すると考えられるサイトカインや MHC クラス I および II 分子などの免疫関連の遺伝子群などが、それぞれ独立な成分として発現変動パターンのクラスターを作ることを経験している。このため、PLSR によって多重共線性を織り込んだ形でモデル化することで、類似した傾向を示す遺伝子群をグループ化して解釈を簡略化できる。さらに、いままで大量の遺伝子の中に埋もれていた、新規の共発現遺伝子群を同定できる可能性もある。
本システムでは、PLSR により臨床病理学的因子などの付随情報を目的変数群、発現データを説明変数に取り、それぞれの関連性を回帰係数および因子負荷量の形で得る(図 1)。これにより、それぞれ多変量のデータである遺伝子間や臨床病理因子間の類似性の評価が可能となる。さらに、その際に得られる因子負荷量の行列は共発現遺伝子群を表す情報となり、肝細胞癌のなかでどのような遺伝子が相関して変動するかの情報も得ることができる。また、より発展的な用途としては、回帰係数を分子ネットワーク上にマッピングすることによる可視化や、ある閾値やスコアを設けることで患者のフェノタイプを説明する上でのコアとなる遺伝子モジュールの抽出を行うことも可能であると考えられる。

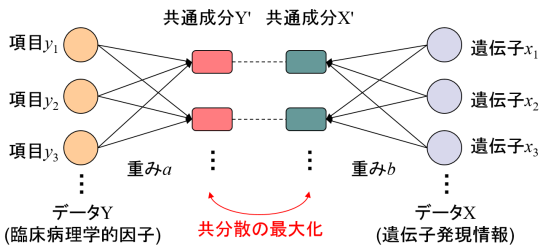

図 1. 部分最小二乗回帰の模式図

(2) 肝細胞癌の遺伝子発現情報
まず、 Affymetrix HG-U133 Plus 2.0 (54,613 プローブセット)により測定された、既存の 168 検体の肝細胞検体に対する遺伝子発現プロファイルのデータの整形を行い、統計処理言語 R 上で扱えるようにした。元となるマイクロアレイの生データ(CEL ファイル形式)は、当時 iCOD で登録利用者向けに公開されていたものを使用した。マクロアレイデータ の正規化は R の affy パッケージに含まれる Robust Multiarray Average (RMA)法を用いて実施した。また、得られた発現量のデータに対し、$\log_2$値の対数変換を行った。マイクロアレイデータ上に配置された各プローブに関わる遺伝子情報(アノテーション)の対応付けには、Affymetrix の NetAffx で入手できるカンマ区切りテキスト(CSV)形式のファイルを一括でダウンロードし、R 上で取り込みとデータ整形を行った。

(3) 臨床病理学的因子の整形と取り込み
臨床病理学的因子については、量的データとして手術時年齢と術前腫瘍マーカー値(AFP および PIVKA-II)を使用した。AFP と PIVKA-II は患者間で大きな開きがあるため、あらかじめ対数変換を行った。またカテゴリーデータとして、性別、HBV (陰性 vs. 陽性)、HCV (陰性 vs. 陽性)、多発性 (腫瘍数 1 個 vs. 2 個以上)、腫瘍最大径(5cm 未満 vs. 5cm 以上)、被膜形成 (fc)、被膜浸潤 (fc-inf)、門脈浸襲(vp)、肝静脈浸襲(vv)、ステージ(II 以下 vs. III 以上)、肝障害度(A/B/C のそれぞれに対する二値化)、肝硬変の有無、ミラノ基準(基準内 vs. 基準外)、手術時の腫瘍露出(sm)、遠隔転移の有無、の項目を用いた。カテゴリーデータを 0/1 データに変換して量的データと統合し、R 上に実数の行列として取り込みを行った。

(3) PLSR の実装の検討
部分最小二乗回帰の実装としてはいくつか存在するが、R の追加ライブラリである"pls"パッケージが、開発の成熟度の面から今回の用途では適切であると判断した。その後、図 1 に示したような各係数の推定を PLSR により実施した。得られた遺伝子および臨床病理学的因子間の相関関係の評価を行い、臨床的な解釈と一致するかどうかの検討を進めた。また R での条件検討における並列計算には、parallel パッケージを用いて引数を変更することで処理することとした。

4. 研究成果
(1) 計算環境の構築
本研究を開始するにあたり、計算システム選定と計算環境の構築を行った。まず、統計処理言語 R で既存の PC を用いて簡易的なプロトタイプを構築し、計算量の見積もりを行った。R の pls パッケージに含まれる PLSR のアルゴリズム自体は並列化されておらず、単独スレッドで動作するため、一回当たりの計算時間には CPU の動作周波数がそのまま影響することになる。一方、条件検討などの場合には複数のカットオフ値などを用いて並列的な評価を行う必要があるため、ある程度の CPU コア数も要求されることになる。しかし、一般に CPU コア数が増えると動作周波数の上限が低下することになるため、両者のバランスを適切に勘案する必要がある。最終的には予算との兼ね合いも検討し、本研究に適した構成のワークステーション (Intel Xeon E5-2680v3 のデュアル CPU 構成、64GB メモリ)を選定した。また計算サーバー用途のため、Linux 環境で用いることとした。

(2) 解析システムの準備と開発
発現データやアノテーション情報を MySQL データベースに投入し、解析環境の整備を行うことを検討した。しかし、これらのデータ(54,613 プローブセット× 168 検体)を SQL のクエリーで取得する際に時間が掛かり、デ

ータの受け渡しやデータ形式の変換などで発生したオーバーヘッドが速度の低下を招いていると考えられた。結果的にはR専用の保存形式のファイルを実行時に直接読み込む場合のほうが 1〜2 秒程度のオーダーで終了し、動作が高速であった。
また、整形した肝細胞癌のデータセットに対し、臨床病理学的因子などを含む患者情報と、DNA マイクロアレイで取得した遺伝子発現を用い、PLSRを用いて多変量データ同士の回帰を実施し、解析・解釈を行った。計算システムの構築を行うとともに、部分最小二乗回帰の実装方法について検討した。統計処理言語R および pls パッケージを用いてプログラムを作成し、開発と評価を進めた。
しかし、解析対象とする臨床病理学的因子の絞り込みや、遺伝子の選択条件などの閾値(四分位偏差などの指標による発現変動が少ない遺伝子の除外、ノイズの原因となりうる低発現遺伝子の除去など)、PSLR で用いるコンポーネントの数によっても得られる結果が変化する。このため、ユーザー自身がいくつかの条件を指定して簡便に再解析が可能な機能の実装を検討した。そこで、R 言語において対話的なウェブアプリケーションを構築するためのライブラリである"Shiny"を用い、ユーザーが指定した条件を用いて候補遺伝子を選定する方法の検討を進めた。また公開にあたって、それなりの計算が実行できるレンタルサーバーもしくはクラウド・コンピューティング環境が必要になるため、Amazon EC2などのサービスの調査と評価を行った。
本研究の当初の目標では平成 28 年度に本システムの公開までを目標としていたが、研究期間内での公開には至らなかった。研究期間終了後も開発を続け、近い将来の公開を目指したい。また、肝細胞癌のみならず、治療が難しい他の疾患にも同様な創薬基盤データベースを展開することにより、新たな治療戦略の開発に寄与できるものと考える。

(3) 本課題に関連した派生研究
本課題を実施する過程で構築したデータセットの活用を試みた。以下のような研究にも携わり、医師のクリニカル・クエスチョンに対して遺伝子発現情報の臨床的意義の解釈を深めることで、本システムで得られる知見へのフィードバックの可能性を検討した。

① 非ウイルス性肝細胞癌のリスク因子[引用論文1]
肝炎ウイルスに起因しない肝細胞癌(非ウイルス性肝細胞癌)のリスク因子としてはメタボリック・シンドロームが知られている。その指標として糖尿病の有無と BMI に着目し、それぞれと発現パターンが相関する遺伝子のスクリーニングを実施した。その結果、結合組織増殖因子(CTGF: connective tissue growth factor)をはじめとするいくつかの遺伝子が有意な相関を示した。また CTGF 高発現群では低発現群より予後が悪く、肝細胞癌の予後予測因子としても有用であることが明らかになった。

② 肝細胞癌の遺伝子発現と造影剤に対する感受性の関連[引用論文2]
近年、肝細胞癌の画像診断において、ガドリニウム造影剤である Gd-EOB-DTPA が用いられているが、患者によって癌部での EOB 造影剤の取り込みの度合いに違いがみられることが知られている。そのメカニズムの解明に焦点を当て、造影剤の取り込み亢進群とそれ以外の群で癌部における遺伝子の発現パターンの変動を検討した。その結果、有機酸のトランスポーターの一つである SLCO1B3 (solute carrier organic anion transporter family member 1B3)と正の相関を示していることが明らかになった。SLCO1B3 は OATP-8 としても知られており、ビリルビン輸送などの胆汁酸の産生に関わることが知られており、特に肝臓での発現が高い遺伝子であるため、造影剤の取り込み亢進群は肝機能を保持した高分化な癌細胞であると考えられる。

③ 肝細胞癌の切除時年齢と関連する因子[引用論文3]
肝細胞患者の切除時の年齢に対し、さまざまな臨床病理学的因子に加え、どのような遺伝子発現パターンの差異が見られるかを検討した。癌部・非癌部のそれぞれにおいて年齢と相関する遺伝子群の解析を行うとともにパスウェイ解析を実施した。その結果、高齢群での癌部での PI3K パスウェイの亢進と、非癌部での線維化に関わるパスウェイの抑制を見出した。

(4) 解析方法の改善に向けた検討
本システムに用いる遺伝子発現情報や臨床病理学的因子などのデータは基本的に正値である。このため近年注目されている、正の値の制約を設けたモデルである非負最小二乗法(non-negative least squares)や非負値行列因子分解(NMF: nonnegative matrix factorization)に基づく回帰をもとにした解析方法についても、今後の応用について調査・検討を進めた。
まだ完全な実装には至っていないが、非負値の制約を課したモデルを用いた場合、ある共通成分に対して逆相関を示すもの(係数の符号を反転すればほぼ一致するような成分)も独立な成分として得られるケースが多く見られた。したがって PLSR 法のほうが正・負の相関が集約された成分を持つモデルが得られることが分かった。
このため、ある条件による亢進遺伝子群・抑制遺伝子群といった発現変動の方向性まで分離して検討する場合にはもちろん有用であるが、考慮すべき要素の数が最大で倍になるため、そのぶん解釈は煩雑になる。したが

って現段階では、本システムの目的としてはPLSRが適していると思われるものの、ユーザーの解析内容に応じてアルゴリズムを切り替えられるようにするなどの手段もあるため、継続して評価を進めたい。

＜引用文献＞

[1] Akahoshi K, Tanaka S, Mogushi K, Shimada S, et al. Expression of connective tissue growth factor in the livers of non-viral hepatocellular carcinoma patients with metabolic risk factors. J Gastroenterol. 2016;51(9):910-22.

[2] Miura T, Ban D, Tanaka S, Mogushi K, et al. Distinct clinicopathological phenotype of hepatocellular carcinoma with ethoxybenzyl-magnetic resonance imaging hyperintensity: association with gene expression signature. Am J Surg. 2015;210(3):561-9.

[3] Katsuta E, Tanaka S, Mogushi K, Matsumura S, et al. Age-related clinicopathologic and molecular features of patients receiving curative hepatectomy for hepatocellular carcinoma. Am J Surg. 2014;208(3):450-6.

５．主な発表論文等

〔雑誌論文〕（計７件）

以下に代表的なもの５件を記載する。

1. Ohata Y, Shimada S, Akiyama Y, Mogushi K, Nakao K, Matsumura S, Aihara A, Mitsunori Y, Ban D, Ochiai T, Kudo A, Arii S, Tanabe M, Tanaka S. Acquired Resistance with Epigenetic Alterations Under Long-Term Antiangiogenic Therapy for Hepatocellular Carcinoma. Mol Cancer Ther. 2017;16(6):1155-1165.（査読あり）
doi: 10.1158/1535-7163.MCT-16-0728

2. Oba A, Shimada S, Akiyama Y, Nishikawaji T, Mogushi K, Ito H, Matsumura S, Aihara A, Mitsunori Y, Ban D, Ochiai T, Kudo A, Asahara H, Kaida A, Miura M, Tanabe M, Tanaka S. ARID2 modulates DNA damage response in human hepatocellular carcinoma cells. J Hepatol. 2017 May;66(5):942-951.（査読あり）
doi: 10.1016/j.jhep.2016.12.026

3. Akahoshi K, Tanaka S, Mogushi K, Shimada S, Matsumura S, Akiyama Y, Aihara A, Mitsunori Y, Ban D, Ochiai T, Kudo A, Arii S, Tanabe M. Expression of connective tissue growth factor in the livers of non-viral hepatocellular carcinoma patients with metabolic risk factors. J Gastroenterol. 2016;51(9):910-22.（査読あり）
doi: 10.1007/s00535-015-1159-8

4. Katsuta E, Tanaka S, Mogushi K, Shimada S, Akiyama Y, Aihara A, Matsumura S, Mitsunori Y, Ban D, Ochiai T, Kudo A, Fukamachi H, Tanaka H, Nakayama K, Arii S, Tanabe M. CD73 as a therapeutic target for pancreatic neuroendocrine tumor stem cells. Int J Oncol. 2016;48(2):657-69.（査読あり）
doi: 10.3892/ijo.2015.3299

5. Katsuta E, Tanaka S, Mogushi K, Matsumura S, Ban D, Ochiai T, Irie T, Kudo A, Nakamura N, Tanaka H, Tanabe M, Arii S. Age-related clinicopathologic and molecular features of patients receiving curative hepatectomy for hepatocellular carcinoma. Am J Surg. 2014;208(3):450-6.（査読あり）
doi: 10.1016/j.amjsurg.2014.01.015

〔学会発表〕（計１件）

1. Mogushi K, Gene expression analysis and cancer signaling networks. 2nd International Symposium on BioComplexity. 2017年1月20日、別府国際コンベンションセンター(大分県別府市)

６．研究組織

(1)研究代表者

茂櫛　薫（MOGUSHI, Kaoru）
順天堂大学・医学部・助教
研究者番号：60569292