様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 6 月 13 日現在

機関番号：62603
研究種目：若手研究(B)
研究期間：2009 ～ 2010
課題番号：21700313
研究課題名（和文） 組織的メタ遺伝子解析の多重検定理論とその実用化

研究課題名（英文） Multiple testing theory for systematic analysis of meta genes and its applications

研究代表者
吉田 亮（YOSHIDA RYO）
統計数理研究所・モデリング研究系・助教
研究者番号：70401263

研究成果の概要（和文）：
ゲノム、トランスクリプトーム、プロテオーム、メタボローム、そして、フェノームなど、今後 10 年のライフサイエンスの基幹を担うであろう網羅的多面多階層計測から生成されるトランスオミックスデータを統合的に解析するための基礎理論と統計的方法論を開発した。本研究では、とりわけ、トランスクリプトームとデータベースを介して入手される網羅的生体内分子間相互作用情報を組み合わせ、細胞の分子機能に関与する転写因子モジュールを同定することを目的に研究開発を推進した。

研究成果の概要（英文）：
This research project aims to develop statistical testing theory and integrated analytic tools for analyzing massively-collected information on cellular regulatory programs, called ‘omics data’, involving genomics, transcriptomics, proteomics, and metabolomics. Specifically, the newly-derived methods enable us to identify key regulatory units of genes characterizing molecular basis of different cellular phenotypes, such as tumor cells demonstrating sensitivity or resistance to anticancer drugs.

交付決定額

（金額単位：円）

	直接経費	間接経費	合 計
2009 年度	1,600,000	480,000	2,080,000
2010 年度	1,500,000	450,000	1,950,000
年度			
年度			
年度			
総 計	3,100,000	930,000	4,030,000

研究分野： 総合領域
科研費の分科・細目：情報学・統計科学
キーワード：バイオインフォマティクス、多重検定、メタ遺伝子、トランスクリプトーム、バイオマーカー探索

１．研究開始当初の背景

DNA マイクロアレイチップの誕生から約十年，産学の垣根を超えた品質改善努力が実を結び，実用上十分な精度で遺伝子発現量の網羅的測定値が入手できるようになり，分子生物学や医学の研究形態は大きく変容を遂げた．ヒト細胞の遺伝子数は約三万から四万個と推計されているが，これら個々の遺伝子の mRNA 転写量（遺伝子発現量）を計測することで，細胞は数万次元の変数で特徴付けられることになる．DNA マイクロアレイ技術の発明当初から現在に至るまで，研究者の期待

は一貫して「バイオマーカー遺伝子の大規模探索」に向けられてきた．例えば，癌細胞群（ケース）と正常細胞群（コントロール）の遺伝子発現状態を計測した後，遺伝子ごとに発現差の仮説検定を行えば，バイオマーカー候補（統計的に有意な発現差を示す遺伝子）の絞り込みに役立つであろう．2000 年代前半のバイオインフォマティクスでは，DNA マイクロアレイの波及期と相まって，この種の素朴な統計解析技術が重点的に研究されてきた．本研究では，従来のケース・コントロール型研究が個々の遺伝子の発現差に着目してきたのに対して，むしろ「一定の生化学機能に関わるバイオマーカー遺伝子群」の同定を目的とした検定理論および方法論の開発を推進した．

２．研究の目的

ここ数年のバイオインフォマティクスは，一つのベクトルとして，メタ遺伝子の知見とマイクロアレイデータの発現情報をいかに統合して、生体系で鍵となる遺伝子の生化学機能ユニットを直接的に発見していこうという問題にシフトしている．例えば，メタ遺伝子に属する個別遺伝子の統計的エビデンスを統合することで「疾患に関与する生化学機能ユニット（メタ遺伝子）」の直接的な発見に結び付くことが期待される．統計的な視点から問題を眺めると，個別遺伝子の仮説検定から得られた統計的エビデンス（例えば P 値や検定統計量）を統合して，いかに仮説集合全体のエビデンスを算出するかという問題に帰着する．統計科学の文脈において，この種の問題はメタアナリシスのそれに最も類似性を見出すことができる．しかしながら，従来のメタアナリシスの統合対象は「研究」であり，仮説自体は同一のものを想定する一方，本研究では「異なる仮説集合全体」の検定が目的であり，似て非なるものである．バイオインフォマティクスの分野では，この問題の重要性に対して既にコンセンサスが形成されつつあり，事実，これまで幾つかの解析手法が提案されてきている．しかしながら，ほぼ全てと言っても過言でないほど，厳密な理論展開の下で確立された手法は存在しない．その問題点は，数理的発展性のみならず，実際のデータ解析でも重大な誤謬を引き起こすことが，指摘されている． また，申請者自身も，Gupta and Yoshida et al. (2007) LNCS において，既存手法が引き起こす潜在的擬陽性の危険性を，数理的および実データ解析を通して示した．このような研究動向を踏まえ，本研究では，組織的メタ遺伝子解析を「正当な」統計数理の方法論に仕上げ，実問題においても頑強な性能を示す解析ツールの構築を目指した．

３．研究の方法

本研究の開発推進事項は大まかに「論理的包含関係によって構造化された仮説群の最適多重検定方式の構築」と「ソフトウェア開発」に大別されるが，平成２１年度は前者，平成２２年度は後者を重点的に推進してきた．なお，研究期間内の応用研究として「実データ解析」を実施したが，これは方法論の有効性を示すこと，及び実データ解析を通して得られる経験を方法論設計にフィードバックすることで，当該研究を実用上の有効性に結び付けることがねらいであった．

平成２１年度は，論理的包含関係を考慮した最適多重検定方式の確立を目指した．DAG によって仮説群が結び付けられた場合の FDR を厳密に計算したもとで，FDR を最小化するための検定関数の導出を行った．申請時の段階で, FDR 最適化原理から検定方式を導くというアイデアが, Storey の以下の論文で提案されていた．

Storey, J.B. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing, Journal of Royal Statistical Society B, 69, 347-368.

Storey の研究は基本的には仮説間の依存関係を陽な形で取り込んでおらず，直接的にメタ遺伝子解析に適用することはできなかった．また，個々の仮説検定を実行した後，計算された検定統計量，もしくは P 値が与えられたもとで, FDR 最小化を達成する多重検定方式が導出される．本研究と Storey（2007）との大きな違いは，多重検定方式を導く前に個々の仮説検定方式を明示的に指定せずに，最適 FDR を達成するための「個別の仮説検定」と「多重検定」の手順を同時に導くことであった．

また，当該年度では，方法論の整備を大きく前進させたともに，ソフトウェア開発に向けてプログラムの作成を行った．さらに，実際のデータ解析では，(1) メタ遺伝子の誤った分類に対する検定の頑強性，(2) 観測ノイズに対する頑強性，(3) 計算速度，などが開発手法の評価尺度として要求されることから，数値実験や解析的アプローチを通じて，適宜これらのポイントの評価，新たな問題の洗い出しを行い，問題の克服に努めた．

方法論の研究と並行して実際の実験データへの適用を実施することで，開発手法の実問題への有効性を検証し，方法論研究へのフィードバックを行った．特に，EGFR シグナル伝達経路下流の抗癌剤感受性バイオマーカー遺伝子の同定を試みた．東京大学医科学研究所の宮野悟教授から実験データ提供を受けた．なお，本研究課題の目的は，あくま

で統計解析の方法論構築であり，応用研究で得られた生物学的知見については成果の対象外である．

ソフトウェア開発に関して，転写モジュールの情報として， JASPAR データベースを援用した．また，タンパク質相互作用については，DIP (http://dip.doe-mbi.ucla.edu/) を用いた．両データベース共，分子生物学の分野で世界的に知名度が高く，オープンデータアクセスという共通点がある．研究実施期間後も，より有益なデータベースが見つかれば，適宜変更・追加を行う予定である．Gene Ontology（GO）データベースに関しても，データベースのバージョンアップがある場合，これを速やかに反映する．

４．研究成果

本研究で対象とした統計的問題は以下の二点である．(1) 個別の仮説が与えられたもとで，仮説集合全体の有意性を評価するための方法論を構築する，(2) 仮説集合間に論理的包含関係がある場合の多重比較検定法の考案を行う．後者の多重比較問題は，多くのメタ遺伝子データベースの情報が特殊なデータ構造を用いて記述されていることに起因する．例えば，GO のメタ遺伝子は入れ子構造，すなわち，下位の遺伝子セットは必ず上位の遺伝子セットに含まれる形式で記述されている．この場合，上位の仮説集合に対する帰無仮説が真であれば，必然的に下位の仮説も真になる．逆に，下位の仮説が偽であれば，上位の仮説は偽になる．GO の仮説集合群の論理的包含関係は DAG によって表現される．また，前述の転写因子と下流遺伝子セットの情報も DAG 形式で記述される．申請時に行った先行調査によれば，このような論理的依存関係を持つ仮説群の「実用的な検定論」の整備は今のところほぼ未着手であった．本研究では解決方策として，DAG 構造を持つ多重検定の性能評価尺度に False Discovery Rate（以下 FDR）基準を導入し，最適な検定・決定方式を導いた．

応用的実用性の観点からは，ソフトウェア開発し，医学，薬学，生物学での実運用を稼働させる共に．より多くの研究者に使ってもらえるよう MetaGP というウェブアプリケーションを公開した．

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 3 件）

(1) Y. Tamada, R. Yamaguchi, S. Imoto, O. Hirose, R. Yoshida, M. Nagasaki, S. Miyano (2011) SiGN-SSM: open source parallel software for estimating gene networks with state space models, Bioinformatics, 15;27(8):1172-1173 (査読有)

(2) R. Yoshida, M. Saito, H. Nagao, T. Higuchi (2010) Bayesian experts in exploring reaction kinetics of transcription circuits, Bioinformatics, 26(18), i589-595. (査読有)

(3) R. Yoshida, M. West (2010) Bayesian learning in sparse graphical factor models via variational mean-field annealing, Journal of Machine Learning Research, 11:1771-1798. (査読有)

〔学会発表〕（計 12 件）

※下線は研究代表者

※＊印は発表者

(1) R. Yoshida*, M. Saito, H. Nagao, T. Higuchi, Bayesian experts in exploring reaction kinetics of transcription circuits , 9th European Conference on Computational Biology (ECCB2010), 2010/9/29, Belgium

(2) 吉田亮*, 長尾大道, 斎藤正也, 長﨑正朗, 長崎正朗, 井元清哉, 山口類, 山内麻衣, 後藤典子, 宮野悟, 樋口知之, 癌細胞シミュレーションと生化学反応系の統計モデリング, 2010 年度統計関連学会連合大会, 2010/9/6, 東京（早稲田大学）

(3) 長尾大道*, 吉田亮, 斎藤正也, 樋口知之, 長﨑正朗, 井元清哉, 山口類, 宮野悟, 山内麻衣, 後藤典子, 遺伝子発現時系列データおよびバイオデータベース情報を基にした活性/抑制型転写因子の同定, 2010 年度統計関連学会連合大会, 2010/9/6, 東京（早稲田大学）

(4) 山口 類*, 井元 清哉, 山内 麻衣, 島村 徹平, 長﨑 正朗, 吉田 亮, 樋口 知之, 後藤 典子, 宮野 悟, 状態空間モデリングによる遺伝子発現制御に関わる動的薬剤効果の推定, 2010/9/6, 東京（早稲田大学）

(5) 吉田亮*, LiSDAS: Life Science Data Assimilation Systems, 生命体統合シミュレーション サマースクール 2010, 2010/7/5, 湘南国際村

(6) 吉 田 亮 *, Life Science Data Assimilation Systems: 生化学反応系の複雑性とその克服へ向けて, 第 1 回「バ

イオモデリングと統計科学」研究会, 2010/6/30, 東京（統計数理研究所）

(7) 吉田亮*, LiSDAS: Life Science Data Assimilation Systems, 第59回理論応用力学講演会, 2010/6/8, 東京（日本学術会議）

(8) 吉 田 亮 *, Life Science Data Assimilation Systems: 生化学反応系の複雑性とその克服へ向けて, グローバルCOE 特別セミナー, 2010/5/14, 東京大学(本郷)

(9) 吉田 亮*, グラフィカルファクターモデルと事後エントロピーにもとづく決定論的アニーリング：システムズバイオロジーへの応用, 御殿場基礎科学研究会「最適化を基軸とする数理的展開」, 2010/1/30 , 静岡(御殿場)

(10) 吉田 亮*, Mike West, 事後エントロピーを利用した焼きなまし法と遺伝子発現データのスパース学習, 2009年度統計関連学会連合大会, 2009/9/7、京都（同志社大学)

(11) 中村和幸*, 吉田亮, 長崎正朗, 宮野悟, 樋口知之, 超多数粒子フィルタによる遺伝子ネットワークデータ同化, 2009 年度統計関連学会連合大会, 2009/9/7, 京都（同志社大学）

(12) 山口類*, 井元清哉, 島村徹平, 山内麻衣, 長崎正朗, 吉田亮, 樋口知之, 後藤典子, 宮野悟, 状態空間モデルからの動的予測に基づく遺伝子発現制御関係の差異の探索, 2009年度統計関連学会連合大会, 2009/9/8, 京都（同志社大学）

〔図書〕（計0　件）
〔産業財産権〕
○出願状況（計 0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計0件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://metagp.ism.ac.jp/

６．研究組織
(1)研究代表者
吉田　亮（YOSHIDA RYO）
統計数理研究所・モデリング研究系・助教
研究者番号：70401263

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：