様式C-19

科学研究費補助金研究成果報告書

平成 22 年　5 月 21 日現在

研究種目：基盤研究（Ａ）
研究期間：2007～2009
課題番号：19200013
研究課題名（和文）　大規模変数次元データの因果構造マイニング手法開発と遺伝子機能関係知識ベースの作成

研究課題名（英文）　Development of Causal Structure Mining Method for Large Scale Dimensional Data and Construction of Gene Function Knowledge Base

研究代表者
鷲尾　隆（WASHIO　TAKASHI）
大阪大学・産業科学研究所・教授
研究者番号：00192815

研究成果の概要（和文）：科学者は多数遺伝子の発現強度変数測定データ（大規模変数次元データ）から、遺伝子発現間の因果関係を把握し、各遺伝子の機能を解き明かそうとする。しかし、人手では数十～数百個もの変数間の因果関係を見出すのは困難である。ところが、最新の計算機データ解析技術でも２０～３０変数間の因果関係解析しかできない。そこでこ、本研究では新たな統計的因果解析原理を開発し、計算機を用いて数十～数百個の変数間の因果関係を明らかにする手法を確立した。さらに、この手法を用いて科学者が参照可能な遺伝子発現機能関係知識ベースの構築を行った。

研究成果の概要（英文）：Scientists attempt to figure out function of each gene through the analysis of causal relations between gene expressions by using measurement data of the many gene expression variables (large scale dimensional data). However, the analysis of causal relations between dozens or hundreds of variables is hardly performed manually. In spite of this problem, the number of variables to which the computer based causal analysis is applicable is limited to 20 – 30 in the state of the art. Accordingly, this work developed a novel principle of the statistical causal analysis, and furthermore constructed a knowledge base of the functional relations among expressed genes for the scientists by using our developed approach.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	12,100,000	3,630,000	15,730,000
２００８年度	9,500,000	2,850,000	12,350,000
２００９年度	8,000,000	2,400,000	10,400,000
総　計	29,600,000	8,880,000	38,480,000

研究分野：データマイニング
科研費の分科・細目：情報学・知能情報学
キーワード：大規模次元、因果推論、データマイニング、遺伝子機能、知識ベース

１．研究開始当初の背景

最近の計測センサ技術やネットワーク通信技術の発展は目覚ましく、それによって人間の膨大な遺伝子発現の様子や社会の様々な出来事を一挙に測定することが可能となった。このような測定データは、多数の測定

項目の集まりによって表される。たとえば、人間の遺伝子発現データでは、各々の遺伝子発現の強さが１項目（１変数）で表される。通常、科学者は一度に数十～数百個もの遺伝子発現の様子から細胞内で起こる現象を捉える必要があるため、１回の実験測定で多くの遺伝子発現強度変数を測定する。このような多数の変数で表されるデータを大規模変数次元データと呼ぶ。実際の遺伝子発現データは、更にこのようなデータを多数回の実験に亘り集めたものである。科学者は、このデータからどの遺伝子の発現が他のどの遺伝子の発現を引き起こすかという遺伝子間の因果関係を明らかにし、更にその連鎖の中で各遺伝子が細胞内で担う役割（機能）を解き明かそうとする。しかし、数十～数百個もの変数データから、人間が直接に変数間の因果関係を正確に見出すのは無理である。そこで、計算機の助けを借りる必要がある。

しかしながら、最新の計算機に現状のデータ解析アルゴリズムを実装し、このような大規模変数次元データの解析を行っても、実行可能な計算量の限界と統計的不確定性により、正確な因果関係を見出すことができる変数は、現状では２０～３０個までであることが知られている。そこで、以下を考えた。

(1) 新たな原理により大規模変数次元データから数十～数百個の変数間の因果関係を明らかにする手法の開発

が必要であるとの考えに至った。さらにその開発手法による成果として、

(2) 科学者が遺伝子の因果関係と機能を明らかにする際に参考とする遺伝子発現機能関係知識ベースを構築すること

２．研究の目的

(1) 第１目的は、計算機を用いて数十～数百個の変数からなる大規模変数次元データから、変数間の因果関係を表す因果ネットワークを完全導出する効率的手法を開発することである．このため，一般に多くの計算を要し統計的不確定性も大きい既存のベイジアンネットワーク手法よりも、少ない計算量でかつ多変数の因果ネットワークを同定しやすい独立成分分析(ICA)に基づく最新の統計的因果推論を更に拡張し、大量変数間の因果ネットワークを高速に探索導出する原理と手法を確立する。

(2) 第２目的は、上記で確立した手法を幾つかの重要な大規模変数次元の遺伝子発現データに適用して情報生化学分野の科学者と共に因果解析を実施し、そこで得られた新たな遺伝子機能と因果関係に関する知見を知識ベース化することである．

３．研究の方法

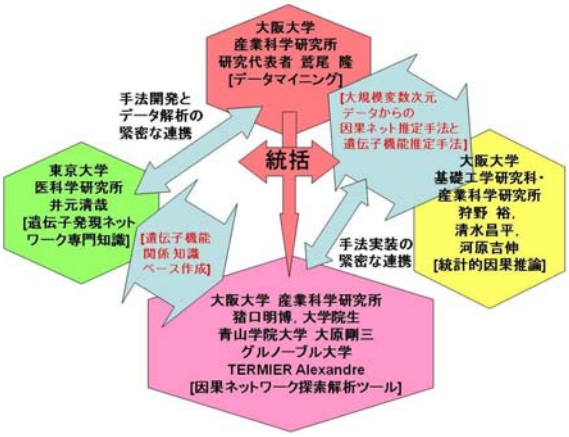

図１　本研究の遂行体制

図１に本研究の遂行体制を示す。本研究は、データマイニング・機械学習研究に取り組む人工知能学者、統計的因果推論研究に取り組む統計数理学者、これら手法を統合した解析ツールを開発する計算機科学者、遺伝子発現因果ネットワーク解析に取り組むシステムバイオロジー学者という４分野の共同が欠かせない学際的かつ専門的研究であった。研究代表者の鷲尾は、データマイニングのみならず、過去に多くの数理的因果推論手法開発の経験を有する。また、高次元の遺伝子発現データから因果ネットワークの発見を試み、分子生物学者と遺伝子機能推定共同作業を経験している。研究代表者は、以上経験に基づいてプロジェクト全体の進行を統括すると共に、大阪大学の清水、河原、狩野と共同して「大規模変数次元データから変数間の因果関係を表す因果ネットワークを完全導出する効率的手法」及び「遺伝子機能推定手法」の開発にあたった。また、元大阪大学（現青山学院大学）の大原、グルノーブル大学のTERMIER，大阪大学大学院生と共同して上記手法を「解析ツール」として実装し、更に東京大学医科学研究所の井元とそのツールを用いて「遺伝子機能関係知識ベース作成」を行い、その知識ベースの内容・品質の吟味を行った。そして、上記部分因果ネットワーク推定手法に必要な更なる改良・拡張点を洗い出し、清水，河原，狩野との共同研究にフィードバックした。このようなスパイラルにより、目的手法と目的知識ベースを実用的完成度に高めた。

１つの遺伝子発現データは、マイクロアレイ上の各遺伝子に対応する格子点で計測される多数の発現変数で表され、これが検体数 n 個分存在する。一般に n は数十程度と変数個数より少ないことが多く、全体因果構造を直接得るには条件不足である。そこで初年度は、中心極限定理の拡張により因果構造上でより上流に位置する変数の方がより非ガウスな揺らぎを示すことが言えるため、変数の

揺らぎの非ガウス性と変数同士の揺らぎの相関関係から因果的に最上流の外生変数と呼ばれる変数を見つけ、そこから順番に部分的な因果ネットワーク構造を同定する方法を開発した。この方法は上流から順に小さな部分因果ネットワークを見つけるため、データ数（検体数）が限られていても安定した推定結果をもたらすことができる。

次年度は、初年度に確立した大規模変数次元データからの因果ネットワーク推定手法を用いて推定した多量の遺伝子発現因果ネットワークについて、その内部の各遺伝子に標準化記述子 Gene Ontology (GO) Term 等によって機能ラベル付けを行った。そして、各遺伝子が機能ラベルに変換された多量の遺伝子発現因果ネットワークの分析によって、推定された因果構造の妥当性を領域専門家と一緒に吟味した。そして、妥当性が高いことが確認された結果について、更に遺伝子発現因果ネットワークに含まれる既知の機能を持つ遺伝子の機能情報から、類似した遺伝子発現因果ネットワークに含まれる未知機能遺伝子の機能を類推した。これにて明らかになった未知機能遺伝子の機能情報を含め，遺伝子発現因果ネットワークから成る遺伝子機能関係知識ベースを作成した。

最終年度は、以上の知識ベースの妥当性を既存の遺伝子関係データベースの内容と照合検討し、問題点については因果ネットワーク構造同定手法にまで遡り、改良を行った。その改良によって得られた因果ネットワーク同定結果を、更に遺伝子機能関係知識ベースに反映することを繰り返した。

４．研究成果

(1)人工データによる検証

はじめに開発手法及び解析ツールの基本的性能評価を行うために、予め既知の因果構造モデルを用いて人工的に作成した模擬遺伝子発現データに適用した。遺伝子発現変数の個数を 1000 とし､その間にランダムに 1000 本の因果関係を表す辺を付与し、合計で 171 個の外生変数を含む人工因果ネットワークを作成した。次に、各外生変数に非ガウス性を模擬する乱数 s_i、それ以外の変数にもよりガウス性に近い乱数 e_i を加え、このネットワークから人工的模擬遺伝子発現データを作成した。こうして模擬遺伝子発現検体データを n 個作成した｡検体数 n は 30, 60, 100, 200 の４種類に変えて実験を行った。

図２に因果ネットワーク導出手法・ツールを検体数 n=60 としたデータに適用した結果を示す。同定した外生変数候補の内、実際に外生である可能性が高いと推定された上位 m 個の変数の中で、本当に外生であった変数の割合（%）を精度として縦軸に取ってある。パラメータ h は e_i の非ガウス性の程度を表し、h が小さいほど非ガウス性が強い。全般に外生である可能性の高い変数に絞ると、ほぼ確実に正しく外生変数を捉えることができる。また、外生変数以外の変数により小さな h で生成された乱数 e_i を加えほど、外生変数とそれ以外の見分けがつき難くなり、外生変数を正しく同定する性能が低くなることも分かった。全体としては、外生変数が特に非ガウス性の強い揺らぎを有する場合には、高い精度で変数間の因果関係を導出できることが分かった。

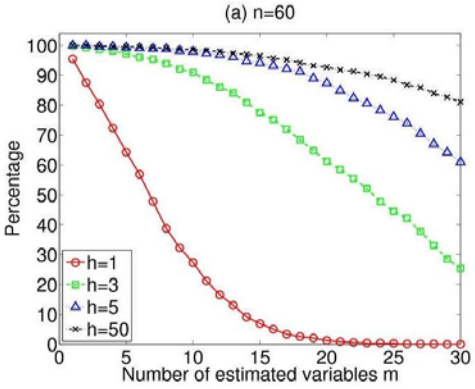

図２　人工データに関する推定精度

(2) 遺伝子発現実データによる検証と知識ベース整備

遺伝子発現実データに関し、開発手法及び解析ツールの性能を示すため、人間の乳がん細胞の遺伝子発現状態に関して実験収集されたデータに適用した結果を取り上げる。この実験データは、各種濃度（0.1, 0.5, 1.0, 10.0 nmol/ℓ）の EGF(上皮成長因子)と呼ばれるたんぱく質を乳がん細胞に投与後、5, 10, 15, 30, 45, 60, 90 分の経過時間毎に、細胞内遺伝子の発現状態を測定したものである。各検体において 22277 個の遺伝子発現状態が測定されているが、その中で実験条件の違いによって遺伝子発現状態が大きく異なる(EG

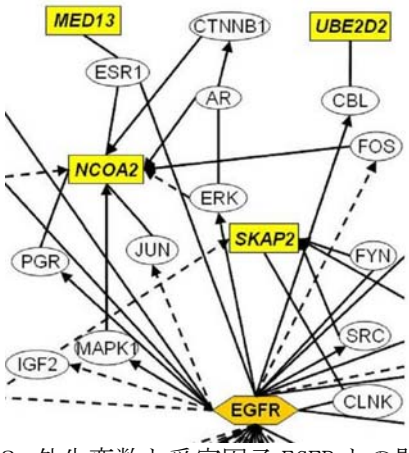

図３ 外生変数と受容因子 EGFR との関係

F 濃度と経過時間の影響を受ける）遺伝子を、統計検定によって 1000 個選択した。

図３に開発手法・ツールの同定結果と生化学専門家が想定する因果ネットワーク情報を総合した外生変数候補と上皮成長因子受容体(EGFR)との因果関係ネットワークを示す。黄色い箱で囲んだ変数が外生変数候補である。それらの多くは EGFR の投与によってほぼ直接影響を受ける遺伝子であると考えられるが、このネットワーク上で因果的に上流の EGFR に近く、開発手法・ツールの同定結果は妥当であると考えられる。この他にも重要な遺伝子発現実験データに関する解析結果をまとめ、知識ベース化を行った。

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１６件）

① S. Shimizu, P. O. Hoyer and A. Hyvarinen, Estimation of linear non-Gaussian acyclic models for latent factors, Neurocomputing, Vol.72, 査読有, 2009, 2024-2027

② 城戸 健太郎, 桑島 洋, 鷲尾 隆: ユークリッド距離の高速高精度推定と範囲問合せへの応用, 情報処理学会論文誌, 査読有, Vol. 50, No. 5, 2009, 1493-1505

③ Y. Tamada, H. Araki, S. Imoto, M. Nagasaki, A. Doi, Y. Nakanishi, Y. Tomiyasu, K. Yasuda, B. Dunmore, D. Sanders, S. Humphries, C. Print, D.S. Charnock-Jones, K. Tashiro, S. Kuhara, S. Miyano, Unraveling dynamic activities of autocrine pathways that control drug-response transcriptome networks, Pacific Symposium on Biocomputing, Vol.14, 査読有, 2009, 251-263

④ K. Takai and Y. Kano, Simple computation of maximum likelihood estimates in latent class model with equality and constant constraints, Communications in Statistics - Simulation and Computation, Vol. 38, No. 3, 査読有, 2009, 654-665

⑤ S. Shimizu and Y. Kano, Use of Non-Normality in Structural Equation Modeling: Application to Direction of Causation, J. of Statistical Planning and Inference, Vol.138, 査読有, 2008, 3483-3491

⑥ V.P. Nguyen and T. Washio: Modeling Dynamic Substate Chains among Massive States, Intelligent Data Analysis, Vol.12, No.3, 査読有, 2008, 271-291

⑦ A.Termier, M.-C. Rousset, M. Sebag, K. Ohara, T. Washio, and Hiroshi Motoda: DryadeParent, An Efficient and Robust Closed Attribute Tree Mining Algorithm, IEEE Trans. on Knowledge and Data Eng. (IEEE-TKDE), Vol. 20, No. 2, 査読有, 2008, 300-320

⑧ R. Yoshida, M. Nagasaki, R. Yamaguchi, S. Imoto, S. Miyano, and T. Higuchi, Bayesian learning of biological pathways on genomic data assimilation, Bioinformatics, Vol.24, No.22, 査読有, 2008, 2592-2601

⑨ O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, D. S. Charnock-Jones, C. Print, and S. Miyano, Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models, Bioinformatics, Vol. 24, No. 7, doi:10.1093/bioinformatics/ btm639, 査読有, 2008, 932-942

⑩ K. Kojima, A. Fujita, T. Shimamura, S. Imoto, S. Miyano, Estimation of nonlinear gene regulatory networks via L1 regularized NVAR from time series gene expression data, Genome Informatics, Vol.20, 査読有, 2008, 37-51

⑪ E. Perrier, S. Imoto, S. Miyano, Finding optimal Bayesian network given a super-structure, Journal of Machine Learning Research, Vol.9, 査読有, 2008, 2251-2286

⑫ R. Yamaguchi, S. Imoto, M. Yamauchi, M. Nagasaki, R. Yoshida, T. Shimamura, Y. Hatanaka, K. Ueno, T. Higuchi, N. Gotoh, S. Miyano, Predicting differences in gene regulatory systems by state space models, Genome Informatics, Vol. 21, 査読有, 2008, 101-113

⑬ K. Takai and Y. Kano, Test of independence in a 2×2 contingency table with nonignorable nonresponse via constrained EM algorithm, Computational Statistics and Data Analysis, Vol.52, No.11, 査読有, 2008, 5229-5241

⑭ T. Washio, K. Nakanishi and H. Motoda: A Classification Method Based on Subspace Clustering and Association Rules, New Generation Computing, Vol.25, 査読有, 2007, 235-245

⑮ 鷲尾 隆, 樋口 知之, 井元 清哉, 玉田 嘉紀, 佐藤 健, 元田 浩: グラフマイニン

グとその統計的モデリングへの応用，特集「予測と発見」，統計数理，Vol.54, No.2，査読有，統計数理研究所， 2007, 315-331

⑯ R. Yamaguchi, R. Yoshida, S. Imoto, T. Higuchi, and S. Miyano, Finding module-based gene networks with state-space models - Mining high-dimensional and short time-course gene expression data, Special Issue on Signal Processing Methods in Genomics and Proteomics, IEEE Signal Processing Magazine, Vol.24, No.1, 査読有, 2007, 37-46

〔学会発表〕(計１７件)

① Y. Kawahara, K. Nagano, K. Tsuda and J. Bilmes, Submodularity cuts and applicastions, Proc. of NIPS2009: Advances in Neural Information Processing, Vol. 22, December 8, 2009, Vancouver (Canada)

② K. Hayashi, S. Shimizu and Y. Kano, Consistency of penalized risk of boosting methods in binary classification, In New Trends in Psychometrics, Post Proceedings of IMPS2007: the 15th International and 72nd Annual Meeting of the Psychometric Society, July 11, 2009, Tower Hall Funabori(Tokyo)

③ K. Takai and Y. Kano, Factor prediction in time series factor analysis, In New Trends in Psychometrics, Post Proceedings of IMPS2007: the 15th International and 72nd Annual Meeting of the Psychometric Society, Psychometric Society, July 10, 2009, Tower Hall Funabori(Tokyo)

④ S. Shimizu, A. Hyvinen, Y. Kawahara, T. Washio: A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model, Proc. of UAI2009: 25th Conf. on Uncertainty in Artificial Intelligence, Causality II & Graphical Models, June 21, 2009, Montreal(Canada)

⑤ A. Inokuchi and T. Washio: A Fast Method to Mine Frequent Subsequences from Graph Sequence, Proc. of ICDM2008: 8th IEEE Int. Conf. on Data Mining, 303-312, December 17, 2008, (Pisa, Italy)

⑥ H. Kashima, K. Yamasaki, A. Inokuchi, H. Saigo, Regression with interval output values, Proc. of ICPR2008: 19th International Conference on Pattern Recognition 2008, 1-4, December 11, 2008, Tampa, Florida (USA)

⑦ Y. Kano, Scaling with SEM: The Role of Effect and Causal Indicators (Invited), IASC2008: the Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, December 7, 2008, Yokohama(Kanagawa)

⑧ K. Ohara and T. Washio: Isomorphism Identification by Using Graph Spectra and Its Application to Graph Mining, Proc. of IASC2008: the Joint Meeting of 4th World Conf. of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, 招待講演, 229, December 6, 2008, Yokohama(Kanagawa)

⑨ K. Takai and Y. Kano, AIC for Missing Data, IASC2008: the Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, December 5, 2008, Yokohama(Kanagawa)

⑩ K. Numata, S. Imoto, S. Miyano, Partial order-based Bayesian network learning algorithm for estimating gene networks, Proc. of IEEE Bioinformatics and Biomedicine 2008, 357-360, November 5, 2008, Philadelphia, Pennsylvania (USA)

⑪ A. Inokuchi and T. Washio, Feasibility of Graph Sequence Mining based on Admissibility Constraints, Proc. of 3rd International Workshop on Data-Mining and Statistical Science, 1-4, September 25, 2008, Tokyo Institute of Technology(Tokyo)

⑫ K. Hayashi, S. Shimizu and Y. Kano, Penalized boosting algorithm for mislabeled data (Invited), IMPS2008: the 73rd Annual Meeting of the Psychometric Society, June 30, 2008, Greensboro(USA)

⑬ K. Takai and Y. Kano, Test of independence in a 2x2 contingency table with nonignorable nonresponse, IMPS2008: the 73rd Annual Meeting of the Psychometric Society, June 29, 2008, New Hampshire(USA)

⑭ Y. Konya, S. Shimizu and Y. Kano, Interval estimations based on normalizing transformations by two approaches, IMPS2008: the 73rd Annual

Meeting of the Psychometric Society, June 29, 2008, New Hampshire(USA)

⑮ Y. Kano, Separability of noisy ICA for high dimensional data, HDM2008: International Conference on Multivariate Statistical Modelling & High Dimensional Data Mining, June 19, 2008, Kayseri(Turkey)

⑯ K. Kido, H. Kuwajima and T. Washio: A Range Query Approach for High Dimensional Euclidean Space Based on EDM Estimation, Proc. of SDM2008: 8th SIAM Int. Conf. on Data Mining, 387-398, April 24, 2008, Atlanta(USA)

⑰ A. Termier, Y. Tamada, K. Numata, S. Imoto, T. Washio, Tomoyuki Higuchi: DIGDAG, a first algorithm to mine closed frequent embedded sub-DAGs, Proc. of MLG Workshop 2007, Mining and Learning with Graphs, 41-46, August 2, 2007, Firenze(Italy)

６．研究組織

(1)研究代表者

鷲尾　隆（GAKUSHIN　TARO）
大阪大学・産業科学研究所・教授
研究者番号：00192815

(2)研究分担者

狩野　裕（KANO YUTAKA）
大阪大学・基礎工学研究科・教授
研究者番号：20201436
担当期間：平成 19 年度

井元　清哉（IMOTO SEIYA）
東京大学・医科学研究所・准教授
研究者番号：10345027
担当期間：平成 19 年度

大原　剛三（OHARA KOUZOU）
青山学院大学・理工学部・准教授
研究者番号：30294127
担当期間：平成 19 年度

ターミエ　アレックサンドル
（TERMIER ALEXANDLRE）
Universite Joseph Fourier・Laboratoire d'Informatique de Grenoble・助教
研究者番号：60435823
担当期間：平成 19 年度

(3)連携研究者

狩野　裕（KANO YUTAKA）
大阪大学・基礎工学研究科・教授
研究者番号：20201436
担当期間：平成 20 年度～21 年度

井元　清哉（IMOTO SEIYA）
東京大学・医科学研究所・准教授
研究者番号：10345027
担当期間：平成 20 年度～21 年度

大原　剛三（OHARA KOUZOU）
青山学院大学・理工学部・准教授
担当期間：平成 20 年度
研究者番号：30294127

ターミエ　アレックサンドル
（TERMIER ALEXANDLRE）
Universite Joseph Fourier・Laboratoire d'Informatique de Grenoble・助教
研究者番号：60435823
担当期間：平成 20 年度

猪口　明博（INOKUCHI AKIHIRO）
大阪大学・産業科学研究所・助教
研究者番号：70452456
担当期間：平成 20 年度～21 年度

清水　昌平（SHIMIZU SHOHEI）
大阪大学・産業科学研究所・助教
担当期間：平成 21 年度
研究者番号：10509871

河原　吉伸（KAWAHARA YOSHINOBU）
大阪大学・産業科学研究所・助教
研究者番号：00514796
担当期間：平成21年10月～平成22年3月