様式 C-19

科学研究費補助金研究成果報告書

平成　21 年　6 月　10 日現在

研究種目：基盤研究(A)
研究期間：2005～2008
課題番号：17200020
研究課題名（和文）　階層的統計モデルに基づく異種ゲノム情報の統合手法に関する総合研究
研究課題名（英文）　Research on information fusion methods for multiple genomic data sources with heterogeneity based on hierarchical statistical modeling

研究代表者
樋口　知之　（ HIGUCHI TOMOYUKI ）
統計数理研究所・モデリング研究系・教授
研究者番号：70202273

研究成果の概要：
一般に生物学的知識は，データベースの形に具現化されることが多い．データベースに登録された情報にも信頼度の属性を与え，つまりそれを確率変数として取り扱うことで，データベースの登録上の過誤やその情報の不確実さをモデル化し，今までのモデルをさらに階層化したベイズモデルを構成した．これにより，マイクロアレイデータからの情報抽出，既存の生物学的知識の有効活用，データベースの信頼性の検証などを統一的に可能にする枠組みを構築した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2005 年度	6,900,000	2,070,000	8,970,000
2006 年度	9,700,000	2,910,000	12,610,000
2007 年度	9,700,000	2,910,000	12,610,000
2008 年度	9,700,000	2,910,000	12,610,000
年度			
総　計	36,000,000	10,800,000	46,800,000

研究分野：ベイジアンモデリング
科研費の分科・細目：情報学・統計科学
キーワード：⑴階層ベイズ，⑵状態空間モデル，⑶グラフィカルモデル，⑷ベイジアンネットワーク，⑸マイクロアレイデータ，⑹遺伝子ネットワーク，⑺カルバックライブラーカーネル，⑻スペクトル判別

１．研究開始当初の背景

(1)経緯：ベイジアンネットワークにもとづく DNA マイクロアレイデータからの遺伝子発現ネットワークの推定において，条件付分布の強健で信頼できる推定のために我々が 2003，2004 年に提案した枠組みは，遺伝子ネットワークのモデル構成プロセスにおいて，マイクロアレイデータからの遺伝子発現量の情報と，データベース化された生物学的知識の情報を結合させる一般的枠組みである．その手法の利点の一つは，マイクロアレイからの情報と生物学的知識間のバランスをどうとるかを情報量規準が定めることができることである．生物学的知識をベイジアンネットワークに付加することにより，マイクロアレイデータからさまざまなノイズによる影響を排除しつつ精密に遺伝子ネットワークを推定でき，結果としてより多くの情報を抽出することに成功している．

(2)動機：統計科学の分野においても，DNA アレイのデータの標準化や検定を基礎にした，結果の吟味等の研究に研究者の寄与がめだつ．しかしながらこれらはデータ解析の入り口部分であり，あまりデータを選別しすぎる

と結果として多くのデータを捨ててしまうこととなる．本研究では，入り口よりも一歩踏み込んだ，新しい知識の獲得をめざしたデータ解析法の研究に重点を置く．すでに，人工知能の分野から多くの研究者がDNAマイクロアレイデータから遺伝子ネットワーク推定研究分野に参入しているが，それらの多くはアルゴリズム（手続き）ベースであり，推定の誤差評価が十分に検討できないことも多い．本研究では結果の解釈が直接的でまたモデル比較手続きが陽に操作できる，モデルベースのアプローチを採用する．そのモデルは，異種複数の情報を確率的に表現するために，複数のグラフィカルモデルをベイズの枠組みをもとに組み合わせた，独創的なものである．

２．研究の目的

本研究では，我々のこれまでの方法論をさらに発展させることを狙っている．一般に生物学的知識は，データベースの形に具現化されることが多い．データベースに登録された情報にも信頼度の属性を与え，つまりそれを確率変数として取り扱うことで，データベースの登録上の過誤やその情報の不確実さをモデル化するわけである．この仮定のもとで，今までのモデルをさらに階層化したベイズモデルを構成し，マイクロアレイデータからの情報抽出，既存の生物学的知識の有効活用，データベースの信頼性の検証などを統一的に可能にする枠組みを考案する．提案する手法は，人工データへの適用による性能評価のあと，実際のデータに適用する予定である．本研究の延長上には，データベースへの登録作業の過誤などが自動的に特定できる，自己修復機能をそなえたデータベースシステムの設計も視野にある．

３．研究の方法

複数のサブテーマ(A-H)をたて，各テーマに主と副の担当者をおく研究体制をとった．各サブテーマの進展チェックを総括者である樋口が行った．

[A]非線形回帰モデルの新展開:遺伝子発現ネットワークのモデル化の基礎となるベイジアンネットワークの矢印の太さを推定する作業は，統計の伝統的研究テーマである非線形回帰問題に帰着できる．この統計モデルの同定に必要な情報量規準，特にGICの適用を進めながら，アレイデータ解析における情報量規準のモデル選択機能の性能評価を行う．

[B]アレイデータ用のブートストラップ新手法の開発:通常の非線形回帰問題は，説明変数に誤差を仮定しないが，本研究ではそこにも観測誤差を仮定した非線形回帰の枠組みですすめる．一つの条件付分布について非線形回帰を行い，そこからの結果をその遺伝子が影響している次の遺伝子に関する非線形回帰に反映させる．その手続きを順次繰り返すことで，全体で一つの同時分布を推定する作業とする．

[C]DNA タイムコースデータの解析:状態空間モデルを利用した DNA アレイ時系列データ（タイムコースデータ）の新しい手法の開発を行う．DNAタイムコースデータは普通10〜30程度しか時点数がない一方，遺伝子数は数千以上になるため，従来の状態空間モデルの枠組みをそのまま利用することはできない．安定したパラメータ推定に関する数値的課題はもちろんのこと，潜在変数で構成される状態ベクトルがどのような意味を持つのか，またその次元推定など克服すべき数理的課題も多い．本研究では，観測モデルに混合因子分析の枠組みを組み込んだモデルを適用する．

[D]ベイズモデルによる DNA アレイデータ情報と生物学的知識の統合:マルコフランダム場モデルとして表された事前情報をどの程度組み込むかはハイパーパラメータが制御するが，その決定を最尤法で行おうとすると組み合わせ爆発をする計算を行う必要がある．従ってこれまでは周辺尤度の上限値，下限値をもとめることで，間接的にこの値を推定していた．よって，この尤度の見積もりがどの程度精密なのか，あるいは全く別の観点からのハイパーパラメータの決定法を考察するなど，ハイパーパラメータの推定法に関する残された数理的課題に取り組む．

[E]確率的最適化手法による遺伝子空間の地形探索:タンパク質工学とは,天然の遺伝子に人工的な改変(塩基配列の一部の削除，挿入および部位特異的突然変異の導入など)を加えた遺伝子を生細胞（特に，酵母など）やこれらに由来する無細胞の合成システムを使って発現させ，目的にかなった性質をもつタンパク質を合成することをいう．タンパク質工学における望みの機能をもつタンパク質の発見の過程を情報科学における最適化とみなし，タンパク質工学への応用のための遺伝的アルゴリズムの一手法を示しつつ，その有効性を計算機実験により確認しウェットな実現の可能性を議論する．

[F]グラフ構造探索のハイパーフォーマンスコンピューティング:本研究では従来よりも複雑な手続きで矢印の太さを決めるので計算量はかなりのものとなる．そのため，並列計算向きの探索計算アルゴリズムの開発を行う．

[G]生物データベースの修復:データベースに登録された情報にも信頼度の属性を与え，つまりそれを確率変数として取り扱うことで，データベースの登録上の過誤やその情報の不確実さをモデル化する．

[H] カーネル法と状態空間モデルの統合：状態空間モデルと，バイオインフォマティクスの分野で活発に研究されているカーネル法を融合させる研究を進める．カーネル法では通常データそのものを入力とする場合が推奨されているが，前処理として状態空間モデルを利用して，一信号をあえて多信号に分解し，それらを入力とするアプローチを採用する．

４．研究成果

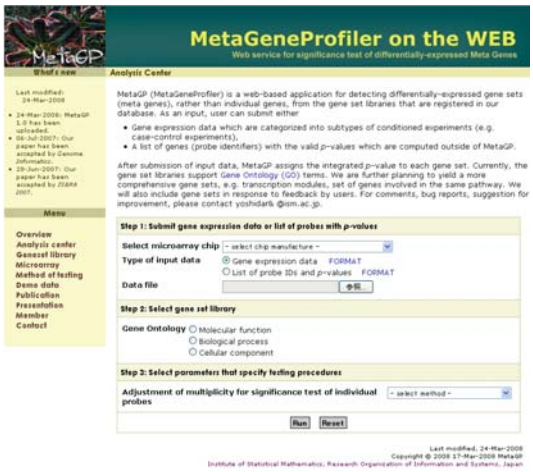
生物学的知識を無向グラフィカルモデルであるマルコフランダム場モデルとして表現し，DNA アレイデータに現れた遺伝子間相互作用を表すベイジアンネットワークを組み合わせた枠組みで研究をすすめた．

ベイジアンネットワークの矢印の太さを推定する非線形回帰問題において，通常は説明変数に誤差を仮定しないが，本研究ではそこにも観測誤差を仮定した．一つの条件付分布について非線形回帰を行い，そこからの結果をその遺伝子が影響している次の遺伝子に関する非線形回帰に反映させる．それらを順次繰り返すことで，トータルで一つの同時分布を推定する作業とした．一つ一つの非線形回帰問題で問題となる，ブートストラップのやり方に関するフリーなパラメータ値は，n 個の条件付分布に関しても一定のものになるような評価でもって，consistent になるように定めた．([A],[B]の成果)

マルコフランダム場モデルとして表された事前情報をどの程度組み込むかはハイパーパラメータが制御するが，その決定を最尤法で行おうとすると組み合わせ爆発をする計算を行う必要がある．従ってこれまでは周辺尤度の上限値，下限値をもとめることで，間接的にこの値を推定していた．よって，この尤度の見積もりがどの程度精密なのか，あるいは全く別の観点からのハイパーパラメータの決定法を考察するなど，ハイパーパラメータの推定法に関する数理的課題に取り組んだ．ベイジアンネットワークを利用した情報処理でもっとも計算負荷の高い作業は最適グラフ構造の探索である．これにはかなりの計算量が必要であるため，工夫したアルゴリズムを高速な並列計算機に実装した．確率的最適化手法やグラフ探索のアルゴリズムのハイパーフォーマンスコンピューティングが予想以上に急進展したことにより、並列度をあげた計算機環境のもとでアルゴリズムのパーフォーマンスの十分な検討も行った．

本研究では Greedy アルゴリズムを用いてグラフの MAP 解を求めているが，MAP 解に至るまでに探索的に得られた途中解を大量に蓄積し，グラフマイニング技術を用いて多くの因果構造ネットワークに共通する不変的構造及び各ネットワークに固有な特徴的構造の発掘を試みた．([D],[F]の成果)

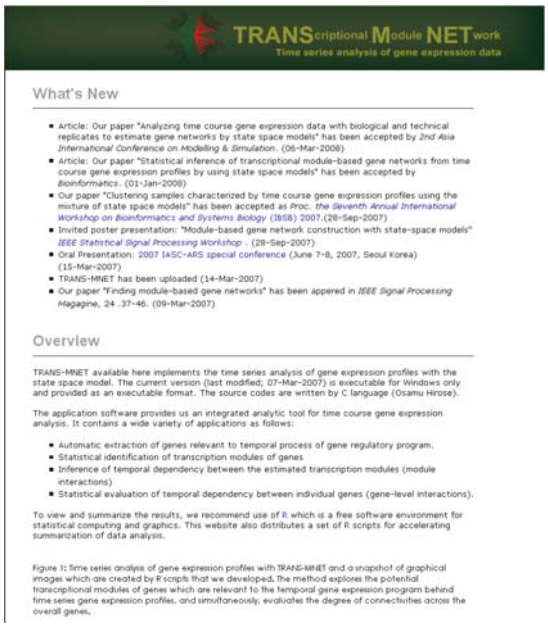
状態空間モデルを用いて，マイクロアレイデータのタイムコースデータからの遺伝子ネットワーク推定も試みた．アレイデータは，通常の時系列データと違い，その時点数（サンプル数）が１０数点といった極めて少ないことが特徴である．一方，観測ベクトルの次元は数千から一万超にもなり，状態推定に関して不定となる問題設定が容易におこる．我々は状態変数ベクトルの次元を超低次元とし，実体的意味のなかった状態変数に遺伝子モジュールの概念を付与することで状態空間モデルの有効性を一連の研究により示し続けている．当初は最尤法によりパラメータ推定を行っていたが，モデルを階層化し，さらに正則条件を加えることで，動的なモジュールネットワークの変遷を推定することを可能にした．同様の枠組みを用いた，既に開発済みのクラスタリングの手法をソフト化し，その解説をソフトウェア公開論文としてトップジャーナルに発表した．開発したアルゴリズムを Web 上でサービスあるいは公開(MetaGeneProfiler on the WEB http://metagp.ism.ac.jp/ 及び，TRANScriptional Module NETwork http://daweb.ism.ac.jp/~yoshidar/software/ssm/)し，継続的に機能拡充とユーザーインターフェースの整備を行った．([C]の成果)



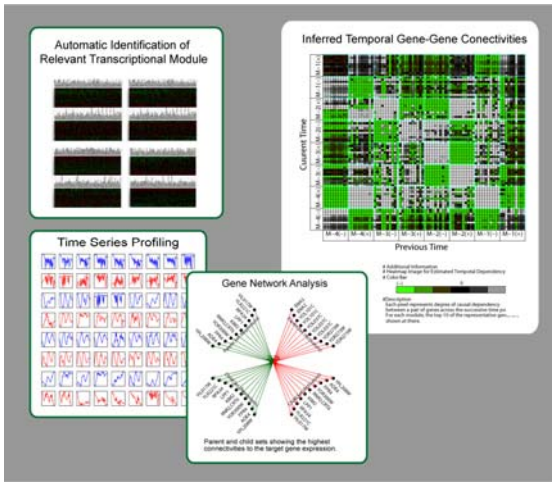
図：MetaGeneProfiler の Web ページ

また，状態空間モデルと，バイオインフォマティクスの分野で活発に研究されているカーネル法を融合する研究も進めた．カーネル法では通常データそのものを入力とする場合が推奨されているが，前処理として状態空間モデルを利用して，一信号をあえて多信号に分解し，それらを入力とするアプローチを採用することの有効性を示せた．また判別においてさまざまカーネルが一般には可能であるが，特にスペクトル判別に注目し，ま

た正定値性が保証されていない KL カーネルをあえて採用することで，微妙なスペクトル判別が精度良くできることを実例を用いながら示した．これらの統合手法の有効性が予想以上のものであることが確認できたため，研究開発した手法の有効性のアピールと拡張性の検討のために必要な器機を導入した．本機器を用いて身近な環境で取得可能なデータにもとづき，手法の有効性の確認実験を行った．（[H]の成果）



図：状態空間モデルによるアレイデータ解析手法の公開ホームページ．



図：開発したソフトウェアによる解析結果の例示．

タンパク質工学における望みの機能をもつタンパク質の発見の過程を情報科学における最適化とみなし，タンパク質工学への応用のための遺伝的アルゴリズムの一手法を示しつつ，その有効性を計算機実験により確認しウェットな実現の可能性を探った．（[E]の成果）

一般に生物学的知識は，データベースの形に具現化されることが多い．データベースに登録された情報にも信頼度の属性を与え，つまりそれを確率変数として取り扱うことで、データベースの登録上の過誤やその情報の不確実さをモデル化した．これまでのモデルをさらに階層化したベイズモデルを構成し，マイクロアレイデータからの情報抽出，既存の生物学的知識の有効活用，データベースの信頼性の検証などを統一的に可能にする枠組みを考案、それを人工データへ適用し性能評価を行った．（[G]の成果）

国際的に研究の進捗状況を傍観すると（研究課題提案時の H16 年時には），数多くの統計学者が，特にアメリカ，イギリスを中心に，生物学者，医者などバイオ関係の研究者と DNA アレイのデータ解析の共同研究をスタートさせていた．当時は単一のデータセットからの解析方法の研究が中心ではあったが，現在は異種のデータベースからの情報統合をシンポジウムテーマとする研究集会も活発に開催されている．本研究課題はまさにこの動向を先取りする形で実施されたもので，今後とも DNA データ解析法の研究に関して先陣を切るように努力していきたい．日本独自の方法論の展開がなければ，ナショナルセキュリティーにかかわる問題となり，国際的動向に俊敏な対応を踏まえた迅速な開発研究が真に望まれる．

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　17　件）

① O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, D. S. Charnock-Jones, C. Print, 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, 932-942, 2008. 査読有

② R. Yoshida, M. Nagasaki, R. Yamaguchi, S. Imoto, S. Miyano, T. Higuchi, Bayesian learning of biological pathways on genomic data assimilation, Bioinformatics Vol.24, No.22, 2592-2601, 2008. 査読有

③ 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, 21:101-113, 2008. 査読有

④ 石垣司, 樋口知之, 渡辺嘉二郎, Kullback-Leiblerカーネルによる正規化周波数スペクトル判別とその圧力調整器劣化診断への応用, 電子情報通信学会論文誌D, Vol. J90-D, No. 10, 2787-2797, 2007. 査読有

⑤ R. Yoshida, T. Higuchi, S. Imoto, S. Miyano, Mixed factors analysis: Unsupervized statistical discrimination with kernel feature extraction, Proceedings of The International Workshop on Data-Mining and Statistical Science (DMSS2007), No.25, 71-88, 2007. 査読有

⑥ 石垣司, 樋口知之, Kullback-Leibler カーネルの正規化スペクトル判別における特性, 第4回人工知能学会データマイニングと統計数理研究会(JSAI SIG-DMSM)予稿集,2007. 査読無

⑦ R. Yoshida, T. Higuchi, S. Imoto, S. Miyano, ArrayCluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles, Bioinformatics, 22, 1538 - 1539, 2006. 査読有

⑧ M. Nagasaki, R. Yamaguchi, R. Yoshida, S. Imoto, A. Doi, Y. Tamada, H. Matsuno, S. Miyano, T. Higuchi, Genomic Data Assimilation for Estimating Hybrid Functional Petri Net from Time-course Gene Expression Data, Genome Informatics(IBSB2006), 17, (1), 46-61, 2006. 査読有

⑨ T. Ishigaki, T. Higuchi, K. Watanabe, Spectrum classification for early fault diagnosis of LP gas pressure regulator based on Kullback-Leibler Kernel, Proceedings of the 2006 IEEE Signal Processing Society Workshop(MLSP2006), 453-458, 2006. 査読有

⑩ 石垣司, 樋口知之, 渡辺嘉二郎, Kullback-Leibler カーネルを用いたSVMによる高圧ガス圧力調整器の早期故障診断, Proceedings of The International Workshop on Data-Mining and Statistical Science, 220-225, 2006. 査読有

⑪ S. Tasaki, M. Nagaski, M. Oyama, H. Hata, K. Ueno, R. Yoshida, T. Higuchi, S. Sugano, S. Miyano, Modeling and Estimation of Dynamic EGFR Pathway by Data Assimilation Approach Using Time Series Proteomic Data, Proceedings of Genome Informatics 2006, 17(2), 226-238, 2006. 査読有

⑫ R. Yamaguchi, T. Higuchi, State-space Approach with the Maximum Likelihood Principle to Identify the System-Generating Time Course Gene Expression Date of Yeast, International Journal of Data Mining and Bioinformatics, Vol.1, No.1, 77-87, 2006. 査読有

⑬ S.Imoto, T.Higuchi, T.Goto, S. Miyano, Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks Statistical Methodology, Statistical Methodology, 3, 1-16, 2006. 査読有

⑭ S. Imoto, T. Higuchi, S. Kim, E. Jeong, S. Miyano, Residual Bootstrapping and Median Filtering for Robust Estimation of Gene Networks from Microarray Data, Proceedings of Computational Methods in Systems Biology, 3082, 149-160, 2005. 査読有

⑮ R. Yamaguchi, S. Yamashita, T. Higuchi, Estimating gene networks with cDNA microarray Data Using State-space models, Proceedings of 2005 International Workshop on Data Mining and Bioinformatics, 3482, 381-388, 2005. 査読有

⑯ R.Yoshida, S. Imoto, T. Higuchi , A Penalized Likelihood Estimation on Transcriptional Module-based Clusterintg, Proceedings of 2005 International Workshop on Data Mining and Bioinformatics, 3482, 389-401, 2005. 査読有

⑰ R.Yoshida, S. Imoto, T. Higuchi , Estimating Time-Dependent Gene Networks from Time Series Microarray Data by Dynamic Linear Models with Markov Switching, Proceedings of Computational Systems Bioinformatics Conference(CSB2005), 289-298, 2005. 査読有

〔学会発表〕(計 65 件：以下最近のもの)

① 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, The 19th International Conference on Genome Informatics (GIW2008), 2008年12月2日, Gold Coast ,

Australia

② O. Hirose*, R. Yoshida, R. Yamaguchi, S. Imoto, T. Higuchi, S. Miyano, Clustering with time course gene expression profiles and the mixture of state spase models, The Seventh Annual International Worksyop on Bioinformatics and Systems Biology (IBSB2007), 2007年8月2日, 東京大学医科学研究所

③ M. Nagasaki, R. Yamaguchi*, R. Yoshida, S. Imoto, A. Doi, Y. Tamada, H. Matsuno, S. Miyano, T. Higuchi, Genomic Data Assimilation for Estimating Hybrid Functional Petri Net from Time-course Gene Expression Data, The Sixth International Workshop on Bioinformatics and Systems Biology (IBSB2006), 2006年7月24日, Boston, MA USA.

④ R. Yoshida, S. Imoto, T. Higuchi*, Estimating Time-Dependent Gene Networks from Time Series Microarray Data by Dynamic Linear Models with Markov Switching, Computational Systems Bioinformatics Conference(CSB2005), 2005年8月9日, Stanford, CA USA.

〔図書〕(計 4 件)

① 樋口知之, シュプリンガー・ジャパン, パターン認識と機械学習 下 - ベイズ理論による統計的予測[翻訳＆監訳], 2008, 総ページ433

② 樋口知之, 東京電機大学出版局, ベイジアンモデリングによる実世界イノベーション 統計数理は隠された未来をあらわにする, 2007, 総ページ136

③ 分担執筆 T. Higuchi (S. Miyano 他 編集), 共立出版, バイオインフォマティクス事典, 2006, 総ページ807

④ 樋口知之 (他 北川源四郎・岸野洋久・山下智志・川崎能典), 共立出版, モデルヴァリデーション (第3章 : 地球科学におけるモデルヴァリデーション), 2005, 総ページ210

〔その他〕

http://www.ism.ac.jp/~higuchi/
http://daweb.ism.ac.jp/
http://metagp.ism.ac.jp/
http://daweb.ism.ac.jp/~yoshidar/software/ssm/

ソフト開発

Mta Gene Profiler Web インターフェース

６．研究組織

(1)研究代表者

樋口 知之 (HIGUCHI TOMOYUKI)
統計数理研究所・モデリング研究系・教授
研究者番号：70202273

(2)研究分担者

川崎 能典 (KAWASAKI YOSHINORI)
統計数理研究所・モデリング研究系・准教授
研究者番号：70249910

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

玉田 嘉紀 (TAMADA YOSHINORI)
統計数理研究所・統計科学技術センター・助教
研究者番号：

(3)連携研究者

上野 玄太 (UENO GENTA)
統計数理研究所・モデリング研究系・助教
研究者番号：40370093

染谷 博司 (SOMEYA HIROSHI)
統計数理研究所・モデリング研究系・助教
研究者番号：00333518

井元 清哉 (IMOTO SEIYA)
東京大学医科学研究所・ヒトゲノム解析センター・准教授 研究者番号：10345027

(4)研究協力者

アレックス ターミエ (ALEXANDRE TERMIER)
情報・システム研究機構・新領域融合研究センター・融合プロジェクト特任研究員
研究者番号：60435823

石垣 司 (ISHIGAKI TSUKASA)
総合研究大学院大学複合科学研究科・博士課程
研究者番号：20469597