

大規模データ解析と人工知能技術による
がんの起源と多様性の解明

文部科学省「富岳」成果創出加速プログラム



Newsletter

No. 5

large-scale data analysis and artificial intelligence technology

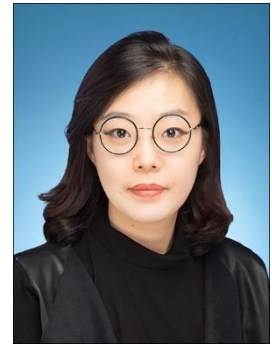
Unravelling origin of cancer and diversity by

2022年11月

「富岳」成果創出加速プログラム 大規模データ解析と人工知能技術によるがん の起源と多様性の解明

PredictiveNetwork: 胃がんの薬剤耐性メカニズム解明を目指す遺伝子ネットワーク解析

課題参加者 東京医科歯科大学M&Dデータ科学センター 教授 Heewon Park



胃がんは、日本を含め世界がん死亡数の順位の中、第3位であり(最新がん統計, 2020; Global Burden of Disease, 2019)、患者さんの数も多くなっています。胃がんの治療法としては抗がん剤を用いた薬物療法(化学療法)が一般的です。しかしながら、がん細胞が薬剤耐性を獲得してしまい、化学療法が効かなくなることが、胃がんの治療において重大な障害となっています。本研究では、胃がん抗がんの薬剤耐性獲得メカニズム解明やそれに寄与する遺伝子マーカー探索を目指して、ネットワーク推定技術を開発し、米国FDAから承認された胃がんの薬 doxorubicin, mitomycin-c, 5-Fluorouracil (5-FU), and docetaxelの感受性・耐性関連遺伝子ネットワーク解析を行いました (Park et al., BMC Bioinformatics. 2022;23(1):342)。

1. ネットワーク推定の機械学習技術開発

1.1. 従来の研究の限界

薬剤耐性のメカニズム解明はこれまで多くの研究で注目されており、計算科学分野においても様々な研究が行われています。従来の研究では遺伝子発現量差に着目してそこから薬剤耐性マーカー探索や

それらの関連メカニズムを調べる研究が主となっています。しかしながら、がんの薬剤耐性に関する複雑なメカニズムを解明するためには、各遺伝子の発現異常に基づく解析だけでは不十分と思われます。そこで、遺伝子間の因果関係を表す遺伝子ネットワークを解析し、がん細胞が薬剤耐性を獲得するメカニズムで重要な役割をしている遺伝子同士の相互作用や、そこでどの遺伝子を活性・抑制すれば薬剤耐性を獲得することを防止できるのか、などに関する情報抽出が重要と考えられます。

そのため、ネットワーク解析に基づく薬剤耐性の予測やメカニズム解明に関する研究が注目されています。これまでの研究では、ネットワークを先に構築し、構築されたネットワークをモデルのインプットとして用いる解析がほとんどでした。その場合、抗がん剤耐性関連メカニズムを含め我々が明らかにしたいがんのメカニズムがネットワークの推定に反映されていないため、そこからがん細胞の薬剤耐性メカニズム解明に重要な情報を抽出するには限界があります。

1.2. 開発した技術: PredictiveNetwork

本研究では、**がん細胞の薬剤感受性・耐性状況を説明するのに最適化された遺伝子ネットワーク推定の機械学習手法 (PredictiveNetwork)を開発**しました。開発したPredictiveNetworkの一つの特徴は、遺伝子ネットワークの推定と薬剤感受性の予測が同時できるということです。遺伝子ネットワークを推定する際、ネットワーク推定の誤差を最小化するだけではなく、薬剤感受性・耐性予測誤差も最小化することによって、がん細胞の薬剤感受性・耐性状況を説明するために最適な遺伝子間の因果関係推定を可能にする技術です(数理的な詳細は5ページのSupplementary Pageをご覧ください)。

2番目の特徴は、Network Biologyの知識を予測モデルに反映させながらモデル推定を行うことです。既存の機械学習モデルは統計・アルゴリズムの観点から開発され、結果に対するバイオロジー的な解釈が上手くできないという問題がありました。本研究では、「多数のリンクを持つHub遺伝子が遺伝子制御や関連するバイオロジープロセスで重要な役割をする」

癌細胞の薬剤感受性・耐性の状況を説明するのに 最適な遺伝子ネットワーク推定

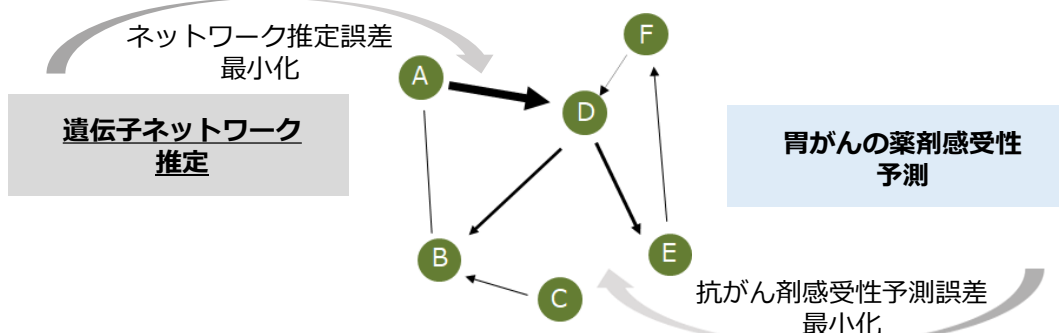


図1: 開発したPredictiveNetworkの概要

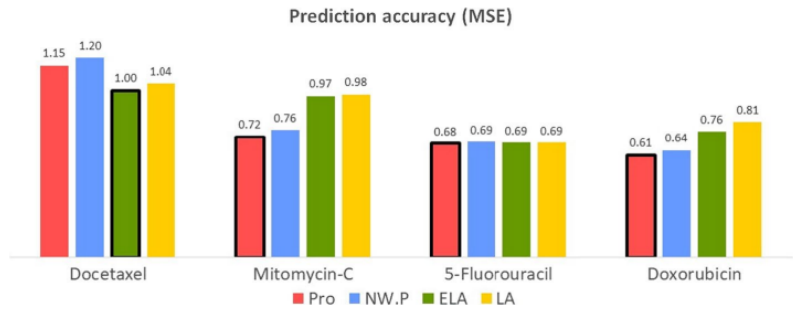


図2: 胃がん薬剤感受性の予測結果

ことと、「ネットワーク上で繋がりがある遺伝子同士は類似のバイオロジー機能を持つ」というネットワークバイオロジーの知見を予測モデルに反映させました。これにより、Hub遺伝子が予測モデルにおいて大きな影響を持つ/繋がりがある遺伝子同士は予測モデルで似たような影響を持つ、ことを可能にするモデリング手法を開発しました(数理的な詳細は5ページのSupplementary Pageをご覧ください)。これにより、あるパスウェイに乗っている遺伝子群(Hub遺伝子とその周りにある遺伝子で構成されたsub-network)で予測モデルが構築され、解析結果に高解釈性を与えることができました。

2. ネットワーク解析による胃がんの薬剤耐性メカニズム解明

開発したPredictiveNetworkを用いて、胃がんの薬剤耐性メカニズムの解明を目指し、胃がんに対する4つの抗がん剤予測やその予測に向けた(薬剤感受性を説明するのに最適な)遺伝子ネットワーク推定を行いました。

【データ】Cancer Genome ProjectのSanger Genomics of Drug Sensitivity in Cancer (GDSC) のデータ。約30種類のがんの約

900の細胞株; 胃がんの抗がん剤データ: 米国FDAから胃がん抗がん剤として承認された18薬剤中、GDSCに含まれている4つの抗がん剤(Doxorubicin, Mitomycin-c, 5-Fluorouracil (5-FU), Docetaxel)の薬剤感受性(IC50値のz-score); 発現量データ: 9,764個の遺伝子から、発現量の分散が大きい976(9,764の10%)遺伝子の発現量。

2.1. 数理的な予測精度

PredictiveNetworkの数理的精度を評価するために、既存の手法と予測精度を比較しました。ネットワークを先に推定し、推定されたネットワークをインプットする予測モデルの結果(NW.P)、ネットワークではなく発現量に基づく予測モデル結果(Elastic net: ELA; Lasso:LA)、PredictiveNetwork(Pro)の予測結果(予測誤差:平均二乗誤差)を図2に示しています。ネットワークに基づく予測モデル(NW.P・Pro)が、発現量に基づくモデル(ELA・LA)より全般的に高精度を達成していることが確認できました。また、提案したPredictiveNetworkは4つの抗がん剤中、Mitomycin-c, 5-Fluorouracil (5-FU), Docetaxelの予測に対して一番良い予測精度を達成しました。薬剤感受性を説明するためには、各遺伝子の発現量だ

けでは不十分であり、遺伝子間の因果関係を表すネットワークの考慮が重要であることと、開発されたPredictiveNetworkは薬剤感受性を説明するのに最適なネットワークを推定したことが示されたと考えられます。

2.2. 胃がん薬剤耐性のメカニズム解明

まず、推定された遺伝子ネットワークから、抗がん剤予測マーカー探索を行いました。4つの抗がん剤の予測に共通な重要マーカーを抽出しました(図3)。今回解析から得られた7個の遺伝子中、5個の遺伝子は胃がんと胃がんの抗がん剤のマーカーとして知見が発表されていましたが、残りの2個(ZNF204P・C1QL1)に対しては、まだ先行研究は全くなく関連メカニズムは不明でした。

胃がんの薬が効く細胞と効かない細胞における予測マーカーの因果関係の特徴を抽出するため、薬剤感受性・耐性細胞に対してネットワークを推定しました。4つの胃がん薬剤が全ての効く・効かない細胞(それぞれ95細胞株)を抽出し、その細胞における予測マーカーのネットワークを推定したものを図4に示しています。

遺伝子	胃がんの抗がん剤	胃がん
AKR1C1	Biochem Pharmacol. 2013 Int J Mol Sci. 2021 BMC Cancer. 2012 Oncotarget. 2016 J Exp Clin Cancer Res. 2019 Pharmacol Rev. 2021 Oncol Rep. 2017	Int J Mol Sci. 2021 Oncotarget. 2016 Sci Rep. 2016 Chem Res Toxicol. 2017
ZG16B1	Oxid Med Cell Longev. 2020 Cancers (Basel). 2021	Oncotarget. 2017
CRYAB	Open Med (Wars). 2020	J Comput Biol. 2019 J Cell Mol Med. 2018 Medicine (Baltimore). 2021
ZNF204P	-	-
C1QL1	-	-
TMEM139	-	Medicine (Baltimore). 2019
PEG10	PLoS One. 2017 Oncotarget. 2017 Clin Invest Med. 2012	PLoS One. 2017 J Mol Endocrinol. 2019

図3: 抽出された胃がん抗がん剤の予測マーカー

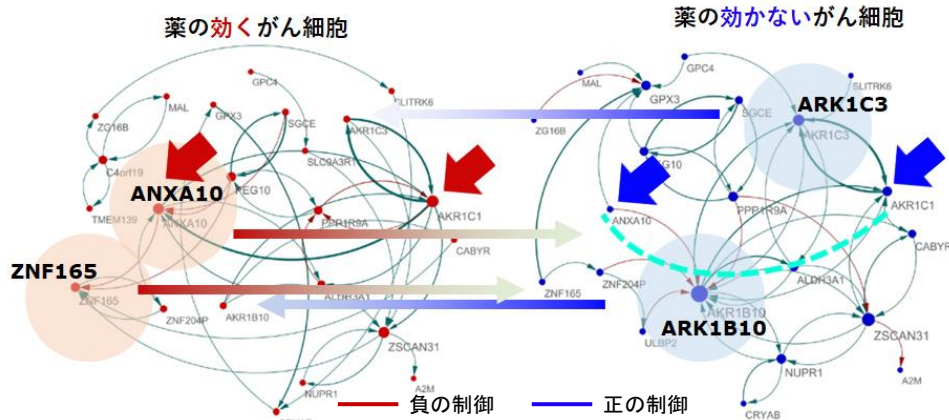


図4:薬が効く・効かない細胞における胃癌薬剤予測マーカーのネットワーク

予測マーカーとして抽出された遺伝子は薬剤が効く細胞と効かない細胞において異なる因果関係を示しています。薬剤が効く細胞においてはAKR1C1 がANXA10を

強く制御していますが、薬剤耐性を持つ細胞においてはその関係がなくなっていることが確認できました。また、ANXA10とZNF165のhubness(多数のリンクを持つ)

が薬剤が効く細胞の特徴として、ARK1C3とARK1B10のhubnessが抗がん剤が効かない細胞の特徴として見つかりました。

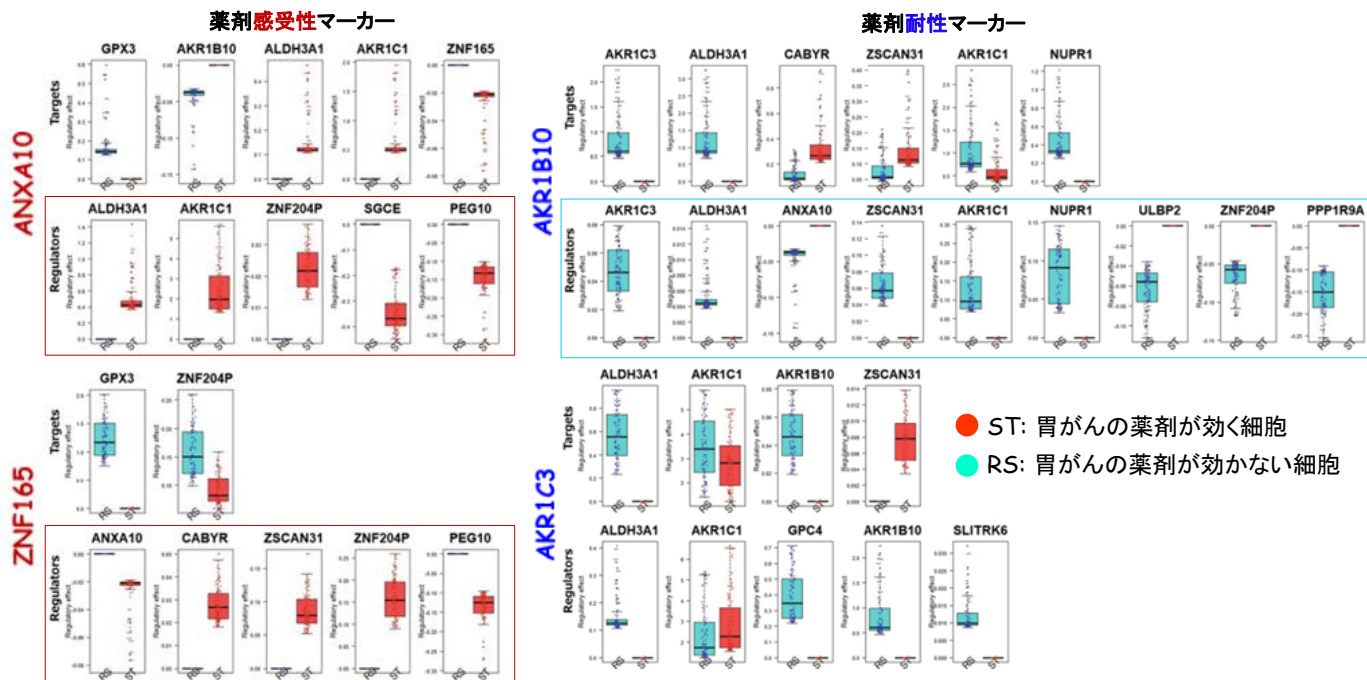


図5: 薬が効く・効かない細胞における薬剤感受性マーカーと耐性マーカーのアクティビティ(regulatory effect)

抽出された薬剤感受性マーカーANXA10・ZNF165と薬剤耐性マーカーARK1C3・ARK1B10の薬が効く・効かない細胞におけるアクティビティをregulatory effect (発現量x制御の強さ)に基づいて確認しました(図5)。各マーカーのターゲット遺伝子に対するregulatory effect(図5のTargets)とマーカーのレギュレーター遺伝子のマーカーに対するregulatory effect(図5のRegulators)を計算しました。薬剤感受性マーカー(図5左)は薬が効く細胞におい

て、薬剤耐性マーカー(図5右)は薬が効かない細胞において、それぞれ高いアクティビティを示していました。特に、薬剤感受性マーカー(ANXA10・ZNF165)のレギュレーター遺伝子は、薬が効かない細胞においてはアクティビティがなくなっていることと、薬剤耐性マーカーARK1B10のレギュレーター遺伝子は薬が効く細胞においてはアクティビティを失っていることが確認できました。

まとめると、AKR family遺伝子の高いアクティビティとANXA10のアクティビティ減少ががん細胞が胃癌薬剤の耐性を獲得するのに寄与しているのではないかと推察されます。

【本研究の解析からの提案】:「ARK遺伝子ファミリーを抑える」とこと「ANXA10・ZNF165を活性化すること」が胃癌薬剤耐性の防止ためにカギだと考えられます。

【数理モデル】

q 個のレギュレーター遺伝子 $R_j = (r_{1l}, r_{2l}, \dots, r_{nl})^T, j = 1, \dots, q$ が l -番目のターゲット遺伝子 $T_l = (t_{1l}, t_{2l}, \dots, t_{nl})^T$ の転写を制御しているとする。

$$T_l = \sum_{j=1}^q \beta_{jl} R_j + \varepsilon_l, \quad l = 1, \dots, k,$$

ただし β_{jl} は l -番目のターゲット遺伝子に対する j -番目レギュレーター遺伝子の影響の大きさを表す回帰係数である。

【従来のネットワーク手法】

$$\hat{\beta}_l = \operatorname{argmin} \left\{ \frac{1}{2} \sum_{i=1}^n (t_{il} - \sum_{j=1}^q \beta_{jl} r_{ij})^2 + P_{\lambda\delta}(\beta_l) \right\}$$

ただし $P_{\lambda\delta}(\beta_l)$ は L_1 正則化ペナルティである、e.g., Elastic net

$$P_{\lambda\delta}(\beta_l) = \lambda \sum_{j=1}^p \{(1-\delta)\beta_{jl}^2 + \delta|\beta_{jl}|\}.$$

■ 従来のネットワーク解析の限界

: ネットワークを先に推定し、推定されたネットワークを癌メカニズムの解明や関連予測に用いる。

→ ネットワークはがん関連メカニズムを反映せずに推定される

: 従来の数理モデル・解析手法はバイオテクノロジー知識は考慮せず、数理・統計観点から開発されている。

→ 結果に関するバイオテクノロジー的な解釈の問題が存在

【開発したPredictiveNetwork】

薬剤感受性の n 個の細胞において観測を $z = (z_1, z_2, \dots, z_n)^T$ とする。

$$\operatorname{argmin}_{B, \theta} \left\{ \sum_{i=1}^n (z_i - \theta_0 - \theta^T B^T x_i)^2 + \sum_{j=1}^k \sum_{i=1}^n (y_{ij} - \beta_j^T x_i)^2 \right\} \quad ①$$

薬剤感受性予測 遺伝子ネットワーク推定

$$+ \lambda_1 \sum_{j=1}^k \|\beta_j\| + \lambda_2 \sum_{j=1}^k \|\beta_j\|^2 \quad ②$$

$$+ \lambda_3 \|\theta\| + \lambda_4 \sum_{q=1}^k \sum_{j=1}^k \left(\frac{\operatorname{sgn}(\theta_q) \theta_q}{\sqrt{d_q}} - \frac{\operatorname{sgn}(\theta_j) \theta_j}{\sqrt{d_j}} \right)^2 w_{qj}$$

ただし $\theta = (\theta_1, \theta_2, \dots, \theta_k)^T$ は予測モデルの回帰係数、 $B = (\beta_1, \beta_2, \dots, \beta_k)$ は遺伝子間の制御関係を表す $p \times k$ 行列、 $w_{qj} = (|B_{qj}| + |B_{jq}|)/2$ 、 $d_j = \sum_{i=1}^p w_{ji}$ は j 番目の遺伝子の degree とする。

■ PredictiveNetwork の特徴

① 遺伝子ネットワーク推定と抗がん剤感受性の予測を同時に行う。

: 遺伝子ネットワークはネットワーク推定誤差を最小化するだけでなく、予測誤差も最小化するように推定される

→ 細胞の薬剤感受性の状況 z を説明するのに最適なネットワーク推定

② ネットワークバイオテクノロジーの知識を予測モデルに反映

: 制御関係を持つ(ネットワーク上で繋がりがある)遺伝子同士(例、 x_q と x_j) は類似のバイオロジー機能を持つ

→ その遺伝子同士は予測モデルにおいて類似な影響($\theta_q \approx \theta_j$)を持つように

: 多数のリンクを持つ hub 遺伝子が遺伝子制御や関連バイオロジープロセスで重要な役割をする

→ Hub 遺伝子 ($d_k \uparrow$) は予測モデルに大きな影響($\theta_k \uparrow$)を持つように

→ ある pathway に乗っている遺伝子群(hub 遺伝子とその周りの遺伝子で構成された sub-network) で予測モデル構築

編集者メモ

- ・ 2022年12月18日(日)14:00-16:30:「富岳」成果創出加速プログラム「大規模データ解析と人工知能技術によるがんの起源と多様性の解明」とNPO法人バイオインフォマティクス・ジャパンの共同開催によるシンポジウム「**がんはどのようににはじまり、そしてどのようにに強敵にばけるのか**」をオンラインで開催予定。
<https://odcla.mddsc.jp/posts/2022/09/2022-09-02-symposium-fy2022/>
- ・ 2022年10月8-12日:ベルリンで開催された第21回International Conference on Systems Biologyで、東京医科歯科大学M&Dデータ科学センターのHeewon Park教授と宮野悟特任教授が、説明可能AIを「富岳」に実装して得られた研究成果について招待講演

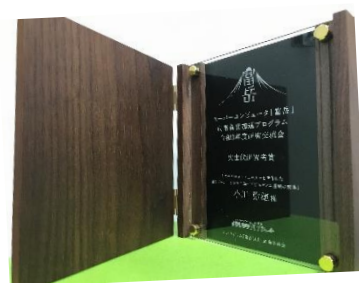
「XAI×Fugaku uncovered EMT mechanisms via cancer characteristic specific gene network analysis (Park H, Maruhashi K, Miyano S)」を行いました。また、Heewon Park教授はこのニュースレターで紹介している研究内容についても講演を行いました。生命システムの解明には人工知能技術とスパコンの融合がカギです。

<https://www.icsb2022.berlin/>

- ・ 2022年3月14-15日:東京医科歯科大学M&Dデータ科学センター統合解析分野の特任助教の小川弥穂先生が、「富岳」成果創出加速シンポジウム・研究交流会で発表した「スーパーコンピュータを活用した全ゲノムシーケンスに基づくがんゲノム医療の現場」が次世代研究者賞を受賞しました(全体で2名が受賞していますがそのうちのー

人です)。インタビュー動画が配信されました。インタビュアーは、早稲田大学総合研究機構グローバル科学知融合研究所の高橋桂子教授で、小川弥穂先生に受賞対象の研究内容についてインタビューし、一般の方々にも分かりやすく解説しました。小川先生が若手研究者、女性研究者という点にもフォーカスし、研究と家庭との両立や、キャリアプラン等、後に続く若手研究者の励みとなる話もお聞きしています。

https://www.youtube.com/watch?v=cnB8m-OC_Ck



1. 代表機関

国立大学法人東京医科歯科大学
研究課題開発責任者
M&Dデータ科学センター 特任教授 宮野悟

2. 協力機関

国立大学法人京都大学
協力機関代表者
大学院医学研究科 教授 小川誠司

愛知県がんセンター

協力機関代表者
研究所システム解析分野 分野長 山口類

3. 連携機関

(株)富士通研究所
連携機関連絡担当者
富士通株式会社 フェロー 岡本青史
国立大学法人東京大学
連携機関連絡担当者
医科学研究所 准教授 片山琴絵
Karolinska Institutet
連携機関連絡担当者
Department of Medicine/Center for Hematology and Regenerative Medicine
Visiting Professor 小川誠司

4. 事業参加者

事業参加者(代表機関)
東京医科歯科大学M&Dデータ科学センター
統合解析分野 特任教授 宮野悟
准教授 長谷川嵩矩
助教 伊東聰
特任助教 角田将典
特任助教 小川弥穂
特任研究員 田中洋子
AI技術開発分野 教授 Heewon Park

事業協力者(協力機関)

京都大学大学院医学研究科
腫瘍生物学講座 教授 小川誠司
特定教授 南谷泰仁
助教 越智陽太郎
他11名

愛知県がんセンター研究所

システム解析分野 分野長 山口類

連携参加者(連携機関)

富士通株式会社 フェロー 岡本青史
富士通株式会社 富士通研究所
プロジェクトマネージャ 丸橋弘治

東京大学医科学研究所

ヒトゲノム解析センター 准教授 片山琴絵
先端医療研究センター 教授 南谷泰仁
(2022年11月1日時点)

スーパーコンピュータ「富岳」成果創出加速プログラムについて
(理化学研究所計算科学研究センターより抜粋)

<https://www.r-ccs.riken.jp/jp/fugaku/promoting-researches>

スーパーコンピュータ「富岳」成果創出加速プログラムは、「富岳」を用いた成果を早期に創出することを目的として文部科学省が設置しました。①人類の普遍的課題への挑戦と未来開拓、②国民の生命・財産を守る取組の強化、③産業競争力の強化、④研究基盤の4領域について課題の選定が行われ、19課題が選定されています。選定された課題は、「富岳」の計算資源を優先的に無償で使用し、速やかな成果創出を目指します(2020～2022年度)。

課題名: 大規模データ解析と人工知能技術によるがんの起源と多様性の解明
課題番号: JPMXP1020200102

課題ID

2022年度: hp220163
2021年度: hp210167
2020年度: hp200138

Information



文部科学省「富岳」成果創出加速プログラム
課題名: 大規模データ解析と人工知能技術によるがんの起源と多様性の解明
ニュースレター No. 5
発行日★2022年11月4日
課題代表者★宮野 悟

- 東京医科歯科大学M&Dデータ科学センター 統合解析分野
- 郵便物宛先: 〒113-8510 東京都文京区湯島1-5-45
- 宅配便宛先: 〒101-0062 東京都千代田区神田駿河台2-3-10
駿河台キャンパス22号館5階
- TEL: 03-5280-8620 FAX: 03-5280-8632
- E-mail: mdsc.dsc@tmd.ac.jp
- HP: <https://odcla.mddsc.jp>