ビッグデータ時代に向けた革新的アルゴリズム基盤

加藤　直樹

本稿では，ビッグデータに関する CREST 研究領域で，筆者が研究代表者を務めている研究プロジェクト「ビッグデータ時代に向けた革新的アルゴリズム基盤」（2014 年 10 月～2020 年 3 月）の概要とその最新の研究成果を紹介する．

キーワード：CREST 研究領域，ビッグデータアルゴリズム，劣線形時間パラダイム

1. はじめに

AI（人工知能），ビッグデータという言葉は，多くの人がメディアを通して聞いており，これらの技術がこれからの社会の変革を担うであろうということも何となく知っている．日本でも，この分野で世界をリードする立場に立つべく，AI，ビッグデータ技術に対する国家を挙げての投資が進んでいる．本稿では，その中でも，科学技術振興機構 (JST) が推進している「戦略的創造研究推進事業」の中核事業の，CREST と呼ばれるプロジェクトの一つである，「ビッグデータ基盤：ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」（研究総括者，国立情報学研究所所長，東京大学教授の喜連川優先生）に採用された，筆者が代表を務める研究プロジェクトを紹介する．

JST は，国の科学技術基本計画を推進するため，研究開発戦略，ファンディング，地域創生，人材育成，国際協力など幅広い事業を実施しており，世界トップレベルの研究開発を行うネットワーク型研究所として，未来共創イノベーションを先導し推進するという大きな使命を有している．

JST の推進する中心的事業に「戦略的創造研究推進事業」がある．国が定める戦略的な目標などの達成に向けた，革新的技術シーズの創出を目指す研究開発プログラムであり，大学・企業・公的研究機関などの研究者からなるネットワーク型研究所（組織の枠を超えた時限的な研究体制）を構築し，その所長であるプログラムオフィサー（研究総括など）による運営の下，研究を推進するものである．そのいくつかの事業の中に CREST という研究推進事業があり，国が定める戦略目標の達成に向けて研究総括の運営の下，独創的で国際的に高い水準の基礎研究を推進し，今後の科学技術イノベーションに大きく寄与する卓越した成果を創出することを目的としている．研究代表者が複数の共同研究グループを組織し実施するネットワーク型研究である．現在，30 以上の研究領域で CREST プログラムが実施されている [1].

「ビッグデータ基盤：ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」は，2013 年に始まった CREST プロジェクトの一つである．その戦略目標は，「分野を超えたビッグデータ利活用により新たな知識や洞察を得るための革新的な情報技術及びそれらを支える数理的手法の創出・高度化・体系化」である．

筆者は，電気通信大学の伊藤大雄氏，京都大学の牧野和久氏，東京大学の渋谷哲朗氏らと 2 年間の準備を経て，ビッグデータ基盤 CREST 研究領域の 2 期目に申請した研究計画が採択され，2014 年 10 月から研究を開始した．われわれの提案する研究課題名は「ビッグデータ時代に向けた革新的アルゴリズム基盤」である．本稿では，その研究課題の目的，研究計画の概要，および得られた主要な研究成果について報告する．

2. 研究概要と目的

CREST 申請書に掲げた目的と概要の一部を再掲すると以下のようになる．

インターネットを中心とする通信技術の発達，観測・測定機器の進歩により，超巨大データが蓄積されている．しかしながら，近年のデータサイズの急速な増大はハードウエアやアルゴリズムの進化の速

かとう　なおき
兵庫県立大学社会情報科学部
〒 651–2197 兵庫県神戸市西区学園西町 8–2–1

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

度を大幅に上回っている．このようなビッグデータを対象とした大規模問題群の解決には，従来のアルゴリズム理論では対応不可能であり，アルゴリズム革新による解決が強く望まれている．これまでは多項式時間アルゴリズムならば「速い」アルゴリズムであると考えられてきたが，ペタスケールやそれ以上のビッグデータに対して $O(n^2)$ 時間アルゴリズムを直接適用するだけでは，計算資源や実行時間などの点で大きな困難に直面する．少なくとも線形時間，場合によっては劣線形時間や定数時間アルゴリズムが求められている．

そのために，本研究では，ビッグデータ時代に向けた，劣線形時間アルゴリズムという新しい計算パラダイムを提唱する．具体的には，ビッグデータ用のアルゴリズム理論と，最新の圧縮・検索技法を基盤とする劣線形データ構造，統計力学的情報粗視化を基盤とするデータモデリングの 3 つの基礎理論から構成されている．ビッグデータのためのアルゴリズムに関しては，重要な基本問題に対して，確率，近似の概念を利用することにより，ペタスケールあるいはそれ以上のビッグデータでも実用時間で動作する，ほぼ線形，あるいは劣線形のアルゴリズム開発を目指す．ビッグデータのデータ構造としては，情報論的下限を突破する劣線形サイズの簡潔索引構造の開発を基礎とした劣線形データ構造パラダイムの構築を目指す．ビッグデータモデリング手法としては，ビッグデータ生成過程の理解を通して得られる統計的な特性や物理モデルを利用した劣線形モデリング技術パラダイムの構築を目指し，ビッグデータ時代に向けた革新的アルゴリズム基盤を構築する．

ここで，「劣線形時間アルゴリズム」という言葉の意味を説明しておく．入力データサイズを n とすると，n に比例するか，それより小さな時間で解を求めるアルゴリズムのことで，ビッグデータを対象とするアルゴリズムは，劣線形時間で動作することが求められ，われわれのプロジェクトでは，データ構造，モデリングも含めて，劣線形サイズで動作するアルゴリズムの開発を目指す考え方を提唱している．

本研究では，理論面だけでなく，具体的な応用分野を定めて，研究を推進していく．重要な応用分野として，大規模災害時における避難計画策定，たんぱく質立体構造と機能解明，シーケンサーデータ解析，経営データ分析の 4 分野に焦点を当てる．図 1 には，「革新的アルゴリズム基盤」概念と目標をまとめている．

図 1　革新的アルゴリズム基盤の構成図

先に述べたように，「劣線形時間アルゴリズム」，「劣線形データ構造」，「劣線形モデリング」の三つのアプローチを同時並行的に実現する必要があるが，これらの三つのアプローチに対応して，三つの研究グループを構成した．「劣線形時間アルゴリズム」グループは研究代表者の筆者がグループリーダーとなり，「劣線形データ構造」グループは東京大学医科学研究所の渋谷哲朗氏がリーダーを務め，「劣線形モデリング」グループは東北大学の田中和之氏が務めている．渋谷氏は，ゲノムデータ解析やデータ圧縮アルゴリズムにおいて顕著な業績を挙げておられ，田中氏は特定領域研究「確率的情報処理への統計力学的アプローチ」（平成 14〜17 年度）の代表者を務めるなど，情報統計力学の専門家である．

3.　劣線形アルゴリズムグループの主要な研究成果

3.1　複雑ネットワークに対する定数時間アルゴリズムの開発

インターネットやソーシャルネットワークなどは複雑ネットワークと呼ばれ，典型的なビッグデータの一つであり，近年の研究により，「直径が小さい」「次数の冪乗法則」「クラスタ性」などの性質があることがわかっている．電気通信大学の伊藤大雄氏は，複雑ネットワークの「冪乗法則」と「クラスタ性」に着目し，「階層的孤立クリーク構造」というある種の階層的なクラスタ性をもつグラフが，任意の性質が検査可能であること（定義の詳細は後述する）を証明した．この結果から，ほんの一部のデータのみを用いて，ネットワークのクラスタ性や中心性を定数時間で検証できる．

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

もう少し説明を加えよう．統計量の計算（平均，分散など）は全データを見ることなしに，ある程度のサイズのサンプルデータから高い精度で，平均，分散が推定できることが統計学では古くから知られている．つまり，定数時間アルゴリズムは，このような計算では可能である．

この延長線上で，グラフ，ネットワークのさまざまな指標の計算も定数時間で行える．これは，驚くべき事実であるが，1990 代からの理論的進歩により可能になってきた．

たとえば，グラフ G は，“連結グラフかどうか”という問いに対しては，ややこの問いを緩和して，

(1) グラフ G が連結グラフである．

(2) G から ϵm 個の辺を加えても非連結である（ここで m は G の辺数）．

の二つの問のいずれかが成り立つことを m に依存せず，ϵ のみに依存する計算時間で答えることができたら，定数時間で検査できるということにする．ただし，(1) でも (2) でもない場合，答えはどちらでもよいとし，この点が元々の問いである，“連結グラフかどうか”という問いを緩和している．これを性質検査 (property testing) と呼ぶ．このような性質検査は 1990 年代から研究が進み，グラフのさまざまな特徴量に対する性質検査アルゴリズムの開発が進んでいる．

日本では，吉田，伊藤を中心に優れた研究成果がこれまでに生まれている研究分野である．これまでの研究では，多くの場合，対象のグラフは密なグラフか，次数が定数 d 以下の粗グラフに限定されていた．伊藤氏は，SNS などの社会ネットワークを対象としてモデル化された階層的スケールフリー性をもつネットワークに対しても性質検査が可能であることを論文 [2] によって明らかにした．この結果は，本 CREST 研究の最大の理論的成果である．これにより，Web グラフなどの巨大グラフのグラフ的特性量の計算がサンプリングなどのテクニックによって定数時間で計算可能となった．これはまだ理論的成果に過ぎず，実用化に向けてはさらなる改良が求められている．

3.2 組合せ剛性理論によるたんぱく質機能解析

本プロジェクトでは，たんぱく質の機能活性化のメカニズムの解明を組合せ剛性理論を用いて解明することも目指している．組合せ剛性理論の研究は，日本では，谷川，加藤の分子構造予想の証明など，強い分野である [3]．分子を構成する原子を剛体と見て，原子間の結合や原子間に働く力を棒材を用いて表現するモデルは，分子フレームワークモデルと呼ばれる．分子構造をフレームワークで表現し，その剛性を調べる研究は，21 世紀に入って盛んに行われるようになっている（図 2）．

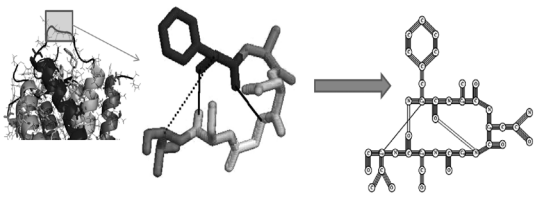

図 2 分子構造グラフ

たんぱく質の構造を分子フレームワークと見て，たんぱく質のどの部分が剛であるのかを高速に調べるのに，組合せ剛性理論に基づくアルゴリズムを用いることができる．たんぱく質は剛性や立体構造によってその機能が決まるため，可能な形状変化を予想することは今日の分子生物学における一大テーマである．

組合せ剛性理論の進展と最近の分子構造予想の解決により，計算機上でたんぱく質の挙動解析を実用時間で行うことが可能になってきた．われわれの CREST の研究員である Sljoka 氏は，北米の生化学者のグループとの連携により，生化学上のいくつかの重要な問題に取り組み，組合せ剛性理論に基づく高速な剛性判定アルゴリズムを適用して，重要な知見を明らかにした．以下に，Sljoka 氏の行った業績について簡単に触れておく．

(1) 酵素の触媒反応のメカニズムは，いまだ解明されておらず，科学の謎の一つである．酵素触媒反応に関する有名な仮説に，酵素の時間変化に伴う動きが酵素反応に重要な役割を果たしているのではないかというものがある．しかし，これまでこの仮説は立証されていなかった．Sljoka 氏は，トロント大学の生化学研究グループと共同で，その解明に迫る重要な発見を行い，その成果が，自然科学分野で最も影響力の高い *Science* に掲載された [4]．具体的には，フルオロ酢酸デハロギナーゼにおける酵素触媒による化学反応促進において，触媒反応時における酵素の動的変化が果たす役割を明らかにした．

Sljoka 氏は，この触媒反応における，たんぱく質動力学とたんぱく質内の遠隔情報伝達の（アロステリー信号伝達）役割を組合せ剛性理論とそれに基づくアルゴリズムを用いて解明した．これにより，従来，分子動力学シミュレーションでは計算量の膨大さゆえ困難であった酵素の動的変化の解明が可能となった．この論文において，剛性理論に基づく数理的アルゴリズム

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

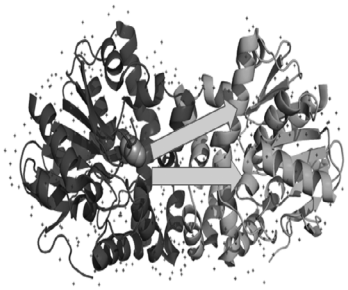

図 3　フルオロ酢酸デハギロナーゼとアロステリー信号伝達左の部分にある丸は酵素と結合したフルオロ酢酸を表している．

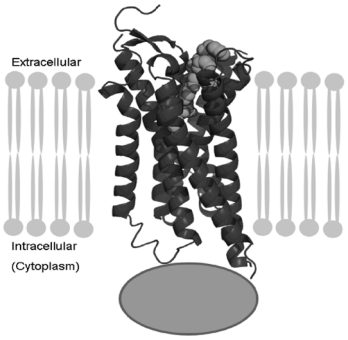

図 4　GPCR の説明図

が，たんぱく質の機能に関する生化学的謎の理解を助ける強力な道具であることが示されたが，さらに原子レベルで生命の謎を理解するための必須な道具であることも明らかにした [4].

たんぱく質における動的変化にアロステリー信号伝達が重要な役割を果たしている．アロステリー信号伝達とは，たんぱく質のある部分に薬などの小物質が結合すると，それが同じたんぱく質内の遠くの場所の構造変化や動的変化を引き起こすという現象であるが，その仕組みはよくわかっていなかった．Sljoka 氏は，アロステリー現象を剛性理論を用いてモデル化し，アロステリー現象が生じるメカニズムを説明することに成功した．また，これにより，アロステリー信号伝達が生じる部位のペアを高速に同定するアルゴリズムを開発した．Sljoka 氏のアルゴリズムに基づいて，酵素における互いに離れた位置にある二つの部分の間のアロステリー信号伝達に関する仮説を立証し，酵素の働きに関する謎の解明に迫る新しい知見を得ることに成功した．Sljoka 氏による剛性理論に基づくアプローチは，近年，アルツハイマー病に対する新薬開発などでも注目されている．離散数学とアルゴリズムの研究が，生化学分野における重要な発見に本質的に寄与したことを示している．

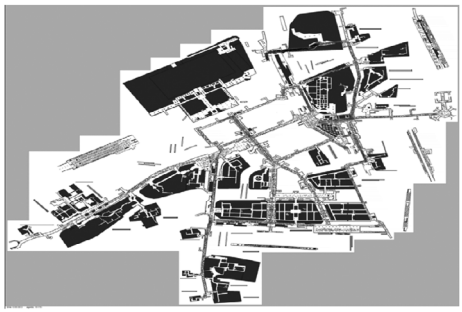

図 5　梅田地下街避難シミュレーション [7]

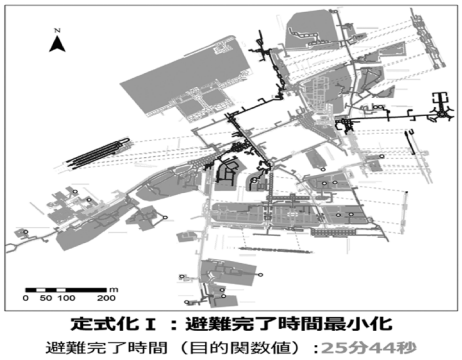

図 6　梅田地下街最適分割 [8]

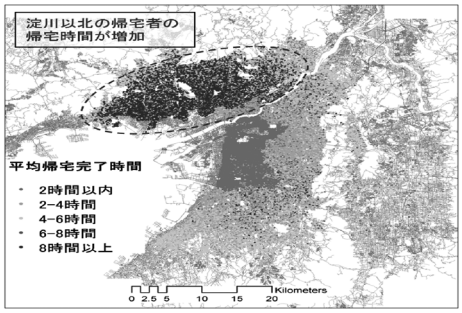

図 7　大阪市の大規模徒歩帰宅シミュレーション [9]

Sljoka 氏らによるほかの主要論文は [5, 6] である．GPCR とは，細胞膜の外部と内部間の信号伝達を行っている重要なたんぱく質で，薬，ホルモン，神経伝達に関与し，市場に出ている約 50%の薬が GPCR をターゲットにしていると言われているくらい重要なたんぱく質である（図 4）．論文 [5] はある種の GPCR に対してその信号伝達の仕組みを解明したものである．

3.3　避難計画問題

大阪市立大学の瀧澤重志氏は，大阪市梅田地下街の避難シミュレーションや，地下街を最適に区域割して，避難時間が最小になるように，地下街に接続する地上

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

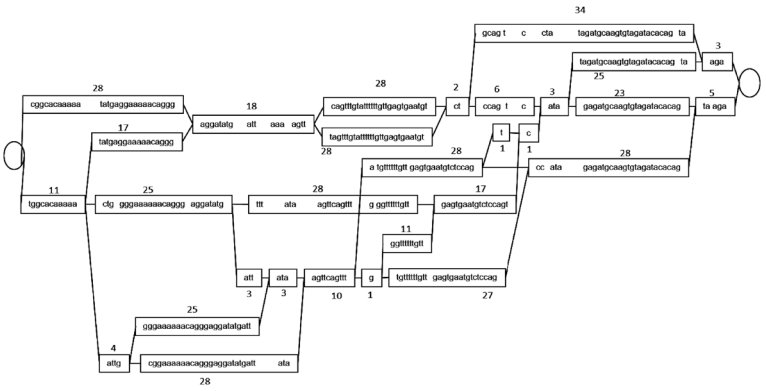

図 8　superbubble グラフの例 [11]

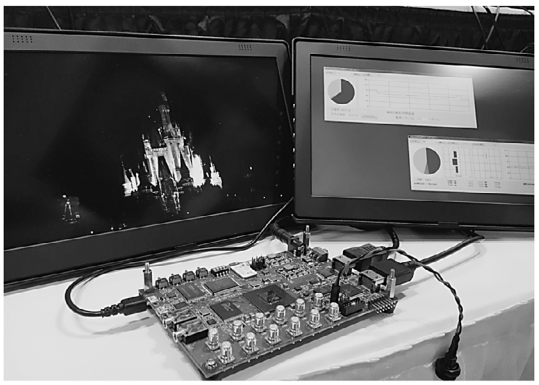

図 9　FPGA による基盤実装
画像を圧縮してリアルタイム転送している．

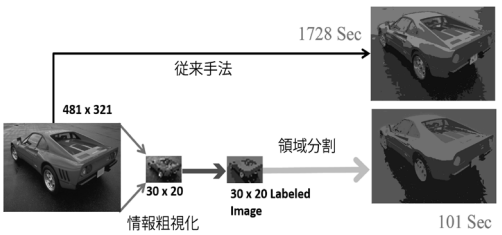

図 10　デジタル画像の領域分割に対する情報粗視化技術の適用例

の避難所（オフィスビルや商業ビル）を割り当てる問題を解くアルゴリズムを開発し，地下街の優れた領域分割を得た．また，大規模地震に伴って，大阪市のオフィス街に勤務する人たちの大規模一斉徒歩帰宅シミュレーションの実装と実施を行い，問題点を明らかにした（図 5，図 6，図 7）．瀧澤氏は大阪市大学に所属するということもあり，以上の研究成果は大阪市との連携の下で展開されたものである．

4.　劣線形データ構造グループの主要な研究成果

簡潔データ構造 (Succinct Data Structure) は東京大学の定兼邦彦氏を中心に開発されたデータ構造で，日本が世界をリードする研究分野である [10]．本グループの研究は，そのような研究基盤を背景に展開されている．

4.1　DNA データ圧縮技術

定兼氏，渋谷氏らは，生物配列においては，DNA のアセンブリ問題と呼ばれる問題において，大規模データを圧縮した状態で解析する新たなアルゴリズムの開発に成功した．そこで現れる大きなサイズの de Bruijn グラフと呼ばれるアセンブリグラフから superbubble（図 8）などの特徴部分グラフを最適な計算量 $O(m \log m)$ で検出するアルゴリズムを開発した [12]．

4.2　ストリームデータに対するリアルタイム圧縮技術

筑波大学の山際伸一氏，九州工業大学の坂本比呂志氏によるストリームデータに対するリアルタイム圧縮技術は，劣線形サイズへの圧縮を可能とする文法圧縮技術を基礎に開発されたもので [13]，世界でも初めてのリアルタイム圧縮技術と言える．この技術は，すでに，医療画像データの高速伝送や，自動車の車載ネットワークの高性能化など，さまざまな産業への応用に展開されている（図 9）．開発したハードウェア化による無遅延なロスレス圧縮技術はビッグデータ時代の通信環境を大幅に改善しうるが，これまでは限られたサイズのメモリしか搭載できない FPGA などのハードウェア上でそれを実用化できるアルゴリズムがなかった．本研究のロスレスデータ圧縮器は，そのような制約のあるハードウェア上でも容易に実装可能なアルゴリズムの実現とそのハードウェア化によって，データの種類を問わずナノ秒オーダの途切れない高速圧縮・復

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

号を低コストで可能にした世界初の新技術であり，論文を発表した国際会議で最優秀論文賞を受賞し，一般財団法人組込みシステム技術協会から ET/IoT アワード特別賞（2015 年）を受賞している．2018 年には大学発ベンチャー企業に対する科学技術振興機構理事長賞を受賞している．

4.3 Succinct ORAM 技術の開発

東京大学の小野寺拓氏，渋谷氏は，データベースアクセスを秘匿する技術として近年脚光を浴びている技術である ORAM に対して，劣線形パラダイムを活用し，従来の ORAM のセキュリティおよびアクセス速度を保ちつつ，必要データ容量を漸近的に最適な容量にする Succinct ORAM 技術の開発に成功した [14]．これまで，実装に必要なデータ容量が実データ容量と比べて極めて大きいために超ビッグデータに適用するのには困難とされてきた．しかし，小野寺氏，渋谷氏の提案した Succinct ORAM によって，ORAM のビッグデータへの実応用が拓けていくことが考えられ，さまざまな，より高度なビッグデータ処理技術の創出へとつながっていくことが期待される．

5. 劣線形モデリンググループの主要な研究成果

当グループは，繰り込み理論を代表とする統計力学分野で古来より開発されてきた伝統的な情報粗視化理論と現代型データサイエンス理論である統計的深層学習理論を二本柱として，ビッグデータの粗視化（ビッグデータの圧縮・データ削減）のためのモデリング理論，すなわち，“劣線形モデリング理論”とそれに附随する高速統計計算法の創出を目指している．また，もう一つの狙いは，統計力学的な物理的近似計算アルゴリズム手法とアルゴリズム理論からの数学的近似計算アルゴリズムを有機的に融合させることによる“新しい側面からの高速な計算処理アルゴリズム設計理論”の創出である．これは当グループの独自の視点からの劣線形時間アルゴリズムへのアプローチであり，世界的にも成功例がほとんど見られない先駆的な研究である．

ここで，そのグループで開発した「繰り込み群の理論を用いた情報粗視化モデル」の画像処理（画像切り出し）の課題に応用した具体例を紹介する [15, 16]．画像は膨大な数のピクセルで構成されており，ビッグデータの一つとして数えられている．当グループは大きなサイズの画像を提案モデルにより粗視化（圧縮）し，粗視化された小サイズ画像に対して画像処理を施すことにより，元々のサイズの画像を処理するために必要な重要パラメータを粗視化後の小さな画像のみから高精度で抽出することに成功した（図 10）．これにより，画像処理にかかる計算時間を十分の一程度に削減している．この成果は，劣線形モデリンググループの最終目標である，ビッグデータの劣線形モデリング理論創出のための基盤的な汎用技術と考えられる．つまり，情報粗視化を行って，小さなサイズのデータでモデリングしてそのモデル上で計算を行い，結果を元のビッグデータに対して適用させようという試みである．

6. おわりに

CREST 研究課題「ビッグデータ時代に向けた革新的アルゴリズム基盤」の理念および目指している研究の方向，主要な研究成果を解説した．紙面の都合上，一部の重要な研究成果のみの紹介となってしまったが，詳細は，ウェブサイト [17] をご覧いただきたい．この CREST 研究グループには OR 学会でも活躍する多くの研究者によって展開されている．OR の研究や教育が，AI やビッグデータという大きな時代のうねりに乗って，大きく飛躍し，新しい社会の構築に貢献することを期待したい．

参考文献

[1] https://www.jst.go.jp/kisoken/crest/research_area/index.html

[2] H. Ito, “Every property is testable on a natural class of scale-free multigraphs,” In *Proceedings of 24th Annual European Symposium on Algorithms*, 51:1-51:12, 2016.

[3] 加藤直樹，“組合せ劣性理論の最近の進展と応用，”電子情報通信学会和文誌 D，**J99-D**, pp. 1055–1068, 2016

[4] T. H. Kim, P. Mehrabi, Z. Ren, A. Sljoka, C. Ing, A. Bezginov, L. Ye and R. Pomes, “The role of dimer asymmetry and protomer dynamics in enzyme catalysis,” *Science*, **355**(6322), eaag2355, 2017.

[5] L. Ye, C. Neale, A. Sljoka, B. Lyda, D. Pichugin, N. Tsuchimura and S. T. Larda, “Mechanistic insights into allosteric regulation of the A_{2A} adenosine G protein-coupled receptor by physiological cations,” *Nature Communications*, **9**, article number: 1372, 2018.

[6] J. R. Jeliazkov, A. Sljoka, D. Kuroda, N. Tsuchimura, N. Katoh, K. Tsumoto and J. J. Gray, “Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification,” *Frontiers in Immunology*, **9**, article number: 413, 2018.

[7] 瀧澤重志，高木尚哉，谷口与史也，“浸水被害を想定した梅田地下街の垂直避難シミュレーション，”都市防災研究論文集，**2**, pp. 35–38, 2015.

[8] R. Yamamoto and A. Takizawa, “Partitioning vertical evacuation areas in Umeda underground mall to minimize the evacuation completion time,” *The Review of Socionetwork Strategies*, 2019, to appear.

[9] 川岸裕，瀧澤重志，“大地震時を想定した大阪市からの一

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

斉徒歩帰宅シミュレーション，”都市防災研究論文集，**4**, pp. 7–13, 2017.
[10] K. Sadakane and R. Grossi, “Squeezing succinct data structures into entropy bounds,” In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 1230–1239, 2006.
[11] T. Onodera, K. Sadakane and T. Shibuya, “Detecting superbubbles in assembly graphs,” In *Proceedings of the 13th International Workshop on Bioinformatics. Lecture Notes in Computer Science*, **8126**, pp. 338–348, 2013.
[12] W.-K. Sung, K. Sadakane, T. Shibuya, A. Belorkar and I. Pyrogova, “An $O(m \log m)$-time algorithm for detecting superbubbles,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **12**, pp. 770–777, 2015.
[13] S. Yamagiwa, K. Marumo and H. Sakamoto, “Stream-based lossless data compression hardware using adaptive frequency table management,” In *Proceeding of BPOE2015: Big Data Benchmarks, Performance Optimization, and Emerging Hardware. Lecture Notes in Computer Science*, **9495**, pp. 133–146, 2015.
[14] T. Onodera and T. Shibuya, “Succinct oblivious RAM,” In *Proceedings of the 35th Symposium on Theoretical Aspects of Computer Science*, **96**, article number: 52, 2018.
[15] K. Tanaka, S. Kataoka, M. Yasuda, Y. Waizumi and C. T. Hsu, “Bayesian image segmentations by Potts prior and loopy belief propagation,” *Journal of the Physical Society of Japan*, **83**, article number: 124002, 2014.
[16] K. Tanaka, S. Kataoka, M. Yasuda and M. Ohzeki, “Inverse renormalization group transformation in Bayesian image segmentations,” *Journal of the Physical Society of Japan*, **84**, article number: 045001, 2015.
[17] http://crest-sublinear.jp/

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.