様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １３ 日現在

機関番号：３２６８９
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ２０３９５

研究課題名（和文）統計的論理関係解析法に基づく機能未知遺伝子の機能推定

研究課題名（英文）Function estimation of functional-unknown genes using statistical logical relationship analysis

研究代表者

福永　津嵩（Fukunaga, Tsukasa）

早稲田大学・高等研究所・講師（任期付）

研究者番号：８０７９１４３３

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究は、統計的論理関係解析法という新しいデータマイニング手法を用いて、ゲノム中の機能未知遺伝子の機能を推定する新しいソフトウェアを開発する事を研究目的とした。結果、統計的論理関係解析法を実装したオミクスデータ解析ソフトウェアであるLogicome Profilerに加えて、統計的に有意な配列モチーフを列挙する手法であるMotiMul、高精度進化系統樹推定ソフトウェアであるMirage、偽陽性を排除した系統プロファイル解析法IPMなど、多くの遺伝子機能推定ソフトウェアを開発することに成功した。

研究成果の学術的意義や社会的意義
ゲノム決定技術の進歩により多様な生物のゲノムが決定されてきている一方で、ゲノムに含まれる遺伝子の多くは機能未知である。これら機能未知遺伝子の中には創薬・医療・カーボンニュートラルなどに関係する重要な遺伝子も多く含まれていると考えられるため、その遺伝子機能を推定するソフトウェア開発は生物学全般において重要な課題であるといえる。本研究は、統計的論理関係解析法と呼ばれるデータマイニング手法をベースに、さまざまな遺伝子機能推定ソフトウェアを開発した。本研究で開発したソフトウェアを広く活用することで、多くの機能未知遺伝子の機能が解明されることが期待される。

研究成果の概要（英文）：This study aimed to develop novel software programs to estimate the function of function-unknown genes in genomes using statistical logical relationship analysis. In this study, we developed many programs for gene function estimation, e.g., (1) Logicome Profiler, an omics data analysis program that implements the statistical logical relationship analysis method, (2) MotiMul, a method for enumerating statistically significant sequence motifs (3) Mirage, which reconstructs a gene-content evolutionary history based on various gene gain/loss models by unsupervised classification of evolutionary patterns among gene families, (4) IPM, an accurate phylogenetic profiling method based on the inverse Potts model.

Translated with www.DeepL.com/Translator (free version)

研究分野：バイオインフォマティクス

キーワード：遺伝子機能推定　確率モデル　ゲノム解析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ゲノムデータの増加に伴い機能未知遺伝子の数も増大しているため、情報科学に基づく遺伝子機能推定はバイオインフォマティクスにおける重要な研究課題である。Guilt-by-assocation 法はそのような手法の 1 つであり、機能未知遺伝子と共起する機能既知遺伝子を探索し、その既知機能を伝播させることで機能未知遺伝子の機能を推定する。この手法は、遺伝子共発現解析や系統プロファイル法など、これまでに多様な生物種・実験条件・比較条件で適用され多くの成功を収めてきたものの、その予測精度は十分でないという問題があった。

Guilt-by-association 法は、多くの場合二遺伝子の関係に注目して機能推定を行う。しかしながら、遺伝子は実際には複数の遺伝子と関連しながら機能するため、より多くの遺伝子の関係性に注目して解析する事で、遺伝子機能が精度よく推定されることが期待される。よって、複数の遺伝子に拡張した Guilt-by-association 法の手法整備を行うことで、より高精度な遺伝子機能推定を行うことができると考えた。

２．研究の目的

申請者はデータセット内で統計的に有意に頻出する三遺伝子間の論理関係を検出する事で遺伝子機能推定を行う、「統計的論理関係解析法」という新たな解析手法を考案した。ここで論理関係とは、たとえば C = (not A) and B (A がなく B がある時に C がある)といったような、命題論理の式で表現出来る関係性を意味する。統計的論理関係解析法は新規のデータマイニング手法であるため、ソフトウェアを整備し、様々なデータセットに適用することで多くの生物学的知見を得る事が期待される。このことから、1. 統計的論理関係解析法のソフトウェア整備、及び技術的な拡張を行う事で手法の汎用性を高める事、及び 2. 開発した手法を多様なデータセットに適用する事で、機能未知遺伝子の機能推定及び生物学的知識発見を行う事を本研究課題の研究目的とした。

３．研究の方法

当初の研究の方法としては、まず「統計的論理関係解析法」をソフトウェア化し、その後に手法の汎用性向上として、「生物種の系統樹情報を考慮した解析手法の開発」「離散的な交絡因子を考慮した検出法の開発」「並び変え検定を利用した高感度検出法の開発」に取り組む予定であった。 また、統計的論理関係解析法の多様なデータセットへの適用として、「系統遺伝子プロファイルデータへの適用による遺伝子論理関係の進化解析」及び「組織別遺伝子発現データへの適用による遺伝子間論理制御関係の発見」に取り組む予定であった。

統計的論理関係解析を行う上で重要な課題として、検定すべき仮説数が莫大であるため多重検定の補正基準が厳しくなり、統計的検出力が著しく低下するという問題が存在する。そのため、本研究では多重検定の補正法として、Tarone 法の応用や Westhall-Young 法の適用を行う予定であった。

４．研究成果

(1) 統計的論理関係解析法である Logicome Profiler の開発

研究目的に述べた統計的論理関係解析法についてソフトウェアを整備し、Logicome Profiler として公開した。Logicome Profiler は、先行研究である LAPP 法(統計的ではない論理関係解析手法)とは異なり、サンプルサイズが増大しても検出力が低下しないという長所を備えていることがシミュレーションから明らかとなった。また、Logicome Profiler を大規模海洋メタゲノムデータセットに適用する事で、光合成と窒素代謝に関する論理関係を始め、新規の三遺伝子間論理関係を多数同定することに成功した。しかしながら、多重検定法として当初予定していた Tarone 法や Westhall-Young 法を統計的論理解析に適用することは、技術上の制約や計算時間上の問題から困難であったため、より簡便な検定法である Bonferroni 法による FWER 調整及び Benjamini-Yekutieli 法による FDR 調整法を採用することとした。本研究結果は、PLOS ONE 誌より既に論文が出版されている(Fukunaga and Iwasaki, PLoS One, 15, e0232106, (2020))

また、離散値を対象とする Logicome Profiler を連続値データに対しても適用できるよう拡張を行った。ここでは、従来使用していた二値論理をファジイ論理の一種に置き換えることでアルゴリズムの拡張に成功した。しかしながら、開発したソフトウェアを適用するにあたって適切なデータセットを選定し生物学的な知見を得ることには現状成功していないため、本研究は論文出版には至っていない。

(2) 統計的に有意な配列モチーフを発見する手法である MotiMul の開発

Logicome Profiler において Tarone 法を適用すべく検討を行っていた際に、Tarone 法は配列モチーフを検出するのに有効であることに気がついたため、その研究を行った。MEME や

MOCCS をはじめとする既存の配列モチーフ検出法では、モチーフ長を指定する必要があり、適切なモチーフ長を事前に選ぶのが難しいという問題点があった。本研究では、Tarone 法と頻出文字列パターン列挙アルゴリズムである PrefixSpan を組み合わせることで、統計的に有意な配列モチーフを全て列挙することでモチーフ長を指定せずにモチーフ検出するアルゴリズムを開発し、それを Motimul として実装した。MotiMul の性能を確かめるために、USFI, GABP 及び SRF という 3 つの ChIP-seq データセットに適用したところ、MEME や MOCCS 単独では捉え切れないモチーフ配列特徴を Motimul が捉えられていることを明らかとした。本研究結果は、国際会議である IEEE BIBM にて報告を行った (Mori, Ozaki and Fukunaga, IEEE BIBM, 186-193, (2020))

MotiMul はモチーフ長を事前に指定しなくて良いという長所がある一方で、有意な配列モチーフを全て列挙するため解釈性に乏しいという欠点がある。そのため、確率論理プログラムである SLIPCOVER を用いて、列挙された配列モチーフを縮約して表現することも試みた。結果、配列情報は大いに圧縮され視認性が向上したものの、PWM など既存の表現方法と比べて優れた特徴を見出すのが難しく、生物学的な知見を得るには至らなかったため、本解析は論文出版には至っていない。

(3) 高精度進化系統樹推定ソフトウェアである Mirage の開発

「生物種の系統樹情報を考慮した解析手法の開発」のために、既存のゲノム進化史推定手法についてサーベイを行ったところ、既存の手法は遺伝子の獲得／欠失速度のモデル化が生物学的に不自然である事を発見した。すなわち、既存の手法は、これらの進化速度が全ての遺伝子間で同一であると仮定するか、あるいは遺伝子間において進化速度のみが異なると仮定するかのいずれかであり、進化のパターンが異なる事を組み込んでいるモデルが存在しないことがわかった。このため、遺伝子間において進化のパターンが異なることをモデルに組み込むことで、より高精度なゲノム進化史推定手法が開発できるのかと考え、この実装を行った。具体的には、ゲノム進化史推定において、祖先のゲノム状態と遺伝子がどのクラスターに属しているか(クラスターごとに進化のパターンが異なるようにモデル化する)の 2 つを潜在状態としてモデル化し、これを EM アルゴリズムによって進化速度パラメータを推定した後、Viterbi-like アルゴリズムでゲノム進化史を推定し、そのアルゴリズムを Mirage として実装した。シミュレーションデータの結果として、Mirage は既存の研究と比べて高精度にゲノム進化史を推定可能である事を示した。また、Mirage をさまざまなクレードのゲノムデータに適用することで、どのクレードにおいても代謝関連遺伝子は頻繁に獲得／欠失しやすい事を明らかにした。本研究結果は、Bioinformatics Advances 誌より既に論文が出版されている(Fukunaga and Iwasaki, Bioinformatics Advances, 1, vbab014, (2021))

また、Mirage は高精度であるものの、計算速度が十分ではなく 100 種程度のゲノム情報にしか適用できないという問題点があった。そのため、Mirage を高速化すべく、分子進化解析で利用されている高速化手法である Partition 法を Mirage に組み込んだ。結果として、既存の手法に比べてわずかに精度がさがるものの、計算速度は大幅に向上し、数千種程度のゲノム情報であっても十分に適用可能であるほど高速に実行可能となった。本研究結果は、Bioinformatics 誌より既に論文が出版されている(Fukunaga and Iwasaki, Bioinformatics, 38, 4039-4041 (2022))

(4) 偽陽性を排除した系統プロファイル解析法 IPM の開発

Logicome Profiler の開発過程において明らかとなった問題点の一つに、得られた検出結果が非常に多く、偽陽性を多く含んでいる可能性が高い点がある。そのため本研究では、系統プロファイル法において偽陽性を排除して、より高精度に遺伝子機能推定を行う手法を開発する事を研究目的とした。偽陽性が検出される要因の一つとして擬似相関があるが、これは相関検出を行う際に二遺伝子のみしか考慮しないで検出を行うことが原因の一つである。そのため、全ての遺伝子を同時にモデルに含んだ上で相関係数に類似した指標を計算することができれば、擬似相関をおよそ低下させることが可能であり、このようなモデルはイジングモデル／ポッツモデルとして知られる。本研究では、系統プロファイル法においてこのポッツモデルを用いて解析を行ったところ、既存の手法に比べて高精度に遺伝子機能推定が可能である事を示した。本研究結果は、Bioinformatics 誌より既に論文が出版されている(Fukunaga and Iwasaki, Bioinformatics, 38, 1794-1800 (2022))。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Fukunaga Tsukasa、Iwasaki Wataru	４．巻 38
２．論文標題 Inverse Potts model improves accuracy of phylogenetic profiling	５．発行年 2022年
３．雑誌名 Bioinformatics	６．最初と最後の頁 1794-1800
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btac034	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Fukunaga Tsukasa、Iwasaki Wataru	４．巻 1
２．論文標題 Mirage: estimation of ancestral gene-copy numbers by considering different evolutionary patterns among gene families	５．発行年 2021年
３．雑誌名 Bioinformatics Advances	６．最初と最後の頁 vbab014
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioadv/vbab014	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Fukunaga Tsukasa、Iwasaki Wataru	４．巻 15
２．論文標題 Logicome Profiler: Exhaustive detection of statistically significant logic relationships from comparative omics data	５．発行年 2020年
３．雑誌名 PLOS ONE	６．最初と最後の頁 e0232106
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0232106	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 福永 津嵩、岩崎 渉
２．発表標題 Mirage: A phylogenetic mixture model to reconstruct gene content evolutionary history using a realistic evolutionary rate model
３．学会等名 第9回生命医薬情報学連合大会
４．発表年 2020年

１．発表者名
Koichi Mori, Haruka Ozaki and Tsukasa Fukunaga

２．発表標題
MotiMul: A significant discriminative sequence motif discovery algorithm with multiple testing correction.

３．学会等名
IEEE BIBM 2020（国際学会）

４．発表年
2020年

１．発表者名
Koichi Mori, Haruka Ozaki and Tsukasa Fukunaga

２．発表標題
統計的有意性を担保可能な系列パターンマイニングに基づく配列モチーフ検出ソフトウェアの開発

３．学会等名
第9回生命医薬情報学連合大会

４．発表年
2020年

１．発表者名
毛利公一、福永 津嵩

２．発表標題
Tarone法を用いた頻出部分文字列マイニングの多重検定補正

３．学会等名
IBIS2019

４．発表年
2019年

１．発表者名
毛利公一、福永 津嵩

２．発表標題
Tarone法を用いた系列パターンマイニングの多重検定補正

３．学会等名
日本バイオインフォマティクス学会2019年年会

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関