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　３ 年　 ６ 月 ２２ 日現在

機関番号：８２６５７
研究種目：若手研究(B)
研究期間：2017～2020
課題番号：１７Ｋ１２７８２

研究課題名（和文）遺伝子制御領域に着目したゲノム進化過程解明のための新しいオーソログ解析手法の開発

研究課題名（英文）Developing a new ortholog analysis method for elucidating genome evolution process focusing on gene regulatory regions

研究代表者

千葉　啓和（Chiba, Hirokazu）

大学共同利用機関法人情報・システム研究機構（機構本部施設等）・データサイエンス共同利用基盤施設・特任助教

研究者番号：６０６２５６４８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：
増え続けるゲノム・プロテオームデータに対応するために、公共データベースから必要なデータをダウンロードして管理するパイプラインを構築した。また、大規模なオーソログクラスタリングに対応させるために、クラスタリングプログラムを改良して、メモリの使用効率を高めるとともに、ドメインが分断しすぎないようにした。さらに、クラスタリング結果をどのように表現するかというデータモデルを策定し、RDFでデータベースを構築して、様々なアプリケーションから利用可能にした。SPARQLによる論理的なクエリ作成も行えるようにするとともに、クエリの結果を可視化するプログラムを開発した。

研究成果の学術的意義や社会的意義
近年のゲノム配列解読技術の進展にともない、これまで多くの生物のゲノム情報が得られてきた。これらの情報を最大限に活用して研究を推進するには、独自に情報を収集して管理し、アップデートする枠組みを作るとともに、これまでの解析手法を発展させて計算を行い、二次的なデータベースの開発、データ統合、公開までつなげることが必要である。本研究では、これまでのドメインレベルのオーソログ検出手法を基盤として、対象生物を拡張するとともに、検出手法の精緻化を進め、データベースを統合的に解析可能なシステムの構築を推進した。

研究成果の概要（英文）：
To cope with the ever-increasing genome and proteome data, we built a pipeline to download and manage the necessary data from public databases. To cope with large-scale orthologous clustering, we improved the clustering program to increase the efficiency of memory usage and to prevent the domain from being too fragmented. We also developed a data model for representing the clustering results, constructed a database in RDF that can be used by various applications, and developed a program to create logical queries in SPARQL as well as to visualize the query results.

研究分野：生命情報科学

キーワード：オーソログ解析　オーソログデータベース

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

種分化によって分岐した遺伝子間に定義されるオーソログ関係は、比較ゲノム解析の基盤となる情報であり、オーソログ検出を正しく行う手法の開発は重要な課題である。申請者らは、主に微生物ゲノムを対象として、オーソログ検出手法の精緻化を行うとともに[1]、データベースの構築を行ってきた[2]。さらに、様々な生命科学系データベースとの連携を見据えてセマンティック・ウェブ技術を採用し、オーソログ情報を再利用可能な形で公開するシステムの構築を行ってきた[3,4]。

２．研究の目的

近年のゲノム配列解読技術の進展にともない、これまで多くの生物のゲノム情報が得られてきた。これらの情報を最大限に活用して研究を推進するには、独自に情報を収集して管理し、アップデートする枠組みが前提となるため、データ管理の重要性も大きい。その上で、これまでの解析手法を発展させた計算を行い、二次的なデータベースの構築、データ統合、公開までつなげるシステムが必要であった。本研究では、これまでのドメインレベルのオーソログ検出手法を基盤として、対象生物を拡張するとともに、検出手法のさらなる精緻化を行い、データベースを統合化して解析可能なシステムを構築するために、データモデルの標準化を進めて、統合解析システムの構築を推進することを目指した。

３．研究の方法

ゲノム・プロテオームデータは、公共のデータベースから取得した。ゲノムについては、NCBI の RefSeq データベースを使用した。プロテオームについては、UniProt のプロテオームを利用した。これらの公共データベースを基盤として、二次的なデータベースを構築するためには、独自のデータ処理パイプラインを構築することが必要である。Python 言語を用いて、データをダウンロードして、管理するパイプラインを構築した。

また、配列の相同性検索および、クラスタリング計算に関しては、BLAST および、DomClust、DomRefine[1]をベースとしてプログラム開発を行った。DomRefine は Perl 言語を用いて構築されてきたため、これを改修・拡張する形で開発を進めた。

ユーザーが利用するアプリケーションについては、Web ブラウザ上での利用を前提とし、JavaScript で開発を行った。

４．研究成果

本研究では、これまでの研究を拡張して様々な開発を行った。まず、増え続けるゲノム・プロテオームデータに対応するために、公共データベースから必要なデータをダウンロードして管理するパイプラインを構築した。また、クラスタリング可能な生物種を抽出するために生物分類を階層的に探索できるシステムを開発した。大規模なオーソログクラスタリングに対応させるために、クラスタリングプログラム[5]を改良して、メモリの使用効率を高めるとともに、ドメインが分断しすぎないようにした。また、オーソログ解析システムに必要なプログラムを集めた Docker コンテナを開発した。さらに、クラスタリング結果をどのように表現するかというデータモデルを策定し、RDF でデータベースを構築して、様々なアプリケーションから利用可能にした。SPARQL による論理的なクエリ作成も行えるようにするとともに[6]、Stanza と呼ばれるフレームワーク[7]を用いてクエリの結果を可視化するプログラムを開発した。特に、特定の遺伝子についての保存プロファイルを表示するアプリケーションを構築し、他のプログラムからも利用できるようにした。

当初は明確な対象としては定義していなかった部分として、真核生物ゲノムへの適用がある。オーソログに関する国際コンソーシアムにおいて、オーソログ検出手法のベンチマーク手法を策定しているが、そこには多くの真核生物が含まれている。我々の手法を適用した結果、他の優れた手法と比較しても優れた性能を示すことが分かってきた。さらに、解析に用いられるデータベースも多様なものが利用可能となってきており、転写因子結合部位のデータも RDF 化されて統合可能な状況になっている。これらの真核生物を対象とした解析については未発表であるが、国際学会においてベンチマークの公開を予定している。

引用文献

[1] Hirokazu Chiba, and Ikuo Uchiyama. Improvement of domain-level ortholog clustering by optimizing domain-specific sum-of-pairs score. BMC Bioinformatics 15(1), 1-16 (2014)

[2] Ikuo Uchiyama, Motohiro Mihara, Hiroyo Nishide, Hirokazu Chiba. MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Research 43(D1), D270-D276 (2015)

[3] Chiba, Hirokazu, Hiroyo Nishide, and Ikuo Uchiyama. Construction of an ortholog database using the semantic web technology for integrative analysis of genomic data. PLOS One 10(4), e0122802 (2015)

[4] Hirokazu Chiba, Ikuo Uchiyama. SPANG: a SPARQL client supporting generation and reuse of queries for distributed RDF databases. BMC Bioinformatics 18(1), 1-6 (2017)

[5] Ikuo Uchiyama, Motohiro Mihara, Hiroyo Nishide, Hirokazu Chiba, Masaki Kato. MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons, Nucleic Acids Research 47(D1), D382-D389 (2019)

[6] Tarcisio M. de Farias, Hirokazu Chiba, Jesualdo T. Fernandez-Breis. Leveraging logical rules for efficacious representation of large orthology datasets. Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4LS 2017) Published on CEUR-WS, Vol-2042 (2017)

[7] Toshiaki Katayama, Shuichi Kawashima, Shinobu Okamoto, Yuki Moriya, Hirokazu Chiba, Yuki Naito, Takatomo Fujisawa, Hiroshi Mori, Toshihisa Takagi. TogoGenome/TogoStanza: modularized Semantic Web genome database. Database 2019, bay132 (2019)

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　1件／うちオープンアクセス　3件）

１．著者名 Ikuo Uchiyama, Motohiro Mihara, Hiroyo Nishide, Hirokazu Chiba, Masaki Kato	４．巻 47
２．論文標題 MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons	５．発行年 2019年
３．雑誌名 Nucleic Acids Research	６．最初と最後の頁 D382-D389
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nar/gky1054	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 Toshiaki Katayama, Shuichi Kawashima, Shinobu Okamoto, Yuki Moriya, Hirokazu Chiba, Yuki Naito, Takatomo Fujisawa, Hiroshi Mori, Toshihisa Takagi	４．巻 2019
２．論文標題 TogoGenome/TogoStanza: modularized Semantic Web genome database	５．発行年 2019年
３．雑誌名 Database	６．最初と最後の頁 1-11
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/database/bay132	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 Tarcisio M. de Farias, Hirokazu Chiba, Jesualdo T. Fernandez-Breis	４．巻 2042
２．論文標題 Leveraging logical rules for efficacious representation of large orthology datasets	５．発行年 2017年
３．雑誌名 Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4LS 2017) Published on CEUR-WS	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 該当する

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Hirokazu Chiba
２．発表標題 Unraveling the microbial gene repertoire by sub-gene level orthologous clustering.
３．学会等名 The 67th NIBB Conference "Quest for Orthologs"（国際学会）
４．発表年 2019年

１．発表者名
千葉啓和

２．発表標題
オーソログ情報の統合化と利活用

３．学会等名
トーゴーの日シンポジウム2019

４．発表年
2019年

１．発表者名
千葉啓和

２．発表標題
オーソログデータベースに基づくドメイン融合解析

３．学会等名
日本進化学会第20回大会

４．発表年
2018年

１．発表者名
Hirokazu Chiba, Ikuo Uchiyama

２．発表標題
Applying the semantic web technology to diversified biological resources for comparative omics

３．学会等名
NIG International Symposium 2017（国際学会）

４．発表年
2017年

１．発表者名
Hirokazu Chiba, Jesualdo Tomas Fernandez-Breis, Ramon Garcia Martinez, Yuki Moriya, Susumu Goto, Ikuo Uchiyama

２．発表標題
Development of orthology ontology and its application to orthology meta-search

３．学会等名
Quest for Orthologs 5（国際学会）

４．発表年
2017年

１．発表者名 Tarcisio M. Farias, Hirokazu Chiba, Jesualdo T. Fernandez-Breis
２．発表標題 Leveraging Logical Rules for Efficacious Representation of Large Orthology Datasets
３．学会等名 Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences（国際学会）
４．発表年 2017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			
スイス	University of Lausanne	Swiss Institute for Bioinformatics		