様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月３１日現在

機関番号：12601
研究種目：研究活動スタート支援
研究期間：2009～2010
課題番号：21810005
研究課題名（和文）複雑な大規模生命科学データを用いた進化解析のための情報処理技術開発

研究課題名（英文）Bioinformatics for Evolutionary Analysis of Complex and Abundant Biological Data

研究代表者
岩崎 渉（IWASAKI WATARU）
東京大学・大学院新領域創成科学研究科・助教
研究者番号：50545019

研究成果の概要（和文）：グラフや自然言語などの形で情報を記述する複雑かつ大規模な生命科学データが蓄積されつつあることを背景に，これらのデータを系統関係を考慮した生物種横断的な解析に用いるための基盤技術を開発した．具体的には，系統トポロジーの曖昧性を考慮した進化系統表現法である車輪樹法，複雑なグラフ構造をマルチスケールかつインタラクティブに解析するためのツール NaviCluster，および遺伝子と表現型を網羅的に結びつけるための基盤となるデータベースの開発を行った．

研究成果の概要（英文）：Abundant and complex biological data represented in graphs or natural languages are rapidly accumulating. In this research, we developed bioinformatics technologies to analyze those data while considering phylogenetic relationships. These technologies are the wheel tree method, a phylogenetic relationship representation method considering vague phylogenetic topologies; NaviCluster, a tool for interactively investigating complex and large biological networks; and a database for comprehensively connecting genes and phenotypes.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,110,000	333,000	1,443,000
２０１０年度	1,010,000	303,000	1,313,000
年度			
年度			
年度			
総　計	2,120,000	636,000	2,756,000

研究分野：バイオインフォマティクス
科研費の分科・細目：ゲノム科学・システムゲノム科学
キーワード：車輪樹法，系統樹，進化，NaviCluster，可視化，グラフクラスタリング，ネットワーク，文献情報処理

１．研究開始当初の背景

生命科学データは加速度的な勢いで蓄積され続けている．特に塩基配列データの増加は著しく，解読済みゲノムの数は今年中に 1000 の大台を突破するほか，いわゆる次世代シーケンサーが急速に普及しつつある．このことは，今後，一層多くの生物種についてゲノム情報が利用可能になるほか，遺伝子発現ネットワークなどより複雑な情報を含むデータについてもさらに蓄積が進むことを示唆している．また，タンパク質相互作用情報や文献情報等についても質・量の両面で充実の傾向が続いている（例えば，[*Science* **322** 104(2008)] やフルテキスト論文のオープン化）．

申請者はこれまで，数百生物種規模のゲノムデータを利用したゲノムおよび代謝ネットワークの進化解析を遂行してきた[*Bioinformatics* **23** i230(2007), *PLoS Genet.* **5** e1000402 (2009)]．ゲノム進化の連続時間マルコフ過程へのモデル化，ゲノム進化推定問題の期待値最大化法による解決，動的計画法による効率的な周辺確率計算などの手法を利用することにより，ゲノム進化の様相は生物系統ごとに大きく偏っていること，さらに，異なる原核生物系統間の遺伝子水平伝播が代謝ネットワークの進化に大きな役割を果たしてきた可能性がそれぞれ示唆された．これらのことは，ゲノム規模での進化解析を系統関係を考慮しつつ生物種横断的に行うことにより，特定の生命現象にも特定の生物系統にも偏らない生物学的な知見を得られる可能性があること，また，そのためには生物学的な現象を情報科学的に扱いやすい形でうまくモデル化することが有効であること，をそれぞれ示唆している．

２．研究の目的

本研究では，グラフや自然言語などの形で複雑な情報を記述する大規模な生命科学データについて，系統関係を考慮した生物種横断的な解析に用いるための基盤技術を開発する．具体的には，(1)系統トポロジーの曖昧性を考慮した進化系統表現法の開発 (2)複雑なグラフ構造をマルチスケールかつインタラクティブに解析するための手法開発 (3)遺伝子－表現型関係の網羅的取得のための技術開発 の３つのサブテーマに取り組む．

(1) 系統トポロジーの曖昧性を考慮した進化系統表現法の開発

系統関係を考慮した解析を行う前段階として，まず一次配列データを用いた分子進化解析等によって生物種間の系統トポロジーを推定することが必要である．しかし近年，ゲノム規模のデータを用いても大規模な系統トポロジー推定には曖昧さが残ることが明らかになってきた（例えば[*Science* **311** 1283 (2006)]）．このことは，系統関係を利用した解析において，これまでは単一の信頼できる系統トポロジーが得られていることを前提としたモデル化が行われてきたのに対し，ブートストラップ法などによって得られた複数の系統トポロジー群全体を考慮することが必須であることを意味する．系統トポロジー群を単一のデータ構造として表現する試みとしては系統ネットワーク法［*Mol. Biol. Evol.* **23** 254 (2006)］が知られているが，この表現法は広く用いられるには至っていない．これは，可能な限り多くのデータを単一のグラフ構造の中に埋め込むために，生物学的な意味が不明瞭な枝がしばしば出現するためである［*Philos. Trans. R. Soc. London Ser. B* **363** 4023(2008)］．そこで，可能な候補解全てを解とするのではなく，代表としてそれらの重心にあたる解を出力するセントロイド推定法［*PNAS* **105** 3209(2008)］の発想を応用し，可能な進化系統トポロジー群全体を考慮しつつも生物学的な意味の明瞭な枝のみを用いて系統関係をモデル化する手法を開発する．

(2) 複雑なグラフ構造をマルチスケールかつインタラクティブに解析するための手法開発

多くの網羅的な生命科学データはいわゆる二項関係の形で得られる（例えば，遺伝子発現ネットワークやタンパク質相互作用データなど）．一般にこれらの二項関係データは，節点とそれを結ぶ枝からなるグラフの形で表現される．しかし，こういったグラフ表現は枝の数がおおよそ100を上回ると極度に複雑な外観を呈するようになり，生物種横断的に複数のグラフを効率よく比較して知識や仮説を引き出す上では効果的な表現手法になっていなかった．本研究では，グラフ構造の密な部分（モジュール）を階層的にクラスターとして表現する階層的グラフクラスタリング法と，生命科学データの持つ意味情報（Gene Ontology アノテーションなど）を用いた意味的クラスタリング法とを組み合わせることで，マルチスケールかつインタラクティブに複雑な生命科学グラフ構造を効果的に比較解析するための手法を開発する．

(3) 遺伝子－表現型関係の網羅的取得のための技術開発

各種の生命科学データを，表現型情報と関連付けつつ生物種横断的かつゲノム規模で解析するためには，各遺伝子がどのような表現型に影響しているかについて大規模に情報を取得することが必要である．しかし，モデル生物以外の生物種では遺伝子－表現型関係についてのデータベースはほとんど整備されておらず，この目的においては自然言語で書かれた論文アブストラクトないしフルテキスト中からテキストマイニング技術を用いて情報を取得することが有用であると考えられる．しかし，生命科学論文は各生物種について偏りなく出版される形とはなっていないため，既存のテキストマイニング手法を特定の生物種に関する文献に単純に適用しても，特に研究報告の少ない生物種については偏った情報のみしか取得すること

ができない．そこで本サブテーマでは，系統関係を利用して近縁種の遺伝子についての情報を伝播し不足する情報を補完することで，生物種横断的に生命科学論文中から遺伝子―表現型関係を網羅的に取得するための技術開発を行う．

３．研究の方法

(1) 系統トポロジーの曖昧性を考慮した進化系統表現法の開発

ブートストラップ法などによって得られる系統トポロジー群中，一定の割合以上の確率で支持された分類群のみによって作られるコンセンサス系統樹［*Evolution* **39** 783(1985)］をベースに，さらに木構造を保ったままトポロジー群についての情報を加えた表現法を開発する．コンセンサス系統樹には１つの節点から４つ以上の枝が生じる多分節点が生じるが，従来の系統解析やグラフ理論においてはこれら多分節点の回りの枝の順序には意味を持たせてこなかった．本手法ではセントロイド推定の考え方を応用し，これらの順序に系統トポロジー群の情報を反映した表現法を開発する．また，ソフトウェアを実装し，ウェブサーバにて公開する．

(2) 複雑なグラフ構造をマルチスケールかつインタラクティブに解析するための手法開発

グラフ構造の中から密なサブグラフ（クラスタ）を階層的に発見する超高速グラフクラスタリングアルゴリズム［*J. Stat. Mech.* P10008 (2008)］を実装し，これをいくつかのグラフ構造情報（*Saccharomyces* Genome Database よりダウンロードできるタンパク質相互作用データ等）に対して適用する．ここで問題となると予想されるのは，十分に少ないクラスタ数にまでクラスタリングが進まないことである．これは，生命科学におけるグラフ構造の多くは枝の数が節点の数の数倍程度にとどまり，クラスタとして抽出できる密なサブグラフが原理的に十分存在しないためである．

そこで，クラスタ数が十分に少なくならない場合に，グラフに含まれる情報をさらにコンパクトに表現するための技術を開発する．具体的には，生命科学で扱うグラフ構造ではしばしば節点が意味的な情報を持っていることに注目し（タンパク質の場合は GO Term など），意味的な類似性を利用してクラスタ同士のクラスタリングを行う手法を開発する．例えば GO Term の場合には，GO Term を成分としたベクトル同士の内積に GO Term の深さを掛けることで，単語の重要度を補正した GO Term セットの類似性を計ることなどが考えられる．先に超高速グラフクラスタリングアルゴリズムを適用して生物学的に意味のあるクラスタ（タンパク質複合体など）を発見しつつクラスタの数を減らした後に，より遅い意味的クラスタリングを適用するという戦略をとることで，10,000 枝程度のサイズのグラフを数秒のうちに生物学的な意味を反映した形で十分に階層的クラスタリングすることができると期待できる．

以上を Java のグラフ描画パッケージである JUNG を用いて GUI 化する．グラフ全体を表示する際には最も粒度の低いクラスタのみを表示し，部分グラフについて表示する際には粒度の高いクラスタを表示することで，ウェブ上の一般的な地図サービスのようにマルチスケールかつインタラクティブにグラフを描画・解析できるソフトウェアとして実装・公開する．

(3) 遺伝子－表現型関係の網羅的取得のための技術開発

生命科学論文アブストラクトおよび論文フルテキストデータに対して，固有表現認識技術を適用し種名，遺伝子名，表現型名をそれぞれ抽出する．さらに，［*BMC Bioinform.* **9** S6(2008)］で提案された論文中の遺伝子名を生物種名に関連付ける技術を，表現型名についても適用できるよう拡張し，三者の関連付けを行う．

続いて，［*Bioinformatics* **23** i230(2007)］で申請者が開発したゲノム進化解析法を拡張し，系統関係を用いた遺伝子―表現型関係情報の補完を行う．具体的には，生物種の系統関係を用いて系統樹上の各位置における遺伝子の存在確率を推定する手法を，遺伝子の存在のもとでの表現型の存在確率を扱えるよう拡張し，これらの確率を期待値最大化法等の枠組みで推定するアルゴリズムとして実装する（ここで系統樹情報としては，サブテーマ(1)で開発する系統トポロジー分布を考慮した表現法を用いることが可能となるようにする）．文献から遺伝子―表現型関係についての情報が抽出できなかった箇所についても，遺伝子の存在のもとでの表現型の存在確率が高く推定された場合には，当該生物種においても同様の遺伝子―表現型関係があるものと推定し，情報の補完を行う．

４．研究成果

サブテーマ(1)について，系統トポロジーの曖昧性を考慮しつつ，進化系統関係を偏りなく・直感的に・かつ効率的に描くための手

法「車輪樹法」の開発に成功した．本手法はバイオインフォマティクス研究者のみならず進化系統学の専門家にも高く評価され，進化系統学分野の権威ある雑誌へ採録された（*Systematic Biology*, **59**, 584）．また，世界のトップ研究者が必読論文を推薦するサイト Faculty of 1000 にて推薦された（http://f1000.com/7330957）．さらに，RNA の二次構造予測やタンパク質のマルチプルアラインメントといった問題との数理的な類似性に着目し，より一般化した手法として拡張した（*PLoS ONE*, **6**, e16450）．車輪樹ソフトウェアは現在ウェブサイト（http://cwt.cb.k.u-tokyo.ac.jp/）上で公開している．

サブテーマ(2)について，生命科学分野における巨大なグラフ・ネットワークデータを，超高速な階層的グラフクラスタリングアルゴリズムに基づいてウェブ上の地図サービスのように自由にズーム・平行移動できるソフトウェア NaviCluster の開発に成功した（*Bioinformatics*, **27**, 1121）．NaviCluster ソフトウェアは現在ウェブサイト（ http://navicluster.cb.k.u-tokyo.ac.jp/）上で公開している．

サブテーマ(3)について，論文データベースに対して網羅的に高精度な固有表現認識を行い，遺伝子−表現型関係を網羅的に取得するための基盤となるデータベースを構築した．その過程で，論文管理と必読論文推薦を統合的に行うシステム TogoDoc を開発した（*PLoS ONE*, **5**, e15305）．PC にインストールするクライアントソフトウェア TogoDoc Client は現在ウェブサイト（http://tdc.cb.k.u-tokyo.ac.jp/）上で公開している．

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４件）

① Thanet Praneenararat, Toshihisa Takagi, and Wataru Iwasaki. **Interactive, Multi-Scale Navigation of Large and Complicated Biological Networks.** *Bioinformatics*, **27**, 1121-1127. (2011) 査読有

② Michiaki Hamada, Hisanori Kiryu, Wataru Iwasaki, and Kiyoshi Asai. **Generalized Centroid Estimators in Bioinformatics.** *PLoS ONE*, **6**, e16450. (2011)査読有

③ Wataru Iwasaki, Yasunori Yamamoto, and Toshihisa Takagi. **TogoDoc Server/Client System: Smart Recommendation and Efficient Management of Life Science Literature.** *PLoS ONE*, **5**, e15305. (2010)査読有

④ Wataru Iwasaki and Toshihisa Takagi. **An Intuitive, Informative, and Most Balanced Representation of Phylogenetic Topologies.** *Systematic Biology*, 59, 584-593. (2010)査読有

〔学会発表〕（計 18 件）

① Wataru Iwasaki and Toshihisa Takagi. **Wheel Tree: Overlooked Information in Phylogenetic Analysis.** *The 2010 Annual Conference of the Japanese Society for Bioinformatics*, Kyushu University School of Medicine, Fukuoka, Japan. (2010/12/13-15)

② 岩崎渉，高木利久．**誰もが分子進化解析をする時代のための系統関係表現**．*第 33 回日本分子生物学会年会・第 83 回日本生化学会大会 合同大会(BMB2010)*，ポートアイランド，神戸．(2010/12/7)

③ Wataru Iwasaki and Toshihisa Takagi. **A Travelling Salesman Approach to Phylogenetic Trees.** *4th Asian Young Researchers Conference on Computational and Omics Biology (AYRCOB)*, Biopolis, Singapore, Singapore. (2010/12/2)

④ Thanet Praneenararat, Toshihisa Takagi, and Wataru Iwasaki . **Interactive, Multi-Scale Navigation of Large and Complicated Biological Networks.** *生命情報科学若手の会第２回研究会*，国立遺伝学研究所，三島．(2010/10/10)

⑤ 岩崎渉．**車輪樹法：系統樹を超えて**．*生命情報科学若手の会第２回研究会*，国立遺伝学研究所，三島．(2010/10/10)

⑥ 岩崎渉．**系統推定結果の表現と認識**．*第 147回農林交流センターワークショップ*，農林水産省農林水産技術会議事務局筑波事務所，つくば．(2010/8/24)

⑦ 岩崎渉. **車輪樹法：曖昧な進化系統解析結果をどう認識するか**. *日本進化学会 第12回 東京大会*, 東京工業大学大岡山キャンパス, 東京. (2010/8/4)

⑧ Thanet Praneenararat, Wataru Iwasaki, and Toshihisa Takagi. **Interactive, Multi-Scale Navigation of Large and Complicated Biological Networks**. *The Annual CBRC Open House Workshop (CBRC2010) on Bioinformatics*, AIST Tokyo Waterfront, Tokyo, Japan. (2010/7/28-30)

⑨ Wataru Iwasaki and Toshihisa Takagi. **Centroid Representation of Phylogenetic Trees Solved as a Travelling Salesman Problem**. *The Annual CBRC Open House Workshop (CBRC2010) on Bioinformatics*, AIST Tokyo Waterfront, Tokyo, Japan. (2010/7/28-30)

⑩ 岩崎渉. **膨大な細菌ゲノム情報が可能にする代謝ネットワーク進化の俯瞰的解析**. *第83回日本細菌学会総会*, パシフィコ横浜, 横浜. (2010/3/28)

⑪ Thanet Praneenararat, Wataru Iwasaki, and Toshihisa Takagi. **Effective Multi-Scale Graph Navigation System Powered by Fast and Biologically Meaningful Hierarchical Clustering**. *3rd Asian Young Researchers Conference for Computational and Omics Biology*, National Cheng Kung University, Tainan, Taiwan. (2010/3/10-11)

⑫ Wataru Iwasaki and Toshihisa Takagi. **Updated Phylogenetic Tree for the Era of 1000 Genomes**. *3rd Asian Young Researchers Conference for Computational and Omics Biology*, National Cheng Kung University, Tainan, Taiwan. (2010/3/10-11)

⑬ 岩崎渉. **1000ゲノム時代に進化を読み解く：分子系統樹では物足りないときどうするか**. *第37回駒場進化セミナー*, 東京大学駒場キャンパス, 東京. (2010/3/8)

⑭ 岩崎渉, 山本泰智, 高木利久. **生命科学研究者のための統合文献情報管理システム／研究者の論文フォルダには何が埋もれているか**. *第43回 人工知能学会 分子生物情報研究会 (SIG-MBI) ／第13回 オープンバイオ研究会*, 北陸先端科学技術大学院大学, 能美. (2010/3/5-6)

⑮ Thanet Praneenararat, Wataru Iwasaki, and Toshihisa Takagi. **Effective Multi-Scale Graph Navigation System Powered by Fast and Biologically Meaningful Hierarchical Clustering**. *The 20th International Conference on Genome Informatics*, Pacifico Yokohama, Yokohama, Japan. (2009/12/14-16)

⑯ 岩崎渉, 山本泰智, 高木利久. **生命科学研究者のための統合文献情報管理システム**. *第32回日本分子生物学会*, パシフィコ横浜, 横浜. (2009/12/10)

⑰ Wataru Iwasaki and Toshihisa Takagi. **Large-scale genome evolution analysis across the tree of life**. *Global COE Workshop on Bioinformatics in the Era of Genome Information Big Bang*, BGI-Shenzhen, Shenzhen, China. (2009/11/6-7)

⑱ 岩崎渉. **複数遺伝子を必要とするシステムの獲得はどのようにして可能になったのか**. *情報生命科学若手の会第1回研究会*, 国立遺伝学研究所, 三島. (2009/4/25)

〔その他〕

系統関係可視化のための車輪樹法公開ウェブサイト：
http://cwt.cb.k.u-tokyo.ac.jp/

生命ネットワークナビゲーションのためのNaviCluster公開ウェブサイト：
http://navicluster.cb.k.u-tokyo.ac.jp/

文献管理・推薦システム TogoDoc Client 公開ウェブサイト：
http://tdc.cb.k.u-tokyo.ac.jp/

６．研究組織
(1)研究代表者
岩崎　渉（IWASAKI WATARU）
東京大学・大学院新領域創成科学研究科・助教
研究者番号：50545019

(2)研究分担者
なし.

(3)連携研究者
なし.