様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 5 月 31 日現在

研究種目：特定領域研究
研究期間：2005 ～ 2009
課題番号：17017002
研究課題名（和文）　知識処理技術を用いた生命システムの再構築とその解析
研究課題名（英文）　Reconstruction and Analysis of Life Systems Using Knowledge-Processing Technology

研究代表者
高木　利久（TAKAGI TOSHIHISA）
東京大学・大学院新領域創成科学研究科・教授
研究者番号：30110836

研究成果の概要（和文）：生命をシステムとして理解するための要素技術として、大量の表現型画像データを用いた多次元 QTL 解析手法、テキストマイニングを用いた大規模実験の解釈手法、実験手法に関する情報のテキストからの抽出手法、論文中の図表の種類を判別するための手法、大量のゲノム配列情報を用いたゲノムおよび遺伝子ネットワークの進化解析手法、複数の生物種の生命ネットワークの比較手法、および複雑なネットワークデータの解釈のための可視化手法をそれぞれ開発した。

研究成果の概要（英文）：To understand life as systems, we developed the following technologies: high-dimensional QTL analysis by using abundant image data, interpretation of omics experiments assisted by text-mining, text mining of methodological information from literatures, classification of figures in research papers, evolutionary analysis of genomes and biological networks by using abundant genome data, methods for comparing networks from different species, and navigation tools for interpreting huge network data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2005 年度	32,000,000	0	32,000,000
2006 年度	32,400,000	0	32,400,000
2007 年度	32,900,000	0	32,900,000
2008 年度	30,300,000	0	30,300,000
2009 年度	30,400,000	0	30,400,000
総　計	158,000,000	0	158,000,000

研究分野：ゲノム情報科学
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：オントロジー、機能解析、パスウェイデータベース、テキストからの情報抽出、自然言語処理、QTL 解析、知識発見、文献クラスタリング

１．研究開始当初の背景

生命をシステムとして理解するためには、ゲノム配列や蛋白質立体構造だけでなく、発現、局在、相互作用、パスウェイ、ネットワーク、表現型などに関するデータおよびそれらの間の関係や生物学的な制約や文脈などに関する知識などを計算機上に統合し、その性質、特徴、振る舞い、などを解析することが不可欠である。そこで本研究では、
(1)表現型情報と種々のゲノム情報の統合および知識発見技術の開発による生命システムの構造解明

(2)医学・生物学文献から知識やその根拠となった実験事実などを抽出し利用する技術の開発
(3)複雑な生物知識の表現法およびそれらの比較解析や検索のための手法の開発
の3つのテーマについて研究を展開することとした。

２．研究の目的
(1)本領域研究において、多くの生物について遺伝子破壊実験などにより表現型に関する機能情報の蓄積が行われる。当該情報と、下記(2)において抽出予定の関係性ネットワーク、細胞内局在情報、遺伝子発現情報などを統合したデータベースを構築するとともに、これからの知識発見技術を開発する。これにより遺伝子間の複雑なネットワークの構造を明らかにする。

(2)蛋白質/遺伝子や化合物だけでなく、生物学的機能、疾患、症状、生理学、免疫学などに関する知識（様々な概念に関する情報と概念間の関係性）の効率的な検索、分類、抽出のための情報技術を開発する。また、それらの知識の根拠となる実験事実や、それらの知識をコンパクトな形で表現している論文中の図等に関する検索・抽出技術も開発する。さらに、そのための用語辞書等の言語リソースを整備する。これに加え、文献より抽出した断片的な知識を組み合わせて、マイクロアレイ等の膨大な実験データを解釈し、新たな知識発見を支援するシステム開発を行う。

(3)生命システムのもつ階層性と制約（細胞内局在性など）を考慮したパスウェイ・ネットワークの表現法・解析法および表現型データの表現法・解析法を研究する。また、ゲノムやそこにコードされたパスウェイやネットワークがどのようにして形作られてきたのか、その進化過程をゲノム情報から明らかにするとともに、これらのパスウェイ・ネットワークに潜む規則性やゲノムと表現型の間の関係を明らかにする手法の開発を行う。

３．研究の方法
(1)表現型データの処理技術開発の一環として、量的形質に関係する遺伝的要因の探索手法を開発する。特に、統計遺伝学や組み合わせ最適化に基づいた探索アルゴリズムを開発し、実際のデータに適用して有用性を評価する。表現型情報と遺伝的要因の関係を評価するため、連続的な「数値」として観測される形質(量的形質)に関連する遺伝子座(QTL)を網羅的に抽出する解析手法を開発する。その際には、複数のQTLの相互作用が構成する遺伝子座間ネットワークの構造が量的形質に及ぼす効果にも注目し、単独の遺伝子座の効果のみでは説明できない遺伝的背景を抽出する。

(2)潜在的知識発見支援、及び、大規模実験結果解釈支援を目的としたテキストマイニングシステムを開発する。既知の概念ネットワークの構築、及び、大量の遺伝子を文献の類似性をベースにクラスタリングするための自然言語処理技術の開発と共に、効率的な知識発見や実験解釈の自動化のための新規な枠組みを考案する。具体的には、フルペーパーからの情報抽出（蛋白質相互作用、疾患情報など）の高精度化を目指すと共に、高度な情報抽出・情報検索システムを構築するための環境整備として、疾患名称、パスウェイ名称、毒性関連用語などのシソーラスの整備などを行う。また、概念間の関係計算手法及び、概念ネットワークの最適化計算の手法の改良、それらの根拠となる文章の検索・抽出手法の改良を行う。RNAiによる表現型変化の解釈、及び、マイクロアレイの発現量変化の解釈のための機能に関しては、文献情報と共に、配列情報を含む他の情報を統合する手法を考案する。
　また、生物学的知識の根拠をテキストマイニングによって効率的に抽出することで、特に、蛋白質の機能に関する網羅的な知見の信頼性を高めることを目指す。具体的には、従来主に文献処理の対象とされてきた論文要旨からは取得することが困難な、蛋白質／ドメイン間相互作用に関する知見の根拠となる実験環境や手法に関する情報をフルペーパーから抽出する技術の開発を行う。
　テキストー画像横断検索支援に関しては、画像とテキストの横断検索システムのための画像の特徴量の量子化とその取り扱いの方法につき、研究開発を行う。また、文献に含まれる多くの図および図の脚注に研究成果に関する重要な情報が含まれていることが多いことに着目し、さまざまなクエリによって関連のある図を検索することのできるデータベースシステムを構築することを目的とする。

(3)1000を超える数のゲノム配列が大規模に読まれつつあることを利用して、それらのゲノムの進化過程を高精度かつ効率的に推定するためのアルゴリズムを開発する。さらに、推定されたゲノム進化過程をもとに、各種データベースに蓄積されているより高次の情報（相互作用や遺伝子機能など）と組み合わせ、パスウェイやネットワークがどのように現在の姿になったのかについて、進化解析を行う。
　生物学研究者が研究を進める際には、論文等の知識を総合的に検討し、新たなターゲットの選定や条件の絞込みを行っている。この

ような判断を支援する際、個々の遺伝子についての情報のみでなく、パスウェイ同士の種間比較を行うことで新たな知見を得ることが期待される。そこで、研究者が興味を持つパスウェイをクエリとし、指定された探索条件によってさまざまな種から類似したパスウェイを抽出し、それらを比較するシステムを構築する。特に、蛋白質間相互作用やシグナル伝達などのパスウェイ情報を対象とする。

４．研究成果

(1)QTL 解析に関しては、量的形質とマーカー座の遺伝子型の関連を任意の遺伝子座での遺伝子型を補間しながら染色体に沿って１次元的にスキャンする手法（区間マッピング）を実装した。実装した区間マッピングを２次元化して、2 つの遺伝子座間の相互作用を評価できるように拡張した。両手法を独自に実装することによって、手法内で用いられている遺伝モデルの変更や内部パラメータの情報の可視化が行えるようになり、より詳細な解析が可能になった。さらに、探索木を用いた組み合わせ最適化アルゴリズムを開発して、3 つ以上のマーカー座間の相互作用を有意さの高いものから順に出力できるように拡張し、一連の手法によって網羅的な QTL 解析が行えるようになった。

開発した手法を用いて、メダカの顔貌形質に関係する QTL 解析を行った。メダカの顔貌形質の解析に関しては、顔貌の「部分」のみでなく、「全体」の数値的な表現方法を検討した。従来は、顔貌の特徴点を手作業で認識し、その特徴点によって定義される多数の形質を別々に評価していたが、本研究では複数の形質を統合して顔貌全体を表す「合成形質」を定義し、その合成形質のマッピングを試みた。その結果、顎付近の比較的少数（3ー5 程度）の形質の加重線形和による近交系集団の効率良い分離と、それに関連する候補遺伝子座が確認できた。

より詳細な解析を行うために、メダカ顔貌の幾何学的モデルを定義し、そのモデルを画像に当てはめることによって特徴点を自動的に抽出する手法を開発した。得られた特徴点の網羅的な組み合わせを評価して、顔貌全体を表す合成形質を多数生成した。得られた合成形質をクラスタリングすることによって、比較的少数(10 以下)の類型に分類した。類型ごとに形質のマッピングを行い、顔貌との関連が示唆される染色体領域を抽出した。網羅的に合成形質を生成することによって、主観に依らない形質の選択を可能にしている。また、互いに相関した合成形質をクラスタリングすることによって冗長性を排除している。

解析で用いた複数の近交系を含む個体群の頭部の画像データ（新屋みのり班員・遺伝研）をデータベース化した（MCTDB データベース)。区間マッピング等の解析機能を備え、解析結果は WEB 画面から参照することが可能である。

QTL 解析に広く用いられている手法は 1 次元的なものであり、複数の QTL 間の相互作用を評価する標準的な手法は確立していない。本研究で開発した手法は従来の 1 次元的な解析手法に加えて、多次元的な解析を可能にしたものである。また、QTL 解析の対象としたメダカの顔貌に関する形質を画像データから自動的に抽出する手法は存在していない。本研究で作成したメダカ顔貌の幾何学的モデルは、メダカに類似した小型魚類の顔貌データへの応用も期待できる。

(2)フルペーパーに対応した固有表現抽出、及び、語彙的曖昧性の除去手法の開発を行った。またこれをベースに相互作用情報や疾患情報などの情報をフルペーパーから情報抽出するプログラムのプロトタイプを作成した。

潜在的な知識発見システムについては、概念間の関係の計算手法の改良及び、概念認識プログラムの改良を行った。また、このシステムへの連鎖解析の結果解釈への適用を試み、その妥当性及び有用性を検討した。大規模実験法の解釈機能については、文献類似性による遺伝子クラスタリングの手法の改良と共に、ユーザーフレンドリーなインターフェースを開発した。

蛋白質相互作用に関する知識の根拠となる実験環境等のフルペーパーからのマイニングに関しては、オープンアクセス可能でかつ RefSeq 等の公共データベースにおいてアノテーションが行われる際の根拠文献として比較的多く採用されている Journal of Biological Chemistry (JBC)および Cell の 2 つの論文誌 についてまず、5 年分のフルペーパーを取得した。そして、これらのフルペーパーから細胞名や実験手法、蛋白質／遺伝子名のテキストマイニング技術による抽出を行った。このデータを利用し、これまで新型シーケンサーによって発現データが報告されている細胞株について、発現の確認された遺伝子のうちどの程度の割合について当該細胞株において機能が報告されているか、調査を行った。

テキストー画像横断検索支援に関しては、まず、生体機能の理解に不可欠なシグナル伝達および代謝パスウェイを表した図に着目し、図がこれらの図であるか否かの自動分類を行うため、文献中の図の脚注および本文中の該当図説明文を用いて、教師つき機械学習法によって判別モデルを構築した。その結果、図の脚注のみを用いた場合で既存の関連研

究を上回る精度を達成し、なおかつ図の脚注と本文を組み合わせることで精度が向上することを示すことができた。さらに、目的の図からの文字認識にも取り組み、シンプルな図を対象に解析を行った結果、高い精度で図中の文字領域を特定することに成功した。

また、対象とする図の種類を複数に増やして図の分類モデルの拡張を行った。具体的には、脳科学研究におけるヒトの脳イメージングの結果を表わした図を、イメージング手法ごとに分類できるようにした。対象としたのは、脳領域と機能・構造間の対応関係を直接的・視覚的に表現した CT、MRI、 PET、fMRI の図、および、脳電図（EEG）・脳磁図（MEG）の波形データを脳内空間にマッピングした図である。こうした図は画像的な特徴が非常に似通っているため、画像処理技術を用いて分類を行うことは極めて困難であることが既存研究によって明らかになっている。よって本研究では、図の脚注・本文に含まれる情報から効率的にマイニングを行うことで分類を実現した。

さらに、画像とテキストの横断検索のために、画像の特徴量を量子化しテキストと同様にインデックス化し、横断検索を行えるようにした。画像とテキストの情報の重み付けを行い、画像の類似度が低く、特定の意味で類似しているもの、画像の類似度が高く、特定での意味は類似していないものなどを自在に検索できるようになった。

フルペーパーには、論文要旨からでは得られない重要な知識が記述されていることや、オープンアクセスが可能である学術誌あるいは論文が増えていることから、近年テキストマイニングや知識抽出技術を適用する対象として注目されている。特に、潜在的知識発見は、ここ 2-3 年で着目されている技術であり、開発が急がれる。また、図の特徴量とテキスト情報を同時に利用するテキストマイニングは、世界でも数少ない研究であり、国際会議 BIBM 2008 にて発表を行った。

(3)膨大なゲノム配列情報と信頼性の高い分子系統樹とをもとに、過去の生物種がどのような全遺伝子のセットを持っていたかを高い信頼度で推定するための手法を開発した。ゲノム進化においては、例えばゲノム重複や寄生生活への移行により遺伝子数の急激な増加や減少がしばしば起こるが、本手法の特長は、系統樹上の各枝における遺伝子の獲得／欠失速度を、最尤法および期待値最大化法の枠組みにより推定することで、ゲノムの非単調な進化過程を反映した解析を行う点にある。さらに、本手法を真核生物・真正細菌・古細菌にまたがる大規模ゲノムデータに適用し、その進化解析を行うことで、実際にゲノムは時に急激な変化を経つつ進化してきたことが示唆された。また、このゲノム進化過程の再構築結果を 160 種の原核生物ゲノムデータに適用し、代謝パスウェイの進化過程の網羅的な再構築・解析を行った。その結果、初期のパスウェイ獲得は異なる系統群間で同時代的に起こったことが示唆された。この結果をもとに、パスウェイの進化が原核生物コミュニティ内での双方向的な遺伝子水平伝播により促進されるという新たな進化モデルを提案した。

ユーザが指定した細かなクエリ・探索条件に基づいて、複数種の蛋白質間相互作用データから、目的のネットワークを生物種毎に抽出するシステムを構築した。細かい類似条件を設定できるほか、ギャップ、ミスマッチについても個々に指定でき、これまでの探索プログラムに比べてネットワークの抽出や絞込みを容易にした。さらに、蛋白質間相互作用に、TRANSPATH および KEGG からのシグナル伝達に関するデータを統合し、パスウェイ情報を複数の種について検索・比較するシステムを構築した。クエリ中の蛋白質と探索対象種の蛋白質間を結びつける情報としては、配列類似性のほか、Gene Ontology の意味類似性、オルソログ情報など複数の条件を利用可能とした。加えて、抽出した類似ネットワークの提示において、抽出ノードをクエリ上のノードも含め各類似条件によってクラスタリングしたり、クエリ上のエッジと対応するパスウェイのみ表示したりする機能を付加し、新たな知見を見出すための利用環境を検討した。

全遺伝子セットの進化過程を効率的かつ高精度で再構築する手法は世界的に高い評価を受け、世界最大のバイオインフォマティクス国際会議での口頭発表に採択された。また、パスウェイの進化が原核生物コミュニティ内での双方向的な遺伝子水平伝播により促進されるというモデルは、パスウェイの進化という長く議論されてきた問題に新たな視点を提供するものであり、全世界のトップ研究者が毎月読んだ論文の中で優れたものを推薦するウェブサイト Faculty of 1000 Biology にて取り上げられた。

蛋白質間相互作用の種間比較については多くの研究が行われているが、種全体にわたって比較し、一定の条件下で有意なパスウェイを抽出する研究が主で、抽出条件・結果が必ずしも生物学研究者の要求を満たすものでない。また、一部、クエリを与えることに対応したものもあるが、パスウェイの構成要素の個々について細かい探索条件を指定できるものはなく、今回構築したシステムが唯一のものである。

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 23 件) ※すべて査読有

(1)Koike, A. and Takagi, T.: Classifying biomedical figures using combination of bag of keypoints and bag of keywords, Proc. of 2nd Int. Workshop on Intell. Inform. in Biol. and Med., 848-853 (2009).

(2)Iwasaki, W. and Takagi, T.: Rapid Pathway Evolution Facilitated by Horizontal Gene Transfers across Prokaryotic Lineages, PLoS Genetics., 5(3), e1000402 (2009).

(3)Ishii, N., Koike, A., Yamamoto, Y. and Takagi, T.: Figure Classification in Biomedical Literature towards Figure Mining, IEEE International Conference on Bioinformatics and Biomedicine, 263-269 (2008).

(4)Iwasaki, W. and Takagi, T.: Reconstruction of Highly Heterogeneous Gene-Content Evolution across the Three Domains of Life, Bioinformatics (ISMB/ECCB2007 proceeding issue), 23, i230-i239 (2007).

(5)Yamamoto, Y. and Takagi, T.: OReFiL: an online resource finder for life sciences, BMC Bioinformatics, 8, 287 (2007).

(6)Yamamoto, Y. and Takagi, T.: Biomedical knowledge navigation by literature clustering, Journal of Biomedical Informatics, 40(2), 114-130 (2007).

(7)Koike, A. and Takagi, T.: Knowledge discovery based on an implicit and explicit conceptual network., J. Am. Soc. Info. Sci. Tech., 58(1), 51-65 (2007).

(8)Dohkan, S., Koike, A. and Takagi, T.: Improving the performance of an SVM-based method for predicting protein-protein interactions, In Silico Biol., 6, 0048, (2006).

(9)Noguchi, H., Park, J. and Takagi, T.: MetaGene: prokaryotic gene finding from environmental genome shotgun sequences, Nucleic Acids Res., 34(19), 5623-5630, (2006).

(10)Ao, H. and Takagi, T.: ALICE: An Algorithm to Extract Abbreviations from MEDLINE, JAMIA, 12(5), 576-586 (2005).

(11)Koike, A., Niwa, Y. and Takagi, T.: Automatic Extraction of Gene/protein Biological Functions from Biomedical Text, Bioinformatics, 21, 1227-1236 (2005).

(12)Yakushiji, A., Miyao, Y., Tateisi, Y. and Tsujii, J.: Biomedical Information Extraction with Predicate-Argument Structure Patterns, Proc. 1st Int. Symp. on Semantic Mining in Biomedicine, 60-69 (2005).

(13)Tateisi, Y., Yakushiji, A., Ohta, T. and Tsujii, J.: Syntax Annotation for the GENIA corpus, Proc. IJCNLP 2005, 222-227 (2005).

〔学会発表〕(計 18 件)

(1)Praneenararat, T.(Takagi, T.), Effective multi-scale graph navigation system powered by fast and biologically meaningful hierarchical clustering, GIW2009, 2009/12/14-16、横浜

(2)Koike, A. (Takagi, T.), Biomedical figure search using combination of bag of keypoints and bag of words, GIW2009, 2009/12/15, 横浜

(3)Ishii, N. (Takagi, T.), Classification of neuroimaging figures toward automatic figure annotation system, GIW2009 2009/12/14, 横浜

(4)岩崎 渉、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、2009/12/10、横浜

(5)鈴志野 康正 (中谷 明弘)、メダカ顔貌形状に関わる遺伝子座の探索に向けた量的形質の定量手法、第 32 回日本分子生物学会年会、2009/12/9、横浜

(6)Iwasaki, W., Reconstruction of highly heterogeneous gene-content evolution across the three domains of life, ISMB/ECCB2007, 2007/7/24, Wien

(7)Fukagawa, H. (Takagi, T.), A prototype platform for flexible comparison analyses of interactions and pathways, ISMB/ECCB2007, 2007/7/22, Wien

〔図書〕(計 2 件)

(1)小池　麻子 (高木 利久)、羊土社、医学生物学分野におけるシソーラスとテキストマイニング技術の開発 (実験医学増刊号 Vol.26, No.7)、2008、6

〔その他〕

データベース

(1)MCTDB: Medaka Craniofacial Trait Database
http://medaka.cb.k.u-tokyo.ac.jp/mctdb/

(2)BioTermNet
http://btn.ontology.ims.u-tokyo.ac.jp/

(3)OReFiL
http://orefil.dbcls.jp/

(4)PIPS

http://prime.ontology.ims.u-tokyo.ac.jp:8081/cgi-bin/PIPS.cgi
(5)McSyBi
http://textlens.hgc.jp/McSyBi/index.html
(6)GENA
http://gena.ontology.ims.u-tokyo.ac.jp
:8081/search
(7)Multiple Ontology Viewer
http://gena.ontology.ims.u-tokyo.ac.jp
:8081/mov
(8)MEDLINE全体のテキストをEnjuで解析した結果。
リクエスト先メールアドレス：
genia@is.s.i-tokyo.ac.jp
(9)GENIA Treebank
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/ Corpus/GTB.html
(10)ALICE
http://uvdb3.hgc.jp/ALICE/ALICE_index.html

６．研究組織
(1)研究代表者
高木　利久（TAKAGI TOSHIHISA）
東京大学・大学院新領域創成科学研究科・教授
研究者番号：30110836

(2)研究分担者
辻井　潤一（TSUJII JUNICHI）
東京大学・大学院情報理工学系研究科・教授
研究者番号：20026313
（H17-H18.7月）

中谷　明弘（NAKAYA AKIHIRO）
東京大学・大学院新領域創成科学研究科・准教授
研究者番号：60301149

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
（H21）