

科学研究費助成事業 研究成果報告書

平成 27 年 6 月 3 日現在

機関番号：12601

研究種目：若手研究(B)

研究期間：2011～2014

課題番号：23710231

研究課題名(和文) オーミクス実験データ解釈におけるボトルネックの解消

研究課題名(英文) Alleviating bottlenecks in interpretation of omics datasets

研究代表者

岩崎 渉 (Iwasaki, Wataru)

東京大学・理学系研究科・准教授

研究者番号：50545019

交付決定額(研究期間全体)：(直接経費) 3,500,000円

研究成果の概要(和文)：爆発的なペースでオーミクス実験データが蓄積し続ける中、研究者がそれらのデータを論文や様々なデータベースの内容に照らし合わせつつ仮説を構築するステップが生命科学研究における深刻なボトルネックとなっている。本研究では、論文データベースおよび近年様々なプロジェクトで整備されてきたデータベース群のデータをネットワークの形式で表現することで、膨大なオーミクス実験データの解釈を支援するための情報技術開発を行った。

研究成果の概要(英文)：As omics data are accumulating at an ever-increasing pace, life science is facing with a fundamental bottleneck in generating hypothesis by referring to existing knowledge in databases and literature. In this study, we created a system that represents knowledge in databases and literature in network structures. The system will help researchers interpret massive omics datasets.

研究分野：バイオインフォマティクス

キーワード：バイオインフォマティクス オーミクス データ解釈 データベース

1. 研究開始当初の背景

近年のオーミクス実験技術の進歩は著しく、中でもいわゆる次世代シーケンサは世界の生命科学研究を大きく変えつつあった。これまで手に入らなかった種類・量のゲノム解読を目指す野心的な国際プロジェクトがすでにスタートしていた(Genome 10K プロジェクトや 25,000 ガンゲノムプロジェクトなど)ほか、ゲノム配列情報にとどまらず、次世代シーケンサ技術を用いた網羅的な DNA メチル化解析やヌクレオソーム構造解析、トランスクリプトーム解析、さらにはゲノム 3 次元構造情報解析などの新たな技術開発も世界的に競って推進されていた。極めて安価に購入可能な解析装置も市場に登場し、一般的な研究室でも多様なオーミクス実験データを扱う時代が到来しつつあった。

このことは、「研究者が実験を行い、その結果を考察して知識とする」生命科学研究の中で、前者について加速度的なハイスループット化が進行していることを意味していた。すなわち、オーミクス実験データを表現型などの高次生命概念に結びつけて解釈を行う過程が生命科学におけるボトルネックとなりつつあった。この“解釈”に関して現在事実上の標準となっている手法としては、Gene Ontology 中で統計学的に有意に関連するタームを発見するものがよく用いられていたが、Gene Ontology で表現できる情報はごく限られたものであり、「すでによく知られた知識のオーミクスデータによる確認」を超えて新たな知識を得る上で限界があることが問題となっていた。一方で、これらのオーミクス実験データを処理しデータベース化するバイオインフォマティクス技術について、我が国を初め各国のデータベースセンターにおいて様々なデータベースを整備し連携・統合する作業が行われつつあった。

2. 研究の目的

本研究では、オーミクス実験データの解釈、特に高次生命知識との結びつけを支援し、将来的には半自動化を目指すための基盤となる情報技術開発を目的とした。例えば、100 を超える数の遺伝子がある表現型に影響していることがオーミクス実験によって明らかになったときに、各遺伝子についての詳細な情報をつづつ研究者が調べることで「どうしてそれらの遺伝子が当該表現型に影響するか」解釈・仮説構築を行うのではなく、様々なデータベース群および論文データベースの内容を直感的にわかりやすい形式で縮約・可視化することで、実験後スムーズに仮説を構築するためのシステムを構築することを目指した。

3. 研究の方法

本研究は大まかに 5 つの段階によって進行した。すなわち、

- (1) 文献データベースを用いた遺伝子・概念ネットワークの作成
 - (2) ゲノムデータベース・発現データベース・相互作用データベース等を用いた上記ネットワークの補完
 - (3) 入力オーミクス実験データと入力キーワード群を結ぶ生物学的背景をネットワークの形式で推定するグラフィアルゴリズムの開発
 - (4) 実際のオーミクス実験を用いた仮説構築・検証実験およびフィードバックによるシステムの完成度向上
- の 4 段階である。

4. 研究成果

(1) 英文で書かれた学術文献中から生物種横断的に概念・遺伝子名を発見するシステムを構築した。英文中からの遺伝子名の抽出(固有表現認識)については多くの研究があるが、複数生物種の遺伝子名辞書への対応および遺伝子 ID との対応付けを行うことができる既存システムについて、複数の生物種の遺伝子が同一文献に出てきた場合への対応、および条件付き確率場(CRF)エンジンの追加などの改良を行った。一方、生命科学概念については、文献集合中から統計的に有意な連語ペアを抽出するソフトウェアを用いて機械的にフレーズを抜き出すプログラムを実装した。以上を世界中の生命科学研究者が執筆した 2000 万件を超える膨大な書誌情報が登録されている生命科学文献データベースである米国 NCBI の PubMed (MEDLINE) に対して適用した。続いて、同一抄録および文中で同時に現れやすい概念・遺伝子ペアを網羅的に取得することで、関連性の高い概念・遺伝子同士が互いに関連づけられたネットワークを構築した。

(2) 一般に文献からテキストマイニング技術によって抽出・作成したネットワークは正確性・網羅性の点で不十分であることが知られている。そこで、各生物種ゲノムデータベースのアノテーションから取得できる遺伝子・概念ペア、および発現データベースや相互作用データベースから取得できる遺伝子同士のペアをそれぞれ遺伝子・概念ネットワークに追加し、補完した。なお一般に生命科学研究においては酵母のホモログ遺伝子に関する知識を利用してヒトの遺伝子に関する仮説生成を行うといったことがよく行われるが、これに関して、比較ゲノムデータベース上に整備されている遺伝子の類似性情報を用いることで生物種ごとに違った名称が付けられている概念同士をネットワーク上で間接的に関連づけることに成功した。

(3) 遺伝子・概念ネットワークの中から入力

オーミクスデータの遺伝子セットと入力キーワード群とを繋ぐサブネットワークを探索するグラフアルゴリズムを開発、実装した。オーミクスデータが遺伝子ごとに割り当てられた数値データの場合にはそれに対応した「重み」を持たせた遺伝子セットへ、またネットワークデータの場合にはネットワーク中でモジュールを作っている遺伝子に「重み」を持たせた遺伝子セットへとそれぞれ変換し、これらの「重み」をグラフ探索アルゴリズムにおいて探索の優先度決定に用いることと、より多くの遺伝子・キーワードを含む密なネットワークを優先した探索を行うことで、仮説生成の点から注目度の高いネットワークを抽出できるようにした。

(4) 完成したシステムについて、実際にオーミクス実験を行っている研究者等からの密接なフィードバックを受けつつ、性能面およびインターフェース面について検証・調整を行った(グラフ探索アルゴリズムにおける探索の優先順位の調整など)。さらに、実際のオーミクス実験データを用いた仮説構築を行い、生物学的な実験を行うことにより本システムの有効性を検証した。

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22件)

福永津嵩, 岩崎涉 .

Computational Ethology: バイオインフォマティクスと動物行動学の融合
領域融合レビュー, 査読有, **4**, e003. (2015)
DOI: 10.7875/leading.author.4.e003

Tsukasa Fukunaga, Shoko Kubota, Shoji Oda, and Wataru Iwasaki.
GroupTracker: Video Tracking System for Multiple Animals under Severe Occlusion.
Computational Biology and Chemistry, 査読有, **57**, 39-45. (2015)
DOI: 10.1016/j.compbiolchem.2015.02.006

尾崎遼, 岩崎涉 .

NGS データ解析
実験医学, 査読有, **32**, 3197-3202. (2014)

Marty Kwok-Shing Wong, Haruka Ozaki, Yutaka Suzuki, Wataru Iwasaki, and Yoshio Takei.
Discovery of osmotic sensitive transcription factors in fish intestine via a transcriptomic approach.
BMC Genomics, 査読有, **15**, 1134. (2014)
DOI: 10.1186/1471-2164-15-1134

Shotaro Hirase, Haruka Ozaki, and Wataru Iwasaki.

Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes.
BMC Genomics, 査読有, **15**, 735. (2014)
DOI: 10.1186/1471-2164-15-735

Sebastian J. Schultheiss, Joshua SungWoo Yang, Wataru Iwasaki, Shu-Hsi Lin, Angela Jean, and Magali Michaut
Crossing Borders for Science.

PLOS Computational Biology, 査読有, **10**, e1003519. (2014)
DOI: 10.1371/journal.pcbi.1003519

Susumu Yoshizawa, Yohei Kumagai, Hana Kim, Yoshitoshi Ogura, Tetsuya Hayashi, Wataru Iwasaki, Edward F. DeLong, and Kazuhiro Kogure.

Functional characterization of flavobacteria rhodopsins reveals a unique class of light-driven chloride pump in bacteria.

Proc Natl Acad Sci U S A, 査読有, **111**, 6732-6737. (2014)
DOI: 10.1073/pnas.1403051111

Hikari Yoshitane, Haruka Ozaki, Hideki Terajima, Ngoc-Hien Du, Yutaka Suzuki, Taihei Fujimori, Naoki Kosaka, Shigeki Shimba, Sumio Sugano, Toshihisa Takagi, Wataru Iwasaki, and Yoshitaka Fukada.

CLOCK-controlled polyphonic regulations of circadian rhythms through canonical and non-canonical E-boxes.

Molecular and Cellular Biology, 査読有, **34**, 1776-1787. (2014)
DOI: 10.1128/MCB.01465-13

Ching-chia Yang and Wataru Iwasaki
MetaMetaDB: A Database and Analytic System for Investigating Microbial Habitability.

PLOS ONE, 査読有, **9**, e87126. (2014)

DOI: 10.1371/journal.pone.0087126

Tsukasa Fukunaga, Haruka Ozaki, Goro Terai, Kiyoshi Asai, Wataru Iwasaki, and Hisanori Kiryu

CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data.

Genome Biology, 査読有, **15**, R16. (2014)
DOI: 10.1186/gb-2014-15-1-r16

Teresa Szczepinska, Wataru Iwasaki, and Thomas Abeel.

The Spirit of Competition: To Win or Not To Win.

PLOS Computational Biology, 査読有, **9**,

e1003413. (2013)
DOI: 10.1371/journal.pcbi.1003413

Seishiro Aoki, Motomi Ito, and Wataru Iwasaki.

From - to -Proteobacteria: the Origin and Evolution of Rhizobial Nodulation Genes *nodJ*.

Molecular Biology and Evolution, 査読有, **30**, 2494-2508. (2013)

DOI: 10.1093/molbev/mst153

Wataru Iwasaki, Tsukasa Fukunaga, Ryota Isagozawa, Koichiro Yamada, Yasunobu Maeda, Takashi P. Satoh, Tetsuya Sado, Kohji Mabuchi, Hirohiko Takeshima, Masaki Miya, and Mutsumi Nishida.

MitoFish and MitoAnnotator: A Mitochondrial Genome Database of Fish with an Accurate and Automatic Annotation Pipeline.

Molecular Biology and Evolution, 査読有, **30**, 2531-2540. (2013)

DOI: 10.1093/molbev/mst141

尾崎遼, 岩崎涉.

ENCODE プロジェクトで明らかになったこと
細胞工学, 査読無, **32**, 101-106. (2013)

Thanet Praneenararat, Toshihisa Takagi, and Wataru Iwasaki.

Integration of Interactive, Multi-scale Network Navigation Approach with Cytoscape for Functional Genomics in the Big Data Era.

BMC Genomics, 査読有, **13(Suppl 7)**, S24. (2012)

DOI: 10.1186/1471-2164-13-S7-S24

佐藤行人, 八谷剛史, 岩崎涉.

水圏生物学における次世代シーケンサー活
用の現状と応用可能性への展望

水産育種, 査読有, **41**, 17-32. (2012)

岩崎涉, 山本泰智, 高木利久.

文献管理・文献推薦

実験医学, 査読無, **29**, 2556-2562. (2011)

〔学会発表〕(計 106 件)

岩崎涉.

定量生物学と情報生物学. 定量生物学の会第
七回年会, 九州大学筑紫キャンパス, 福岡県
春日市. (2015/1/11)

岩崎涉. 大規模遺伝子情報から見えてく
る生物進化. 地球システム変動研究の新戦略
-知識連携から見えてきた地球環境と生命の
進化-, 東京大学本郷キャンパス情報学環・
福武ホール, 東京都文京区. (2014/12/10)

岩崎涉. 分子系統学とデータ表現. 第 188
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, 茨
城県つくば市. (2014/11/6)

岩崎涉. Fundamental bottlenecks in
big-data biology. 第 52 回日本生物物理学
会年会, 札幌コンベンションセンター, 北海
道札幌市. (2014/9/27)

岩崎涉. バイオインフォマティクスから
時空間解析へ. 日本学術会議学術フォーラム
生命情報ビッグデータ時代における新しい
生命科学, 日本学術会議講堂, 東京都港区.
(2014/8/29)

〔図書〕(計 1 件)

岩崎涉, ゲノムからひもとく生命進化の
道のり, 小久保英一郎・嶺重慎編著『宇宙と
生命の起源 2 素粒子から細胞へ』(岩波ジュ
ニア新書 777), 181-200. (2014)

〔産業財産権〕

出願状況 (計 0 件)

取得状況 (計 0 件)

〔その他〕

ホームページ

<http://iwasaki lab.bi.s.u-tokyo.ac.jp/>

6. 研究組織

(1) 研究代表者

岩崎 涉 (IWASAKI, Wataru)

東京大学・大学院理学系研究科・准教授

研究者番号: 50545019