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

平成　３０　年　　６　月　２５　日現在

機関番号：　８２６５７
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１２５３１

研究課題名（和文）全人類のゲノム変異を包括的に扱うリファレンス・ゲノム・グラフのデータ構造の研究

研究課題名（英文）Study on the data structure of the reference genome graph for handling all human genome variations

研究代表者

片山　俊明（Katayama, Toshiaki）

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：６０３９６８６９

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：当研究では、多数のゲノムをグラフ構造として束ねることにより、今後大幅に増加する個人ゲノムの解析の効率化や、個人差・個体差の大きいゲノム領域やリファレンス配列にない領域を扱えないといった問題を解決するため、ゲノムグラフのデータ構造と、そのデータベースとの連携のための仕組みの開発に取り組んだ。ゲノムのアノテーション情報とゲノムグラフの構造をどちらもセマンティック・ウェブ技術のResource Description Framework (RDF)で表現するための仕組みを開発し、これを用いたアノテーショントランスファーなどの実験を行うことで実用性の検証を行った。

研究成果の概要（英文）：The genome graph is expected to represent multiple genomes as a single graph for accelerating personal genome analysis by dealing with insertion sequences which doesn't exist in the current linear reference genome sequence and handling complex variations of the individual genome sequences. In this study, to integrate genome annotations in the knowledge base and the structure of genome graphs, it is tested to convert both of data into the Resource Description Framework (RDF) and applications such as annotation transfer are made.

研究分野：　バイオインフォマティクス

キーワード：　ゲノムグラフ　セマンティックウェブ　データベース　アノテーション　ヒトゲノム

１．研究開始当初の背景

ゲノム解読にかかるコストが低下したことにより、今後は個人ゲノムの解析が大幅に増加していくことが予想されているが、これまでの西洋人を代表とするリファレンスゲノム配列と比較するだけでは、人種や個人によるゲノムの違いや、それに起因する疾患を包括的に解析することができない。このため全てのゲノムをグラフとして統合するゲノムグラフの技術開発が始まっている。この技術とゲノムデータベースの知識を統合することができるシステムを構築すれば、ゲノム情報の互換性を向上させつつ解析の効率化を図ることができ、ゲノム研究のパラダイムが変わると期待される。

リファレンスゲノムをグラフとして扱うこと自体が挑戦的な課題であるが、2013 年より始まった Global Alliance for Genomics and Health (GA4GH)という国際的な枠組みの中で、そのデータ構造や検索方法の提案が進み始めていた。この際、情報科学的に標準的とはいえない技術の提案もなされており、長期的な観点での持続可能性に懸念があったため、グラフに基づくセマンティック・ウェブ技術の導入をプロトタイプを作成して働きかけていた状況であった。

２．研究の目的

個人ゲノムの配列決定が進むなか、ガンや認知症から、希少疾患や未診断病まで、それに関わる遺伝子メカニズムの解明が求められている。このためには多数の個人ゲノムと疾患の関係を統合的に解析する必要がある。全人類のゲノムを１つのグラフとして統合することで、個々人のゲノム変異情報を取りこぼしなく取り扱うことが可能となる。

一方で、ゲノム変異を解釈するためには、整備が進められている様々な医科学データベースとの連携が必須である。これらを統合的に利用するために、申請者の所属するライフサイエンス統合データベースセンターでは、セマンティック・ウェブ技術を用いた標準化を進めてきている。ここで用いられる Resource Description Framework (RDF)もグラフに基づく知識表現形式であり、異種の情報を同一形式で記述し、一元的に扱うことができる。

本研究では、ゲノムグラフと知識のグラフを統合し、個人ゲノムの解釈に必要な情報を個々人のゲノムと連携させるシステムの構築に繋がる技術開発を行うことで、次世代のゲノム研究の高精度化・効率化を目指す。

３．研究の方法

ゲノムグラフを RDF に対応させるためには、主な開発グループである米国カリフォルニア大学サンタクルツ校(UCSC)と、欧州 Sanger Institute の中心メンバーの協力が必要となる。このため、これらの研究機関を訪問して打ち合わせを行うとともに、ライフサイエンス統合データベースセンターが日本で毎年開催してきている国際開発者会議 BioHackathon や RDF summit、GA4GH の国際会議などの機会を活用しつつ、セマンティック・ウェブの専門家とも連携して開発を進める。

４．研究成果

ゲノムグラフの技術はゲノムのアセンブルに繋がるものであり、ルーツを辿ればゲノムプロジェクトと並んで古くから発展してきた歴史があるといえるが、リファレンスゲノムをグラフに置き換えるという計画は当初 2013 年ごろから GA4GH で取り上げられてきたものである。一方で、ライフサイエンス統合データベースセンターでは 2010 年から様々なデータベースのセマンティック・ウェブによる統合を進めてきていた。2015 年の GA4GH のハッカソンで、当時提唱されていた Genomics API の標準化を、独自の仕様で進めるのではなく、セマンティック・ウェブの標準技術で行えることを示すプロトタイプを作成したところ、開発メンバーから一定の理解を得られた。Genomics API は今後ゲノムグラフにも対応していく必要がある状況であったため、ゲノムグラフをセマンティック・ウェブでも表現できる技術を開発することで、これまで何度も問題となっていた生命科学で独自技術を採用することによるオーバーヘッドを削減し、標準技術による解析システムの構築と普及ができると期待された。

当時すでに国内でも日本人ゲノムの解読が進んでおり、西洋人を代表とするリファレンスゲノムにない日本人固有のゲノム領域の解析も課題の一つとなっていた。これまでの方法では、リファレンスゲノムと対応が取れる大部分の領域と、個人や民族固有の領域を分けて解析せざるを得ず、固有領域については国際的な共通の解析ツールがそのままでは利用できないことなどによる、解析パイプラインの複雑化や評価基準のダブルスタンダード化などが懸念された。そこでリファレンスを人種にかかわらず１つに統一できるゲノムグラフの利用が普及することで、解析手順やツール開発が一元化され、研究のクオリティの担保や将来にわたる効率化が期待できると考えられた。

このため、ライフサイエンス統合データベースセンターが 2015 年に開催した国際開発者会議 BioHackathon にゲノムグラフの開発者として UCSC から Benedict Paten 氏らを招聘し、日本人ゲノムに対するゲノムグラフの適用の可能性や、ゲノムグラフのセマンティック・ウェブへの対応について検討を開始した。この間に本研究の提案を行い、とくにゲノムグラフの RDF による表現方法について研

究を開始した。2016 年にはゲノムグラフ自体の開発促進とセマンティック・ウェブとの融合を目指した小規模なハッカソン RDF summit2 を開催し、先に招聘した Benedict Paten 氏らと共にゲノムグラフの研究コミュニティの中心的なメンバーである Erik Garrison 氏、Juoni Siren 氏、Maciek Otto 氏、Adam Novak 氏らと国際的な人的ネットワークの形成を行った。ハッカソンには、スイスのバイオインフォマティクス研究所 SIB で代表的なアミノ酸配列のデータベース UniProt の RDF 化を行っている専門家 Jerven Bolleman 氏と欧州バイオインフォマティクス研究所 EBI でゲノムデータベース Ensembl の開発に携わる Kieron Taylor 氏も参加し、ライフサイエンス統合データベースセンター、東北大学東北メディカル・メガバンク機構、東京大学、産業総合研究所などの国内の研究者と共にゲノムグラフの RDF 化を開始した。

ゲノムグラフの構築には、上記のメンバーが中心となって開発している vg というツールが使われており、vg からゲノムグラフのデータ構造を RDF に変換するための機能が実装された。ここでは、ゲノムグラフに含まれるノードと、ゲノム上での繋がりを表すパスを、ステップで結びつけて順番を表現するモデルが採用された(図 1)。これにより、vg で作成されたゲノムグラフは RDF に必ず変換できること、また RDF からゲノムグラフが再構成できることになった。

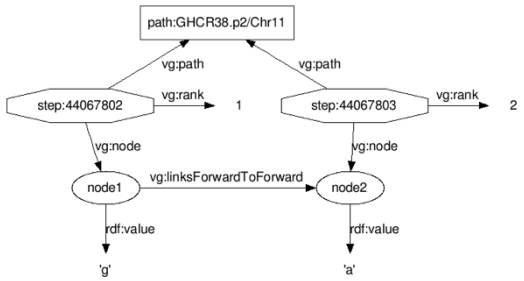

図 1 ゲノムグラフの RDF によるモデル

また、RDF 化されたゲノムグラフは RDF データベースの検索言語 SPARQL によって問い合わせを行うことができる(図 2)。

```
PREFIX rdf: <ttp://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vg: <http://example.org/vg/>
SELECT ?near_node ?sequence
WHERE {
  ?step vg:path path:GRCh37.p13/Chr11;
     vg:node ?node ;
     vg:position ?position ;
     vg:rank ?rank.
  ?node rdf:value ?sequence .
  FILTER (?position > 44067802 && ?position < 44067820)
  ?node vg:linksForwardToForward|^vg:linksForwardToForward ?near_node .
}
```

図 2 ヒトゲノムグラフの SPARQL 検索例

これをもとに、2016 年には UCSC を訪問し、Benedict Paten 氏、Adam Novak 氏、David Steinberg 氏らと、ゲノムグラフにおけるセマンティック・ウェブ技術の応用と、その GA4GH における利用について打ち合わせを行った。この時点では、セマンティック・ウェブ技術が、当時 GA4GH の Genomics API 標準化における様々な課題を解決するポテンシャルを持つことが確認でき、2015 年に作成したプロトタイプを改良して提案を行っていくための協力を進めていくことになった。

しかし、2017 年から当面の GA4GH の開発体制が短期的な目標の実現にシフトされた結果、ゲノムグラフの開発はいったん GA4GH の活動とは切り離されることになった。このため、当初期待していた Genomic API でのゲノムグラフの採用と、その際にセマンティック・ウェブ技術を利用した実装を行っていくという研究プランは変更を余儀なくされた。一方で、UCSC でゲノムグラフの研究リーダーを務める David Hausslar 氏に問い合わせたところ、Hausslar 氏をはじめゲノムグラフの開発メンバーは継続して GA4GH と関わりを持ち続けるとともに、ゲノムグラフ自体の開発は研究としてこれまで以上に取り組んでいくということが示された。

並行して、2016 年の RDF summit2 の前後より、東北大学東北メディカル・メガバンク機構との共同研究において日本人ゲノムの解析をゲノムグラフを用いて行うための検討を開始していたため、ゲノムグラフの中に現在標準のリファレンスゲノムの他に日本人の標準ゲノムを融合し、さらに個人ゲノムにおける変異を統合する方法を確立する必要があった。

このため、2017 年に Sanger Institute を訪問し、Richard Durbin 氏、Erik Garrison 氏と打ち合わせを行った。この結果、ヒトゲノムの染色体レベルでのゲノムグラフ構築にはスケーラビリティにおいて実装上の課題が残っていることが明らかとなってきたほか、将来的に必要となるハプロタイプの扱いについてもまだ開発途上であることが分かってきた。一方で、検証実験としてヒト遺伝子の複数のトランスクリプトとチンパンジーにおけるオーソログ遺伝子を含むグラフの構築を行い、ここに 1000 人ゲノムのデータからゲノム変異情報を取り込むことは実現できた。これをもとに、vg ツールの使い方について Garrison 氏とともにチュートリアルを作成した。

2017 年後半からは、ゲノムグラフに関心のある国内の若手研究者との研究ミーティングが始まり、がんゲノムなどの構造変異をゲノムグラフで扱う方法とその可視化技術の開発や、メタゲノムをゲノムグラフを用いて解析する技術の開発に取り組んでいる。

2018 年に入って、この活動はゲノムグラフ研究会の設立に繋がった。また京都大学との連携で、ゲノム変異のセマンティック・ウェブにおける標準化を検討するハッカソン RDF summit3 を開催した。この中で、ゲノムグラ

フ研究会のメンバーおよび、Erik Garrison 氏、David Steinberg 氏、Jerven Bolleman 氏らとともにゲノムグラフ技術の開発と RDF によるゲノムアノテーション情報の融合、可視化に取り組んだ。

特に、大腸菌 K-12 株のゲノムアノテーションを、アノテーションが不足している O157 株へ適用する、アノテーショントランスファーを、ゲノムグラフとセマンティック・ウェブ技術の組み合わせで行えることを示せたことで、本研究の目的を実現することができた。ここではまず、大腸菌 2 ストレインのゲノム配列をパスとして持つグラフを vg msga により構築し、そこに BED 形式の遺伝子座標を vg annotate によって対応させ GAM ファイルを作成、この GAM ファイルを vg mod によってゲノムグラフに取り込んで vg index によりインデックスを作成した。遺伝子座標のマッピングを 2 ストレイン分適用し、完成したゲノムグラフを vg view -t で RDF 形式に出力した。次に、Ensembl の遺伝子アノテーション情報の RDF を取得し、ゲノムグラフの RDF とともに RDF データベース Virtuoso にインポートした。これをもとに、ゲノムグラフ上で共通のノードを持つ 2 ストレインの各トランスクリプトについて、それぞれアノテーション情報を取得し相互に補完する SPARQL 検索例を作成することができた(図 3)。

```
PREFIX vg: <http://example.org/vg/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterm: <http://purl.org/dc/terms/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
prefix up: <http://purl.uniprot.org/core/>

SELECT
  DISTINCT ?transcript1 ?uniprot1 ?comment1 ?transcript2 ?uniprot2 ?comment2
WHERE
{
  ?path1 skos:closeMatch ?transcript1 ;
    ^vg:path ?step1 .
  ?step1 vg:node ?node  .
  ?transcript1 a obo:SO_0000234 .
  ?path2 skos:closeMatch ?transcript2 ;
    ^vg:path ?step2 .
  ?step2 vg:node ?node  .
  ?transcript2 a obo:SO_0000234 .
  FILTER (! sameTerm(?path1, ?path2))
  OPTIONAL {
    ?uniprot1 rdfs:seeAlso ?transcript1 ;
      up:annotation ?annotation1 .
    ?annotation1 a up:Function_Annotation ;
      rdfs:comment ?comment1 .
  }
  OPTIONAL {
    ?uniprot2 rdfs:seeAlso ?transcript2 ;
      up:annotation ?annotation2 .
    ?annotation2 a up:Function_Annotation ;
      rdfs:comment ?comment2 .
  }
}
```

図 3 ゲノムグラフとアノテーションの検索

本研究により、当初目的としていたゲノムグラフとアノテーションなど知識のグラフを、セマンティック・ウェブの技術を用いて RDF において標準化・統合し、利用できることを示せた。一方で、ヒトゲノム等大規模なゲノムへの応用においては、ゲノムグラフを構築する vg ツールの高性能化がまだ必要とされるほか、ハプロタイプを扱う解析に対応させていくなどの課題が残されていることが分かった。これらは現在も vg 開発チームによって日々対応が進んでいる状況であり、今回設立したゲノムグラフ研究会のメンバーらとともに進展を見守りつつ応用研究を進めていきたいと考えている。

５．主な発表論文等

〔雑誌論文〕（計 1 件）

1. 片山俊明
BioHackathon 2017 報告、情報管理、査読無、Vol.60、2018、pp.744-747
DOI:10.1241/johokanri.60.744

〔学会発表〕（計 6 件）

1. Toshiaki Katayama
Standardization and utilization of pathogenic variant data、3rd International Symposium on BioComplexity、国際学会、2018
2. Toshiaki Katayama, Shuichi Kawashima, Yasunori Yamamoto
Semantic Web technology accelerates integration of genetic and phenotypic information in biomedical databases 、American Society of Human Genetics、国際学会、2017
3. Toshiaki Katayama
Underlying technologies for integration, sharing, and analysis of data in life sciences 、 Advanced Genome Science International Symposium、招待講演、国際学会、2017
4. 片山俊明
生命医科学データの RDF 化の現状と課題
第 6 回生命医薬情報学連合大会、2017
5. 片山俊明
生命医科学 RDF データの機械学習・人工知能への応用、第 31 回人工知能学会全国大会、2017
6. 片山俊明
リファレンスのスタンダードはグラフになるか！？、第 5 回 NGS 現場の会、2017

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
ゲノムグラフ研究会
http://genomegraph.jp/

６．研究組織
(1)研究代表者
片山 俊明 (KATAYAMA, Toshiaki)
大学共同利用機関法人　情報・システム研究機構、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：60396869

(2)研究分担者
なし
(　　)

研究者番号：

(3)連携研究者
なし
(　　)

研究者番号：

(4)研究協力者
なし
(　　)