1N5-OS-10c-3

ヒト遺伝子データベース H-InvDB の RDF 化と Endpoint の公開

RDFization of human gene database H-InvDB and its SPARQL Endpoint

村上 勝彦[*1] Katsuhiko Murakami
山崎 千里[*1] Chisato Yamasaki
今西 規[*1, 2] Tadashi Imanishi

[*1] 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology (AIST)
[*2] 東海大学 Tokai University

The semantic web technology is prevailing in ife science rapidly. H-Invitational Database (H-InvDB) is a public integrated database of human genes, transcripts, and proteins, providing the curated annotations. Here we prepared H-InvDB dataset in RDF scheme, and constructed a SPARQL Endpoint server. We employed several widely used ontologies, such as Gene Ontology (GO) and Sequence Ontology (SO). Since H-InvDB include our original ortholog gene set for the 13 other species ranging from Chimpanzee to Fugu, the server provides hub-like functionality for a cross-species search, giving the wider applications in researches. H-InvDB SPARQL Endpoint is freely available at http://h-invitational.jp/sparql/hinv/.

1. はじめに

ライフサイエンスにおいては，総合的に知識をつなげていくことでユーザー自身のデータと比較したり，重要な知見が得られたりして，発見のためのアイデアを得ることが多い．このためにデータベースを連携させて使用できることが生命科学系では特に重要である．さまざまなプロジェクトによってデータが生成，公開されているが，これらを統合するための技術としてセマンティック Web が注目されている[山口 11]．現在稼働中のライフサイエンス分野のデータベースは、名前が判明しているだけでも国内外で 1,126 件存在する（http://integbio.jp/dbcatalog/）．しかし、セマンティック Web に対応しているデータベースはまだ非常に少なく，その数を増加させることが重要である．

その中でも初期に RDF データを公開しているものでは、UniProt [Uniprot 13] (http://beta.sparql.uniprot.org/) が有名である．これは単一データベースであるが，一方で複数の独立なデータベースを集めて RDF 化したものとして Bio2RDF [Belleau 08] (http://bio2rdf.org/) や NeuroCommons [Ruttenberg 09] (http://neurocommons.org/page/Main_Page) が挙げられる．他にもライフサイエンス系データベースで，RDF データを公開するようになったものは少しずつ増加しており，新旧の Linked Data は相互作用的にその価値を高めている．

産業技術総合研究所の創薬分子プロファイリング研究センター（旧バイオメディシナル情報研究センター）では，ヒト遺伝子データベース H-Invitational database （H-InvDB, http://h-invitational.jp/）を構築している [Imanishi 04]. H-InvDB は，ヒトのすべての遺伝子・転写物・タンパク質の分子配列情報を様々な計算機的手法で解析した結果を格納したものである．この全データは XML 形式とフラットファイル形式でダウンロード可能である（http://h-invitational.jp/hinv/dataset/download.cgi）．

今回、XML データをもとにして、H-InvDB のコンテンツの RDF データを生成し、RDF ストアに格納して SPARQL Endpoint として公開した (http://h-invitational.jp/sparql/hinv/)．本稿では、その内容と意義について報告する．

連絡先：村上勝彦，産業技術総合研究所 創薬分子プロファイリング研究センター，135−0064 東京都江東区青海二丁目4番7号，k−murakami@aist.go.jp，aaaccc.k@gmail.com

2. ヒト遺伝子データベース H-InvDB

2.1 H-InvDB とは

H-InvDB はヒトの全「遺伝子」について，遺伝子の転写物という分子の情報を主体にし，その他さまざまな情報をのせたデータベースであり，DNA 配列構造，選択的スプライシングバリアント，機能性 RNA，タンパク質機能，機能ドメイン，細胞内局在，代謝経路，立体構造，疾病との関連，遺伝子多型（SNP，マイクロサテライト等），遺伝子発現プロファイル，分子進化学的特徴，タンパク質間相互作用(PPI)，および遺伝子ファミリーなどの情報を提供し，これまで毎年更新している．

2.2 Linked Data となる意義

H-InvDB のデータが Linked Data となる意義，期待できる点は何だろうか．生物学的な意義と情報科学的な意義をそれぞれあげてみよう．生物学的には以下の2つの点がわかりやすい．1つめは H-InvDB のさまざまな独自データが RDF で扱えることである．例えば，遺伝子の状況証拠はあるけれども確定的な機能が未だ確認されてないゲノム上の領域，すなわち新規遺伝子候補といえる「仮説的タンパク質 (hypothetical proteins) 」カテゴリのデータが 2012 年の H-InvDB (Release 8.0) には 13,320 件含まれている．過去のバージョンもあわせると 19,309 件の候補が同定されていたが，そのうち 233 件は後になってタンパク質であることが実際に判明している[Takeda 12]. Linked Data の枠組みに載せることでリンクする他のデータ（状況証拠）が増える．すると，「仮説的タンパク質」が特定の機能をもつかどうかや，その種類を判断しやすくなる．そのため，より適切な候補を選定して確認実験を行うことが出来，その結果，機能をもつ「遺伝子」の発見が促進されることが期待される．

2 つ目の意義として，H-InvDB は，ヒトとその他の 14 脊椎動物（チンパンジー，オランウータン，マカクザル，マウス，ラット，イヌ，ウマ，ウシ，オポッサム，ニワトリ，ゼブラフィッシュ，メダカ，ミドリフグ，トラフグ）の間の分子進化情報を独自に解析して保持している．このため，マウスなど実験動物との遺伝子情報をつなげて比較できる．ここでいう分子進化情報とは，ヒトのある遺伝子に相当する遺伝子は，別の生物では進化上どの遺伝子かと

いう情報であり，この対の遺伝子は互いのオーソログ（オルソログ，ortholog）といわれる．オーソログ遺伝子のペアは，種が異なっても基本的に同様の機能を果たしていると考えられている．オーソログデータは姉妹データベースである Evola [Matsuya 08] から提供されている．他にオーソログを提供しているリソースには先に紹介した Bio2RDF があるが，マカクザル，イヌ，ウマ，ウシ，オポッサム，メダカ，テトラオドン，およびフグの 8 種のオーソログ遺伝子セットが提供されていない．これらは H-InvDB にしかないものである．

また情報科学的には，遺伝子特有のデータ構造とその規模である．遺伝子のデータベースにおいての特徴的かつ大量にあるデータのタイプは，塩基・アミノ酸残基の配列である．そのため，この一次元の文字列データを扱えることが重要であるが，これが少なくとも転写物配列のレベルで可能となる．分子配列を扱う場合に，セマンティック Web や RDF の枠組みが，どこまで有効利用できるのかは，テキストや Web 上の音声などと大きく異なる点であり興味深い問題であろう．また，H-InvDB 自体には大量(>1TB)というほどのデータはないが，例えばヒト遺伝子情報が解析にかかせない 1000 人ゲノムプロジェクト (http://www.1000genomes.org/) のゲノムデータは 200TB という大規模なデータであり，これをどこまでどのように取り込めるかということも含め，合わせて考えたい課題である．

このように H-InvDB の SPARQL Endpoint の公開は、Linked Data の世界を広げる様々なポテンシャルを持っており，統合的な検索を利用した高度な解析によって生命科学の研究期待できる．

3. RDF 化とシステム構成

3.1 RDF 化とオントロジー

H-InvDB は遺伝子，転写物，タンパク質をそれぞれ主キーとした関係データを保持している．今回はこのうち最もアノテーション情報の多い「転写物」を中心に RDF 化した．H-InvDB にあるダウンロード可能なデータは遺伝子，転写物，タンパク質についてそれぞれ XML で提供されている．このためのデータはファイル名が”XCDNA_[1-5]”（正規表現）という5つのファイルで，XML で書かれたもの（Release 8.3）を利用した．ここでは 207 個の要素(element)が定義されている．RDF 化を行うだけでも恩恵が多いことから，まずはこれらの XML 要素名をそのまま用いて DTD に書かれた 207 個の要素をすべて RDF/XML へ変換した．転写物（主語）244,619 個についての総トリプル数は，84,325,293 個となった．ファイル数は 6.9GB であった．

オントロジー作成については，H-InvDB の特徴を考慮する必要がある．リソースをオントロジーで記述するとき，分子を扱うデータベースの場合，多くの目的語(object)は ID で記述される（データベースのみで使用される用語が ID で管理されたもの，例えば Gene Ontology タームも含む）か，自然言語のフレーズや名前を示す記号が列挙されることが多い．このうち，前者（データ ID）の場合，外部に URI がすでに提供されていたり Bio2RDF で扱えるものについてはその ontology を利用して外部データとの連携を図った．一方，後者の自然言語による表現の場合は全タームをいきなり既存オントロジーで表現するのは調査が困難なので，最初のステップとしてリテラル（定数）または独自リソースとして扱い，次のステップとして世の中でオントロジーが利用できるものを選定しながら，既存オントロジーとの共通化を図ることにした．

遺伝子のデータベースにおいて、特徴的なデータは「塩基配列」である．そこで配列を扱うためのオントロジーを利用することは H-InvDB のコンテンツを有効利用するために欠かせない．このためのオントロジーとしては，Sequence Ontology (SO, http://http://www.sequenceontology.org/) が有名であるため H-InvDB の一部であるが SO に対応した．図1は，生体内における遺伝子，転写物，タンパク質への情報の流れをしめしたものである．遺伝子解析では，どの領域の文字情報が別分子に伝搬したかを扱うことがよくある．例えば，ゲノム配列のある位置の塩基に変異があった場合に，それがタンパク質の機能に及ぼす影響は何かを知るための検索を行ったり，またはその逆に，あるタンパク質の何残基目が何かに変わるような DNA 変異が何かを知るための検索を行うためのものである．

H-InvDB SPARQL Endpoint のオントロジーは，開始したばかりで，まだまだ発展途中である．H-InvDB データを RDF 化した利点を生かすためにも今後益々充実させていく予定である

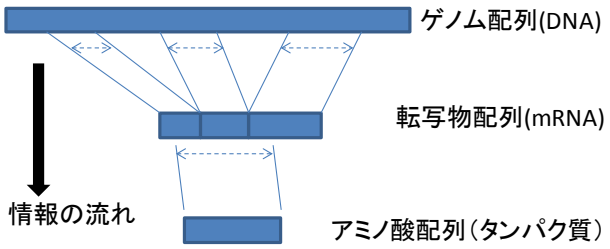

図1 H-InvDB で中心的に扱う生体分子と情報伝搬における上流下流の関係．長方形は1本鎖の分子を示す．各分子は 4 文字種（A, C, G, T)または 20 文字種の文字列データで表現される．棒線で囲まれた点線矢印の領域は，情報が伝搬する領域とその対応を示す．ゲノムと転写物では一般に飛び飛びの領域が対応する．データとして，何番目から何番目の文字までの領域の情報が伝わるか（配列上の位置）を表現・検索できる必要がある．

3.2 RDF ストアシステム

構築した RDF/XML に変換したデータを、RDF ストアに格納した．RDF ストアとしては Sesame version 2.6.9 (http://www.openrdf.org/) に bigdata version 1.2.1 (http://www.bigdata.com/) を結合した形を採用した．bigdata を採用した理由は，将来的なデータの拡張に備えるためである．構築したサーバーは，http://h-invitational.jp/sparql/hinv/ からアクセス可能である．

4. おわりに

ヒト遺伝子データベース H-InvDB のアノテーションデータを RDF 化し，遺伝子関連オントロジーを作成した．また，SPARQL 検索を可能とするため Endpoint の公開を行った．ヒト遺伝子・タンパク質を中心に様々なアノテーションを豊富に持つ H-InvDB のデータを UniProt 等外部のデータベースと合わせるような複合的な探索が Linked Data として可能となった．今後は，特にオントロジーについて拡張していく．さらに H-InvDB には、ヒトと 14 脊椎動物種にわたるモデル生物間のオルソログ・データがあ

り、これによって実験動物などのデータとヒト疾患情報をつなげることが出来る. このように Endpoint の公開により, H-InvDB の特徴的なデータを利用できるほか, H-InvDB を 1 つのハブとして外部データベースと連携させることができ, Linked Data の世界において一粒万倍の効果が期待出来る.

参考文献

[Uniprot 13] The Uniprot Consortium: Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic acids research, Vol. 41, pp.43-47. 2013

[Belleau 08] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P. and Morissette, J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics, Vol. **41**, pp.706-716. 2008

[Ruttenberg 09] Ruttenberg, A., Rees, J.A., Samwald, M. and Marshall, M.S.: Life sciences on the Semantic Web: the Neurocommons and beyond. Briefings in bioinformatics, Vol. 10, pp.193-204. 2009

[山口 11] 山口敦子, 片山俊明: データベース統合利用基盤としてのセマンティックウェブ技術, 細胞工学、学研メディカル秀潤社, Vol. 30, No. 11, pp.1210-2011. 2011

[Imanishi 04] Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M. et al.: Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS biology, Vol. 2, No. 6, pp.e162. 2004

[Takeda 12] Takeda, J., Yamasaki, C., Murakami, K., Nagai, Y., Sera, M., Hara, Y., Obi, N., Habara, T., Gojobori, T. and Imanishi, T.: H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. Nucleic acids research, Vol. 41, pp.D915-919. 2013

[Matsuya 08] Matsuya, A., Sakate, R., Kawahara, Y., Koyanagi, K.O., Sato, Y., Fujii, Y., Yamasaki, C., Habara, T., Nakaoka, H., Todokoro, F. et al.: Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. Nucleic Acids Res, Vol. 36, pp.D787-792. 2008

[Whetzel 11] Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T. and Musen, M.A.: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research, Vol. 39, pp.W541-545. 2011