

多領域生物情報リソースの 遺伝子集約型モデルによる統合

政策・メディア研究科 M2

大下 和希

要旨

バイオインフォマティクス分野では数多くのデータベースや解析Webサービスがオンラインで公開されており、多くの研究者がそれらWebリソースから生物学リソースを取得し解析を行っている。これらのリソースを用いてより効率的な解析を行うため、解析Webサービスの連携による複雑かつ高度な解析フローの構築や、多領域生物学データベースおよびWebサービスの効率的な統合と運用を行うシステムの構築が求められてきた。そのため、本論文では解析・データアクセスWebサービス群と各種データベースを対象に、それぞれを効率的に統合し運用することを目的としたシステム的设计・構築を行った。G-Linksは生物学Webリソースを効率的に統合し、そこからユーザが必要な生物学データセットを高速かつ自動的に抽出するシステムである。G-Linksでは多領域生物学情報に対して遺伝子集約型のデータ統合モデルとID変換をベースとした統合を行っており、URLにアクセスするだけでユーザが対象とする遺伝子に関する生物学情報セットを高速に収集し、得られた情報セットからユーザが必要な情報だけを抽出、任意のフォーマットへ変換というプロセスを高速かつ自動で行うことができる。本システムは <http://link.g-language.org/> より利用できる。これらのデータ統合プラットフォームを用いることで、研究者は多領域に渡る大量の生物学Webリソースから、生命システムに関する知識をより効率的に導出することが可能となる。

1 序論

1.1 バイオインフォマティクスにおける Web リソース

DNA およびタンパク質の最初期のデータベースが世に公開されて以来 (Dayhoff *et al.*, 1976), バイオインフォマティクスにおけるデータベースは急速な発展を遂げている。次世代シーケンサに代表される分子レベルの実験技術の飛躍的向上は、研究者が得る事の出来るデータ量や研究対象とする事が出来るデータの種類の増加などをもたらしており、それに伴う形での生物学データベースの数、扱うデータの種類の増加、および内包するコンテンツのデータ量の増加が著しい。これらの生物学データベースの多くは Web 上にフリーで公開されており、研究者はそのデータ群を自由に用いてより大規模かつ複雑な研究解析を行うことが可能である。多領域かつ複雑な生命現象を大きな一つのシステムとみなし理解しようとするシステムバイオロジーでは、そのシステムを構成する遺伝子およびタンパク質などの翻訳産物に代表される分子情報や、それらの機能および相互作用といった機能アノテーションの統合が重要な課題の一つとされている (van den Berg *et al.*, 2010)。しかしながらこの生物学データベースにおける爆発的なデータ量の増加は、研究者にメリットと共に運用コストというデメリットをもたらしている。この肥大化したデータリソースを効率的に扱う有効なアプローチの一つがデータベース検索ツール Application Programming Interface (API) である。ユーザのクエリを解釈してそれに適した結果を抽出し高速で取得することができる検索 API は、メンテナンスやセットアップコストが不要という利点も併せ持つ。これらの理由から生命情報解析のための Web サービスが数多く存在することもバイオインフォマティクス分野の特徴の一つである。

上記の理由からバイオインフォマティクス分野では数千の生物学データベース (Fernandez-Suarez and Galperin, 2013) や 1200 を超える解析 Web サービス (Brazas *et al.*, 2012) が Web 上でオープンに提供されており (Bhagat *et al.*, 2010), それらを組み合わせることでより複雑な解析を行うことができる。しかしながら、複数のデータベースに分散して存在する生物学的データの爆発的増加に伴って、このデータ統合プロセスにおける労力の増加が研究者にとってのネックとなっている。バイオインフォマティクス研究ではその作業のほとんどが 1. 研究対象に関連する大量のエントリーを複数の生物学データベースから収集し、2. そこから得られたエントリーを統合し、3. その大量のデータから研究者が必要とするデータだけを抽出する、という 3つの作業に湿られている。さらに近年の解析 Web サービスの台東により、Web サービスによる解析結果もデータベースと同じく Uniform Resource Identifier (URI) にて指定可能な生物学リソースの一つとしてみなすことが出来る。真に生物学情報を統合するにはデータベースと合わせて生物学 Web リソース全体をシームレスに統合し、効率的に運用するためのプラットフォームの開発が必要不可欠である (Stein, 2002, 2008)。

1.2 データベースの統合的利用

生物学データベースの単純統合にはデータ量とスキーマ定義という大きな問題が存在する。データ量と種類の爆発的増加は巨大データアーカイブに対する検索や閲覧など再利用性確保のための膨大な計算資源を要求する他、生物学で扱われるデータの種類が増加する度にデータベース全体のスキーマを変更し更新する必要がある。これらの問題を解決するため生物学ではこれまで様々なアプローチがとられてきた。複数のデータベースの検索ツールによる結果を統合する Federated Query (Jacso, 2004) 型データ統合は主に SOAP などの検索ツール Web API を用いたサービス統合による問題解決を目指しており (Wilkinson *et al.*, 2003), BioMoby (Wilkinson *et al.*, 2008) や myGrid プロジェクトに代表され

る生物情報解析 Web サービスの連携による解析フロー構築の研究へと発展している。ユーザが必要なデータベースだけを単一システムに落とし込んだ統合型データベース構築のアプローチの筆頭である BioMoart (Kasprzyk, 2011) は複数のデータセットを一つのスキーマにまとめる作業を支援することで、複数のデータベースから自身の用途にあったリソースのスライスを容易に取り出すことができる。

1.3 ID 変換によるアプローチ

この生物学データ統合問題におけるもう一つの主要なアプローチが ID 変換である。多くの生物学データベースはそれぞれのエンタリー間の Link によってデータベース間の関係性を表現する Linked Data モデルであり、ユーザはハイパーリンクを辿るだけでそのリソースに関連するリソースを収集できる。データベースには複数のデータ群について関係性の情報を管理することでより複雑なデータ構造を表現する Relational Database (RDB) (Codd, 1969) というアーキテクチャが存在するが、Linked Data モデルでは新規概念に対応したデータベースに Link を張るだけでスキーマの変化に対応できる。さらに Link によるデータベース間の関係性抽出は各エンタリーを示す ID とそれに関連する ID の変換作業と同値である。このため、Linked Data による関連性ネットワークを用いて ID 変換を行い、複数のデータリソースから特定の生物学オブジェクトに関連する ID を横断的に収集することで生物学リソースの擬似的統合が可能となる。

この ID 変換システムを構築する上で問題点とされてきたのが、異なる種類のデータベースを統合する際のスキーマの問題とネットワークの大規模化に伴うレイテンシである。遺伝子情報に特化した SOURCE (Diehn *et al.*, 2003) やタンパク質情報に特化した Protein Identifier Cross-Referencing (PICR) (Cote *et al.*, 2007) は遺伝子やタンパク質など基準をおいた ID 整理を行うことでスリム化された高速なシステムとして動作する。bioDBnet (Mudumuri *et al.*, 2009) はユーザから受け取った ID の解決部分に関連データ取得部分と切り離し、ID の Link ネットワークのみ抽出したスリムなデータベースを構築する事で横断検索部分の高速化を実現している。

このように ID 変換では各エンタリーを示すポインタとその間の Link のみを取り扱うため、データアーカイブの全統合と比較してデータベースの高速な統合的利用が可能である。しかしながら ID 変換によって得られるデータは ID のリストであり、実際に生物情報解析を行う際はその ID 群が指し示すリソース群を別途取得し統合する必要がある。また、Link は「関連している」という状態は容易に表現できる一方でその Link が持つ意味を表現できないため、自動処理を行う場合は大量に集まった Link 情報からユーザが必要とする Link だけを選別する必要がある。

1.4 Semantic Web

これらの問題の解決策として現在着目されているのが Tim Berners-Lee によって提唱された World Wide Web (WWW) の利便性を向上するためのプロジェクト、Semantic Web である。Semantic Web ではリソース内に含まれる個々のオブジェクトにまで URI を割り振り、そのリソース自体や Link のセマンティクス自体を Web Ontology Language (OWL) によって記述する。このように意味情報の形式化を行うことで、WWW の全てのドキュメントに対する意味情報を加味した自動的な情報収集や分析が可能になる。また、Semantic Web では Resource Description Framework (RDF) にて全てのリソース関係グラフを直接記述するため、テーブル型でないスキーマレスなフォーマットでデータを管理できる。しかしながら Semantic Web には、リソース細分化による Link ネットワークの複雑化とそれを扱う計算資源の問題や、RDF の生成に必要な労力の高さ、意味情報を表現する語彙集であるオントロジーの統一化の必要性などの大きな問題が存在する。そのため、Semantic Web の技術をベースとした統合データベースで実用段階にあるプロジェクトは生物学では未だ数えるほどしか存在しない。

2 要求分析

本論文ではこれらのデータ統合の問題を解決するために、バイオインフォマティクス研究の作業の大半を占める以下のデータ統合プロセスを自動的かつ効率的に行うシステムの構築を行った。

- 多数の生物学データベースや Web サービスから得られるデータの統合
- 研究者が対象とする生命現象に関する情報の網羅的な取得
- 実際の解析で利用するデータの抽出

このシステムを構築する上で非常に大きな問題が生物学情報の領域の多様性である。バイオインフォマティクス研究では生命システムの複雑さ故に多領域に渡るデータを用いて多方面からのアプローチを採る必要があるが、表現するデータの増加によるデータモデルの複雑化は生物学リソースの統合を非常に難しくしていた。これに対抗する形で生まれたのが Link を張るだけでデータベース間の関係性を表現する Linked Data モデルと ID 変換のアプローチである。本システムではレイテンシの問題の解決や、密な Linked Data ネットワークを構築しているという生物学データベースの特徴などから ID 変換をベースにしたシステムを構築を行った。このシステムを構築を行うにあたって、第一に本システムを実現するにあたって要求される要素についての分析を行った。

- 出力可能な情報の網羅性
対象の生命現象に関連する多領域に渡る情報を効率的に統合し解析作業を行う必要があるため、研究者が入力したクエリに対して、関連する生物学情報を広い範囲から網羅的に取得できる必要がある。
- 汎用的な入力系
より利便性の高いリソース取得を行うためには、ユーザがどのような形の入力を行ったとしてもその入力に対して適切な生物学データセットを出力する必要がある。
- ID の持つロケーション問題の解決
ID 変換をベースとした本アプローチにおいても結果として ID をユーザに提供するだけでなく、その ID が示すリソースもしくはそれに対応した URI をユーザに提供する必要がある。
- ID 情報以外のリソースの取得
より利便性の高い生物学データセットの生成を行うためには、ID 変換を用いたリソース間の関連情報の解決を行った上で、その ID から取得することができるリソースまで含めた状態でユーザに提供できる必要がある。
- リソースの厳選
研究者がより正確な解析を行うためには、情報量の高いリソースだけを統合することでこれらのノイズ情報を除去し、かつそこから研究者が必要な情報だけを抽出できるシステムを実装することで、バイオインフォマティクス解析においてより価値の高い生物学データセットを取得できる必要がある。

- データ統合から抽出までのプロセスの自動化と高速化
上記の統合・取得・抽出というバイオインフォマティクス分野において作業の大半を占めるプロセスについて、この大きな労力が必要な作業を自動的かつ高速に行うことができるシステムである必要がある。
- 他サービスとの相互運用性
本システムで得られた出力は様々な環境やプログラミング言語から容易に利用でき、かつ既存ソフトウェアや各種技術とシームレスに連携できる必要がある。

3 設計と実装

3.1 アーキテクチャ

G-Links は、生物学の多領域に渡るリソースを高速かつ網羅的、自動的に収集するためのゲートウェイサーバである。多数の生物学データベースに対して ID 変換を用いることでデータを収集し、ユーザのクエリに関連する分子情報や機能性アノテーションを高速かつ自動的に提供する。汎用的な生物学情報サポートによるレイテンシの問題に対して G-Links では Primary Key を設定し、Linked Data ネットワークを整理することで解決を試みた。Primary Key の選定において、全ての遺伝情報は遺伝子から伝播するというセントラルドグマの考え方から、全ての生物学的情報は遺伝子を中心に統合できると考え、多数の遺伝子 ID の中から、UniProt ID を採用した。UniProt はタンパク質をコーディングしている遺伝子を中心としたデータ構造で (The UniProt Consortium, 2012)、非常に品質の高く Linked Data ネットワークにおいてハブになりうる数のクロスリファレンスを持つ。Link 情報を用いた ID mapping サービス (Huang *et al.*, 2011) を提供しているなど Primary Key として非常に理想的である。G-Links では遺伝子を表すクエリを入力として想定している。内部データベースは bioDBnet と同様に、ユーザからのクエリを UniProt ID に変換する ID 解決部と、その UniProt ID に関連するアノテーションの取得部という 2 種類のテーブルを使用することで高速化を行った。本システムのメイン部分および内部データベースの更新用スクリプトは Perl 言語で構築されており、各データベースでは MySQL 5.0 を用いた RDB を利用している。内部データベースは UniProt の更新頻度と同じく毎月 1 回の更新作業が行われる。G-Links のアーキテクチャ図を図 3.1 に示す。

3.2 ユーザクエリの ID 解決

クエリ解決部に求められるのがユーザの入力に対する汎用性である。G-Links では遺伝子を表す ID に対する単純な ID 変換のアプローチだけではなく、遺伝子セットを示す ID の入力や塩基/アミノ酸配列に対する配列類似性検索による ID マッピングという 3 種類の入力に対応した。ID 変換については、UniProt が提供している ID 変換サービス用データセットをベースに独自の拡張を加えて作成した。KEGG Orthology に代表されるような遺伝子セットの ID が入力された場合はその遺伝子セットに対応する UniProt ID 群を検索し、得られた UniProt ID 群全てについて関連する生物学情報を提供する。カンマ区切りによって複数の遺伝子 ID を渡された場合にも同様である。また、生物種を示す ID を入力した場合にはその生物種の持つ遺伝子を示す UniProt ID のセットへと変換する。対応表の元データは UniProt が提供する Taxonomy Search (<http://www.uniprot.org/taxonomy/>) を用いている。入力として扱える生物種 ID としては NCBI Taxonomy (Federhen, 2012) および RefSeq (Pruitt

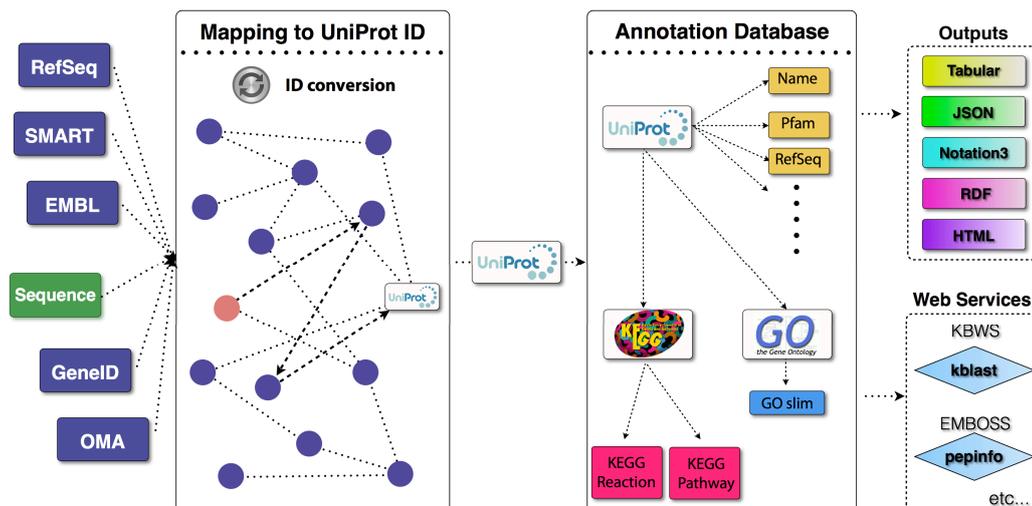


図 3.1: G-Links 全体のアーキテクチャ図

G-Links では遺伝子を示す ID および配列情報をユーザからクエリとして受け取り、それを UniProt ID へと ID 変換および配列類似性検索を用いて変換する。その後、当該 UniProt ID に関連する他データベースの ID 情報、クロスリファレンスおよびそこから取得したリソース、外部 Web サービスの解析結果を示す URL などを含んだ結果をユーザに任意のフォーマットにて提供する。

et al., 2012) をサポートしている。また、G-Links では配列類似性検索を用いてユーザからの入力された配列情報の UniProt ID への変換を行う。配列類似性検索がもたらすレイテンシ問題への対策として、ユーザから入力された配列が塩基配列だった場合は European Molecular Biology Open Software Suite (EMBOSS) (Rice *et al.*, 2000) の transeq を用いてアミノ酸配列へ翻訳を行い、BLAST Like Alignment Tool (BLAT) (Kent, 2002) による類似性検索を Swiss-Prot をターゲットとして行う。塩基配列をアミノ酸配列に変換する際はフレームずれの可能性を考慮し、翻訳開始点を +0, +1, +2 した 3 パターンについて、Watson 鎖と Click 鎖両方に遺伝子がコードされていることを想定した計 6 パターンのアミノ酸配列へ変換を行い、全てをクエリとして配列類似性検索を行っている。また、より精度の高い変換を行うため BLAT を行う際の E-value と Identity の 閾値の初期値を高く設定することで類似性検索をできるだけ ID 変換の精度に近づけている。さらに確実な変換を行うため、G-Links に配列情報を与えた場合、ユーザは候補となる UniProt ID とともに E-value や Identity, 生物種名や遺伝子の名前、その UniProt ID を入力とした G-Links の結果 URL のテーブルを得ることができる。その結果から正しい UniProt ID をユーザ自身が選択することで、より正確な ID 変換を実現している。

3.3 アノテーション

ID 変換によって得られた UniProt ID に関連するアノテーション情報を収集するため、G-Links では UniProt ID に紐付けされた外部データベースの ID リストを内部データベースから取得する。ここで用いている内部データベースは UniProt の情報をベースに Link を辿ることで拡張を行う他、GO slim (Harris *et al.*, 2004) のような事前計算が必要なリソースに関しても予め計算を行うことで取得している。さらに G-Links では ID 情報のみならず、その遺伝子に関連するドメインや組織に関する情報など「人が読むための情報」も保存されている。これらの情報もユーザに提供することで、ID 情報だけでは理解できないその遺伝子に関する知識を容易に取得することができる。このテーブルは UniProt ID を主キーとした転置インデックスによってデータを格納しているためスケーラビリティが高い設計となっている。生物種を示す ID がクエリであった場合は大量の遺伝子についての処理が必要があるが、生物

種に対するクエリについては Perl の Storable モジュールでシリアライズされたキャッシュを事前生成することで高速な処理を実現している。

3.4 アウトプット

G-Links では、ユーザから与えられた遺伝子および遺伝子セットに関連するアノテーション情報を収集した後、それらをユーザに対して利便性の高い形で出力を行う。G-Links が出力する全てのリソースは RESTful に一意の URL で指定することが可能であり、その出力結果を既存技術と容易に連携することが可能である。また、どのフォーマットであっても ID 情報とその ID が利用できるデータベース名、その ID が示すリソースを指し示す URL の 3 情報を基本的に含んでいる。

G-Links は出力データフォーマットとして、Programmable なフォーマット、研究者が読むことを前提とした Human-Readable なフォーマット、Semantic Web 上で利用するためのフォーマットの 3 種類への対応を行なっている。Programmable 出力フォーマットとしては JSON と Tabular のサポートを行なっている。Human-Readable である HTML 出力はブラウザから 1 クエリに対する情報を人が閲覧するために利用されることを想定しており、ID 情報や UniProt などに登録されている記述情報だけでなく、KEGG Pathway のパスウェイマップや COXPRESdb (Obayashi *et al.*, 2013) の共発現遺伝子ネットワーク図などの画像情報をユーザに提供する。この画像情報の表示は PHP: Hypertext Preprocessor (PHP) にて実装されたスライドギャラリー ImageFlow (<http://imageflow.finnrudolph.de/>) にて行われており、記述情報と ID 情報は JavaScript にて実装された tableorter (<http://tableorter.com/docs/>) によって各カラムが自由に並び替え可能なテーブルとして表現されている。Semantic Web 用のフォーマットとしては、RDF/XML および Notation 3 のサポートを行なっている。Notation 3 の出力は Perl 言語による独自実装を行なっており、RDF/XML は RDF::Notation3 ライブラリを用いて Notation 3 から変換している。RDF を生成する際のオントロジーとして G-Links では EDAM Ontology と UniProt Core Ontology を採用している。

4 結果

4.1 利用方法

G-Links は RESTful なインタフェースで提供されており、ユーザが目的とする遺伝子 ID および遺伝子セットを示す ID、塩基/アミノ酸配列を含んだ一意の URL にアクセスするだけで、当該遺伝子に関連する情報を高速に取得することが可能である。本サービスは <http://link.g-language.org/> から利用することができる他、詳細なドキュメントおよび利用サンプルが <http://g-language.org/wiki/glinks> から利用できる。サービス自体のソースコードは <https://github.com/cory-ko/G-Links> にて公開されており、内部データベース内に登録されているデータは月 1 回の頻度で更新が行われる。また、以下に G-Links のシンタックスを示す。[] はユーザからの必須クエリの入力部、() は任意入力オプション部を示す。各オプションの機能と利用方法については本章にて記述する。

(1) 遺伝子 ID, 遺伝子セットの ID, 生物種名をクエリとした場合

```
http://link.g-language.org/[GENE or GENE SET ID]
(/filter=[FILTER])(/extract=[EXTRACT])(/format=[FORMAT])
```

(2) 配列情報をクエリとした場合

```
http://link.g-language.org/[SEQUENCE]
(/value=[E-VALUE])(/identity=[IDENTITY])(/direct=[0 or 1])
```

G-Links では入力として 85 のデータベースから得られた 205,829,185 の ID (205,811,947 の遺伝子 ID および 17,238 の生物種 ID) および塩基/アミノ酸配列に対応しており, 132 のデータベースから得られた 315,481,016 のエントリーから, ユーザのクエリに関連する情報を高速に取得し, 利用しやすい各種フォーマットでユーザに提供する. 遺伝子 ID を入力する際にはデータベースの情報は不要であり, ID のみを入力すればその ID が利用できるデータベース名を推測し適切なリソースをユーザに提供することで汎用的な入力系を実現している. これらのリストの最新情報は http://link.g-language.org/input_list および http://link.g-language.org/output_list から利用できる.

4.2 ブラウザ経由での動作

G-Links は REST サービスとして実装されており, 何らかの ID を入力するだけでブラウザから容易に利用することができる. この時にデータベース名の入力は必要なく, [http://link.g-language.org/\[GENEID\]](http://link.g-language.org/[GENEID]) のように何らかの遺伝子 ID が含まれた簡単な URL にアクセスするだけで, ユーザは自身が対象とする遺伝子もしくは遺伝子群についての網羅的な情報を確認することができる. そのため G-Links は, 研究者が着目している遺伝子について調べている際などにブラウザに簡単な URL を入力するだけで, ユーザはその遺伝子がどのような遺伝子かという「その遺伝子に関する知識」情報を容易に閲覧することが可能になる. 例として, *Homo Sapiens* の BRCA1 遺伝子 (Serova *et al.*, 1997) を示す UniProt のエントリー, BRCA1_HUMAN について情報を取得するには http://link.g-language.org/BRCA1_HUMAN にアクセスをすればよい. この出力結果に含まれるデータ量及びデータ取得速度を表 4.1 以下に示す.

表 4.1: G-Links の実行結果の詳細

実行時間	0.03 秒 (TSV), 1.98 秒 (HTML)
画像データ	25 種類 (KEGG Pathway, PDB, COXPRESdb など)
記述情報	184 エントリー (48 種類)
ID 情報	443 エントリー (68 データベース)

図??と同様に, http://link.g-language.org/BRCA1_HUMAN へアクセスした際の出力結果についての詳細情報を示す. G-Links を用いることで, ユーザは簡単な 1URL にアクセスするだけで大量の情報を高速に取得し閲覧することができる.

4.3 遺伝子セットに対するデータ取得

G-Links では単一の遺伝子を示す ID や配列だけではなく, 複数の遺伝子セットに対してのデータ取得も 1URL の指定で行う事が出来る. ユーザは複数の遺伝子 ID をカンマ区切りで指定するだけで, そ

これらの遺伝子に関連する情報を取得することが可能である。このときデータベースが異なる ID が複数混在していたとしても、それぞれの ID に関してデータベース名を自動推測しデータ収集を行う。例えば UCSC ID の uc003hui および、GeneID の 93986 の両遺伝子についての情報を収集するには、<http://link.g-language.org/uc003hui,93986> へアクセスをするだけでよい。また、KEGG Orthology に代表される遺伝子セットをしめす ID を入力した場合も、その ID リソースに含まれる全ての遺伝子についての情報を収集する。この概念の拡張として、生物種を示す ID を指定した場合はその生物種が持つ遺伝子全てについての生物学情報セットを提供する。このときの生物種と遺伝子のマッピングは UniProt taxonomy をベースに行っている。

4.4 汎用的な出力フォーマット

以上のようにして指定されたリソースについて、G-Links ではユーザが利用しやすい複数のフォーマットで出力することができる。以下に G-Links で利用できる各種フォーマットと当該フォーマットの指定方法について表 4.2 に示す。

表 4.2: G-Links で利用可能なフォーマット

指定する値	出力形式	補足情報
tsv	タブ区切り	デフォルト値
slim	タブ区切り	URL など一部情報を削除
json	JSON	
html	HTML	ブラウザからのアクセス時のデフォルト
rdf	RDF/XML	
n3	Notation3	

G-Links にて出力として使用できるデータフォーマットの一覧を示す。これらの値を format オプションで指定することで、ユーザは 6 種類のフォーマットから自身の目的に最適な形式で出力を得ることができる。例として、BRCA1 遺伝子に関する出力を JSON フォーマットで取得する場合は、http://link.g-language.org/BRCA1_HUMAN/format=json へアクセスをするだけで JSON を取得できる。また、ブラウザからの閲覧の場合は HTML、それ以外からのデータ取得の場合は tsv など、ユーザが利用しているコンテキストに合わせて出力フォーマットのデフォルト値を自動的に変換することで、ユーザに対してより利便性の高い出力を行うことができる。

G-Links では大きく分けて 3 種類のフォーマットを提供している。HTML フォーマットによる Human-readable な出力は画像情報の付与など人が目で見て理解することを目的としており、ID 情報や記述情報は利用可能なハイパーリンクとともに並び替え可能なテーブルに格納されている。また、Programmatic な出力としては JSON や Tab Separated Value (TSV) をサポートしている。これらは各プログラミング言語や UNIX コマンドラインツールなどで容易に処理することができるフォーマットであり、フォーマットの指定も含めて簡便な URL を指定するだけで取得できる。そのため、研究者は G-Links を解析用のデータ収集を行うためのデータソースとしてユーザ自身のプログラムから容易に利用することができる他、Web アプリケーション開発時の高速なバックエンドデータアグリゲータとしても利用が可能である。各種 Semantic Web 技術と連携を行うため RDF/XML や Notation3 といった RDF 出力も可能である。Semantic Web における大きな問題点の一つであった RDF リソース高速出力が可能である他、そのリソースを一意的 URL で直接指定し利用できる。G-Links の RDF ではオントロジーとして基本的に EDAM Ontology を使い、カバーできない部分に関して UniProt Ontology を用いている。EDAM Ontology はバイオインフォマティクスを行う上で必要な情報の広範囲をカバーしており、データ収集と Web サービス解析の双方を備えた本サービスには非常に適したオントロジーであると言える。

4.5 必要なデータの抽出

G-Linkはその容易さおよび高速性から解析のためのデータセット収集の段階で非常に有用であるが、そのデータ量に起因する通信速度の問題とノイズ情報による情報量低下の問題が発生する。より研究者にとって価値の高いリソースを提供するパイプラインを構築するには、関連情報を網羅的に全て提供するのではなく研究者が必要とする情報のみで構築されたより平均情報量の高いリソースへと昇華する必要がある。G-Linksでは、遺伝子自体に対するフィルタリングと取得される生物学情報に対する情報抽出という2つのアプローチをオプションとして提供することでこの問題の解決を試みた。

filter オプションではユーザによって指定された遺伝子セットのうち、本オプションで指定された条件に合致した遺伝子に関する情報だけを抽出する。filter の条件指定はデータベース名および”DISEASE”といったG-Linksで使われている情報カテゴリを示す「情報のセクション名」と「フリーワード」の2種類が利用可能であり、「セクション名:フリーワード」の様に”:”を用いてその区別を行う。セクション名フリーワードはそれぞれ個別に指定することも可能である。例えば、”DISEASE”セクションの情報を持っている遺伝子は”filter=DISEASE”，がん関連の情報を持っている遺伝子は”filter=:cancer”，がんに関する”DISEASE”セクションの情報を持っている遺伝子”filter=DISEASE:cancer”と指定することで、その条件に合致した遺伝子の情報だけを抽出できる。また、filter オプションは”|” (パイプ) によって複数条件を記述、またはfilter オプションを複数回用いることで絞り込み条件を追加することが可能である。これら複数条件を指定した場合、G-LinksではAND条件として解釈する。もう一つのフィルタリング方法であるextract オプションでは、ユーザが指定した「情報セクション名」を元に情報抽出を行う。データレベルでのフィルタリング方法である。情報抽出に利用できるのはデータベース名およびセクション名で、例えば”DISEASE”セクションの情報のみが必要な場合、”extract=DISEASE”と指定すればよい。extract オプションもfilter オプションと同様に”|”を用いることで複数条件を同時に指定することができる。なお、extract オプションにおける複数の条件指定はOR条件として解釈される。これらのオプションを組み合わせることで、ユーザは多数存在する生物学データベースの統合、その大規模なリソースから自身の研究対象に関連のある情報の収集、そこで得られた生物学情報セットから研究者自身が必要とする情報の抽出という複雑かつ労力のかかるデータ統合プロセスを簡単なURLにアクセスするだけで容易かつ高速、自動的に行うことができる。両オプションの利用例を以下に示す。

filter オプションと extract オプションによるリソース抽出の例

Homo Sapiens の全遺伝子のうち、がん関連遺伝子の情報をタブ区切りで

<http://link.g-language.org/9606/format=tsv/filter=DISEASE:cancer>

さらに胸部と子宮に関連し、かつ SNP と遺伝子多型を持つ遺伝子に絞り込み

[http://link.g-language.org/9606/format=tsv/filter=DISEASE:cancer
/filter=:breast|:ovarian](http://link.g-language.org/9606/format=tsv/filter=DISEASE:cancer/filter=:breast|:ovarian)

そこから dbSNP と SNPedia の情報を抽出

[http://link.g-language.org/9606/format=tsv/filter=DISEASE:cancer
/filter=:breast|:ovarian|:snps|:polymorphisms
/extract=dbSNP|SNPedia](http://link.g-language.org/9606/format=tsv/filter=DISEASE:cancer/filter=:breast|:ovarian|:snps|:polymorphisms/extract=dbSNP|SNPedia)

filter と extract を用いて、G-Links から得られたリソース群からユーザが必要とするリソースのみを抽出した例。このように filter と extract を組み合わせることで、「子宮頸癌と乳がんに関連する *Homo Sapiens* の遺伝子のうち、SNP 情報と遺伝子多型の情報があるものについて、全 dbSNP と SNPedia の情報」を一つの URL にアクセスするだけで取得することができる。

5 議論

本論文では、バイオインフォマティクス Web サービスおよび生物学データベースなど、多領域に渡る生物学リソースの効率的な統合モデルに関しての議論およびシステム設計を行った。生物学研究者は数千ものオープンに公開されたデータベースを自由に用い自身が対象とする生命現象に関する解析を行うことができる。しかし生命システムは多レイヤーから構成される複雑な系であり、それをより深く理解するためには多数の生物学データベースの情報を統合することで多領域に渡る生物学情報を収集し、それらを用いたより詳細かつ大規模な解析を行う必要がある。バイオインフォマティクス研究ではその作業のほとんどが、研究対象に関連するデータセットの収集・統合・抽出の作業に占められており、この作業を高速かつ自動的、効率的に行うシステムの構築が求められてきた。G-Links ではユーザが与えた遺伝子を示す ID について、その ID が含まれた簡単な URL にアクセスするだけで関連する生物学情報を 130 以上のデータベースおよび解析 Web サービスから網羅的かつ高速に収集しユーザに提供する。また、遺伝子 ID だけではなく、遺伝子セットを示す ID や生物種を示す ID 配列類似性検索を用いることで塩基/アミノ酸配列の直接入力を行うこともできるため、遺伝子を表すオブジェクトに対して汎用的な入力系を実現している。さらに、本システムは複数データベースに対するデータセットの統合と取得だけではなく、得られたリソース抽出プロセスについても遺伝子レベルと情報レベルの2つの抽出方法を組み合わせることでサポートする。これらのオプションを利用することで、ユーザは一意的 URL にアクセスするだけで、対象の遺伝子セットに関連するデータセットを複数の生物学データベースから網羅的かつ高速に取得し、そこから自身が必要なデータセットだけを抽出し取得するというプロセスが実行可能となる。これらの特徴に加え G-Links では G-language EMBOSS REST サービスとそれに含まれる KBWS REST サービスと連携を行うことで、単純なデータベース統合では得られなかった、解析ツールによって導出される生物学リソースをも統合して利用することができる。両サービスとも URL にて解析結果リソースが指定できるため、G-Links が持つ他の出力と同レベルでシームレスな統合が可能である。また、KBWSを採用することで新たなサービスへの容易な拡張も可能である。生命システムという多領域の情報による複雑な関係ネットワークの上に構築されている現象を理解するための解析を行うには、多領域にわたる生物学情報を効率的に統合し解析を行う必要がある。しかしながら生物学リソースの多領域性とデータ量の規模ゆえに、全ての生物学リソースを統合しそこから自身の研究対象と関連のあるリソースを抽出・取得するプロセスは多大な労力を必要とする。本論文ではこの問題を解決するシステムの実装を行い、ユーザは自身の研究に用いる多領域生物情報のデータセットを高速に、必要なデータを必要なだけ、自動的かつ容易に取得することを可能にするサービスの提供を行った。多領域生物学リソースの効率的統合は生物学の大きな課題の一つであるが、この統合モデルを用いることで、生物学で求められてきたリソース統合のためのサイバーインフラのベースとなりうるシステムの構築を行うことが可能となると言える。

謝辞

このような研究の場与えてくださった富田勝教授に感謝申し上げます。また、本研究を行うにあたって様々な助言をくださった荒川和晴特任講師、および G-language Project の全てのメンバーに心より感謝申し上げます。

参考文献

- Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orłowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., and Goble, C. A. (2010). BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res.*, **38**(Web Server issue), W689–694.
- Brazas, M. D., Yim, D., Yeung, W., and Ouellette, B. F. (2012). A decade of Web Server updates at the Bioinformatics Links Directory: 2003–2012. *Nucleic Acids Res.*, **40**(Web Server issue), W3–W12.
- Codd, E. F. (1969). Derivability, redundancy and consistency of relations stored in large data banks. *IBM Research Report, San Jose, California*, **RJ599**.
- Cote, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., and Hermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, **8**, 401.
- Dayhoff, M. O., Barker, W. C., Schwartz, R. M., Orcutt, B. C., and Hunt, L. T. (1976). Data base for protein sequences. In *Proceedings of the June 7–10, 1976, national computer conference and exposition, AFIPS '76*, 261–266, New York, NY, USA. ACM.
- Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J. C., Hernandez-Boussard, T., Rees, C. A., Cherry, J. M., Botstein, D., Brown, P. O., and Alizadeh, A. A. (2003). SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**(1), 219–223.
- Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**(Database issue), D136–143.
- Fernandez-Suarez, X. M. and Galperin, M. Y. (2013). The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, **41**(D1), 1–7.
- Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**(Database issue), D258–261.
- Huang, H., McGarvey, P. B., Suzek, B. E., Mazumder, R., Zhang, J., Chen, Y., and Wu, C. H. (2011). A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics*, **27**(8), 1190–1191.
- Jacso, P. (2004). Thoughts about federated searching. *Information Today*, **21**(9), 17–20.
- Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.
- Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**(4), 656–664.
- Mudunuri, U., Che, A., Yi, M., and Stephens, R. M. (2009). bioDBnet: the biological database network. *Bioinformatics*, **25**(4), 555–556.
- Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Motoike, I. N., and Kinoshita, K. (2013). COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Res.*, **41**(D1), D1014–1020.
- Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**(Database issue), D130–135.
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**(6), 276–277.
- Serova, O. M., Mazoyer, S., Puget, N., Dubois, V., Tonin, P., Shugart, Y. Y., Goldgar, D., Narod, S. A., Lynch, H. T., and Lenoir, G. M. (1997). Mutations in BRCA1 and BRCA2 in breast cancer families: are there more breast cancer-susceptibility genes? *Am. J. Hum. Genet.*, **60**(3), 486–495.
- Stein, L. (2002). Creating a bioinformatics nation. *Nature*, **417**(6885), 119–120.
- Stein, L. D. (2008). Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat. Rev. Genet.*, **9**(9), 678–688.

- The UniProt Consortium (2012). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*
- van den Berg, B. H., McCarthy, F. M., Lamont, S. J., and Burgess, S. C. (2010). Re-annotation is an essential step in systems biology modeling of functional genomics data. *PLoS ONE*, **5**(5), e10642.
- Wilkinson, M., Gessler, D., Farmer, A., and Stein, L. (2003). The BioMOBY Project Explores Open-Source, Simple, Extensible Protocols for Enabling Biological Database Interoperability.
- Wilkinson, M. D., Senger, M., Kavas, E., Bruskiwich, R., Gouzy, J., Noirot, C., Bardou, P., Ng, A., Haase, D., Saiz, E. d. e. A., Wang, D., Gibbons, F., Gordon, P. M., Sensen, C. W., Carrasco, J. M., Fernandez, J. M., Shen, L., Links, M., Ng, M., Opushneva, N., Neerincx, P. B., Leunissen, J. A., Ernst, R., Twigger, S., Usadel, B., Good, B., Wong, Y., Stein, L., Crosby, W., Karlsson, J., Royo, R., Parraga, I., Ramirez, S., Gelpi, J. L., Trelles, O., Pisano, D. G., Jimenez, N., Kerhornou, A., Rosset, R., Zamacola, L., Tarraga, J., Huerta-Cepas, J., Carazo, J. M., Dopazo, J., Guigo, R., Navarro, A., Orozco, M., Valencia, A., Claros, M. G., Perez, A. J., Aldana, J., Rojano, M., Fernandez-Santa Cruz, R., Navas, I., Schiltz, G., Farmer, A., Gessler, D., Schoof, H., and Groscurth, A. (2008). Interoperability with Moby 1.0—it's better than sharing your toothbrush! *Brief. Bioinformatics*, **9**(3), 220–231.