Title	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
Sub Title	
Author	大下, 和希(Oshita, Kazuki)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2012
Jtitle	生命と情報 No.19 (2012.) ,p.29- 40
Abstract	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の設計・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。
Notes	慶應義塾大学湘南藤沢キャンパス先端生命科学研究会 2012年度学生論文集 修士論文ダイジェスト
Genre	Technical Report
URL	http://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO92001004-00000019-0029

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の設計・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(Dayhoffetal., 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, 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(vandenBergetal. , 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ApplicationProgrammingInterface(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-SuarezandGalperin, 2013)や1200を超える解析Webサービス(Brazasetal. , 2012)がWeb上でオープンに提供されており(Bhagatetal., 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UniformResourceIdentifier(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FederatedQuery(Jacso, 2004)型データ統合は主にSOAPなどの検索ツールWebAPIを用いたサービス統合による問題解決を目指しており(Wilkinsonetal. , 2003), BioMoby(Wilkinsonetal. , 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トリー間のLinkによってデータベース間の関係性を表現するLinkedData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RelationalDatabase(RDB)(Codd, 1969)というアーキテクチャが存在するが，LinkedData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トリーを示す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，LinkedData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(Diehnetal. , 2003)やタンパク質情報に特化したProteinIdentifierCross-Referencing(PICR)(Coteetal. , 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(Mudunurietal. , 2009)はユーザから受け取ったIDの解決部分を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変換では各エント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 SemanticWeb

これらの問題の解決策として現在着目されているのがTimBerners-Leeによって提唱されたWorldWideWeb(WWW)の利便性を向上するためのプロジェクト，SemanticWebである．SemanticWeb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WebOntologyLanguage(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，SemanticWebではResourceDescriptionFramework(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SemanticWeb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，SemanticWeb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LinkedData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LinkedData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はなく，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PrimaryKeyを設定し，LinkedDataネットワークを整理することで解決を試みた．PrimaryKey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，UniProtID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(TheUniProtConsortium, 2012)，非常に品質の高くLinkedData

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IDmappingサービス(Huangetal. , 2011)を提供しているなどPrimaryKey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UniProtIDに変換するID解決部と，そのUniProtID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MySQL5.0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．KEGGOrthology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UniProtID群を検索し，得られたUniProtID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UniProtID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TaxonomySearch(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NCBITaxonomy(Federhen, 2012)およびRefSeq(Pruitt

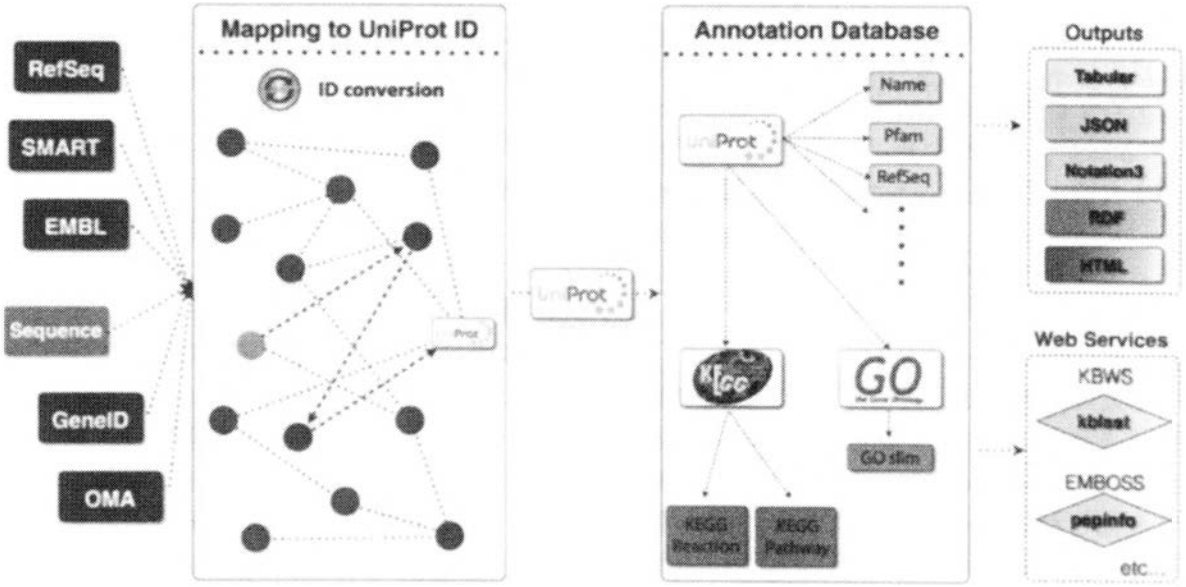

図3.1: G-Links全体のアーキテクチャ図G-Linksでは遺伝子を示すIDおよび配列情報をユーザからクエリとして受け取り，それをUniProtIDへとID変換および配列類似性検索を用いて変換する．その後，当該UniProtIDに関連する他データベースのID情報，クロスリファレンスおよびそこから取得したリソース，外部Webサービスの解析結果を示すURLなどを含んだ結果をユーザに任意のフォーマットにて提供する．

etal. , 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UniProtID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EuropeanMolecularBiologyOpenSoftwareSuite

(EMBOSS)(Riceetal. , 2000)のtranseqを用いてアミノ酸配列へ翻訳を行い, BLASTLikeAlignmentTool (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し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UniProtIDとともにE-valueやIdentity, 生物種名や遺伝子の名前, そのUniProtID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UniProtID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UniProtID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UniProtID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, GOslim (Harrisetal. , 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, その遺伝子が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UniProtID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, SemanticWeb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. Programable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はなく, KEGGPathwayのパスウェイマップやCOXPRESdb(Obayashietal. , 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:HypertextPreprocessor(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JavaScripにて実装されたtablesorter(http://tablesorter. 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. SemanticWeb用のフォーマットとしては, RDF/XMLおよびNotation3のサポートを行なっている. Notation3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Notation3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EDAMOntologyとUniProtCoreOntology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, ()は任意入力の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.

G-Links Syntax

(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]
(/evalue=[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]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，HomoSapiensのBRCA1遺伝子(Serovaetal. , 1997)を示すUniProtのエントリー，BRCA1HUMAN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3.1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UCSCID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, KEGGOrthology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UniProttaxonomy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遺伝子に関しての出力を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TabSeparatedValue(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SemanticWeb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．SemanticWeb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な他，そのリソースを一意のURLで直接指定し利用できる．G-LinksのRDFではオントロジとして基本的にEDAMOntologyを用い，カバーできない部分に関してUniProtOntologyを用いている．EDAMOntology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"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

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

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 にアクセスを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，ユーザは一意の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-languageEMBOSSRESTサービスとそれに含まれるKBWSREST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冨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-languageProject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., Orlowski, 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). EMBOSS: 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., Kawas, E., Bruskiewich, 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.