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	Gene-centric integration of multi-domain biological resources
Author	大下, 和希(Oshita, Kazuki) 冨田, 勝(Tomita, Masaru)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2013-05
Jtitle	優秀修士論文
Abstract	バイオインフォマテイ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。解析Webサービスのインタフェース統一による相互運用性の向上を行ったツールであるKeio Bioinformatics Web Service(KBWS)は, 多数の生物情報解析Webサービスに対して統一されたインタフェースを持つSOAPプロキシサーバを提供することで, サポートする42のWebサービスに対して一元化された手法でのアクセスを提供する。さらに, クライアントツールをローカルツールとの相互運用性の高いEMBOSS追加パッケージとして実装することで, Webサービスだけでなくローカルツールも交えた解析フローの容易な運用を可能にした。本ツールはhttp://www.g-language.org/kbws/より利用できる。さらに本論文では, 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G-Linksを構築した。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	2012年度先端生命科学プロジェクト
Genre	Thesis or Dissertation
URL	http://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO92001003-2013-001-0001

ISBN 978-4-87762-266-4
SFC-MT 2013-001

優秀修士論文

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

2012年度

大下 和希　政策・メディア研究科 修士課程

先端生命科学プロジェクト

慶應義塾大学湘南藤沢学会

Keio SFC Academic Society

推薦のことば

生命システムは多様な要素がそれぞれに影響しあうことで非常に複雑な系であり、より深い理解を行うためには多数のデータベースから関連情報を収集し、解析を行う必要がある。しかしながら生物学の発達により生物学データベースの数やデータ量、および種類は爆発的に増加しており、データの統合という解析準備のためのプロセスに研究者が非常に多くの労力を費やさなければならないという自体が起きている。本論文では「全ての遺伝情報は遺伝子から辿ることができる」というセントラルドグマを元にしたデータ統合モデルをベースとし、多数の生物学 Web データベース群や解析 Web サービスから研究者が必要とするデータセットを網羅的かつ高速に取得するシステムの構築を行った。多領域にわたる上に様々なデータスキーマを持つなどの問題から、生物学 Web リソースを効率的に統合・運用するためのシステム構築はバイオインフォマティクス分野における大きな課題の一つである。単純に研究者が解析準備に必要とする労力を大幅に削減できるだけでなく、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なデータ統合プラットフォームの設計・構築を行った意義は大きいと考えられる。

慶應義塾大学
環境情報学部教授
冨田　勝

修士論文　2012年度（平成24年度）

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

慶應義塾大学大学院 政策・メディア研究科

大下 和希

先端生命科学プロジェクト

2013年1月

修士論文　2012年度（平成24年度）

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

論文要旨

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．解析Webサービスのインタフェース統一による相互運用性の向上を行ったツールであるKeio Bioinformatics Web Service (KBWS) は，多数の生物情報解析Webサービスに対して統一されたインタフェースを持つSOAPプロキシサーバを提供することで，サポートする42のWebサービスに対して一元化された手法でのアクセスを提供する．さらに，クライアントツールをローカルツールとの相互運用性の高いEMBOSS追加パッケージとして実装することで，Webサービスだけでなくローカルツールも交えた解析フローの容易な構築・運用を可能にした．本ツールは`http://www.g-language.org/kbws/`より利用できる．さらに本論文では，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G-Linksを構築した．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．

キーワード

Bioinformatics, Database Integration, Web Service, Semantic Web, Systems Biology

慶應義塾大学大学院 政策・メディア研究科

大下 和希

Abstract Of Master's Thesis Academic Year 2012

Gene-centric integration of multi-domain biological resources

Summary

In bioinformatics, multitude of biological databases and analysis web services are freely available online, and researchers routinely access these multi-domain biological resources for computational analyses in order to understand the complex biological systems. The key to such tasks is the efficiency in the processes of aquisition, integration and management of biological resources and web services, and thus an environment that enables complex analysis workflows by seamless interoperation of biological web services. Therefore, this thesis describes a design and implementation of web-based software systems that integrate multi-domain biological web-resources to achieve higher efficiency in bioinformatics researches. Firstly, Keio Bioinformatics Web Service (KBWS) (`http://www.g-language.org/kbws/`) provides 42 bioinformatics web services through a unified user interface by providing SOAP proxy server. Users are able to utilize these web services easily via simple and standardized accessing methods. KBWS client tools implemented as EMBOSS associated package further enables easy construction of analysis workflows comprised of web services and local analysis tools. G-Links (`http://link.g-language.org/`), a RESTful gateway server that effectively integrates the multi-domain biological web resources and allows rapid information retrieval with facet queries from over 100 databases. G-Links internally integrates multi-domain biological resources as an ID conversion system, using a gene-centric data integration model. With G-Links, users are able to retrieve biological resources related with given gene sets and extract necessary datasets, by simply accessing a simple URL, and the results can be obtained in a variety of formats, including graphical HTML5 presentation, flat files for computational access, and Semantic Web compliant formats. This set of biological data integration platforms collectively enable efficient retrieval of dispersed multi-domain knowledge about the biological systems of interest.

Keywords

Bioinformatics, Database Integration, Web Service, Semantic Web, Systems Biology

Graduate School of Media and Governance
Keio University

Kazuki Oshita

目次

第1章 序論 1

1.1 バイオインフォマティクスにおけるWebリソース 1
1.2 Webサービスの相互運用 2
1.3 データベースの統合的利用 4
1.4 生物学におけるリソースの効率的統合 6

第2章 Keio Bioinformatics Web Services 7

2.1 背景 7
2.1.1 Webサービスの一般化と相互運用性 7
2.1.2 Service Discoveryとオントロジー 7
2.1.3 EMBOSSによるローカル環境での相互運用性と発見可能性の実現 8
2.2 設計と実装 9
2.2.1 アーキテクチャ 9
2.2.2 KBWSプロキシSOAPサーバ 9
2.3 結果 11
2.3.1 Availability 11
2.3.2 利用可能なサービス 11
2.3.3 KBWSを利用した解析ワークフローの構築 11
2.3.4 Tavernaを用いた解析フローの構築と運用 14
2.3.5 プログラミング言語からのSOAPプロキシサーバの利用 14
2.3.6 GUIからの利用 15
2.4 議論 15
2.4.1 Webサービス利用のための統一的インタフェースの実装 15
2.4.2 より高度なService Discoveryの実現 18

第3章 G-Links 19

3.1 背景 19
3.1.1 生物学データベースの統合的利用 19
3.1.2 データベースの統合とそれに伴う問題 19
3.1.3 データベース単純結合の問題点 20
3.1.4 先行研究 20
3.1.5 ID変換によるアプローチ 21
3.1.6 ID変換の持ちうる問題点 23
3.1.7 Semantic Web 24
3.1.8 Semantic Webが抱える問題点 25
3.2 要求分析 26
3.3 設計と実装 28
3.3.1 アーキテクチャ 28
3.3.2 ユーザクエリのID解決 30
3.3.3 アノテーション 31
3.3.4 アウトプット 34
3.4 結果 35

3.4.1 利用方法 35
3.4.2 ブラウザ経由での動作 35
3.4.3 配列入力時の動作 38
3.4.4 遺伝子セットに対するデータ取得 40
3.4.5 汎用的な出力フォーマット 41
3.4.6 必要なデータの抽出 42
3.4.7 外部 Web サービスとの連携 44
3.4.8 GENIE - a Virtual Biological Research Assistant 45
3.5 議論 49
3.5.1 G-Links による遺伝子中心型の統合モデル 49
3.5.2 既存サービスとの比較 51
3.5.3 G-Links で扱うべきデータ 52
3.5.4 ロケーション問題の解決 53
3.5.5 外部 Web サービスとの連携と出力データセットの拡充 54
3.5.6 オントロジーの問題 55

第 4 章 結論 **57**
4.1 インタフェースの画一化による Web サービス相互運用性の向上 57
4.2 遺伝子ベースの ID 変換を用いた生物学リソースの効率的統合 58
4.3 総括 59

謝辞 **61**

参考文献 **62**

図目次

1.1 CORBA のアーキテクチャ 3
1.2 SOAP のアーキテクチャ 4

2.1 KBWS のシステムアーキテクチャ図 10
2.2 KBWS を用いた解析フローによって作成されたシーケンスロゴ 13
2.3 Taverna の実行例 14
2.4 KBWS SOAP プロキシサーバの rpc/encoded による利用例 16
2.5 EMBOSS Explorer からの利用例 17

3.1 G-Links 全体のアーキテクチャ図 29
3.2 ID 変換部のアーキテクチャ 32
3.3 アノテーション取得部の ER モデル概略図 33
3.4 Web ブラウザ経由での動作例 39
3.5 配列を直接入力した際の実行例 40
3.6 Genie のインタフェース 46
3.7 Genie の実行例 48

表 目 次

2.1 KBWS にてサポートしているサービス一覧 12

3.1 G-Links で入力対応しているデータベースリスト 36
3.2 G-Links で出力対応しているデータベースリスト 37
3.3 G-Links の実行結果の詳細 38
3.4 G-Links で利用可能なフォーマット 41
3.5 G-Links でサポートしている外部解析 Web サービスのリスト 45

第1章　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，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．

しかしながら，International Nucleotide Sequence Database (Nakamura *et al.*, 2013) に代表される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，メリットと共に研究者に運用コストというデメリットをもたらしている．大規模データを扱うためには十分な計算資源や労力が必要であることに加え，インターネットにおける通信速度の限界からインターネットを経由して転送し利用するべきデータ量を超えてくることが予測される．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ツールAPIは研究者にとって非常に有用なツールであり，現在に至るまで多くの検索ツールAPIが開発・整備され続けてきた．さらに，これらの検索ツールAPIの多くはWebサービスAPIとして提供されている．計算資源が豊富なWebサーバで検索ジョブを実行し，データベース全体のダウンロードやメンテナンスの必要がなく，検索結果という必要最低限のデータだけを取得すれば良いこの形態は生物学におけるデータベース肥大化の問題を解決する上で非常にメリットの大きい手法であり，現在ではほとんどのデータベースが検索Webサービスを備えている．

また，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．ツールのメンテナンスやセットアップコストが不要で豊富な計算資源を持つWebサーバを利用可能で，かつユーザのOSなどのローカル環境に非依存的に利用できるなど，Webサービスには大きな利点が存在する．特にBasic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990, 1997) のように解析のための大規模データベースを利用する解析ツールでは，検索ツールAPIと同様の理由からWebサービスである利点はより大きい．

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，その数は現在も加速度的に増加している．これらのリソース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生物学において求められてきたが (Stein, 2002, 2008)，これらのWebサービス統合およびデータベース統合の問題は情報学の分野でも一般的に議論されている大きな問題である．

1.2 Webサービスの相互運用

多くのバイオインフォマティクス研究では個別の解析ツールを単純に利用するのではなく，複数の解析ツールやアルゴリズムをつなぎあわせることで一つのワークフローのように取り扱って解析が行われている．解析Webサービスや検索Webサービスの数および種類の増加に加え，それらにアクセスするためのAPIの整備が進んだ現在では，非常に広範囲の解析をWebサービスを連携させることで実行可能になっている他，豊富な計算資源などの利点を生かしてローカルの解析ツールとWebサービスを連携させることで効率のよいワークフローをも構築可能である．このように複数の計算資源やサービスをシームレスに連携させるためのプラットフォームの開発は情報学でも古くから研究されており，それを達成するためのWeb-APIの仕様が多く開発されてきた．

Common Object Request Broker Architecture (CORBA) (Object Management Group, 1991) は，Object management Group (OMG) が策定した分散処理環境のための国際標準アーキテクチャである．このアーキテクチャではプログラムコードとそのインタフェース情報をカプセル化したオブジェクト General InterORB Protocol (GIOP) メッセージを各サーバに送る方式を採り，そのオブジェクト経由で各サーバに実装されているプログラムを起動するRemote Procedure Call (RPC) の一種である．サーバ上の実装コードとGIOPメッセージとのインタフェースはインタフェース記述言語 (IDL) にて記述されており，この情報を元に各プログラミング言語とのマッピングを行うことで異なる環境間での正しい通信を保証している他，クライアントは自身のプログラムとCORBAによって連携された外部サービスを高い相互運用性の下で利用することが可能となる．このようにCORBAは非常に高機能な通信方式であったが，高機能であるが故の実装の複雑さや難しさも特徴の一つとして挙げられる．例えば，GIOPメッセージのシリアライズおよびデシリアライズを行うCORBA Object Request Broker (CORBA ORB) というミドルウェアやそれが動作する環境をサーバとクライアント双方に設置する必要があり，またCORBA ORB間の通信にはInternal InterORB Protocolという独自の通信プロトコルを利用する必要がある．これらの問題点から，eXtensible Markup Language (XML) にてエンコードを行いHypertext Transfer Protocol (HTTP) にて通信を行うという最低限の仕様を定義したXML-RPCなども登場してきたが，XML-RPCを拡張したSOAP (`http://www.w3.org/TR/soap12-part0/`) や，より簡便なRepresentational State Transfer (REST) (Fielding, 2000) といったアーキテクチャに徐々に移行しつつある．

SOAPはCORBAと同じく異環境間での確実な通信の達成を目的としたプロトコルであるが，一部を簡略化することでCORBAの複雑さの問題を解決している．CORBAと同様に通信メッセージのカプセル化は行うのだが，ORBのようなミドルウェアは必要なく，SOAPメッセージを生成・解釈するためのライブラリさえ実装されていればどのようなプログラミング言語や環境でも利用可能である．SOAPメッセージはXMLで構造化されたテキストベースのメッセージのためライブラリの実装も比較的容易であり，かつHTTPやSimple Mail Transfer Protocol (SMTP) などの一般的なプロトコルの上での通信が可能であるほか，複数の結果を返すことも可能であった．より正確な通信を行うためには双方がメッセージの定義を共有している必要があるが，SOAPではWeb Services Description Language (WSDL) というファイルに標準のスキーマの下で記述されており，ユーザはそれを読み込むことで厳密な通信を行うことができるほか，入出力のマッピングを行うことで複数サービスを連携して利用することも可能である．このようにSOAPは非常に利便性の高い仕様であったが，その役割と内容を一度大きく変更していることも特徴の一つである．SOAPはもともとXML-RPCをベースとして構築されてた”Simple Object Access

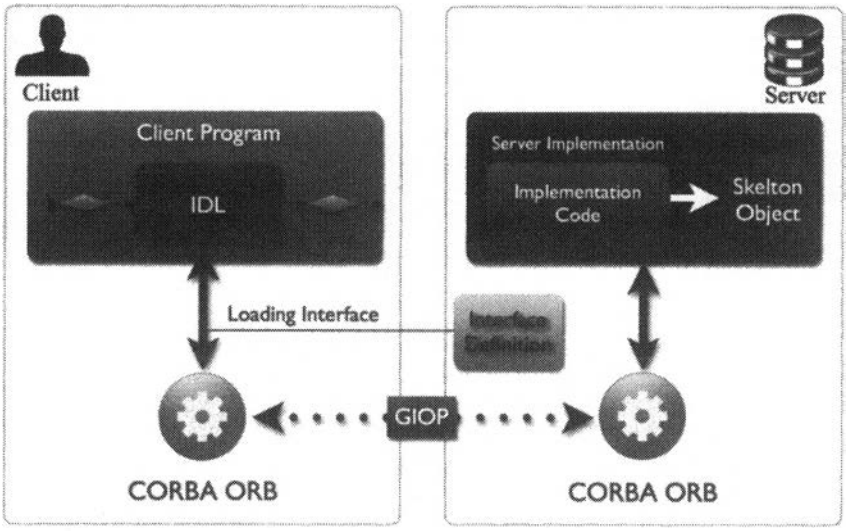


図 1.1: CORBA のアーキテクチャ

CORBA 通信の基本アーキテクチャを示す．CORBA は相手サーバ上に実装されているコードや機能を起動・実行するための通信アーキテクチャである．ユーザは通信相手のサーバ上で実装されている機能についてのインタフェース情報をもとに，その機能や呼び出し方の方法を IDL で記述．この情報をカプセル化したものを CORBA ORB に渡すとそれを GIOP というプロトコルで送信し，相手サーバの CORBA ORB は受け取ったメッセージに含まれる呼び出し情報を元に，サーバ上の実装コードのスケルトンを生成する．この自動生成されたプログラムを実行することで，ユーザは相手サーバの機能を直接実行することが可能になる．CORBA ORB や IDL などを中継することで，CORBA 通信の環境さえと整えれば異なる環境下での確実な通信が可能になる．

Protocol”の略称であり，オブジェクトを XML 形式に符号化して送信し，受信側で復号化することでオブジェクトのメソッドを呼び出していた (SOAP1.1)．しかしながら現在では，XML-RPC ベースの通信を行う rpc/encoded に変わって document/literal という利用方法が SOAP の標準となっている (SOAP1.2)．この方法では SOAP を，送信するドキュメントを単純に XML に埋め込んで通信を行うプロトコルとして再定義しており，符号化・復号化などのオーバーヘッドがない，より簡便な手法となっている．

SOAP にも見られた簡略化の流れが更に推し進められた通信アーキテクチャの一つで，現在非常に普及しているものが REST である．REST の大きな特徴として挙げられるのがリソースの概念である．REST では全てのリソースを Uniform Resource Identifier (URI) によって一意のアドレスで表現することが可能であり，そのリソースを指定することでデータを取得するというアーキテクチャである．通信プロトコルは HTTP に限定され，これまでの通信プロトコルのように通信相手のサーバに実装されたメソッドを起動するのではなく，HTTP で定義されたメソッドセットである GET，POST，PUT，DELETE にて操作を行う．このように，REST は通信プロトコルが HTTP で完結しており，データの解釈機などを準備する必要がなく，開発や利用が非常に容易なアーキテクチャである．特に Web サービスの場合は各リソースを URI のサブセットである Uniform Resource Locator (URL) にて指定することができる．URL ではそのリソースが存在する場所を直接表現できるため，ユーザはそのリソースを Web ブラウザさえあれば取得し利用することができる．通信されるデータがどのようなデータなのかを通信するコンピュータ間が相互に知っている必要があるというデメリットはあるが，XML などでリソースをマークアップすることである程度解決することが可能である．

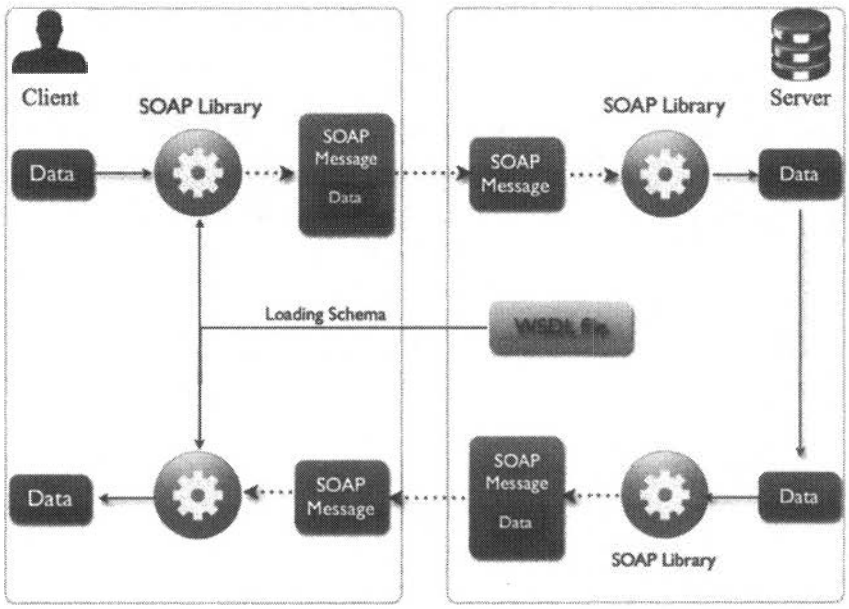


図 1.2: SOAP のアーキテクチャ

SOAP 通信の基本アーキテクチャを示す．SOAP では通信したいデータを送信元が通信に必要な各種パラメータが記述された XML ベースの SOAP メッセージ内に封入し，それを送受信することで確実な通信を行う．SOAP 通信のインタフェースや必要なパラメータ情報は WSDL ファイルという定義ファイルに厳密に指定されているため，それを元に SOAP メッセージの生成・翻訳を行うライブラリを各プログラミング言語から用いることで，ユーザは OS やプログラミング言語などの環境が異なるマシン間での確実な通信を保証することができる．

1.3 データベースの統合的利用

複数のデータベースを統合する際の大きな問題は，そのデータ量とスキーマ定義，シンタックスの差異の問題である．バイオインフォマティクスでは一つひとつのデータベースについてデータ量が増加しており，それらを統合するとなると非常に大きなデータアーカイブを扱う事になる．データベースである以上は内包するデータアーカイブに対して容易に検索や抽出などの再利用が行える必要があるが，そのためには莫大な計算資源が必要となる．また，生物学の技術進歩による生物学で扱うべきデータの種類の増加により，統合データベースはそのデータスキーマが絶えず変化することとなる．さらにこれらの問題を解決した状態で，個々のデータベースのアップデートに合わせて内容を同期する必要がある．データ構造自体が頻繁に変わり，更新の頻度も高い大規模データベースの実用レベルでの運用コストは非常に大きな問題である．また，各データベースが内部に保管しているデータのシンタックスが必ずしも統一されていないことも問題の一つである．データのシンタックスはそのデータが扱う情報をより効率的に表現する方法の一つであり，内包されるデータの種類の増加に伴った変化が起こりうる要素である．完全なデータ統合を行うためには，統合対象となる全てのデータベースのシンタックスを正確に解釈し，シンタックスの異なるリソース間のマッピングを解決した上で，シンタックスの変化に常に対応する必要がある．

この問題の解決策の一つがデータベース検索サービスを用いた擬似的な統合である．先に述べた通り生物学データベースでは内部データの検索 Web API が非常に発達しており，個別のデータベースに対してならば必要なデータを容易に取得することが可能である．そのため，複数のデータベースに対する検索 Web API を連携し結果をマッピングすることで，擬似的に複数のデータベースに対する横断検索を実現する試みがなされてきた (Wilkinson *et al.*, 2003)．

検索 Web ツールの連携による横断検索を実現するためには，ユーザが必要とする結果を返すことができる適切な検索 Web API を見つける必要がある．このために開発された Web サービス用検索システムが Web Service Interoperability (WS-I) による Universal Description, Discovery and Integration (UDDI) (`http://www.w3.org/2001/03/WSWS-popa/paper08`) である．UDDI はベンダーが提供している SOAP サービスのプロトコルや入出力情報などを登録するレジストリであり，ユーザはこのレジストリを利用することで必要なサービスを検索し，容易に連携させることが可能となる．しかしながら，UDDI は普及度の低下などの問題によって現在はサービスが停止されている．

SOAP サービスの衰退や UDDI の停止などから，データベース間の連携手法の主流は REST ベースへと移りつつある．REST に基づいた実装を行うことでデータベース内に登録されている全てのリソースのアドレスを一意の URI で指定できる他，それらのリソースは HTTP や HTTPS にて容易に利用することが可能である．また，URI でリソースを指定できることから，複数のデータベースに跨ったデータの関連性をハイパーリンクによるクロスリファレンスで簡便に表現できることも利点の一つである．ユーザはそのリソースに含まれるハイパーリンクを辿るだけで，そのデータに関連するリソースを容易に収集することが可能である．データベースには，複数のデータ群に対してそれらの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データモデルである Relational Database (RDB) (Codd, 1969) というアーキテクチャが存在するが，このデータベース間の Link 情報を持ってデータの関係性を表現する Linked Data の概念は，複数のデータベース間の関係情報をも容易に扱うことが可能である．

このハイパーリンクによるデータベース統合の利便性をより高めた概念として，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 が存在する．WWW 上のコンテンツの多くは Hyper Text Markup Language (HTML) にて構造化された形で記述されているが，その中に含まれる個々の単語やドキュメント，コンテンツなどに対する詳細な意味情報を記述することはできない．Semantic Web では，これまで扱ってきたリソース単位だけでなくリソース内に含まれる個々のオブジェクトにまで URI を割り振りリソースとして扱う．さらに Web Ontology Language (OWL) によってリソース間の Link やそのリソース自体に意味情報 (セマンティクス) を付与しドキュメントの構造化と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のアプローチが可能になる．これらの特徴から，SOAP などで行われていたメタデータによるリソースのマッピングを可能にしつつ，REST の簡便な手法で利用することが可能な，新たなリソース統合の概念として注目を浴びてきた．また，Semantic Web ではリソース間の関係性を Resource Description Framework (RDF) というフォーマットで記述する．RDF ではリソース間の関係情報について，リソースを表す *Subject* と *Object*，*Subject* に対する *Object* の関係性を表す *Predicate* の 3 要素 (Triple) で表現を行う．そのため，Triple による表記では全てのリソース関係グラフを直接記述することで複雑なデータグラフの表現が可能になるほか，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という利点も存在する．RDF アーカイブに対して検索・操作を行う SQL ライクなクエリ言語である SPARQL Protocol and RDF Query Language (SPARQL) も World Wide Web Consortium (W3C) にて勧告が既に行われており (`http://www.w3.org/TR/sparql11-query/`)，ユーザは RDF に対して柔軟な検索を行うことも可能である．

1.4 生物学におけるリソースの効率的統合

バイオインフォマティクス分野では多くのデータベースや生物情報解析 Web サービスがオンライン上にフリーで公開されている．生命システムという複雑系を理解するためにはこれらの生物学 Web リソースを統合し解析に用いる必要があるが，多領域に渡る生物情報リソースや解析サービスを自身の解析に効率的に組み込むためには，上記のような情報学における事例と同じ問題を解決し，それらの Web リソースを自動的・高速・シームレスに連携を行う必要がある．この問題に対して，本論文では解析 Web サービスの効率的連携と生物学 Web リソースの効率的運用という 2 つのアプローチを行った．第 2 章では，数多く存在するバイオインフォマティクス Web サービスをシームレスに相互運用するためのシステム Keio Bioinformatics Web Service (KBWS) について述べる．KBWS は解析 Web サービスに対して統一的なインタフェースでアクセスをするためのラッパーサービスを SOAP プロキシサーバにて提供することで，アクセス方法やメソッドが異なる 42 の Web サービス全てに対して同一の方法でのアクセスを可能にし，他の解析 Web サービスと組み合わせ複雑な解析フローを容易に構築・運用することを可能にした．また，この SOAP サービスに対してアクセスを行う UNIX コマンドラインツール群をローカルツールとの高い相互運用性を持つ European Molecular Biology Open Software Suite (EMBOSS) の追加パッケージとして実装することで，ローカルツールと Web サービスを交えたシームレスかつ複雑な解析フローを容易に構築することが可能となった．第 3 章では生物学 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抽出・取得するプロセスを高速かつ自動的に行うシステム G-Links を構築した．G-Links では Linked Data の概念に基づく ID 変換をベースとしたデータ収集アプローチに遺伝子集約型のデータ統合モデルを適応することで多領域生物学情報に対して効率的な統合運用を実現した．ユーザは任意の遺伝子 ID を含む簡単な 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でそのリソースを取得することができる．さらに多数の Web サービスについてインタフェースの標準化を行った KBWS を内包することで，生物学的データベースから得られる情報だけではなく，解析 Web サービスによって得られる結果までも含んだ生物学 Web リソースをシームレスに統合するシステムの構築を行った．

第2章　Keio Bioinformatics Web Services

2.1　背景

2.1.1　Webサービスの一般化と相互運用性

バイオインフォマティクス解析は分野の発展と共に複雑化してきており，複数の解析ツールやアルゴリズムを組み合わせて一つの大きな解析ワークフローを構築・実行する必要がある．また，Webサービスはセットアップおよびメンテナンスコストが不要で，かつ豊富な計算資源を持つWebサーバにて利用できるなどの利点から既に一般的な解析手法の一つとなっており，その数や種類は急速に増加している (Bhagat *et al.*, 2010)．そのため，生物情報解析Webサービスを組み合わせた解析フローを構築することで，研究者はより複雑な解析を行うことが出来る．このWebサービスの連携による高度かつ効率的な解析環境を構築するため，Webサービスの相互運用性が盛んに議論されるようになってきた (Stein, 2002, 2008)．

Webサービス間の高い相互運用性を実現するにあたって重要視されている内容に，ユーザが目的に合致したサービスを効率的に発見するためのプロセスであるService Dsicovery (サービス探索) (Al-Masri and Mahmoud, 2008) と入出力の標準化，サービスインタフェースの問題がある．解析フローをスムーズに動作させるには，フロー内のあるステップの出力データ型と次のステップの入力データ型が一致することが全てのステップ間で必要になるため，実際に解析フローを作成する際に，特定のステップに当てはまるべき最適なツールを確実かつ容易に発見できることは非常に重要な要素である．また上記のような入出力型のマッピングを行う際に，各Webサービスが特定のデータ型を独自の表記方法で記述していた場合その関係性を自動的にマッピングすることは難しく，このマッピングを容易かつスムーズに行うためにそれぞれのサービスの入出力型が標準化された単語セットで表現される必要がある．さらに，同じ内容のWebサービスであるにもかかわらず提供機関ごとにインタフェースが異なるケースが多く，ユーザはベンダごとに利用方法を新規習得しなければならない．バイオインフォマティクスには非常に多くの解析Webサービスが存在し，ユーザはそれらの多様な特徴や特性を加味してより適切なサービスを利用する必要がある．大規模な解析フローを効率的に構築するためには，それぞれのサービスが統一されたインタフェースで提供されることで習得コストを軽減し，新規サービスであろうと容易にフローに組み込めることが重要である．

2.1.2　Service Discoveryとオントロジー

Service Discoveryや入出力型の効率的な管理を実現するためのWebサービス用検索システムはこれまでも多く議論されており，バイオインフォマティクス版UDDIといえるシステムが開発されてきた．BioMoby (Wilkinson *et al.*, 2008) は生物情報版のUDDIとして代表的なプロジェクトの一つである．このシステムでは，入出力データ型に代表される，Webサービスに関連したメタデータを開発者自身に登録してもらい，そのレジストリを管理することでサービス検索能とWebサービスのデータオントロジーの標準化を目指していた．しかしながら，BioMobyではサービス

の開発者が独自の記述方法やオントロジーを用いてメタデータを登録していたため，特定の生物学的概念に複数の表記方法が存在するというというオントロジーの氾濫問題が発生してしまった．この問題に対しては，後継の MOWServ (Ramirez *et al.*, 2010) が BioMoby に登録されているオントロジーの管理とキュレーションを行い整理されたオントロジーセットとレジストリを構築することで解決を行なっている．

myGrid プロジェクト (http://www.mygrid.org.uk) は Service Discovery を実現するための検索プラットフォームに加え，そこに登録された Web サービスを連携して利用するためのワークベンチやフローの共有による再利用などの概念を含めた統合環境を提供するプロジェクトである．BioMoby のオントロジー問題を受け，myGrid ではコアとなるオントロジーセットを定義することで拡散を抑えたリポジトリ Feta (Lord *et al.*, 2005) を開発した．その後継となる BioCatalogue (Bhagat *et al.*, 2010) では詳細なアノテーションの登録や登録サービスのキュレーションなども行うなど，より良質なリポジトリを提供している．また，myGrid による Taverna (Hull *et al.*, 2006; Missier *et al.*, 2010; Oinn *et al.*, 2006) はグラフィカルユーザインタフェースからバイオインフォマティクス Web サービスワークフローを作成し実行することができるソフトウェアであり，複数の Web サービスを簡単かつシームレスに連携させ，解析フローとして扱うことが可能である．このソフトウェアは BioCatalogue と連携しフローに組み込むべき最適な解析サービスを容易に探索・追加することができる他，作成した Taverna ワークフローを共有するサービス myExperiment (Roure *et al.*, 2009) とも連携し，フローの共有の他に投稿されたワークフロー自体を組み合わせることでより巨大かつ複雑な解析フローを少ない構築コストで運用することが可能である．

複数の Web サービスのインタフェースを統一することによる利便性の向上を行った例として挙げられるのが European Bioinformatics Institute (EBI) による EBI Web Services (McWilliam *et al.*, 2009) である．このサービスでは 50 以上の主要なバイオインフォマティクスソフトウェアについて SOAP と REST 双方の API および Web ブラウザから扱えるインタフェースを提供しており，ユーザはそれぞれのツールに対応したアクセス URL やパラメータ名に書き換えるだけで，全てのツールを基本的にほぼ同じ使い方で利用することができる．

2.1.3 EMBOSS によるローカル環境での相互運用性と発見可能性の実現

解析ツールにおける相互運用性の問題に関して，ローカル環境という条件の下で高い相互運用性と発見可能性を実現しているプロジェクトが European Molecular Biology Open Software Suite (EMBOSS) (Rice *et al.*, 2000) である．EMBOSS は配列解析のための UNIX コマンドラインツール群のパッケージであり，Character User Interface (CUI) においてパイプ機能による出力渡しなどの機能を用いることで EMBOSS 内のツールのみならず他の UNIX コマンドラインとも高い相互運用性を示す．また，EMBOSS の特徴の一つとして Ajax Command Definition (ACD) ファイルによって入出力データの型が厳密に定義されていることが挙げられる．ACD 内で定義されたデータ型は，その全てが EMBRACE Data and Methods Ontology (EDAM Ontology) (`http://edamontology.sourceforge.net/`) にマッピングされている．このオントロジーは生物学におけるデータフォーマットだけではなくデータベースや解析手法などにまで対応したオントロジーであり，ツールの入出力だけでなくそのツール自体の分類まで対応できる．この特徴によって，EMBOSS は内包している全てのツールを統一的なオントロジーの下で運用することが可能になる．それに加えて，"*tfm*" (The Fine Manual) による詳細なドキュメント表示やツール検索ユーティリティ"*wossname*"による，最適なツールの高い発見可能性も備えている他，EMBOSS Explorer (`http://embossgui.sourceforge.net/`) や JEMBOSS (Carver and Bleasby, 2003),

wEMBOSS (Sarachu and Colet, 2005), SoapLab (Senger *et al.*, 2003) といった多数のユーザインタフェース拡張を持つという利点も存在する．本研究ではバイオインフォマティクス Web サービスを扱うことの出来る UNIX コマンドラインツール群を，内包されるソフトウェアについて高い相互運用性を持つ EMBOSS のシステムに取り込むことで，Web サービスならびにローカルツールをよりシームレスに連携させるためのシステム Keio Bioinfromatics Web Service (KBWS) (Oshita *et al.*, 2011) を構築した．

2.2 設計と実装

2.2.1 アーキテクチャ

KBWS は既存のバイオインフォマティクス Web サービスに対してアクセスするための中継 SOAP サービスを提供しているプロキシサーバと，そのプロキシサーバに SOAP プロトコルを用いてアクセスを行うことで Web サービスを実行する UNIX コマンドラインツール群という 2 つのパートにて構築されている．プロキシサーバで提供されているラッパー SOAP サービスでは，サポートしている全てのサービスについてインタフェースを統一している．そのため，ユーザはその中継 SOAP サービスを中継することで，42 のバイオインフォマティクス Web サービスに対して SOAP による厳密な入力定義かつ統一されたインタフェースからアクセスすることができる．また，クライアントツールである UNIX コマンドラインツール群は EMBOSS の追加パッケージである EMBOSS associated software (EMBASSY) package として実装されており，EMBOSS の持つ高い相互運用性およびサービス発見能と共に利用することができる．これらのツールは C 言語で実装されており，gSOAP toolkit (van Engelen and Gallivan, 2002) によって SOAP 通信を行なっている．システムのアーキテクチャについて図 2.1 に示す．

2.2.2 KBWS プロキシ SOAP サーバ

KBWS では実際に解析 Web サービスを提供している Web サーバに対してクライアントツールが直接アクセスを行うのではなく，オリジナルの Web サービスに対するラッパーサービスを提供するプロキシサーバを経由している．このようなプロキシモデルを採用することで，REST や SOAP，ブラウザから利用できる Common Gateway Interface (CGI) プログラムなど異なるプロトコルで提供されている Web サービス群に対して，ラッパーサービスという統一的なインタフェースを提供することが可能となる．

また，クライアントユーザに対してツールのメンテナンスコストを低下させるという利点もある．オリジナル Web サービスのインタフェースが変更になったり提供 URL が変更になった場合でもプロキシサーバ側でその変更点に対応することで，ユーザ側がクライアントのアップデートやメンテナンスを行う必要なく常に最新の Web サービスを利用することが出来る．プロキシサーバの動作状況は定期チェックを行なっているため，オリジナルの Web サービスが何らかの理由でダウンしていたとしてもアクセス先の Web サービスをプロキシサーバ側で変更することで，ユーザは同様の解析を同じインタフェースから安定して利用することが可能である．このプロキシサーバは Perl 言語および SOAP::Transport::HTTP モジュールにて実装されている．

また，ユーザはプロキシサーバに対して SOAP プロトコルを用いて直接アクセスし，各種バイオインフォマティクス Web サービスを画一的なインタフェースから利用することが可能である．各サービスのインタフェースは WSDL ファイルにて定義されており，rpc/encoded スタイルおよ

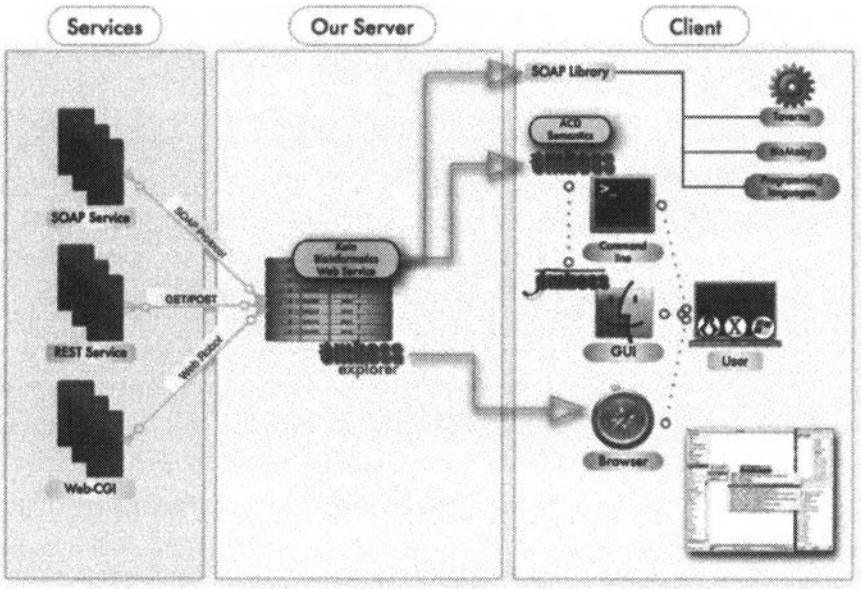


図 2.1: KBWS のシステムアーキテクチャ図

KBWS のシステムアーキテクチャ図を示す．KBWS ではプロキシサーバモデルを採用しており，オリジナルの解析 Web サービスへアクセスするための SOAP プロキシサーバにて，サポートしている全 Web サービスに対して統一的なインタフェースを持つ SOAP サービスを提供する．このサーバへアクセスすることでユーザは多数の Web サービスを SOAP による厳密な入出力定義かつ統一されたインタフェースから利用することができる．さらにこのプロキシサーバにアクセスする UNIX コマンドラインツールを EMBASSY パッケージとして実装することで，KBWS の全サービスを CUI や Graphical User Interface (GUI)，Web ブラウザなど EMBOSS が利用できる様々なインタフェースから EMBOSS の相互運用性とサービス発見能と共に活用できる．

び document/literal 双方の WSDL ファイルが利用可能である．サービスは SOAP1.1 にて利用する事ができ，ユーザは KBWS にて提供されている Web サービスを様々なプログラミング言語，もしくは SOAP サービスを扱うことのできるソフトウェアから利用することが可能である．

2.3 結果

2.3.1 Availability

本ソフトウェアは `http://www.g-language.org/kbws/` にて公開されており，ソースコードは `http://github.com/cory-ko/KBWS` から GNU GPL version 2 ライセンスに基づいて利用することができる．WSDL ファイルは `http://soap.g-language.org/kbws.wsdl` (rpc/encoded) および `http://soap.g-language.org/kbws_dl.wsdl` (document/literal) から利用できる．KBWS プロキシサーバの詳細なドキュメントは `http://www.g-language.org/wiki/kbws/` にて公開されており，プロキシサーバのソースコードは `http://github.com/cory-ko/KBWS-Server/` から GNU GPL version 2 に基づいて利用する事ができる．

2.3.2 利用可能なサービス

本サービスでは 42 のバイオインフォマティクス Web サービスをサポートしており，8 つの REST サービス，3 の SOAP サービス，33 つの CGI サービス，Web サーバにインストールされた 2 つのツールに対して統一的なインタフェースからアクセスが可能である．BLAST サービスに関しては 3 箇所のプロバイダが提供しているサービスについてサポートすることで，より広いターゲットデータベースのサポートと共に特定のサービスがダウンした際のカバーを行なっている．また，後者の理由により幾つかの Web サービスにおいてもアクセス先として複数のベンダの Web サービスがプロキシサーバ側でサポートされているものが存在する．提供している Web サービスのリストを表 2.1 に示す．

2.3.3 KBWS を利用した解析ワークフローの構築

本システムによって実際に複数の Web サービス及び UNIX コマンド，EMBOSS のローカルツールを連携させた解析ワークフローの例としてシーケンスロゴを作成するワークフローを以下に示す．以下のフローではヒトにおける FOXP2 遺伝子 (Enard *et al.*, 2002) の配列について Swiss-Prot (Bairoch *et al.*, 2004) をターゲットデータベースとして BLAST Web Service を実行し (*kblast*)，そこで得られた類似配列の ID のリストを Uniform Sequence Address (USA) 形式に整形 (*sed, uniq*)，MUSCLE を用いて配列のアラインメント (*kmuscle*)．アラインメントされた配列から特定の領域を抽出 (*extractalign*) した後に，その配列を用いてシーケンスロゴを作成する (*kweblogo*)．このように KBWS を用いることで，ユーザは複数の解析 Web サービスやローカルツールを連携させ，より高度な解析フローを簡単に作成し実行することが可能である．ワークフローの例と出力結果を図 2.2 に示す．なお，このワークフローを利用するためのデータベース定義ファイルは `http://soap.g-language.org/kbws/embossrc` より利用することができる．

表 2.1: KBWS にてサポートしているサービス一覧

Category	Name	Reference
ALIGNMENT LOCAL	BLAST	(Altschul *et al.*, 1990, 1997)
	SSEARCH	(Mackey *et al.*, 2002a)
ALIGNMENT MULTIPLE	ClustalW	(Mackey *et al.*, 2002b)
	MAFFT	(Katoh *et al.*, 2009)
	Kalign	(Lassmann and Sonnhammer, 2006)
	MUSCLE	(Edgar, 2004a,b)
	T-Coffee	(Notredame *et al.*, 2000)
NUCLEIC COMPOSITION	WebLogo	(Crooks *et al.*, 2004)
NUCLEIC GENE FINDING	GeneMarkHMM	(Lukashin and Borodovsky, 1998)
	GLIMMER	(Delcher *et al.*, 1999)
	tRNAscan-SE	(Lowe and Eddy, 1997)
PROTEIN LOCALIZATION	PSORT	(Nakai and Kanehisa, 1991)
	PSORT2	(Nakai and Kanehisa, 1991)
	PSORTb	(Yu *et al.*, 2010)
	WoLF PSORT	(Horton *et al.*, 2007)
PROTEIN MOTIFS	Phobius	(Kall *et al.*, 2004)
PROTEIN PROFILES	dbFetch	(Labarga *et al.*, 2007)
RNA 2D STRUCTURE DISPLAY	Centroid Fold	(Hamada *et al.*, 2009; Sato *et al.*, 2009)
	RNAfold	(Hofacker *et al.*, 1994)
MAP TO PATHWAY MAP	Pathway Projector	(Kono *et al.*, 2009)
PHYLIP Tools	PHYLIP	(Lim and Zhang, 1999)

KBWS でサポートをしている全サービスについての一覧を，EMBOSS 内でのカテゴリおよびオリジナルの解析ツールの参考文献情報とともに示す．PHYLIP など一部のサービスは複数のメソッドを含んでおり，10 カテゴリ 42 メソッドの Web サービスについて統一的かつ安定なインタフェースをユーザに提供する．このリストの詳細は `http://www.g-language.org/kbws/` より利用する事ができる．

KBWS による解析フローの例

```
# BLAST を用いて Swiss-Prot に対して検索を行う
% kblast swissprot:FOXP2_HUMAN -d swissprot -format k1 -eval 1e-50 -outfile kblast.out

# BLAST の結果から ID を抽出する
% sed 's/^\(.*\)\.[1-9]/swissprot:\1/g' kblast.out | uniq > match_list.out

# MUSCLE を用いてマルチプルアラインメントを行う
% kmuscle @match_list.out -outfile kmuscle.fasta

# アラインメントされた配列から特定の領域を抽出する
%extractalign -regions '420-430' kmuscle.fasta -outseq extractalign.fasta

# シーケンスロゴを作成する
% kweblogo extractalign.fasta
```

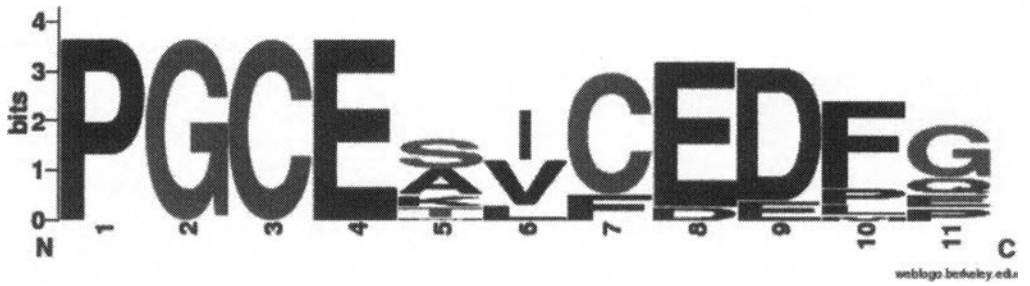


図 2.2: KBWS を用いた解析フローによって作成されたシーケンスロゴ
上記解析ワークフローを用いて生成したシーケンスロゴの例を示す．ユーザは KBWS の各ツール (*kblast, kmuscle, kweblogo*) および EMBOSS のツール (*extractalign*)，UNIX コマンドラインツール (*sed, uniq*) などを適切かつシームレスに組み合わせることで複雑な解析を容易に実行する事が出来る．

2.3.4 Taverna を用いた解析フローの構築と運用

KBWS プロキシサーバは SOAP Web サーバであるため，KBWS にて提供しているサービスは全て Taverna から利用し連携することが可能である．Taverna にて KBWS を用いたワークフローの例として，上記のワークフローと同内容の解析を行うワークフローの実行例を図 2.3 に示す．当該フローは Taverna のワークフロー共有サービスである myExperiment にて公開されており `http://www.myexperiment.org/workflows/1477.html` より利用することができる．

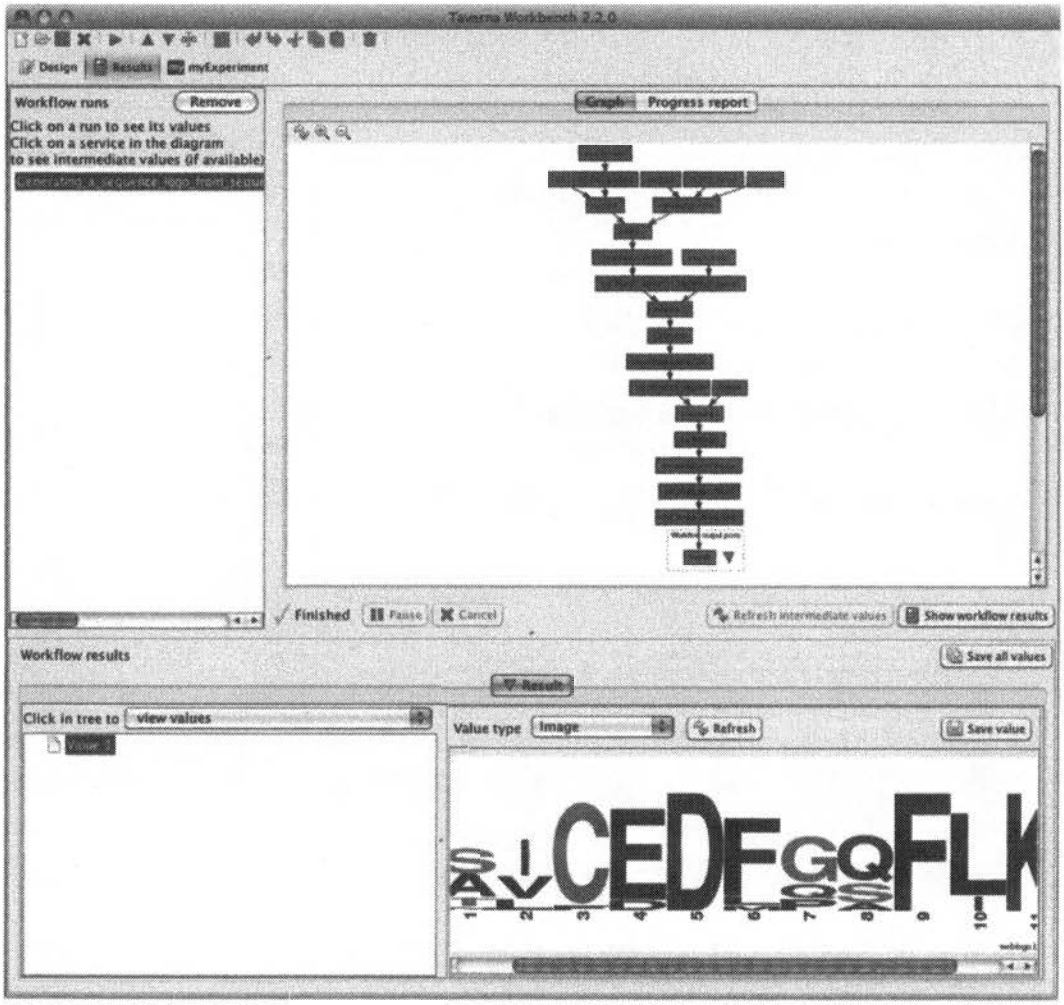


図 2.3: Taverna の実行例
Taverna から KBWS を持ちいてワークフローを構築した例を示す．ユーザは Taverna 内部の解析機能および KBWS, TogoWS を連携することで上記のワークフローと同様内容を解析フローを GUI から簡単に構築・実行する事が出来る．

2.3.5 プログラミング言語からの SOAP プロキシサーバの利用

ユーザは KBWS プロキシサーバに対して SOAP プロトコルでアクセスを行い，必要なサービスを独自のプログラムから言語や環境非依存で利用することができる．利用できる通信方式は SOAP1.1 の rpc/encoded と document/literal の 2 通りであり，ユーザはどちらかに対応している

プログラミング言語およびソフトウェアからアクセス可能である．本プロキシサーバではサポートを行っている全Webサービスについて共通のインタフェースをWSDLによって提供しているため，SOAPサービスのアクセス方法さえ知っていれば各Webサービス間および様々なプログラミング言語からほとんど同じ利用方法でサービスへアクセスすることが可能である．プログラミング言語からの利用例を，バイオインフォマティクスにて一般的に広く用いられているPerl，Ruby，Python，Javaの4プログラミング言語からrpc/encodedを用いてアクセスを行う例を図??に示す．document/literalの利用例を含んだサンプルコードは`http://www.g-language.org/wiki/kbws`から利用することが出来る．

2.3.6 GUIからの利用

KBWSはEMBASSYパッケージであることから，EMBOSSのツールをGUIから利用できるソフトウェアであるJEMBOSSやWebブラウザから利用することが出来るEMBOSS Explorerなどから利用することが出来る．例として，KBWSを登録したEMBOS Explorerを`http://soap.g-language.org/kbws/emboss_explorer/`より利用することができる．実行例を図2.5に示す．

2.4 議論

2.4.1 Webサービス利用のための統一的インタフェースの実装

バイオインフォマティクス解析を行うにあたってWebサービスとローカルツールにはそれぞれ特徴と解析内容に対する適正が存在し，研究者はその特徴を踏まえて最適な手法を選択し，組み合わせて解析フローを構築する必要がある．KBWSでは内部のツール同士や外部のコマンドラインツールとの相互運用性が十分整い，かつサービス検索が容易であるEMBOSSというプラットフォーム上にWebサービスへアクセスできるUNIXコマンドラインツールを載せることで，ローカルツールとWebサービスの垣根を超えた発見可能性の確保およびそれらの融合を実現することができた．

KBWSを用いることで，ユーザは42のバイオインフォマティクスWebサービスについてEMBOSSにパッケージングされたツールと同様に扱うことができる．EMBOSSは既に広く使われている配列解析用ソフトウェアパッケージであり，その非常に高い相互運用性はEMBOSSに内包されるツールのみならずその他のUNIXコマンドラインツールとも非常に高い．EMBOSSに内包されるツール群から目的のツールを検索するシステムやそのツールの使用例まで含めた詳細なドキュメントを確認するためのユーティリティや整備されていることに加え，JEMBOSSやEMBOSS Explorerなどの GUIアプリケーション拡張が多く存在するという利点も持っている．また，EMBOSSおよびWebサービスの採用によるインストールコストの低さに加えて，BLASTなど一部の大規模データベースが必要なサービスに関するセットアップコストの削減，Webサーバという豊富な計算資源の利用など，ユーザに対してより効率的な解析環境を提供している．

KBWSでは中継用SOAPサーバを用いたプロキシモデルを採用している．このモデルの採用により，サポートしているWebサービスの安定性の向上と共に，それら全てを統一的なインタフェースの下で提供することができた．本アーキテクチャを基盤として対応サービスを増加させることでその利便性はより向上し，バイオインフォマティクス解析フローをより効率良く構築することが可能となる．

Perl

```perl
use SOAP::Lite;

# input sequence data (fasta format)
my $seq = `cat nc_000908.fasta`;

# create WSDL driver
my $soap  = SOAP::Lite->service("http://soap.g-language.org/kbws.wsdl");
my $inputParams = SOAP::Data->name("params")->type(map=>{ });

my $jobid = $soap->runGlimmer($seq, $inputParams);

sleep 3 while ( $soap->checkStatus($jobid) == 0 );
print $soap->getResult($jobid);
```

Ruby

```ruby
require 'KBWSDriver.rb'

endpoint_url = ARGV.shift
obj = KBWS.new(endpoint_url)

file = open("nc_000908.fasta")
parameters = {"in0" => file.read, "params" => {}}

jobid = obj.runGlimmer( parameters )

print jobid.s_gensym3

while obj.checkStatus( jobid ).s_gensym3 == 0
  sleep 3
end
puts obj.getResult( jobid ).s_gensym3
```

Python

```python
from SOAPpy import WSDL

wsdl = 'http://soap.g-language.org/kbws.wsdl'
serv = WSDL.Proxy(wsdl)

for line in open('nc_000908.fasta','r')
   file + line

jobid = serv.runGlimmer(file,'')

status = serv.checkStatus(jobid)
while status == 0:
   sleep(3)
   status = serv.checkStatus(jobid)

results = serv.getResult(jobid)
print results
```

Java

```java
import java.io.*;
import kbws.*;

class runBlast {
    public static void main(String[] args) throws Exception {
        KBWSServiceLocator  locator = new KBWSServiceLocator();
        KBWS_PortType serv    = locator.getKBWS();

        FileReader in = new FileReader("/home/kbws/nc_000908.fasta");
        BufferedReader br = new BufferedReader(in);

        String in0 = "";
        String line;
        while ((line = br.readLine()) != null) {
            in0 = in0 + line + "\n";
        }

        BlastInputParams parameter = new BlastInputParams();

        System.out.println (serv.runGlimmer( in0, parameter ) );
    }
}
```

図 2.4: KBWS SOAP プロキシサーバの rpc/encoded による利用例

KBWS の SOAP プロキシサーバに対して Perl, Ruby, Python, Java から rpc/encoded を用いて GLIMMER サービスへアクセスするためのサンプルコードを示す．これらのコードは SOAP1.1 を用いて通信を行なっており，Perl では SOAP::Lite のバージョン 0.60, Ruby では soap4r, Python では SOAPy, Java では Apache Axis のバージョン 1.4 以降をモジュールとして利用している．本サンプルコードに加え document/literal の利用例を含んだ全サンプルコードは **http://www.g-language.org/wiki/kbws** から利用することが出来る．

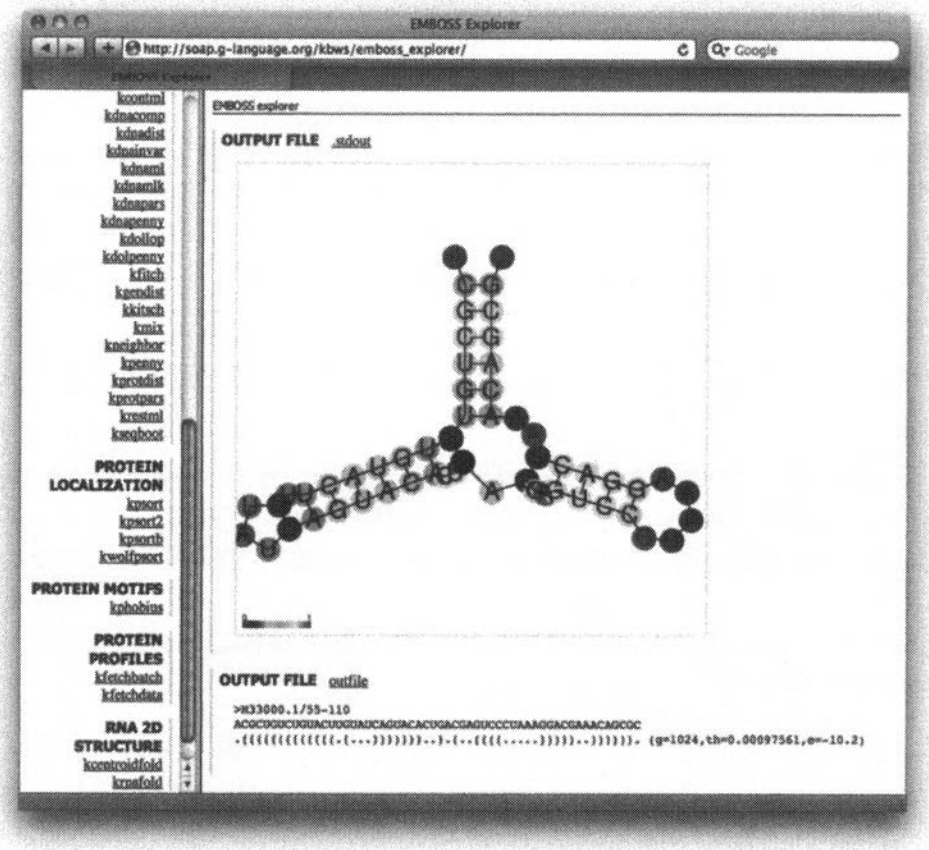


図 2.5: EMBOSS Explorer からの利用例

KBWS を EMBOSS Explorer から用いた例を示す．EMBOSS の GUI 拡張である EMBOSS Explorer や JEMBOSS, wEMBOSS などを用いることで，ユーザは各種 Web サービスを GUI から他の EMBOSS ツールと同様の利用方法簡単に用い，目的の解析を行う事が出来る．KBWS の機能を実装した EMBOSS Explorer は `http://soap.g-language.org/kbws/emboss_explorer/` より利用する事が出来る．

2.4.2 より高度なService Discoveryの実現

EMBOSSの基本的なインタフェースはUNIXコマンドラインである．このインタフェースの特徴は対話型であることであり，ユーザは直前のツールによる出力結果の内容を確認し，その上で次に使うべきツールを決めたり適切な結果が出力されるようにパラメータを調節することで，試行錯誤を繰り返しながら適切なワークフローを構築することができる．そのためKBWSも同じく，フローの再利用よりもWebサービスとローカルサービスを統合した解析フローのトライ・アンド・エラーによる効率的な構築を重視したツールである．この試行錯誤型のフロー作成をより効率的に行うためには，直前のツールの出力データ型を入力として受け付けることが可能かつユーザの目的を達成できるツールを発見し，その解析に最適なパラメータなどを指定する必要がある．KBWSでもこれらの機能はサポートしているが，*tfm*によるドキュメント閲覧や*wossname*によるキーワードベースのツール検索のみであり，最終的には自力で最適なツールとパラメータの組み合わせを見つけ出す必要がある．この問題に対して，より強力なサービス検索能にて解決を行ったソフトウェアも存在する．Seahawk (Gordon and Sensen, 2007) はGUIからWebサービスを導出できるソフトウェアであり，入力として指定しているデータのタイプや属性情報といったコンテキストからそのデータに対するツールを検索し推薦してくれるシステムを備えている．そのため，ユーザは自身が興味のあるデータに対して行うことができる解析リストから興味のあるツールを選択するだけで解析フローを作成できる．Magallanes (Rios *et al.*, 2009) やjORCA (Martin-Requena *et al.*, 2010) はサービス推薦機能をより向上させたプロジェクトで，入力データタイプと最終的にユーザが求めている出力データタイプを指定するだけで，その入出力データの間で利用すべき解析ツールを自動的に判断しワークフローを作成する，経路探索的なWebサービスワークフロー自動構築システムである．KBWSの利点として*sed* (Stream EDitor) や*cut*のようなUNIXコマンドを用いたバッチ処理やローカルツールとの高い相互運用性はこれらのツールにない利点ではあるが，多くのツールやデータタイプのセマンティクスやメタデータを一元的に管理し，指定されたデータタイプを入力とするツールを抽出し推薦するといった，より高度な発見可能性の確保は重要な課題の一つである．

第3章　G-Links

3.1　背景

3.1.1　生物学データベースの統合的利用

生物学における実験技術や機器の急速な高度化によって，バイオインフォマティクス分野ではそれらから得られる大規模なデータを効率的に管理するためのデータベースが広く発展してきた．生物学で扱われるデータの量や種類の増加に呼応するようにデータベース自体の数や内包するデータ量も爆発的に増加しており，現在では数千にもおよぶ生物学データベースがオンラインでオープンに公開されている (Bhagat *et al.*, 2010)．利便性向上のために多くのデータベースで検索ユーティリティの整備なども行われているほか，外部データベースのエントリーと Link によって結ばれている Linked Data の形式であることが知られている．そのため，研究者は複数の大規模データベースから検索ツールで必要なデータを取得，自由に利用できるデータアーカイブで自身のプログラムによる必要なデータ処理を行うことで複数のデータベースを横断的に利用し，より詳細かつ大規模な解析を行うことが可能である．特に Linked Data 形式であるという点はデータベースの横断利用において大きな利点である．研究者はハイパーリンクおよびクロスリファレンスというエントリー間の分かりやすい関係情報を辿るだけで，対象の生物学オブジェクトに関する分子情報やアノテーションなどを取得することができる．Link 情報が有用であるもう一つの理由が生命情報の多層性である．生命システムはゲノムからトランスクリプトーム，プロテオームと様々なレイヤーからなる非常に複雑なシステムであり，それらの情報が様々なデータベースに分散して存在してる．特に WWW 上にオンラインで公開されているデータベースが多い生物学分野では，データベース提供者はそのエントリーに関連する情報をハイパーリンクを張るだけで表現し，ユーザはそのリンク先へジャンプするだけでその情報を閲覧することが可能である．

このように，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) ．その最も先駆的な例が，Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2012) における dbget (Fujibuchi *et al.*, 1998) である．dbget は生体内の代謝パスウェイのデータベースに対して遺伝子の情報を結びつけることで 2 層からのアプローチを可能にした KEGG における非常に優秀な検索ユーティリティであり，特定の遺伝子に関連する代謝経路などを検索することでその遺伝子の生体内での機能についてのデータを直接を得ることが可能である．このように，多領域の生物情報をつなぎ合わせることで研究者は複雑な生命システムに対してより有効なアプローチをとることが可能である．

3.1.2　データベースの統合とそれに伴う問題

生物学データベースを統合的に利用することによって，研究者は多くのメリットを得ることができる．第一に，複数のデータリソースから広くデータを取得できるため，より多くのデータを

用いた解析を行うことができる．解析に用いるデータ数が増加することで，解析自体の正確性が向上する以外にも，生物種などの解析対象が増えることでより広範囲の生命現象についての解析が可能になる．それに加えて，遺伝子に関するデータベースとタンパク質に関するデータベースなど扱っている生物情報のレイヤーが異なる複数のデータベースを統合することで，研究者が対象としている生物学的現象に対して多面的な解析を行うことも可能である．遺伝子間相互作用がもたらす表現系への影響などより創発的な生命システムの分析が可能であるため，その利点は大きい．

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になっている．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数千にも及ぶ生物学データベースから収集し，2. シンタックスおよびセマンティクスの異なる複数のデータベースから得られたエントリーを研究者が使いやすい形で統合し，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つの作業に占められており，ここで作成されたデータセットを用いて生物学的な知見を生み出す必要がある．さらに近年の解析Webサービスの台頭，特にRESTサービスの台頭により，データベースのみならず，Webサービスによる解析結果もデータベースの各エントリーと同じくURIにて指定可能な生物学リソースの一つと見なすことができるようになってきている．TogoWS (Katayama *et al.*, 2010b) のRESTサービスはその代表例の一つである．いくつかの主要な生物学データベースについてデータベース名とそのデータベースにおけるIDを含んだURLを指定することでそのIDが示すリソースを取得できるほか，そのリソース内の必要なセクションだけを一意のURLで指定しデータの抽出を行うことができる．これらの理由から，真に生物情報を統合するためには，データベースとあわせて生物学Webリソース全体の統合が必要であるといえる．この現状を解決するための，データ統合の作業を自動的かつ効率的，高速に行うシステムの構築はバイオインフォマティクスにおける重要課題の一つである．

3.1.3 データベース単純結合の問題点

これらのデータ統合問題の一番単純な解決法は統合型データベースの構築であるが，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ことは序論で触れた通りである．生物学におけるデータ量と種類の爆発的増加は，それを統合した巨大なデータアーカイブに対する検索や抽出，閲覧などの再利用性を確保するための膨大な計算資源を要求する．また，実験技術の進歩や新規概念の発見などによって生物学で扱われるデータの種類が増加していることも大きな問題である．現在の生物学ではLinked Dataモデルにより，新規概念に対応したデータベースやその概念を表す新規データベースに対してLinkを張ることでスキーマの変化に柔軟に対応しているが，統合データベースを作成した際にはデータベースのスキーマをその都度変更する必要がある．それに加え，スケジュールがバラバラな各データベースの情報更新にあわせてメンテナンスを行う必要があるなど運用コストが高い．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．

3.1.4 先行研究

生物学データベースを単純に統合したデータアーカイブを扱おうとすると，その規模や複雑さに起因する多数の問題が起きてしまう．そのため，生物学ではユーザが目的とするデータを持つデー

タベース群に対して同時に検索ツールによるデータ取得を行いその結果を統合する，Federated Query (Jacso, 2004) 型データ統合の試みがなされた．データ全てを結合して用いるのではなく，各データベースに対する検索サービスを連携させ結果だけを結合することでデータベースの統合的利用を実現するため，基本的な検索クエリ自体は個別のデータベースに対するものと同等になる．そのため，全てのデータベースを単純統合した際におこる計算資源やメンテナンスコストの問題は比較的小さいといえる．複数の異なるデータベースから得られる異なる出力の統合のため，この手法ではSOAPにて提供された検索ツールのWeb APIを連携対象として用いるケースが多い (Wilkinson *et al.*, 2003)．環境やプログラミング言語に非依存で確実な通信ができるSOAPサービスを用いることで複数のデータベースと安定した通信が可能になるほか，入出力がXMLでマークアップされており通信されるデータの属性情報を扱うことができるため，メタデータによる入出力データのマッピングを行うことで他検索サービスの結果とシームレスな連携が可能となる．このようにSOAP APIベースの検索ツールによる統合は，各データベースに対する検索サービスを生物情報解析Webサービスの一種であるととらえ，それらを連携させた解析ワークフローを構築・運用することで複数データベースにまたがった複雑なデータ取得クエリを実現している．現在，この研究は第二章で述べたBioMoby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おり，オントロジーの管理によるデータスキーマの統一や，Webサービスのレジストリ検索によるユーザの目的に合致したデータベース検索サービスの発見可能性の確保などが実現されている．

この他に，ユーザが必要なデータを持つデータベースだけを単一システムに落とし込み統合型データベースを構築するアプローチをとったプロジェクトも多く存在する．National Center for Biotechnology Information (NCBI) (authors listed, 2013) のEntrez (Schuler *et al.*, 1996) はNCBIに内包される全てのデータに対する統一的な検索インタフェースである．EntrezではNCBIの全ての検索結果に対して人が理解しやすい大きなデータカテゴリを定義し，その枠の中で検索結果を提供する．これはデータ統合において必要な作業であるデータの”抽出”にあたり，カテゴリを用いて大きくデータ分類を行うことで多領域にわたる生物学情報の中からユーザが必要とするデータを直感的に絞り込むことを可能にしている．また，複数のデータベースを一つのスキーマにまとめる作業を支援する環境を提供するプロジェクトとしてBioMart (Kasprzyk, 2011) が存在する．主要データベースに関するID対応表を取得するなど既存のデータベースの連携を支援するだけでなく，手元のデータを既存のデータベースに統合するなどの機能を持つため，統合して運用したいデータセットがある程度しぼられている場合には，それら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優れたプロジェクトである．

3.1.5 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．バイオインフォマティクス分野におけるデータベースはその多くがオンラインかつ公的に提供されており，それぞれのエントリー間がクロスリファレンスを用いてデータベース間の関係性の表現をしているLinked Dataの形式をとっている．そのため研究者はこのLinkによって構成させるネットワークを辿ることで，現在着目しているエントリーと関連性のある，異なるデータベース上の別エントリーにたどり着くことが可能となるのだが，この行為は各エントリーを指し示すIDについて関連性のあるIDへの変換作業と同値であるといえる．このため，Linked Dataによって形成された関連性の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，ユーザは生物学リソースを統合することが

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の問題である．この問題に対して単純に全てのデータネットワークを扱うのではなく特定の情報を中心として構造化したネットワークを再構築することで対応したプロジェクトが多数存在する．SOURCE (Diehn *et al.*, 2003) やMatchMiner (Bussey *et al.*, 2003)，DAVID gene ID conversion tool (Huang *et al.*, 2008, 2009) などは遺伝子情報に特化したID変換システムであり，逆にタンパク質を中心としたID変換システムとしてProtein Identifier Cross-Referencing (PICR) (Cote *et al.*, 2007) があげられる．これらのシステムは遺伝子やタンパク質それぞれに基準をおいたID整理を行っているため，比較的スリム化され必要な計算資源が少ない高速なシステムとして動作することが可能である．また，もう一つの形として *Homo sapiens* に特化したHyperlink Management System and ID Converter System (Imanishi and Nakaoka, 2009) のように特定の生物種に特化したID変換システムなども存在する．また，横断的データ取得というタスクをより純粋なID変換の問題に落とし込んだシステムがbioDBnet (Mudunuri *et al.*, 2009) である．bioDBnetはユーザから受け取ったIDをそのIDに関連する他IDに変換する部分と，得られたIDに関するデータ取得部分のネットワークを分離を行った．そして前者を各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として構築し，後者のデータ取得部分と切り離す事でデータベース横断検索部分の高速化を実現している．生物学データベースの統合問題をLinked Dataネットワーク上での単純な経路探索問題に落とし込むことに成功したシステムである．また，遺伝子中心型のID変換をベースにしたデータ収集用システムの例として挙げられるのがMyGene.Info (`http://mygene.info`) というWebサービスである．このWebサービスは遺伝子IDや遺伝子を表すシンボルをクエリとして，その遺伝子に関連する情報をID変換を用いて高速に収集し取得することができるRESTfulなWebサービスである．RESTfulなインタフェースのため全てのリソースを一意のURIで指定可能であるほか，JavaScript Object Notation (JSON) など多数のフォーマットで出力が可能なため，単純にデータ取得のプロセスで利用するだけではなくCGIでの利用やWebページへの埋め込みなどWebアプリケーションを開発する際のデータリソースとして活用することも可能である．パラメータを設定することでリソースから必要な情報のみを切り出して出力することも可能であるため，複数のデータベースからの収集，統合及び抽出というプロセスをID変換のアプローチにて解決した研究の一つであると言える．現在，MyGene.Infoは9種類のモデル生物に関するデータ取得に対応しておりBioGPS (Wu *et al.*, 2009) では実際にデータ取得のためのバックエンドとして利用されている．

これ以外のアプローチとして，全ての生物学リソースに対してユニバーサルなIDを付与するというプロジェクトも存在する．上記のようなID変換をベースにしたリソースの結合・取得は生物情報の特徴に非常に合致したアプローチであったが，幾つかの問題が存在した．第一にリソースの存在保証の問題が挙げられる．ID変換のアプローチで扱われるデータは各データベース内のIDであり，そのIDが指し示す生物学情報にアクセスするためにはそのデータが存在するロケーションの情報を含んだURIを生成する必要がある．リソース自体をID単体で指し示す事ができないため，そのIDが示すエントリーが確実に存在しかつそこに含まれる情報が正しいものかどうかの保証をすることができず，各データベース毎にIDを用いてリソースを表すためのURIスキーマが異なりそれらを管理する必要があったり，一部のデータベース間においてID重複が起こっているなどの問題点があった．これを解決するため，全ての生物学情報に対してユニバーサルなIDを付与する目的で開始されたプロジェクトがLife Science Identifier (LSID) (Clark *et al.*, 2004) である．LSIDでは，既存のデータベースにおけるIDにそれぞれ対応するユニバーサルな

IDを用いることでIDの一意性を保証すると同時に，ロケーション情報の解決をDomain Name System (DNS) によって解決していた．LSIDは情報の場所を直接示すURLではなく，情報がある場所の名前を用いて表現するUniform Resource Name (URN) の形式を用いることで，ID内にロケーションの情報を含んでいる．ここで指定された名前に対して，その名前が指し示すロケーション情報とURLのスキーマをDNSサーバにて解決することで，DNSサーバさえ最新であればユーザは正しい情報を確実に取得することが出来るようになる．ユーザはLSIDを扱うシステムにてLSIDを指定するだけで，常に正確な情報を確実に取得することができた．

3.1.6 ID変換の持ちうる問題点

上記のLSIDの項で触れたように，IDはデータベース内のエントリーを指し示すデータであり，生物学データそのものではないという大きな特徴が存在する．ID変換のアプローチでは各エントリーを示すポインタとその間のLink情報のみを取り扱うため，データベース全体を取り扱うケースと比較して高速にデータの統合的利用のための処理を行うことができる．また，Linkによる関連情報によってデータベース間の関係性を表現できるため，複数のデータベースをデータスキーマの問題なく横断的に利用することが出来ることも大きな利点として挙げられる．しかしながら，IDが生物学データそのものではないという点は大きな欠点にもなりうる．ID変換によって得られるデータはあくまでIDのリストであり，実際に解析を行う際にはそのID群が指し示す生物学リソースを取得し統合するプロセスが必要となる．また，Linkが内包する意味情報を機械的に扱う事ができないという点も問題の一つである．ハイパーリンクやクロスリファレンスによるLink情報は2つのIDに何かしらの関連が存在することを表現してはいるが，その関係性がどのような関係なのかというLinkの意味情報までは表現できていない．Web上で情報を閲覧する際にはそのLinkが表す意味 (セマンティクス) を人が判断してLink先に飛ぶことができるが，機械的な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情報だけを何らかの基準をもって選別する必要がある．また，データベースのメンテナンスやURLの変更などによってIDが指し示す情報への到達能が必ずしも保証されておらず，それを保証するシステムが別途必要という点も大きな問題点として挙げられる．このような，URLのようなロケーション情報を直接保持していないという問題が大きく関わったのが，上記のLSIDである．LSIDでは全てのリソースに統一的なIDを指定した際に，そのロケーションの問題をURNとDNSによって解決を行った．IDに名前情報を含んだURNを採用しそこで指定された名前情報とロケーション情報の変換をDNSにて行うという解決方法は非常に有用であったが，DNSを経由しなければならずHTTPのみで完結しないという複雑さが非常に大きな問題となった．この問題に対する代表的な解決策が，Online Computer Library Centerが開発を行ったPersistent Uniform Resource Locators (PURL) (Shafer *et al.*, 1996) である．PURLはHTTPのみでロケーション問題を解決し永続的なURIをユーザに提供するためのアーキテクチャであり，PURLサーバに予約されたURLへのアクセスがあった際に，任意のURLへリダイレクトを行うことで正しいURIをユーザに提供する．例としてUniProt (The UniProt Consortium, 2012) は自身のWebページにおける外部データベースへのハイパーリンクを示すためにPURLサーバを提供しており，UniProtから利用できる全ての外部リソースはそのデータ取得保証がなされている．

3.1.7 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 Semantic Web (`http://www.w3.org/2001/sw/`) と呼ばれる概念である．Semantic Web は WWW の利便性を向上するためのプロジェクトであり，従来の ID が示していたエントリーについて，そのドキュメント内に含まれる個別のリソースに対しても URI を割り振ることでより粒度の小さいレベルでリソースの関連情報を扱うことができる．また Semantic Web では，リソースを表す *Subject* と *Object*，*Subject* に対する *Object* の関係性を表す *Predicate* の 3 要素でリソース間の関係情報を表す Triple というメタデータモデルを用いており，その抽象構文である RDF を用いて全てのリソース関係グラフを直接記述できる．そのため複雑なデータグラフの表現が可能である他，スキーマレスであるという大きな利点が存在する．さらに全ての Link に関するセマンティクスを機械的に扱うことができるため，WWW 上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のアプローチが可能となる他，WWW は主に HTTP にて通信を行うため，その上で実現される Sematic Web では PURL を用いることでロケーションの問題も解決できるなど非常に利点が多い技術であると言える．

これらの，主にスキーマレスという特徴は生物学にとって非常に有用な利点であり，現在は国際的に Semantic Web への移行の流れが生まれている．Bio2RDF (Belleau *et al.*, 2008) はその代表的な例であり，全ての生物学リソースを RDF にて表現し統合することを目的としたプロジェクトである．RDF リソースのホスティングのみならず主要なデータベースに関する統合を行っており，多くのデータリソースについて RDF 形式にてデータを取得することが可能である．RDF クエリ言語である SPARQL に対する endpoint なども利用可能であるなど，バイオインフォマティクス Semantic Web における非常に重要なプロジェクトの一つである．また，BioHackathon (Katayama *et al.*, 2010a, 2011) も Semantic Web への流れを象徴するものの一つである．BioHackathon はバイオインフォマティクスにおけるデータおよび Web サービスを統合するための議論および開発を行うための国際開発会議である．多数のソフトウェア開発者が Semantic Web の技術を用いたアプリケーション開発や要素技術の検証，バイオインフォマティクスが目指すべき方向性などについての議論やその場での実装などを行っており，これまでも多くの成果を上げている．また，Semantic Web による利点をよりうまく利用しデータ統合を行なっているプロジェクトが UniProt (The UniProt Consortium, 2012) である．UniProt はタンパク質に関する包括的なデータベースであり，タンパク質配列を中心としてそれに関連する分子情報や機能性アノテーションを，外部データベースの ID およびその ID が示すエントリーへの Link の形で管理している．これらの内部データは UniProt 独自の Ontology である UniProt Core Ontology (`http://purl.uniprot.org/core/`) でセマンティクスを表現した RDF の形式で管理されているのだが，UniProt ではこの RDF アーカイブに対して意味推論を用いて矛盾したアノテーションがないかの自動検出を行い，そこで矛盾が生じた部分に関して人が目で見てアノテーションの修正を行うという半自動化パイプラインを構築している．また，全ての外部データベースへのリンクを PURL によって表現することで Link 先のデータの存在も保証されている．このようなシステムを用いることで，UniProt では大量のデータに対して少ない労力で質の高いアノテーションの提供・維持を行うことに成功している．他にも，生物学シミュレーションモデルに対するアノテーション情報の標準化を行っている Minimal Information Required In the Annotation of Models (MIRIAM) (Le Novère *et al.*, 2005) は，外部データベースのリソースへのリファレンスに対して URN を用いることで指定されたリソースの永続性を保証することに成功しているプロジェクトである．MIRIAM では生物学シミュレーションモデルを構成する各要素について，それらの生物学オブジェクトを表しうる主要なデータベースのリソースを指し示す MIRIAM URN を構築．そのロケーション情報を解決するためのフ

レームワークである Identifiers.org (`http://identifiers.org`) (Juty *et al.*, 2012) を用いることで，その URN が指し示すリソースを取得することができる．Identifiers.org が提供する Web API では PURL と同様 HTTP 通信のみで名前解決を行うことが可能であり，MIRIAM URN から容易に生成できる URI でリソースを指定できる他，その ID が利用できるデータベースが複数存在する場合はその全てをユーザに提供することでより確実な名前解決を行う．ガイドラインにそって提供される整理されたアノテーションを，その永続性を保ったまま URI で記述できるため，RDF 内でそのまま活用できるなど Semantic Web や Linked Data に適した形で永続的なリソースを提供するサービスである．上記のような生物情報の統合ではなく Web 解析サービスの統合に Semantic Web の技術を押し進めたプロジェクトが SADI (Wilkinson *et al.*, 2010, 2011) である．SADI では各解析 Web サービスについて入出力データ型やデータの種類，対象 Web サービスのカテゴリなどのメタデータ情報を RDF で管理し，解析 Web サービスについての Linked Data ネットワークを構築．それに対して意味推論を行うことで入出力データ型のマッチングによる経路探索に対して，各ステップで扱っているサービスの意味情報まで加味した，より高度で複雑な解析ワークフローを生成・実行する．SADI ではこの仕組みを実装するための意味推論エンジンやワークフローを実行するためのフレームワーク，サービスのメタデータを管理するためのレジストリなどの環境，これらの機能を扱う複数プログラミング言語用のライブラリなどの環境が整っており，Semantic Web をベースにした解析フロー自動生成の先駆者的な研究である．

3.1.8 Semantic Web が抱える問題点

このような利点がありながら，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ものは未だ生物学では数えるほどしか存在しない．これには Semantic Web が抱えるいくつかの問題が関係している．第一の問題が処理ノードの増加と計算資源の限界である．Semantic Web ではそれぞれの ID が示すリソースに含まれる個々の情報も 1 リソースとして扱う必要があるため，ID 変換のアプローチに対して扱うべきリソースの粒度が非常に細かくなる．扱うべきノードの増加はノード間に貼られる Link によるネットワークの複雑性の爆発的増加をもたらすため，そのネットワークに対する経路探索や意味情報処理に必要な計算資源は莫大なものになる．ID 変換の時点で計算時間によるレイテンシの問題が提起されていたことを考えると，現在全ての情報を単純に Semantic Web の上に乗せて運用することは現実的ではないと言える．また，Semantic Web 上でデータセットを扱うためにはそのリソースの関係性を RDF で記述している必要があるが，既存のデータセットを RDF 化するための労力の大きさも問題の一つといえる．より効率的な意味推論を行うためには，データの提供者自身が自身のデータセットの Link 情報の整理を行い，その Link を表現するために適切なオントロジーのセットを選定し，そのオントロジーを用いて全ての Link に意味情報の付与を行い，それを正確な RDF のフォーマットに整形するというプロセスを経る必要がある．この RDF 化をより効率的に行うための技術開発および議論は BioHackathon などでも進められているものの，全てのデータ提供者に正確かつ適切なオントロジーで記述された利用価値の高い Link 情報を持った RDF の生成を求めることは難しい．さらにデータセットによるオントロジーの非一致も非常に大きな問題の一つとなっている．Semantic Web で意味情報を扱うためには，特定の概念を表現する語彙のセットであるオントロジーから適切な用語を用いて各リソースおよびリソース間 Link のメタデータを記述する必要があるが，ここで表現された概念情報に対して機械的な処理を行うためには，特定の概念を表現する単語を統一する必要がある．例として塩基配列という概念を表現する場合，その概念を表現する単語として *Nucleotide* と *Nuc* という 2 つの単語での表現が混在しているとコンピュータはこの 2 つの単語で示されたリソース

が同一概念のリソースであると判断することができず，より正確な処理のためにはどちらか一つの用語に統一する必要がある．そのため，統合する複数の RDF では使用するオントロジーが統一されている必要がある．オントロジー間の関係性を示すオントロジーを用意する，意味推論でオントロジー間の関係性を推測するなどの手法でオントロジーの問題は理論的には解決できると言われているが，その推論に必要な計算量を考慮すると，計算資源とレイテンシの問題がすでに指摘されている Semantic Web において現段階では現実的であるとは言えない．

3.2 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．生命システムはゲノムやトランスクリプトーム，プロテオームといった多層構造になっており，これらの情報が複雑に相互作用し合うことで構成されている．この複雑さ故にバイオインフォマティクス研究では一つの生命現象について解析を行う際にも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のだが，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，単純に Link を張るだけでデータベース同士の関係性を表現する Linked Data とその Link ネットワークを用いた ID 変換のアプローチ，ならびにその Link に意味情報を付加した Semantic Web の技術である．本システムでは Semantic Web が持つレイテンシの問題の解決や，生物学データベースが元々密な Linked Data ネットワークを構築しているという特徴，データ収集をネットワーク上の経路探索という問題に落とし込めるなどの理由から，ID 変換をベースにしたシステムを構築することを目的としている．

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．

(1) 出力可能な情報の網羅性

生命システムはゲノムやトランスクリプトーム，プロテオームといった多層構造になっており，これらの情報が複雑に相互作用し合うことで構成されている．バイオインフォマティクス研究ではこの複雑なシステムに対して分析を行うため，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．そのため，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．

(2) 汎用的な入力系

生物学データベース自体の数の増加に伴って，研究者が扱わなければならない ID の総数は爆発的に増加している．また，例えば同じ遺伝子を扱う ID であっても選択的スプライシングによって複数の転写物が生成される遺伝子やオーソログなど複数の遺伝子セットを扱う ID など，他の

IDと一対一対応がとれず単純な相互変換が不可能なIDも存在する．ID変換のみのサポートでは，そもそも対応するIDに変換されていないリソースなどについてユーザが関連情報を取得することができないという問題も存在する．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．

(3) ID の持つロケーション問題の解決

LSIDでも問題になった項目の一つがこのロケーション問題の解決である．一般的にIDだけではそのIDが示すリソースを取得することができず，そのIDに対応するURIへ何らかの方法で変換を行わなければならない．そのため，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，何らかの方法でその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．

(4) ID 情報以外のリソースの取得

ID変換は高速に関連リソースを収集できるアプローチの一つであるが，得られる情報がIDに限られるという大きな問題がある．研究者が自身の解析にて用いるのはID情報ではなくそれが示す生物学リソースであり，基本的にIDの集合を取得しただけでは，ユーザはそのリソースに関する生物学的な知見を得ることはできない．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．

(5) リソースの厳選

生物学データベース間のLinkネットワークは非常に密であり，そのLinkを辿ることで広範囲のリソースを収集することが可能である．広いリソースを単純にLinkを辿って得られた情報全てを収集するとより多くのリソースを解析に用いることができるため，サンプル数の増加や多種類のリソースを用いた解析を行うことができるというメリットを得る事が出来る．しかしながら，その中には重複情報やユーザにとって必要のない情報，メンテナンスがされていない古い情報などが多く含まれている可能性が高い．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．

(6) 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．さらに，このデータ統合プロセスをより高い利便性のもとで実行するため，本システムはどのような環境からでも容易に利用できる必要がある．

(7) 他サービスとの相互運用性

本システムで得られた多領域生物学データセットについて，研究者はそれに対する解析作業を行うことで生物学的な知見を抽出する．そのため，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．

3.3 設計と実装

3.3.1 アーキテクチャ

これらの問題を解決するため本論文で構築したシステムである 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 変換のアプローチ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．このシステムを構築するための大きな問題がレイテンシと汎用入力系の両立である．生物学データベースにおけるクロスリファレンスの Link ネットワークは各データベースにて扱われるデータの種類や Link 自体の数の多さ故に非常に複雑であるため，汎用性の実現のために多数のデータベースの ID をサポートした場合，すべての Link ネットワークに対して単純に経路探索をすることによるデータ収集を高速に行うことは難しい．そのため，ID 変換をベースとした高速かつ自動的な関連エントリーの収集を実現するには，何らかの方法でネットワークを整理し効率的な処理基盤を生成する必要がある．G-Links ではこの問題に対し，Linked Data ネットワークにおいて Primary Key を設定し，その Primary Key を中心としたネットワークを整理することで解決を試みた．

Primary Key として扱うべき情報を選定する上で大きな問題点となるのが情報の多領域性であり，Primary Key はそれを中心とすることでゲノムやトランスクリプトーム，プロテオームなどの多層に渡る情報空間を効率的に統合できる必要がある．そのため，G-Links ではセントラルドグマの考え方をベースにすることで遺伝子を示す ID を Primary Key として用いることとした．セントラルドグマの考え方では，遺伝子に保存される全ての遺伝情報が転写翻訳などのプロセスを経ることで，メタボロームや代謝経路ネットワークといった層へと伝播していく．このことから，全ての生物学的情報は遺伝子情報を中心に統合できるのではないかと考えた．また，多数存在する遺伝子を表す ID から実際に運用する Primary Key を選定するにあたって，G-Links では UniProt ID を Primary Key として採用した．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を持っており，2012 年 12 月現在で 28,934,417 エントリーという十分量のエントリー数を保有している (The UniProt Consortium, 2012)．さらに，内部データを RDF で管理しているため Semantic Web 技術との親和性も非常に高く，上記で触れた RDF に対する意味推論によるアノテーション管理によって非常に品質の高いクロスリファレンス情報を保有しているという大きな利点も持っている．Linked Data ネットワークにおいてハブになりうる十分な量のクロスリファレンスを保持している他，その 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と言える．

この Primary Key への情報集約型の ID 変換システムを構築するにあたり，内部データベースは bioDBnet (Mudunuri *et al.*, 2009) と同様に，遺伝子を示す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．このような分離を行うことで ID 変換部分で扱うデータ量を減らす事ができるため汎用的な入力に対する ID 解決のレイテンシを小さく保つことがで

きるほか，アノテーション取得テーブルにおいてUniProt IDをインデックスとした転置インデックスを用いることで，内部で扱うアノテーションデータの量の増加に対するスケーラビリティを持たせることが可能になる．また，本システムのメイン部分および内部データベースのメンテナンス用スクリプトはPerl言語で構築されており，ID変換部とアノテーション取得部ではMySQL 5.0を用いたRDBを利用している．内部データベースの全データは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．

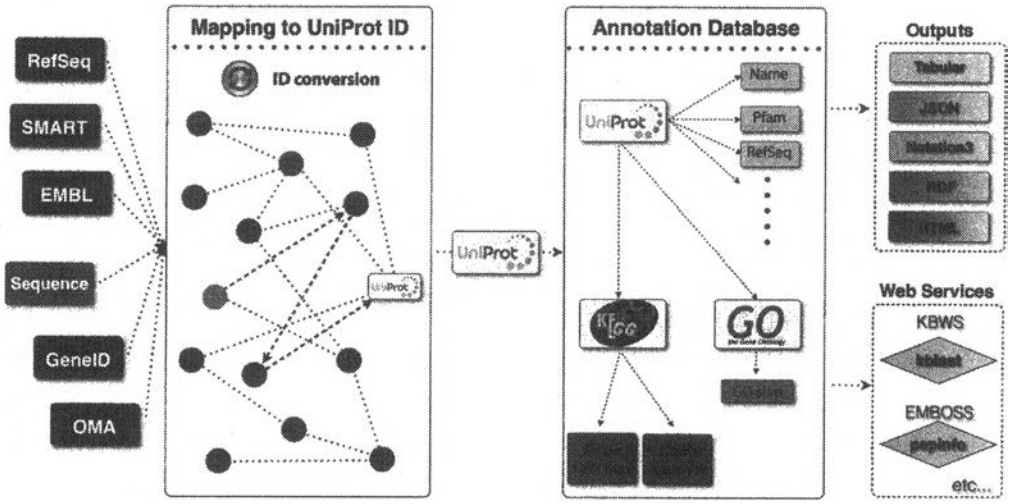


図 3.1: G-Links全体のアーキテクチャ図

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．このように，ID変換部分とアノテーション部分を分割しそれぞれに特化したアーキテクチャを構築することでシステム全体の高速化ならびに効率化に成功した．配列類似性検索などの本ID変換部分の詳細は本章の3.2.2，アノテーション取得用のデータベースについては本章の3.2.3にて詳細を述べる．

遺伝子集約型のデータ統合を行うことで多領域に渡る生物学データを高速に取り扱うことが可能であるが，この手法では遺伝子情報に関連するLinkだけを取り扱うため，それ以外のLinkが持つ情報が欠損するというデメリットが存在する．この問題に関しては上記のセントラルドグマの考え方により全ての生物学データが遺伝子に結びつくという考え方に加え，生物学Linked Dataネットワークの特異性によって対処が可能である．生物学データベースにおけるLinkネットワークは非常に密でスモールワールド性を持つことが知られており，INSD (Nakamura *et al.*, 2013) やUniProtなどクロスリファレンスを大量に持つハブデータベースが存在している．さらに，Linkを辿ることによる関連データ取得の流れからこれらのデータベースへのLinkを持つことが非常に有意義な利便性をもたらすため，Rich get Richerのようにハブデータベースに Linkが集まりやすいという特徴も兼ね備えている．そのハブデータベースの多くが遺伝子を扱っていることを踏まえると，Link情報の収集源としてハブデータベースを選定することで，遺伝子中心型のデータモデルにLinkネットワークをスリム化することによるLink欠損がもたらす情報到達能と網羅性の欠損をほぼをカバーできると考えられる．ハブデータベースのみを取り扱うというアプローチは情報源として扱うデータベース数の削減によるメンテナンスコストの低減に加え，安定して信頼度の高い情報を提供できるというメリットも存在する．一般的にID変換によってLinkをたどり関連情報を収集するアプローチでは特定の情報へたどり着くために複数のLinkを経由しなけれ

ばならないケースが存在するが，各データベースの更新頻度の違いやデータベース自体の質の問題から，一般的にLinkを多く経由すればするほど得られるアノテーションの信頼度を保証することが難しくなる．ハブデータベースを中心に利用することで少ない距離で多くのリソースへ到達可能になる他，ハブになりうるデータベースはアノテーションの品質や量などを踏まえて一般的にそれだけの価値があるデータベースである可能性が高いため，信頼度の低下を抑えつつ多くの関連情報をユーザに提供することが可能になる．

3.3.2 ユーザクエリのID解決

G-Linksでは，第一にユーザから与えられた遺伝子もしくは遺伝子セットを示す情報をUniProt IDに高速に変換する必要がある．この機能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．ここでの汎用性確保のため，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，生物種や遺伝子セットなど複数の遺伝子情報を含有するIDへの対応，および塩基/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．単一遺伝子IDに対するID変換は，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，外部データベースのIDとUniProtの対応テーブルを用いて行っている．また，ロケーション問題の解決方法としてG-Linksではデータベース名が含まれるUSA形式が入力として与えられた場合，USAはデータベースの名前を含むURNなのでデータベース名を用いた絞り込みを用いてユーザの入力に対応するUniProt IDを探索し，データベース名が含まれていない単純なIDだった場合は全テーブルから対応するUniProt IDの検索を行う．例としてEcoGene (Zhou and Rudd, 2013) のIDであるEG10986についてのリソースを取得する際，ユーザがEcoGene:EG10986というUSAではなくEG10986だけを入力として用いたとしても，G-LinksはEcoGeneというデータベース名を自動解決し適切なリソースを提供する．このようにデータベース名がない場合に自動判別を行うことで，ユーザが目的のIDのロケーション情報を知らなかったとしてもそのIDをG-Linksの入力として与えるだけでロケーション問題を解決できる．なお，データベース名が与えられない状態で対応するエントリーの候補が2つ以上存在した場合，G-Linksでは両候補の遺伝子IDに関連する情報をユーザに提供する．

KEGG Orthologyに代表されるような遺伝子セットを表現するIDが入力された場合，G-Linksで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ユーザに提供する．また，G-Linksでは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の動作をする他，遺伝子セットを表すIDについての対応表も単一遺伝子を示すIDと同じテーブル上に実装されており，データベース名がない際にも同様の動作をする．生物種を示すIDを入力した場合にはその生物種が持っている遺伝子を示すUniProt IDのセットへとまず変換を行い，遺伝子セットを示すIDを入力された場合と同様の動作を行う．例として*Homo sapiens*のtaxonomy IDである9606を用いた場合，G-Linksは20063個のUniProt IDのリストへと変換を行い，それら全てのUniProt IDに関する生物学情報をユーザに提供する．生物種IDとUniProt IDの変換では1クエリに対して対応するUniProt IDが非常に多いため，個別のテーブルを用意することで対応している．対応表の元データはUniProt IDへ変換するという制約上，UniProtが提供するTaxonomy Search (`http://www.uniprot.org/taxonomy/`) のID対応表を用いている．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 *et al.*, 2012) の2種類のデータベースのIDをサポートしており，こちらもデータベース名を指定せずIDを直接入力するだけで利用可能である．遺伝子を示すリソースに対して汎用的に対応することを考えた場合，問題となるのがユーザが特定の遺伝子のIDが分からな

いケースである．配列情報は遺伝子を構築する情報そのものであり，遺伝子を示す最も単純かつ的確なリソースであるといえる．そのため，本質的に汎用的な入力系を構築するためには，なんらかの配列が入力された場合にその配列に関連する的確な生物学情報を返す必要がある．このため，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を UniProt ID へ変換を行うことで問題を解決した．ここで問題になる点の一つが配列類似性検索の計算処理がもたらすレイテンシの問題である．レイテンシ問題への対策として，G-Links ではユーザから入力された配列が塩基配列だった場合は EMBOSS 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 の配列のみをターゲットとして行っている．高速な BLAST を配列の長さが短いアミノ酸配列で，かつ Swiss-Prot のみという小規模なデータベースをターゲットとすることで，ユーザからの入力を高速に UniProt ID に変換することが可能である．なお，塩基配列をアミノ酸配列に変換する際はフレームがずれていることなどを考慮し，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，配列入力時の UniProt ID への変換でもう一つ大きな問題として与えられるのが変換の正確性である．このプロセスの目的は類似配列の遺伝子を発見することではなく UniProt ID への変換であるため，候補数の多さではなく変換の精度の高さが要求される．そのため，G-Links では BLAT を行う際の E-value の閾値の初期値を 1e-70，identity の閾値の初期値を 0.98 (98%) とを非常に高いものに設定し，人の目でアノテーションがつけられその存在が確認されている Swiss-Prot のみを対象データベースとして用いることで，類似性検索をできるだけ ID 変換の精度に近づけている．また，BLAT の各種パラメータの閾値を高く設定することは動作の高速化にも貢献することが可能である．さらに，確実な UniProt ID への変換を行うため G-Links に配列情報を与えた場合，ユーザは候補となる UniProt ID とともに当該 UniProt ID に対する 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 をテーブルとしたものをユーザに提供する．そのため，ユーザはその結果を閲覧し正しい 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．また，後述する direct オプションを用いることで与えられた配列をトップヒットの UniProt ID に直接変換を行い，その UniProt ID に関連する生物学情報を取得することも可能である．

3.3.3 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，アノテーション情報が格納された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には UniProt から得られた Swiss 形式フラットファイルに記載されている各レコードの情報をベースにして，そこで得られた ID 群からの Link を辿ることで情報範囲の拡張を行っている．Link を辿る対象のデータベースとして Link Data ネットワークのハブとなりうるクロスリファレンス数の多いデータベースを中心に選択することで，データ収集の際に Link を辿る距離を最小限に保ちつつより多くのデータベースへの Link をサポートしている．この選択方針には，G-Links 内部で扱うデータセット数を押さえることによるメンテナンスコストの低減と，ハブとなるレベルの高品質なリソースのみを提供することによる Link の品質の担保という目的も存在する．また，この拡張では Gene Ontology (GO) (The Gene Ontology Consortium, 2013) のように UniProt からの Link を辿るだけで取得できる ID 情報のみならず，GO slim (Harris *et al.*, 2004) のような計算を行わないと

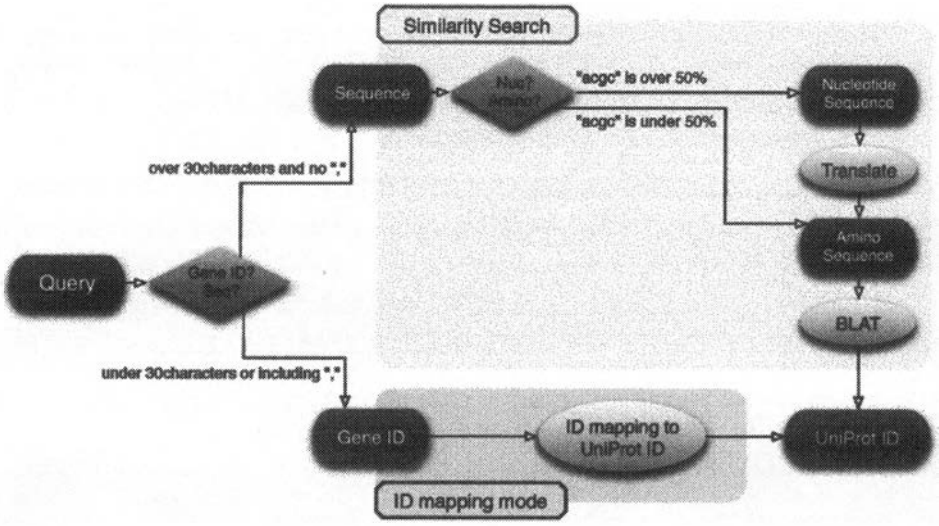


図 3.2: ID 変換部のアーキテクチャ

本サービスではまずユーザからの入力を UniProt ID へ変換する．ユーザのクエリは文字数と複数遺伝子指定時のセパレータであるカンマの有無で判断しており，30 文字以下もしくはカンマがあるときは遺伝子 ID，それ以外の場合は配列情報であると判断する．ユーザのクエリが遺伝子 ID だった場合は UniProt の提供する ID mapping サービスをベースに ID 変換を行い，クエリが塩基もしくはアミノ酸配列だった場合，Swiss-Prot に対して BLAT による配列類似度検索を行うことで UniProt ID へ変換し，ヒットした UniProt ID のリストを E-value や生物種などの補足情報と共に候補リストとしてユーザに提供する．このリストから当該する UniProt ID を選択することで，ユーザは自身の持つ配列をより確実に ID へと変換することができる．

求められないデータに関してもデータ投入時に予め計算を行うことでその ID を取得し，データベースに格納することでより有意義な情報をユーザに提供している．また，G-Links では ID 情報のみならず，UniProt フラットファイルに登録されている *OS* レコードの生物種情報や *CC* レコードのその遺伝子が関連するドメインや組織，病気に関する情報，GO の各タームに対する説明文など，自然言語で記述された「人が読むための情報」も別途保存されている．これらの情報も一緒にユーザに提供することで，ユーザは自身がクエリとした遺伝子について「どういう遺伝子か」という，ID 情報だけでは理解できない知識を容易に取得することが可能となる．これらの情報は MySQL の上で UniProt ID を主キーとして Swiss 形式ファイルのレコード名と同様のカラムに加えて ”ex” という UniProt クロスリファレンスから独自拡張で追加されたアノテーション情報が格納されているカラムを持つテーブルによって保存されている．このテーブルは主キーである UniProt ID にインデックスを張っ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，それぞれの UniProt ID に関連する情報の量の増大に対しても一定の検索速度が維持できる，スケーラビリティが高い設計となっている．アノテーション取得用データベースの ER モデル概略図を図 3.3 に示す．

また，ユーザからのクエリが生物種を示す ID であった場合，例えは *Homo sapiens* であれば 2 万以上という大量の遺伝子に関するアノテーションを取得する必要がある．この作業を高速に行うため，生物種に対するクエリに関しては事前にキャッシュを生成することで対応している．このキャッシュは Perl の Storable モジュールでシリアライズ化されたファイルをデータメンテナンス時に生成することで行なっており，Perl プログラム内でメモリに格納されている形そのままのデータを扱うため高速な I/O を実現できる．

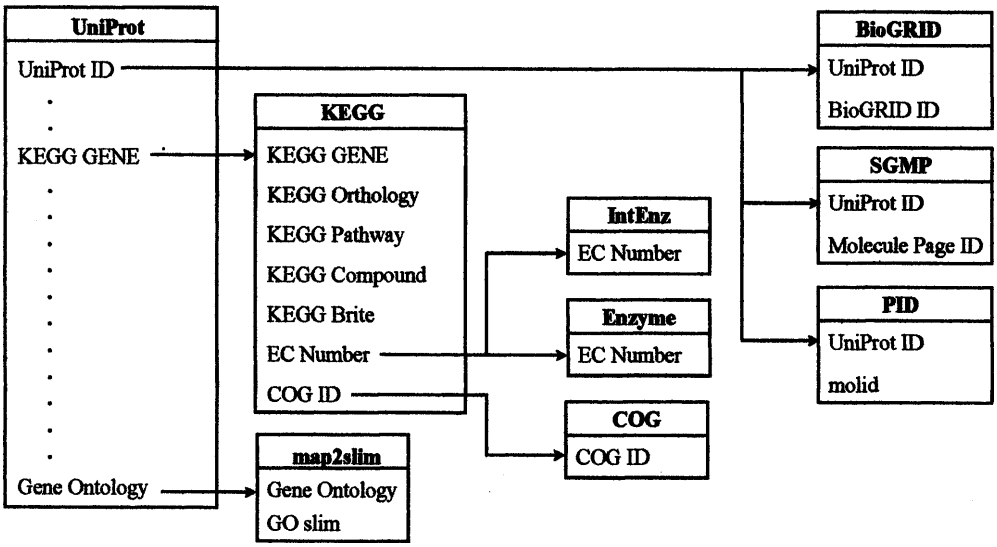


図 3.3: アノテーション取得部の ER モデル概略図

アノテーション収集用 RDB の ER モデルの概略図を示す．本データベースでは，UniProt の RAW ファイルから得られるレコードに記述されるクロスリファレンスに加えて，上記図のモデルに従って関連情報をデータベースをまたがる形で取得し UniProt ID に関連づける形で取り扱っている．UniProt 以外のデータベースについて，GO slim は Gene Ontology が提供しているツール map2slim を利用して算出，他のデータベースに関しては各データベースが保持しているリンク情報を元にモデルを構築している．現在の G-Links でリソース元として UniProt の以外に追加したデータベースとしては BioGRID (Chatr-Aryamontri *et al.*, 2013)，Clusters of Orthologous Groups of proteins (COG) (Tatusov *et al.*, 2001)，IntEnz (Fleischmann *et al.*, 2004)，Enzyme (Bairoch, 2000)，KEGG，Pathway Interaction Database (PID) (Schaefer *et al.*, 2009)，UCSD-Nature Signaling Gateway Molecule Pages (SGMP) (Saunders *et al.*, 2008) があげられる．

3.3.4 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G-LinksはJSONとTabularでの出力を行なっており，Tabular出力についてはG-Linksが持つ全てのアノテーションデータを含むフォーマットと，URL情報などを覗いた軽量バージョンの2種類を提供する．Human-Readableな出力としてはHTMLフォーマットで出力が可能である．この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，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．Semantic Web用のフォーマットとしては，G-Linksでは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から変換することで実装を行なっている．RESTfulなG-LinksにてRDF出力をサポートすることで，G-Linksはユーザが対象としている遺伝子IDに関するRDFを，1URLを指定するだけで生成し利用する事が可能である．なお，RDFを生成する際のオントロジーとしてG-LinksではEDAM Ontologyをベースとしており，EDAM Ontologyに当該する語彙が存在しないリソースに関してはUniProt Core Ontologyを，UniProt Core Ontologyにも適切な語彙が存在しなかった場合は *rdfs:seeAlso* を *Predicate* として指定している．

また，G-LinksではID情報や記述情報のみではなく解析Webサービスによって出力される解析結果へのリンク情報も結果として出力することができる．ここで得られる解析結果の取得にはG-language EMBOSS REST Service (`http://www.g-language.org/wiki/emboss`) を用いており，このRESTサービスを利用するためのURLをリンク情報として結果に含むことで解析Webサービスの結果とデータベースに内包される生物学リソースをシームレスに統合することが可能となる．EMBOSS REST Serviceは第二章で論じたKBWSも内包しているため，本RESTサービスをG-Linksに統合することで，EMBOSSに内包されるツールだけではなくBLASTに代表される様々な生物学解析Webサービスについて共通のシンタックスやインタフェースの下で容易に連携させることが可能になる．各解析Webツールに対して指定される入力クエリについては，EMBOSSを内部で用いているためG-Linksによって得られたID情報からUSAを生成することで対応することが可能になる．

3.4 結果

3.4.1 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．G-Links で入出力として利用できるデータベースのリストを表 3.1 および表 3.2 に示す．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．

本サービスでは出力として各生物学データベースから取得できるリソースだけではなく，そこで得られるリソースを解析 Web サービスに与えた際の出力結果も扱うことが可能である．これらの解析結果リソースとしては G-language EMBOSS REST Service に対応しており，EMBOSS および KBWS に含まれるツールから利用価値が高いものを抜粋して実装している．

3.4.2 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

表 3.1: G-Links で入力対応しているデータベースリスト

AGD	GenomeReviews	REBASE
Aarhus/Ghent-2DPAGE	GermOnline	RGD
ArachnoServer	H-InvDB	Reactome
BioCyc	HGNC	RefSeq
CGD	HOGENOM	RefSeq_genome
CYGD	HOVERGEN	SGD
CleanEx	HPA	TAIR
ConoServer	HSSP	TCDB
DIP	IPI	TIGR
DisProt	KEGG	TubercuList
DrugBank	LegioList	UCSC
ECO2DBASE	Leproma	UniGene
EMBL	MEROPS	UniParc
EMBL-CDS	MGI	UniProtKB-ID
EchoBASE	MIM	UniRef100
EcoGene	MINT	UniRef50
Ensembl	MaizeGDB	UniRef90
EnsemblGenome	NCBI_tax	VectorBase
EnsemblGenome_PRO	NMPDR	World-2DPAGE
EnsemblGenome_TRS	NextBio	WormBase
Ensembl_PRO	OMA	WormBase_PRO
Ensembl_TRS	Orphanet	WormBase_TRS
EuPathDB	OrthoDB	Xenbase
FlyBase	PDB	ZFIN
GI	PeroxiBase	affymetrix
GeneCards	PharmGKB	dictyBase
GeneFarm	PptaseDB	eggNOG
GeneID	ProtClustDB	euHCVdb
GenoList	PseudoCAP	

G-Links で入力として利用できる ID として対応しているデータベース一覧を示す．ユーザはこれら 85 種類のデータベースにおける ID ならばどの ID でも，G-Links を用いて簡単に関連情報を取得できる．これらのデータベースへの対応は UniProt が提供する ID マッピング用サービスを独自拡張することで行なっている．本リストの最新情報は `http://link.g-language.org/input_list` から利用できる．

表 3.2: G-Links で出力対応しているデータベースリスト

AGD	GO_function	LegioList	RefSeq
ArachnoServer	GO_process	Leproma	SGD
ArrayExpress	GOslim_component	MGI	SMART
BioCyc	GOslim_function	MIM	SMR
BioGRID	GOslim_process	MINT	SNPedia
CGD	Gene3D	MRROPS	STRING
COG	GeneCards	MaizeGDB	SUPFAM
COXPRESdb	GeneFarm	NMPDR	SWISS-2DPAGE
CYGD	GeneID	NextBio	SignalingGateway
CleanEx	GeneTree	OMA	TAIR
ConoServer	Genevestigator	Orphanet	TCDB
DIP	GenoList	OrthoDB	TIGR
DOI	GenomeReviews	PANTHER	TIGRFAMs
DisProt	GermOnline	PDB	TubercuList
DrugBank	H-InvDB	PDBsum	UCSC
EC_number	HAMAP	PID	UniGene
EMBL	HGNC	PRINTS	UniParc
EMBL-CDS	HOGENOM	PROSITE	UniProtKB
ENZYME	HOVERGEN	PeroxiBase	UniProtKB-AC
EchoBASE	HPA	Pfam	UniRef100
EcoGene	HSSP	PharmGKB	UniRef50
Ensembl	IPI	PhosSite	UniRef90
EnsemblBacteria	InParanoid	PhosphoSite	VectorBase
EnsemblGenome	IntAct	PhylomeDB	WoLFPSORT
EnsemblGenome_PRO	IntEnz	PomBase	World-2DPAGE
EnsemblGenome_TRS	InterPro	PptaseDB	WormBase
Ensembl_PRO	Jabion	ProtClustDB	WormBase_PRO
Ensembl_TRS	KEGG_Brite	ProteinModelPortal	WormBase_TRS
EuPathDB	KEGG_Disease	PseudoCAP	Xenbase
FlyBase	KEGG_Gene	PubMed	ZFIN
G-Links	KEGG_Orthology	REBASE	dbSNP
GI	KEGG_Pathway	RGD	eggNOG
GO_component	KEGG_Reaction	Reactome	euHCVdb
GO_function	LegioList	RefSeq	neXtProt

G-Links で出力として利用できる ID として対応しているデータベース一覧を示す．ユーザは URL にアクセスを行うだけで，自身のクエリに関係のある ID 情報これらのデータベース群から収集し，かつその情報を用いた Web サービスの解析結果を集積したリソースを高速かつ自動的に取得できる．本リストの最新情報は `http://link.g-language.org/output_list` から利用できる．

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，その遺伝子はどのようなアノテーションを持っているのかという「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．

ユーザがブラウザ経由でアクセスした場合に対象とする遺伝子の機能を容易に閲覧し確認できる出力を提供するため，他のフォーマットが明示的に指定されていない場合は，G-Linksは自動的にHTML形式による利便性および閲覧性の高い出力をユーザに提供する．この出力ではターゲットの遺伝子に関連する情報として，ユーザが目で見て直感的に理解できる画像情報，人が読んで理解するための自然言語による記述情報，他データベースにおけるIDとそのIDに対応するリソースを示すURLの3種類の情報を自動で収集し高速にユーザに提供する．画像や記述情報など単純なID変換へ取得できない人が閲覧するための情報を高速に収集できる他，そのIDが示すリソースのURLから利用可能なハイパーリンクを自動生成し提供することで，ID変換のアプローチが持つロケーションの問題を解決することができる．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．このクエリにアクセスした際の出力例を図3.4，そ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3.3以下に示す．

表 3.3: 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3.4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．

3.4.3 配列入力時の動作

G-Linksでは入力として，遺伝子や遺伝子を示すIDだけではなく，遺伝子自体を指し示す塩基/アミノ酸配列を直接入力することも可能である．配列を直接入力した場合は，入力された配列対してSwiss-ProtへBLATによる配列類似性検索を行うことでユーザから受け取った配列をUniProt IDへID変換を行い，候補となるUniProt IDに加えてその遺伝子の名前，生物種名，BLAT時のE-valueおよびIdentity，そのUniProt IDをクエリとしたG-Linksの出力URLのテーブルをユーザに提供する．ユーザはそれらの情報から自身が目的としていた遺伝子を選択することで，配列からUniProt IDへの確実な変換を実現している．また，入力された配列が塩基配列の場合はアミノ酸へ翻訳を行なってから配列類似性検索を行うのだが，その際にフレームずれと翻訳方向について考慮し，翻訳開始点を+0, +1, +2した3パターンそれぞれに対して両鎖を読まれた場合の6パターンの配列へ翻訳．全配列をクエリとして配列類似性検索を行ない，E-valueおよびIdentityのスコアが高い配列を持つUniProt ID群をユーザに提供する．配列を直接入力した際の出力例を図3.5に示す．

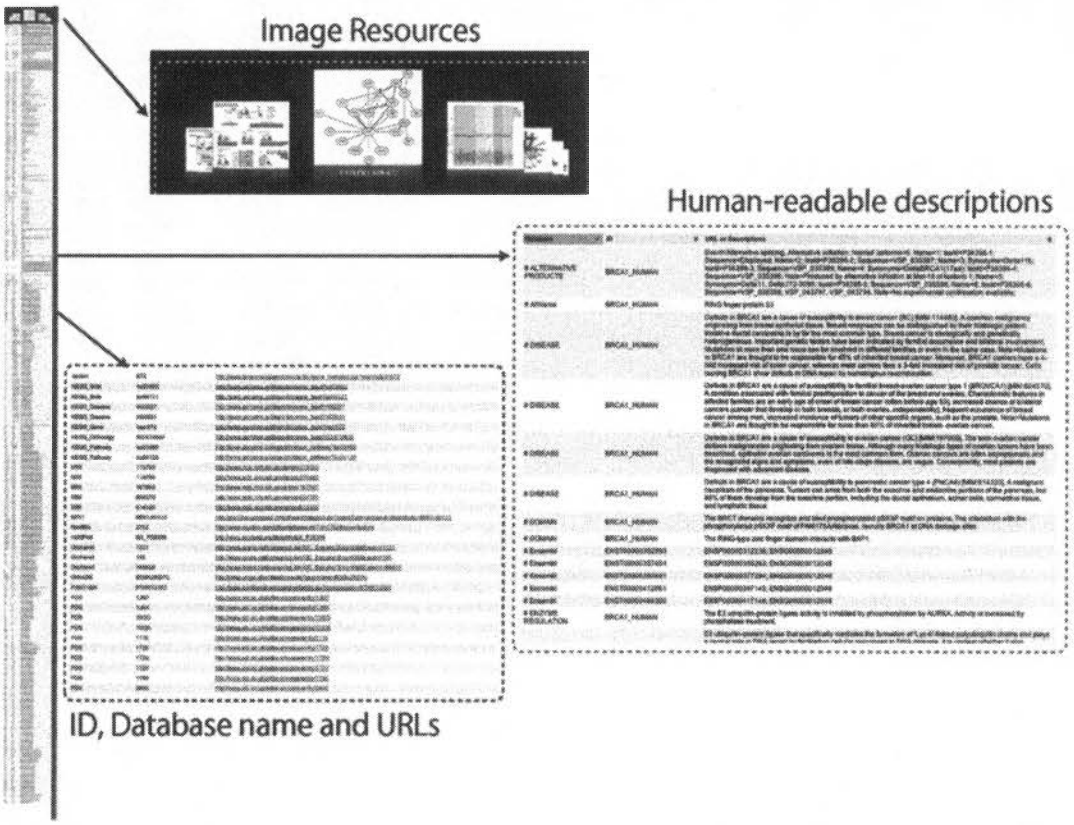
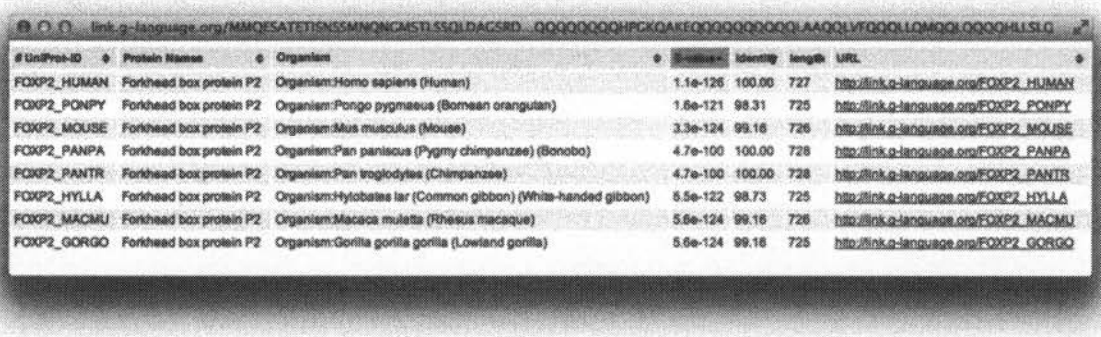


図 3.4: Web ブラウザ経由での動作例

BRCA1 遺伝子に関する情報を G-Links を用いて取得した例．左側にブラウザ経由で `http://link.g-language.org/BRCA1_HUMAN` へアクセスしリソースを取得した場合の全結果のスクリーンショットを，右側にそれぞれのセクションで得られる出力の一部を示す．G-Links を用いることで，ユーザはこれだけのリソースを高速かつ容易に取得し，画像情報や記述情報を閲覧することで自身が対象とする遺伝子についての概要を確認し，URL を辿ることでより詳細な情報を取得することができる．



# UniProt-ID	Protein Names	Organism	E-value	Identity	length	URL
FOXP2_HUMAN	Forkhead box protein P2	Organism:Homo sapiens (Human)	1.1e-126	100.00	727	http://link.g-language.org/FOXP2_HUMAN
FOXP2_PONPY	Forkhead box protein P2	Organism:Pongo pygmaeus (Bornean orangutan)	1.6e-121	98.31	725	http://link.g-language.org/FOXP2_PONPY
FOXP2_MOUSE	Forkhead box protein P2	Organism:Mus musculus (Mouse)	3.3e-124	99.16	726	http://link.g-language.org/FOXP2_MOUSE
FOXP2_PANPA	Forkhead box protein P2	Organism:Pan paniscus (Pygmy chimpanzee) (Bonobo)	4.7e-100	100.00	728	http://link.g-language.org/FOXP2_PANPA
FOXP2_PANTR	Forkhead box protein P2	Organism:Pan troglodytes (Chimpanzee)	4.7e-100	100.00	728	http://link.g-language.org/FOXP2_PANTR
FOXP2_HYLLA	Forkhead box protein P2	Organism:Hylobates lar (Common gibbon) (White-handed gibbon)	5.5e-122	98.73	725	http://link.g-language.org/FOXP2_HYLLA
FOXP2_MACMU	Forkhead box protein P2	Organism:Macaca mulatta (Rhesus macaque)	5.6e-124	99.16	726	http://link.g-language.org/FOXP2_MACMU
FOXP2_GORGO	Forkhead box protein P2	Organism:Gorilla gorilla gorilla (Lowland gorilla)	5.6e-124	99.16	725	http://link.g-language.org/FOXP2_GORGO

図 3.5: 配列を直接入力した際の実行例

G-Links にアミノ酸配列データを直接与えた際の実行例を示す．G-Links では塩基/アミノ酸配列を与えることで，塩基配列の場合はフレームシフトや遺伝子鎖の表裏を考慮した 6 種類のアミノ酸配列へ翻訳した後，配列類似性検索を用いて UniProt ID へと変換を行う．このとき確実な ID への変換を行う方法として，G-Links では配列入力時の結果として当該配列データを表す UniProt ID の候補を，タンパク質の名前，生物種名，E-value，Identity，配列の長さ，その ID を用いた G-Links のクエリ URL をユーザに提供する．これらの情報を見て自身が意図している遺伝子を用いた G-Links のクエリへアクセスすることで，ユーザはより確実に自身がターゲットとする遺伝子の情報を取得することができる．上記結果は http://link.g-language.org/MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQ QQQALQAARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVL SPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQ QQQQQQQQQQQQQQQQQQQQQHPGKQAKEQQQQQQQQQLAAQQLVFQQQLLQMQQLQ QQQHLLSLQ にアクセスすることで取得することができる．

BLAT による配列類似性検索を行う際，G-Links ではレイテンシの問題と正確な ID 変換を提供するため，E-value の閾値は 1e-70，identity の閾値は 0.98 (98%) という非常に高い基本値を設定しているが，evalue オプションと identity オプションを用いることで，E-value および Identity の閾値を独自に指定することが可能である．また，direct オプションを有効にすることで UniProt ID のテーブルを提示せず，トップヒットの配列を持つ UniProt ID へ配列情報を直接変換しアノテーションを取得することも可能である．

3.4.4 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 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，複数の遺伝子を直接指定するのではなく，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 が示す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．さらに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ユーザに提供する．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．以下に複数遺伝子指定時のクエリ例を示す．

複数遺伝子指定時の実行例

(1) http://link.g-language.org/uc003hui,GeneID:93986,RECA_ECOLI

UCSC:uc003hui, GeneID:93986, UniProtKB:RECA_ECOLI の 3 種類の ID についての情報を取得する例．データベースが異なる ID の混在や ID と USA の混在があったとしても，G-Links はそれぞれを適切に処理する．

(2) http://link.g-language.org/K10605

KEGG Orthology の K10605 (KO:K10605) に所属する 7 つの遺伝子についての生物学情報を取得する例．遺伝子セットの ID を指定する場合も単一遺伝子 ID の指定と同様にデータベース名が不要であるほか，カンマ区切りによって複数の遺伝子を指定可能である．

(3) http://link.g-language.org/NC_000913

Escherichia coli str. K-12 substr. MG1655 (NC_000913) が持つ 4417 個の全遺伝子に関して生物学情報のセットを取得する例．生物種を示す ID としては主に真核生物用として NCBI taxonomy ID と，主にバクテリア用として RefSeq の NC 番号が利用できる．遺伝子指定時と同様にデータベース名は不要だが，カンマ区切りによる複数生物種の指定はできない．

3.4.5 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 3.4 に示す．

表 3.4: 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，G-Links では他のソフトウェアや Web サービスと容易な連携を行うための Programmatic 出力として JSON や Tab Separated Value (TSV) による出力もサポートしている．これらのフォーマット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，かつ G-Links のリソースは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．全てのリソースが URL にて指定可能であり JSON での出力も可能であるため JavaScript との親和性も非常に高い他，URL 情報などを除いた軽量かつ高速な TSV 形式である slim フォーマットなど用途に応じた適切な出力を直接得ることができる．また，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 リソースの準備を容易かつ高速に行うことができる他，その RDF リソースは一意の URL で指定できるため，Semantic Web の上で扱えるリソースとして直接指定することも可能である．G-Links での RDF および Notation3 形式ではオントロジとして基本的に EDAM Ontology を用い，EDAM Ontology でカバーできない部分に関して UniProt Ontology を用いている．EDAM Ontology は，ツールによる解析 (Operation)，データの種類や性質 (Data)，バイオインフォマティクスにおけるカテゴリ (Topic)，およびフォーマット (Format) の 4 つの大きなサブクラスを持つ構造を採用しており，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いる．語彙の数は 2500 以上で，配列解析のための統合環境である EMBOSS でも内部のメタデータ管理に利用されているなど，データ収集と Web サービスによる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．

本サービスで出力フォーマットを明示しなかった場合，HTTP/1.1 に定義されているコンテンツネゴシエーション (`http://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html`) の概念に基づきブラウザ経由のアクセスには HTML を，プログラムなどからの場合は TSV をデフォルトの出力として利用する．HTTP Accept Request Header Field によるデータイプの指定も可能であるため，ユーザはより単純な URL だけで自身が必要とする最適なフォーマットでのデータ出力が可能である．

3.4.6 必要なデータの抽出

複数遺伝子セットや生物種などを指定することで，G-Links はユーザが対象とする遺伝子セットに関する網羅的な生物学情報を高速かつ容易に提供することができる．このパイプラインは遺伝子情報の閲覧や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．G-Links ではユーザが対象生物学情報を TSV 形式で取得した場合，ユーザは得られた大量のデータセットをインターネット経由で取得し，自身が必要とするリソースの抽出を行い，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．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リソースが提供されるべきである．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．

遺伝子レベルのフィルタリング方法として，G-Links では filter オプションを実装している．こ

の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出力することができる．本オプション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．このように複数の条件を組み合わせることで，ユーザが必要な遺伝子に関する情報だけを効率的に取得することが可能である．以下にfilterオプションの使用例を示す．

filterオプションを用いた遺伝子抽出の例

G-Linksを用い以下のようにトライアンドエラー的に条件を追加することで，ユーザが目的とするリソースを適切に出力するクエリを探索し，そのリソースを得ることができる．*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

上記クエリに加え，さらに胸部と子宮に関連する遺伝子という条件を追加

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

さらに，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|:snps|:polymorphisms

もう一つのフィルタリング方法として，G-LinksではextractオプションをSに実装した．このオプションでは，ユーザが指定したデータベース名や情報セクション名を元に抽出作業を行うことで，得られた全生物学情報からユーザが必要な情報だけを抽出する，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，一意の簡便な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．extractオプションの利用例を以下に示す．

extract オプションを用いたリソース抽出の例

filter オプションの例のクエリで得られた情報から dbSNP の情報のみを取得

```
http://link.g-language.org/9606/format=tsv/filter=DISEASE:cancer
        /filter=:breast|:ovarian|:snps|:polymorphisms
                /extract=dbSNP
```

dbSNP だけではなく，SNPedia の情報も追加

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

extract オプションを用いて，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．

3.4.7 外部 Web サービスとの連携

バイオインフォマティクス分野では多くの解析ソフトウェアが Web サービスとして提供されており，オンラインで公開されているデータベースへ検索を行うのと同様に，Web サービスを用いることで各種生物学リソースを取得することは，多くの研究者によって一般的に行われている．G-Links では生物学データベースに格納されているリソースの統合だけではなく，これらの Web サービスによって得られる生物学リソースの統合を行うことで，より多領域の生物学リソースの統合を実現している．G-Link で扱う事のできる解析サービスのリストを表 3.5 に示す．

出力結果の統合対象としては，解析 Web サービスのより容易かつ効率的な統合・管理のため G-language EMBOSS REST サービスを選択した．本サービスは EMBOSS という入出力が統一された 400 以上の配列解析用ツールを RESTful なインタフェースから利用できる Web サービスであり，使用する配列を USA を用いて指定を行うことで，各ツールの出力リソースを一意の URL で指定することが可能である．例えば Swiss-Prot の FOXP2_HUMAM に対してアミノ酸残基の統計情報を出力する EMBOSS のツールである *inforesidue* の結果を取得するためには `http://rest.g-language.org/emboss/inforesidue/siwssprot:FOXP2_HUMAN` へアクセスするだけでよい．G-Links の出力結果としてこれらのツールの解析結果を示す URL を統合することで，ID 情報や各データベースに格納されている生物学データセットだけではなく，それらが示すリソースに対する解析結果をも統合した生物学リソースセットをユーザに提供することが可能となる．

また，本サービスを採用した大きな理由の一つが KBWS の各種メソッドのサポートである．第 2 章で論じた KBWS は EMBOSS の追加パッケージの形で実装された UNIX コマンドラインツールであり，配列解析を行うバイオインフォマティクス Web サービスについて，サーバプロキシモデルおよび EMBOSS の入出力管理により統一的なインタフェースをユーザに提供する．この KBWS を G-language EMBOSS REST サービスから利用することで，42 の配列解析 Web サービスの出力結果を簡単な URL の指定だけで取得することが可能となる．また，例えば BLAST サービス用ツールである *kblast* の利用時に，対象データベースの指定がない場合に自動的に Swiss-Prot を対象とし，かつ *blastp* や *blastn* の使用など推論可能な部分の自動推論を

表 3.5: G-Links でサポートしている外部解析 Web サービスのリスト

Nucleotide	Protein	
banana	antigenic	charge
chaos	epestfind	garnier
dan	helixturnhelix	hmoment
density	iep	inforesidue
einverted	octanol	pepcoil
equicktandem	pepdigest	pepinfo
geecee	pepnet	pepstats
isochore	pepwheel	pepwindow
palindrome	sigcleave	
prettyseq	kblast	kphobius
	kpsort	kpsort2
	kpsortb	kwolfpsort

G-Links がサポートを行っている Web サービスのリストを示す．G-Links では G-language EMBOSS REST サービスとそこに含まれる KBWS の REST サービスを利用しており，上記の表にて *k* から始まるツールは KBWS，それ以外は EMBOSS に実装されているツールの機能を用いる事が出来る Web サービスである．EMBOSS ならびに KBWS という統一されたインタフェースを持つ Web サービスを連携することで機械的かつ容易な連携が可能であるほか，他 ID リソースなどと同様に EDAM オントロジーを用いて関係情報を管理できる．解析に必要な配列情報は G-Links の出力から得られた Swiss-Prot ならびに Ensembl の ID を用いた USA で配列を指定しており，それを用いて各ツールの結果リソースを示す URL を G-Links の出力に加えることで統合を行っている．

行うなど，ユーザがより簡単に当該サービスを利用できる．EMBOSS REST サービスの例と同じく Swiss-Prot の FOXP2_HUMAN について配列類似性検索を行う場合，*kblast* の結果は，`http://rest.g-language.org/emboss/kblast/siwssprot:FOXP2_HUMAN` という URL を指定するだけで取得することが可能である．

これらのサービス統合にあたり問題となるのが，それぞれのツールの入力として与えるリソースの指定である．G-Links では自身の出力に含まれる ID 情報を用いて EMBOSS で利用できる USA 形式の入力を生成することでこれに対応している．USA は生物学データベースの ID にデータベース名の情報を加えた URN の一種であり，EMBOSS では USA を入力として与えることでその USA が持つデータベース名と ID 情報からリソースを示す URL へ自動的に変換し，そのリソース対して解析を行う．解析対象のリソース元として，アミノ酸配列に対しての解析ツールの場合は Primary Key である UniProt (Swiss-Prot) ID を，塩基配列に対しての解析ツールには Ensembl (Flicek *et al.*, 2013) ID を用いた USA をそれぞれ生成することで，ユーザが対象としている遺伝子に対する解析結果を適切に提供する．

3.4.8 GENIE - a Virtual Biological Research Assistant

ユーザは G-Links を用いることで，ブラウザから特定の遺伝子に関する生物学情報セットを容易かつ高速に閲覧できるほか，バイオインフォマティクス解析に利用するための大規模なデータセットを高速かつプログラミングから利用しやすいフォーマットでの取得，高速かつ簡便な RDF 出力による既存の Semantic Web 技術との連携とレイテンシ問題の解決，といった用途で利用することが可能となる．しかしながら，これらに加えた大きなユースケースとして想定されるのが，Web

アプリケーション開発時のバックエンドとしての利用である．RESTful なインタフェースでユーザが必要とする生物学情報を高速に，かつプログラムから利用しやすいフォーマットで取得できる G-Links は，BioGPS に対する MyGene.info のように生物学 Web サービスを提供するにあたって非常に有用なデータソースになることができる．この利用例として G-Links をバックエンドに開発された Web サービスが Genie である．Genie は生物学者に対してのバーチャル研究アシスタントとして開発されており，Apple 社の Siri (`http://en.wikipedia.org/wiki/Siri_(software)`) のようなインタフェースで，自然言語を用いてコミュニケートを行うことでユーザが知りたい生物学的な知見や情報を取得するサポートを行う．Genie のインタフェースを図 3.6 に示す．

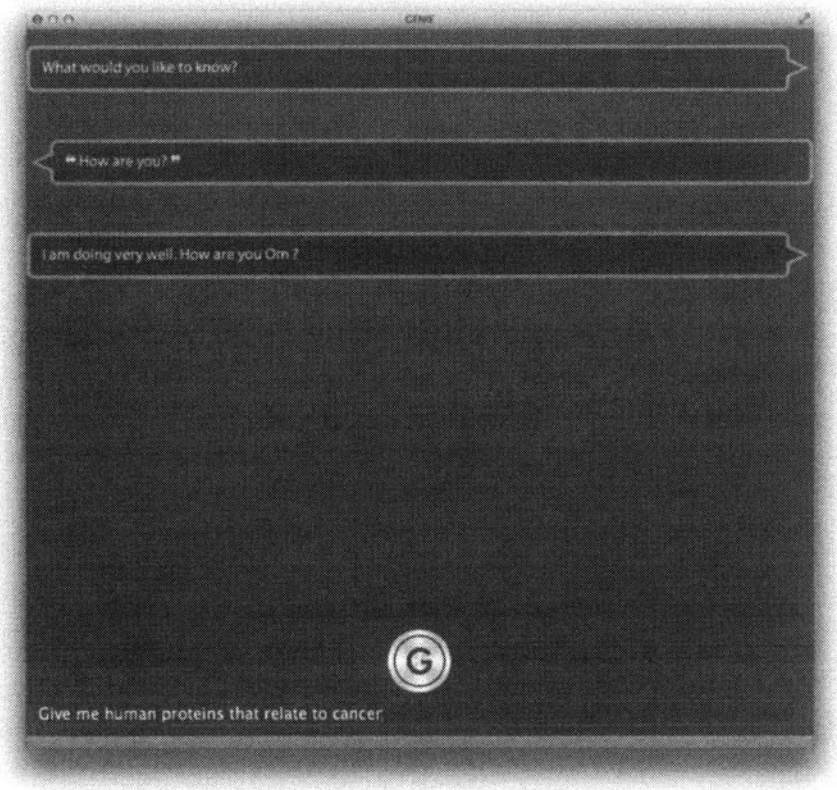


図 3.6: Genie のインタフェース
Genie のインタフェースを示す．ユーザが下部のテキストボックスに自身が知りたい内容を自然言語で入力することで，Genie はそのクエリに対して適切な答えを検索・解析し音声読み上げなどの機能でユーザに提供する．Mac OS X Mountain Lion のディクテーションや HTML5 の音声認識機能を用いて話しかけることでテキスト入力を行うことで，Genie と会話によるコミュニケーションを行いながら複雑な検索および解析を行うことができる．

バイオインフォマティクス研究では，人がそのまま解釈することが難しい生物学情報を統合し，人が理解できるような処理を行うことで，大規模なデータから生物学的に意味のある知見を抽出する．G-Links はこのデータ統合と収集，抽出部分を高速かつ自動的に行うことができるシステムであるが，これをもとにより有意義な研究を行うためには，G-Links から得られた結果を用いて更なる解析を行いそこに含まれる知識をより分かりやすい形で導出するプロセスが必要である．Genie の大きな目的はこの生物学研究に対するアシストをスムーズに行うことであり，そのためには以下の要素を持ち合わせる必要がある．

(1) アシスタントなので，操作が容易な音声入力を行い，結果も音声で返ってくる

(2) 入力を自然言語で行うことができ，かつ結果も自然言語やグラフなど，コンピュータに不慣れな利用者が理解しやすい形で返ってくる

(3) 多領域に渡る生物学情報を効率的に統合し運用できる

(4) これらの処理を短時間で行い，結果を高速にユーザに返す

Genieはこれらの要素を達成するサービスとして，BioHackathon2012にて慶應義塾大学大学院政策・メディア研究科の荒川和晴特任講師および慶應義塾大学環境情報学部 板谷英駿氏と共に共同開発を行った．Genieでは必要要素の3つ目および4つ目を達成するためにG-Linksを採用しており，そのための速度向上および連携のための機能拡張を私が，メインインタフェースとグラフ出力などを板谷氏が，プロジェクト全体の統括とシステム全体の設計，入力に対する自然言語処理，外部サービスとの連携部分などを荒川氏が担当した．

GenieはG-Linksが扱うことのできる全生物学データセットから，1. 配列や機能，細胞内局在，パスウェイ，関連する病理情報やSNP，相互作用，翻訳制御，遺伝子発現レベルといった遺伝子に関する情報，2. マウスのがんに関連する遺伝子の全SNP情報というような，複数の基準を用いた遺伝子セットに関する情報，3. GC skew (Lobry, 1996) や転写開始位置の予測，コドン使用バイアスの計算などのゲノムに関する情報，といった3種類の生物学情報について自然言語で問い合わせることで，その答えを人が分かりやすい形で得ることができる．例えば，Genieに対して"Give me human proteins that relate to cancer"と問い合わせることで，Genieは*Homo sapiens*の持つ全遺伝子のうち，がんに関連する遺伝子が全体の何パーセント存在するかという情報とともに，それらの遺伝子群が持つGO slim functionの割合を円グラフで提供する．さらにより自然なアシスタントを行うため，Genieではそれまでの質問の文脈をふまえた絞り込みをG-Linksのfilterオプションとextractオプションを用いることで行う．上記の"Give me human proteins that relate to cancer"というクエリのすぐ後に，"And relate to breast and ovarian"というクエリを入力することで，Genieは*Homo sapiens*のがんに関連する遺伝子のうち，卵巣と乳房に関連する遺伝子をさらに絞り込む．このようにしてコミュニケーションをGenieと行うことで，ユーザは自身が知りたかった子宮頸癌や乳がんに関連する遺伝子群についての情報を，132の生物学データベースから収集し，分かりやすい形で取得することができる．上記のクエリの実行例を図3.7に示す．また，上記のクエリを実際に実行したデモンストレーションの動画は`http://y2u.be/V4jsuIOAwyM`より閲覧することができる．

Genieは入力できる自然言語として英語に対応しており，ユーザから英語テキスト情報をクエリとして受け取る．この際，MacOSX Mountain LionやiOSの音声認識機能かHTML5の発話認識機能を利用して自身が発した言葉をテキストへ変換することで，ユーザは話しかけるだけで任意のクエリを入力可能になる．ここで得られたクエリに対して，Genieでは生物種，遺伝子名，生物学的なキーワードの3種類の単語を認識する．生物種名については事前に作成した辞書を用いて判定を行っており，1クエリに1生物種が指定されていると定義している．なお，生物種が判定できない場合は*Homo sapiens*を基本値として用いる．その後，G-Linksを用いて当該生物種の全遺伝子名を取得．それらを用いてリアルタイムに遺伝子名判定用の辞書を作成し，ユーザが対象としている遺伝子を特定する．こうする事で，全生物種の全遺伝子名という大規模な辞書を用いず高速に遺伝子名判定の処理が可能になる．なお，遺伝子名が存在しない場合は当該生物種の全遺伝子をターゲットとする．対象となる生物種と遺伝子の情報の確定後，それ以外のクエリに対して簡単な自然言語処理を用いることでクエリ内に含まれる生物学的なキーワードを取得する．ここ

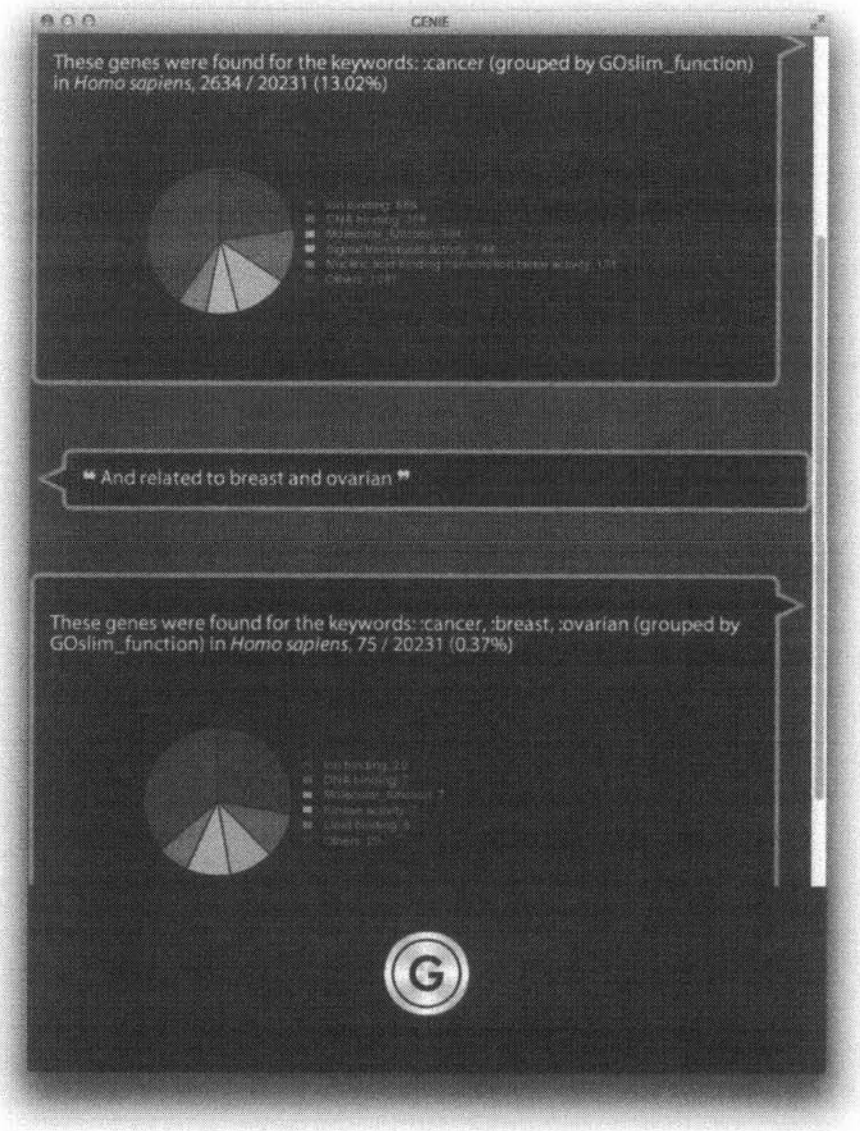


図 3.7: Genie の実行例

Genie の実行例を示す．ユーザは Genie に " Give me human proteins that relate to cancer"などとの自然言語で問いかけるだけで，Genie はユーザが意図するクエリを認識し，G-Links を用いてリソースの取得を行い，それを分かりやすい GO slim function による円グラフなどのサマリで提供する．このサマリは，候補遺伝子数が十分少なくなった場合は遺伝子のリストを返すなど，ユーザが分かりやすい結果を常に提供する．また Genie はコミュニケーションで発生する会話の流れを覚えており，前回のクエリを記憶することでユーザが本当に意図する結果を返す．本図の例では，" And relate to breast and ovarian"というクエリに対して直前のクエリを認識し，人のがん関連遺伝子の中からさらに候補遺伝子を絞り込んでいる．本クエリはデモンストレーションのムービー (http://y2u.be/V4jsuIOAwyM) で実際に使用しているサンプルクエリの一つである．

で得られるキーワードはGenieでサポートしているメソッド (*gcskew*など) およびG-Linksで絞り込みに利用する生物学的なキーワード (上記の例の”cancer”や”breast”) を想定しており，前者の場合は対象となる解析Webサービスを実行し結果を出力，後者の場合はG-Linksのfilterオプションやextractオプションを用いたクエリによるデータ取得と絞り込みを行い，そのサマリをユーザに提供する．解析用のWebサービスとしてはG-language Genome Analysis Environment (G-language GAE) (Arakawa and Tomita, 2006; Arakawa *et al.*, 2003, 2008) のWebサービスであるG-language REST Service (Arakawa *et al.*, 2010) による高速なゲノム解析メソッド，G-language Maps (Genome Projector (Arakawa *et al.*, 2009b)，Pathway Projector (Kono *et al.*, 2009)，Chaos Game Representation REST Service (Arakawa *et al.*, 2009a)) によるゲノム情報の可視化，KBWSによる配列解析用バイオインフォマティクスWebサービス，EMBOSSによる400以上のゲノム解析用メソッドをサポートしており，ユーザはG-Linksによる高速なデータ取得だけではなく，これらの豊富なWeb解析サービスによって取得できる生物学的知見もGenieから取得することができる．また，Genieでは直前のクエリで指定された生物種，遺伝子，絞り込み条件をJavaScript側で保存しており，次のクエリで生物種が変わらなかった場合，そのクエリが解析メソッドを呼び出すなら直前のクエリのリソースに対して解析サービスを呼び出し，絞り込み条件の追加であった場合はその条件を直前のクエリに追加してG-Linksからリソース取得を行う．このように，直前のクエリを保存し利用できる場合はその内容を利用することで，会話のコンテキストという，人がスムーズなコミュニケーションを行う際に自然と利用している情報を解釈し，より適切なアシスタントを行うことができる．このようにして得られた生物学的結果を，解析Webサービスを呼び出した場合は当該サービスによって出力された画像データや解析結果 (例：“Please show me the GC skew of ecoli”)，遺伝子の絞り込みが行われた場合はGO slim functionによる分類分けとそのサマリ (例：Please tell me human proteins relate to cancer”)，特定の遺伝子に関する情報を取得した場合にはG-Linksから得られた自然言語による生物学情報 (例：“Please tell me about BRCA1”) といった，IDなどの単純なデータではない，研究者がその概要を把握しやすい結果を提供する．さらに，GenieではGoogle Translate APIおよびMacOSXの発話サービスを利用することで，これらの生物学情報を合成音声でユーザに提供する．このようなシステムを構築することで，研究者はゲノムや遺伝子に関連する自身が知りたい生物学的知見を，自然言語を用いて話しかけるだけで，132以上の多くの生物学データベースから網羅的に探索し，そこで得られた結果を利用者が分かりやすい形へ加工し，自然言語の音声読み上げや画像データの形で取得する事が出来る．

3.5 議論

3.5.1 G-Linksによる遺伝子中心型の統合モデル

G-Linksは多領域に渡る生物学情報を遺伝子中心モデルにて統合を行うことで，バイオインフォマティクス研究のためのデータセット生成段階における生物学データの統合・取得・抽出のプロセスを自動的かつ効率的，高速に行うシステムである．このサービスを利用することで，ユーザは85種類のデータベースにおける遺伝子ID，及び塩基/アミノ酸配列，複数の遺伝子セットを示すIDのどれからでも，それが指し示す遺伝子に関連する情報を132種類の生物学データベースから網羅的に探索し，そのデータセットを取得することができる．キュレーションされた生物学データベース群のID間におけるLink情報をベースにしているため信頼性も高く，そのLinkネットワークの密度の高さから多領域のリソースをユーザは取得できる．遺伝子や遺伝子IDの

セットを示すID，生物種を示すIDを含んだURLにアクセスするという非常に簡便な方法でデータの取得が可能であり，filterやextractオプションを用いることで自身が欲しい情報だけを自由に抽出し，より情報密度の高いリソースを取得することができる．さらに，データ処理に適したTSVやSemantic Webと親和性が高いRDFやNotation3といったフォーマットによる出力をRESTfulなインタフェースから取得することが可能であるため，Semantic Web技術やその他のWebサービス，解析ソフトウェアと高い相互運用性の元で利用することが可能である．本サービスは`http://link.g-language.org/`よりフリーで利用することができる．

G-Linksでは主に3種類の利用方法を想定してシステム構築されており，これらを満たすために十分な要素を兼ね備えている．

(1) Webブラウザ経由での利用

G-LinksはRESTfulなサービスであるため，遺伝子のIDを含んだURLへアクセスするだけでWebブラウザなどから容易に利用することが可能である．Webブラウザという情報を閲覧するツールをから利用されるケースに対して，G-Linksでは単純なID情報やURL情報だけではなく，人が目で見て直感的に理解しやすい画像データや自然言語による記述情報を含んだHTML形式の出力を行う．与える遺伝子IDはデータベース名などの指定も不要であるため，ユーザは自身が対象としているリソースに対して高いアクセスビリティのもと，その関連情報を閲覧できる．このフォーマットにより，G-Linksは対象の遺伝子についての情報を閲覧しようとしたユーザに対して単純なデータだけではなく人が見るための情報を提供することで「その遺伝子はどのような遺伝子か」という生物学的な知識をユーザに提供することが可能となる．

(2) プログラムなどからの利用

G-Linksでは，TSVやJSONなど他の解析ソフトウェアやユーザが作成した解析プログラムから利用されることを想定したフォーマットもサポートしている．これらのフォーマットは各種プログラミング言語やソフトウェアなどでも広くサポートされている形式であり，これらのフォーマットを用いることで，ユーザは自身の解析ワークフローや解析プログラムなどにおいてG-Linksから得られる出力をシームレスに処理し，解析のためのデータセットとして用いることができる．多数の遺伝子セットについてそれらの遺伝子群に関するリソースセットを小さいレイテンシで得ることができる上，filterおよびextractオプションを用いることでそこからユーザが必要な情報だけを抽出する作業まで1つのURLへアクセスするのみで取得することが可能である．この両者のオプションは複数の条件を追加することが可能であるため，ユーザは得られる結果を見ながらトライアンドエラー的に，自身の必要な情報を取得するために適切なクエリを生成することができる．

プログラムなどから利用されるケースとして，MyGene.InfoのようにWebサービスなどの開発者がバックエンドのデータソースとしての利用が考えられる．その実例として，G-Linksをバックエンドとして活用した生物学者のための研究アシスタントGenieを開発した．Genieはゲノムや遺伝子に関連するユーザが知りたい内容を，自然言語で問い合わせることでそれを解釈し，G-Linksや各種Webサービスを用いることで当該内容のデータや解析結果を取得し，ユーザが分かりやすい自然言語や画像データでその答えを返す．ユーザが最も入力しやすい自然言語による問いに対して，単純なデータではなくそれらのデータを人間が解釈しやすい形に直した上でユーザに返すことで，生物学的な知見を探索することができるWebアプリケーションである．このように，

G-Links という効率的な生物学データ統合システムを中心に様々な解析 Web サービスを統合し，それらを効率的に扱うことのできるインタフェースを構築することで，Semantic Web などによって実現されるとしていた意味情報を加味したリソースの取得のためのシステムの原型が，遺伝子やゲノムに対してのオペレーションに限ってはいるが達成することが可能である．

(3) Semantic Web 技術との連携と RDF リソース出力のための基盤

Semantic Web は WWW 上の全てのドキュメントについて，そのドキュメントやドキュメント間の意味情報を加味した自動的な情報収集や分析のアプローチが可能になるプロジェクトであり，この技術を用いることで多領域に渡る生物学データベース上のリソースについて同様に，全リソースやリソース間の Link おける意味情報を加味した解析・分析のアプローチが可能になると期待されている．この Semantic Web 技術で現在問題とされていることの一つが，本技術でリソースの記述に用いられる RDF によるリソース準備の難しさである．G-Links では出力フォーマットに Semantic Web 上で利用できる RDF (RDF/XML) と Notation3 をサポートすることで，ユーザが対象とする遺伝子または遺伝子セットに関連する網羅的な生物学リソースの RDF を高速かつ容易に出力することが可能である．*Homo sapiens* の持つ全遺伝子に関する RDF リソースなどといった大規模な RDF の出力も 1URL へのアクセスだけで行うことができる他，ここで生成される RDF は必ず一意の URL を持つことから，既存 Semantic Web 技術とシームレスに連携を行うことが可能である．生物学情報は多領域に渡る複雑な構造をしており，その複雑性と広域性が RDF のような関係性を記述したリソース生成の難度を高くしている．これに対して，G-Links は遺伝子中心型というリソース統合モデルの特異性ゆえ，どのような生物学リソースであったとしても遺伝子の情報との Link を生成しさえすれば G-Links に取り込むことができるという性質を持っている．ユーザは自身が Semantic Web 上で利用したいリソースについて遺伝子情報への Link を設定することで，G-Links の RDF を介してより広いリソースとの Link を得ることができる．このように G-Links は生物学 Linked Open Data ネットワークにおけるハブノードとして利用することができるほか，そのスケーラビリティ故に，新規リソースを容易に統合し高速に運用することができる．生物学で Semantic Web を活用する上で，RDF の準備や計算コストなどの問題を大きく肩代わりできるプラットフォームとなりうる．

3.5.2 既存サービスとの比較

G-Links と同様に，遺伝子情報を中心とした ID 変換を行うことでユーザが指定した遺伝子に関連する生物学リソースを収集するサービスの代表例が MyGene.Info である．MyGene.Info は出力フォーマットが JSON で JSONP によるコールバック用のパラメータが利用できるなど主にプログラムからデータリソースとして利用されることを前提としたサービスであり，代表的なモデル生物 9 種の持つ遺伝子についてサポートを行なっている．本サービスは高速に動作する RESTful な Web サービスとして提供されており，Entrez Gene ID と Ensembl Gene ID を含んだ URL へアクセスするだけでユーザはその遺伝子に関連する 31 種類のリソースを取得することができる他，フィルタリング用のオプションを用いることでユーザがそこから必要なリソースだけを抽出し取得することができる．出力される情報は当該遺伝子に関連する ID 情報だけではなく GO term や遺伝子のシンボル名などの具体的なリソースも複数含まれており，Web サービスのバックエンドとして非常に利便性の高いサービスであると言える．これに対して G-Links は，より広い利用方法の想定や広範囲の生物種および遺伝子情報のサポートを行うことで，汎用性の高い ID 変換シス

テムとして実装されている．UniProt にてサポートしている全生物種についての遺伝子を対象としている他，入力として 85 のデータベースの ID を利用し，そこから得られる ID 情報や記述情報をユーザに提供する．TSV や JSON といった汎用的なフォーマット出力による外部プログラムからの連携の他，HTML 出力による情報閲覧や RDF 出力などのサポートも行っているため，より広いデータ統合およびリソース収集の用途に利用することができる．

3.5.3 G-Links で扱うべきデータ

G-Links ではユーザの入力を UniProt ID へ変換する ID 変換部分と，その UniProt ID に関連する情報を取得するアノテーション部分の 2 カ所でデータベースを構築している．この 2 つのデータベースを構築するにあたって複数の生物学データベースの情報を統合しているが，これら両者では内部に格納すべきリソース選定の基準が大きく異なる．

ID 解決部分はユーザの入力を UniProt ID へ直接変換する箇所であり，この部位でどれだけ多くの ID をサポートしているかどうかがサービス全体の入力に対する汎用性となるため，このデータベース部分では対応できるデータは多ければ多いほどよいといえる．それゆえに取り込むべきリソースとしては Link 情報を中心に保存しておりその Link の精度が高いといえる既存の ID 変換サービスの ID 変換テーブルが候補としてあげられ，これらの取り込みは急務であると考えられる．しかしながら汎用性を持たせるために取り扱うデータが増えることはレイテンシの問題を引き起こすと考えられる．研究者が頻繁に利用すると考えられる主要データベースの ID について個別のテーブルを作成して先に検索を行う，複数のテーブルに全 ID 情報を分割保存し並列に検索処理を行う，アンダースコアの有無などの ID の構造の特徴ごとにテーブルを作成することで一度の検索に利用するテーブルのサイズを小さく保つ，などの対策が今後必要になる．

アノテーション部分のデータベースでは複数のデータベースから抽出した Link 情報および生物学情報を UniProt ID に結びつけた形で格納している．ここでも多くのデータをユーザに提供すること自体には一定の価値があるものの，その Link 情報を抽出するために用いるデータベースおよびリソースは多ければ多いほどよい訳ではなく，一定の基準で選定するべきだと考えられる．1 つ目の理由は生物学リソースの数およびデータ量の大きさと計算速度の問題である．G-Links では記述情報など ID 情報以外のデータセットも提供するため，対象となるリソースの種類が多くなると出力するリソースのデータ量が非常に大きくなる．生物学データベースの現存数とその増加速度を鑑みると，それら全ての持つ ID 情報やコンテンツまで統合し出力するのは，そこから出力された巨大なデータセットのフォーマット変換などに必要な計算資源の問題やネットワークの転送速度を鑑みると現実的ではない．2 つ目が生物学 LOD ネットワークの持つスモールワールド性である．生物学データベースにおける各リソース間の Link ネットワークは非常に密であり，スモールワールド性を持っていることが知られている．さらに，良質で利用者が多いデータベースは多くの Link が集まることでネットワークのハブとして働いており，Rich get richer の性質からより多くの Link が集中することも考慮すると，ハブデータベースは外部データベースからの Link を集めるだけの豊富かつ正確な生物学情報や他データベースの ID を持っているため，情報源としての価値が高いデータベースとなる．G-Links のメンテナンスコストや Link を多数辿ることによる情報信頼性の低下などの理由から，上記の性質を持つハブデータベースの情報を中心に統合することで，少ない運用コストの元でより広範囲で大量かつ正確な生物学情報および Link 情報を取得することが可能となる．この点において，十分なエントリー数を持っている上でデータの更新頻度も高く，品質の高いクロスリファレンスを多数保持している UniProt を Primary Key として採用したことには非常に大きな価値があると考えられる．3 つ目の問題点が出力データ量

の増加による情報密度の低下である．生物学情報を保持するデータベースは世の中に多く存在し，それらのデータの品質はデータベースによってバラバラである．アノテーションの正確性を判定する基準の差異や，更新頻度の差によって情報の正確性に差異が生まれるケース，そもそもデータの更新が停止しているケースもある．そのため，闇雲に全生物学データベースの統合を行うことによる解析用データ量の増加は，解析の幅の増加やサンプル数増加による解析精度の向上などのメリットもあるが，ユーザの解析に必要ない種類のデータの存在やデータの重複，更新遅れによる不正確なデータなどのノイズ情報が混入する可能性の増加にもつながる．ユーザに対してより価値の高いデータセットを提供するためには，これらのデメリットによってデータセット自体の情報量の低下がおこることは好ましくないため，G-Links では闇雲に全てのデータベースの情報を取り込まず，研究者にとって意味のある情報を多く持つと考えられるデータベースにしぼって統合を行い，そこから得られた情報をユーザに提供している．

これらの理由から，G-Links では UniProt から得られる情報を中心に UniProt が持っていない Link を持っているデータベースの情報を統合していくことで，より価値の高いデータセットの構築を行っている．このようにして生成されたデータセットに対して filter や extract オプションを用いることで，ユーザはより自身の解析に適した情報量の高いデータセットを G-Links から取得し，利用することが可能になる．

3.5.4 ロケーション問題の解決

従来の ID 解決によるデータ統合問題における大きな問題点の一つが，その ID が示すリソースのロケーション解決問題であった．非 URN な ID に対してロケーション解決を行うための取り組みはこれまで多くされており，LSID において用いられた DNS によるロケーション解決，PURL のリダイレクトによる情報の存在保証，Identifires.org などのプロジェクトがその解決を行ってきた．G-Links では，主に ID 入力時のロケーション解決と出力される情報に関するロケーション解決の 2 つの面から解決を行っている．

一つ目の問題点が，ID だけではその ID が示すリソースまでたどり着けないという問題である．ID 自体にはその ID が利用できる場所の情報は存在しておらず，何らかの方法でその ID が利用できるデータベース名やその ID と対応しているリソースを示す URI を取得しなければならなかった．G-Links ではこの問題に対して，全ての ID を入力として受け付けることで解決を行っている．遺伝子を示す ID であればどのような ID でも G-Links は入力として受け付け，ユーザから受け取った ID を UniProt ID に変換するにあたって，全ての ID と UniProt ID の対応表が入ったデータベースを用いることで，その ID が所属するデータベース名を自動的に解決する．このように全ての入力に対して所属するデータベース名というロケーション情報を自動で推測し解決を行うことで，全ての遺伝子情報を示す ID を擬似的な LSID のように運用することが可能である．ユーザは自身が持っている ID についてどのデータベースの ID なのかというロケーション情報を考慮する必要なく，G-Links のベース URL という共通のロケーションを用いることで，その ID が示すリソースとそのリソースの関連情報を含めてアクセスすることが可能になる．このとき考慮しなければ行けないのが複数のデータベースで同じ ID が使われているケースであるが，G-Links では ID 解決で得られた UniProt ID 全てについてのデータセットをユーザに提供する．ユーザが任意に入力した ID についてこれ以上のロケーションの自動解決は不可能であるため，候補となる遺伝子群やそれらに関する生物学リソースのセットを提供を行い，その生物学リソースセットの情報をもとにユーザに自身が目的とした遺伝子に関する情報を選択してもらうことで，より正確な情報を提供する．なお，G-Links では ID だけではなく生物学における URN の一種である USA

形式での入力も可能であるため，ユーザからのより正確なリソース指定を行うこともできる．また，G-Linksから出力されるリソースについてはIDだけではなくそのIDが示すリソースに対応するURLも共に提供する．このようにIDとロケーション情報の双方を提供することで，ユーザはロケーションの解決を行わなくてもリソースへのURIを利用できる．

二つ目の問題点がリソースの存在保証である．分散して存在するリソースをID変換によって統合する場合，各IDに対してロケーションを解決しURIを取得した後，そのURIが指し示す場所にリソースが存在するかどうか，そのURIが持つロケーション情報が正確で最新のものかどうかを保証する必要がある．G-Linksの出力によって得られる各種URLについては，その正確性を毎月のデータアップデート時にテストを行うことでロケーション情報の正確性および存在の保証を行っている．これらのロケーション情報はデータベースではなく定義ファイルとして別に管理されており，緊急の修正に対しても容易に対応できる形になっている．しかしながら，この手法では過去にG-Linksから得られた結果に含まれるURLについてリソースの存在保証ができないという問題点が存在する．この問題に関しては，今後G-Linksで扱っている全URLに対するPURLサーバを提供することで解決を行う予定である．

3.5.5 外部Webサービスとの連携と出力データセットの拡充

G-Linksでは，ユーザに入力された遺伝子IDに関連する他データベースID群だけではなく，それらのIDが示すリソースのURLやその遺伝子に関する記述情報などのID以外の生物学リソース，それに加えて外部Webサービスによって得られる解析結果へのURLを出力する．解析に利用しているWebサービスは全てRESTサービスであり，G-Linksによって得られたID情報を用いたUSAを入力とすることでURLによる解析結果のリソース指定が可能になる．このようにID情報以外の情報を提供することで，Semantic Webと同様に各IDが示すドキュメントの中に存在する個別の生物学情報や，Webサービスによって得ることができる生物学的なデータをもユーザに提供することが可能になる．各Web解析サービスの出力はその出力リソースを示すURL情報として提供されるため，G-Linksの出力自体には各サービスの動作時間などのレイテンシの問題はほぼないと考えられる．現在出力として利用できるWebサービスはG-language EMBOSS RESTサービスおよびそれに含まれるKBWSの一部メソッドであり，KBWSはG-language EMBOSS RESTサービスを経由することでRESTインタフェースから結果を取得する．KBWSは様々なゲノム解析用のWebサービスについて標準化されたインタフェースのもとでアクセスすることが可能であるため，KBWSがサポートしている全てのツールについてG-Linksの出力から機械的に連携を行うことができる．このWebサービスとの連携について，現在のG-LinksではBLASTなどの単一遺伝子についてのメソッドのみのサポートとなっており，MAFFTのような遺伝子セットに対してのアプローチを行う解析ツールや，G-language REST Serviceのメソッドである*gcskew*のようなゲノムに対してアプローチを行うメソッドなどを取得することができない．また，KBWSでは複数配列に対しての解析やゲノムに対しての解析を行うサービスが多くサポートされているため，G-Linksと連携できているサービス数は少ない．今後，KBWSを拡張することでより多くの解析ツールのサポートを行うと同時に，単一遺伝子以外について解析を行うサービスへの対応も行うことで，より有益なデータセットを提供する必要がある．

3.5.6 オントロジーの問題

G-Links はその結果を RDF で出力することで，既存の Semantic Web 技術と高い親和性の上で連携して利用することが可能である．オントロジーは基本的に，各生物学データに関する語彙だけではなく解析ツールや解析手法を表す語彙までサポートを行っている EDAM Ontology を用いており，EDAM Ontology に適切な語彙が存在しないものについてのみ UniProt Core Ontology を用いることで，各リソース間の Link の意味情報を提供している．外部 Web サービスの出力をサポートしている G-Links にとって EDAM の語彙の広さは最適であり，データの中心を UniProt においていることからサポートとして UniProt Core Ontology を用いることでより適切な意味情報付けを行うことができると考えられる．また，EDAM Ontology は EMBOSS における内部メタデータの管理にも用いられているため，EMBOSS の拡張パッケージである KBWS と非常に相性がよいという利点も存在する．

しかしながら，G-Links の RDF にはいくつかの問題点が存在する．第一に，G-Links のオントロジーは開発者である私自身が適切であると判断したものを各リソースに定義しているため，その定義に厳密性がないと言える．さらに EDAM Ontology は現在，その語彙の定義情報を Web で安定的に提供できておらず，その EDAM Ontology の単語が示す意味情報を参照するには EDAM Ontology の定義アーカイブファイルを参照する必要がある．これらの問題に対しては，G-Links が提供する全てのリソースについてそれらを定義するためのオントロジーを独自に定義し，その定義情報のホスティングならびに既存のオントロジーとのマッピング情報を提供することで解決することが可能である．既存のオントロジーにマッピング可能な独自オントロジーのみを用いて意味情報を定義することで，より柔軟かつ統一性のある RDF リソースをユーザに提供することができる．また各生物学リソースが遺伝子情報にのみ結びついているため，網羅的な関係性ネットワークを構築することができず，G-Links の出力のみで Semantic Web による意味推論を行うことは難しいという問題点も存在する．しかしながら G-Links の RDF がサポートしているデータベースは非常に多く，その多領域にわたる生物学情報をサポートしている RDF を容易かつ高速に生成することができるという特徴から，直接の Link や共通の ID を持たず統合することができなかった RDF リソース間の Link の解決を行うなどの Semantic Web のバックエンドとして利用することが可能である．

第4章　結論

4.1　インタフェースの画一化によるWebサービス相互運用性の向上

第2章では複数のバイオインフォマティクスWebサービスに関するインタフェースの画一化とローカルツールを含めた形で各解析サービスをシームレスに連携させるためのシステムの設計と実装を目的としている．この目的に対して本論文では，様々なプロバイダが提供している解析Webサービスに対するプロキシモデルの採用と，そのプロキシサーバが提供するWebインタフェースに対してアクセスすることができるUNIXコマンドラインツール群をEMBOSSの追加パッケージとして実装することで問題の解決を行った (Oshita *et al.*, 2011)．本システムKBWSでは，全てのサービスに対して統一的なインタフェースでアクセスできるプロキシSOAPサービスを提供することで，REST・SOAP・CGIといったプロトコルの違いやサービスプロバイダごとに異なるパラメータ名などの差異をプロキシサーバで吸収することが可能である．また，それらのWebサービスへのアクセスを行うツールをEMBOSSという入出力が厳密に管理されたパッケージングシステムの一部として作成することで，Webサービスを直接利用するユーザもコマンドラインツールを利用するユーザも，サポートしている全てのWebサービスをどれも同じ方法で利用することができる．

バイオインフォマティクス解析を行うためには複数のアルゴリズムや解析ツールを組み合わせ，一つの解析フローとして扱う必要がある．その解析フローを効率よく構築・実行することを考えた場合に重要な要素が，各ステップ間がスムーズに連携するための相互運用性，特定のステップに必要なサービスを容易に発見できるサービス探索能，発見したサービスを各ステップに自在に差し替えや追加を行うためのインタフェースの画一化の3点である．この問題についてWebサービスのみであればこれまでも多く議論されていたが，ローカルツールまで含めた議論はほとんどされていなかった．そこで，内包している400以上の解析ツールについてこの問題を解決しているEMBOSSのアーキテクチャに各Webサービスを落とし込むことで，ローカルツールまで含めたWebサービスの相互運用性や高いサービスの発見可能性の確保，インタフェースの画一化が可能となった．

本システムではオリジナルのWebサービスに対するプロキシサーバモデルを採用することで，ユーザに対して画一的なインタフェースを持つWebサービスのセットを提供している．ユーザはどれか一つのサービスの利用法さえ習得するれば他の全てのサービスを同様のアクセス方法で利用することが可能になるため，解析フローを構築する際に本システムがサポートしている全てのサービスを容易に差し替え・追加することが可能になる．また，プロバイダによって利用できるデータベースやパラメータが異なる同一内容のWebサービスを複数サポートし，ユーザが入力したパラメータやメソッドにあわせて，そのパラメータセットが利用できるプロバイダへアクセス先を判断・変更を行うことで，対象の解析Webサービスについてユーザからのより広い要求をサポートすることができる．

KBWSは42もの既存Webサービスをサポートしており，それら全てに対してプロキシサーバが提供しているSOAPサービス，およびCUIからアクセスすることが可能である．プロキシ

サーバはSOAPサーバとして実装されているため，ユーザは本システムのサービスを様々な環境やプログラミング言語から利用することができる．さらにTavernaなどのGUIワークベンチを用いることで，CUIやプログラミングが苦手なユーザであっても他のバイオインフォマティクスWebサービスと高い相互運用性のもとで容易に連携させ，複雑な解析フローを構築することができる．本システムはEMBOSSの追加パッケージとしても実装されていることも大きな特徴である．EMBOSSの基本インタフェースであるCUI上で各種Webサービスを運用することで，ユーザは動作結果を逐次見ながら最適なパラメータやツールを探索する試行錯誤やパイプ機能によるツールの連携を，ローカルツールとWebサービスが混在した形でシームレスに行うことができる．また，EMBOSSが対応する各種インタフェースからも全てのサービスが利用可能である．

4.2 遺伝子ベースのID変換を用いた生物学リソースの効率的統合

生物学における実験技術や実験機器の急速な高度化によるハイスループット化は，バイオインフォマティクスという大規模データに対して計算資源を用いることで高速かつ網羅的な解析処理を行う分野の発達を促した．実験によって得られた大規模な生物学データを効率的に管理・運用するためにデータベースが広く発展し，現在では数千ものオープンに公開されたデータベースを研究者は自由に用い，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．しかしながら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を多領域に渡る複数のデータベースから収集し，そこで得られたデータセットを研究者が利用しやすい形で統合し，そのデータセットから自身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はバイオインフォマティクスにおける大きな課題の一つであった．

この問題に対して，第3章では遺伝子情報を中心に置いた多領域生物学リソースの統合モデルの提唱と実装を行うことで解決を行った．生物学データベースでは各データベースのID間に多くのLinkが張られており，そのLinkを辿ることでIDの変換を行い，各データベース間の関係性情報を取得することができる．このLinkネットワークに対して，遺伝情報は全て遺伝子に結びつけることができるというセントラルドグマの考え方をベースにしたデータ統合モデルを採用することで，遺伝子を示すIDを中心としてその他の情報全てが遺伝子情報へと紐付いている遺伝子集約型のLinkネットワークに再構築した．このLinkネットワークはユーザは多領域情報への経路を維持したままでLink数を減らすことに成功しており，132にも渡る生物学データベースからユーザが指定した遺伝子に関連する情報を網羅的かつ高速に取得し提供することが可能となっている．

本章で実装を行ったシステムG-Linksでは，ユーザが与えた遺伝子を示すIDについて，その遺伝子に関連する情報を132のデータベースから網羅的かつ高速に収集しユーザに提供する．G-Linksでデータを取得するためには遺伝子IDを含んだURLにアクセスをするだけでよいため，ユーザはどのような遺伝子IDであってもG-LinksのベースURLさえ知っていればその遺伝子IDが示すリソースを取得することができる．更にG-Linksは遺伝子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．また，G-Links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，それに対する抽出プロセスをも条件に合致する遺伝子のフィルタリングと出力する情報のフィルタリングという2

種類の抽出機能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．これらの特徴を踏まえてG-LinksはWebアプリケーション構築時の高速バックエンドとして利用することが可能であり，その一例としてGenieという生物学用研究アシスタントサービスの構築を，慶應義塾大学大学院政策・メディア研究科の荒川和晴特任講師および慶應義塾大学環境情報学部 板谷英駿氏と共同で行った．Genieは遺伝子情報やゲノム情報に関する情報収集および解析を行うことができるシステムであり，ユーザが自然言語で問いかけると，その内容を自然言語処理によって解釈し，対応する検索クエリやWebサービスを呼び出し，その結果を返す．生物学研究ではユーザが解析作業を行った結果に対して解釈を行うことで生物学的な知見や知識を導出する必要がある．Genieは出力として単純な検索クエリではなく人が理解するための自然言語による記述情報や画像データ，検索結果のサマリとグラフ，Webサービスによる解析結果などを提供することで，データに対する解釈のプロセスまで自動化し，研究者をサポートすることができる．

これらの特徴に加えG-Linksでは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．G-LinksではG-language EMBOSS RESTサービスとそれに含まれるKBWS RESTサービスというURLを用いることで解析結果リソースを指定できるサービスを採用することで，URIにてリソースを指定するID変換ベースのデータ統合において容易に統合することを可能にした．また，多数の解析Webサービスに対して画一的なインタフェースを提供するKBWSを経由することで，高いメンテナンス性のもと，G-Linksへの新たなサービス拡張を容易に行うことが可能となっている．

G-LinksではこれらのリソースをRDFで出力することが可能であり，既存のSemantic Web技術とシームレスに連携させることが可能である．Semantic WebはWWW上の全てのドキュメントに対して意味情報を加味した推論や解析のアプローチを行うことができるという非常に画期的なプロジェクトであり，現在も多くのプロジェクトが技術開発を行なっている．しかしながら，計算資源の問題やSemantic Webで扱うためのRDFリソースの生成難度などの問題から，現段階では限定的な技術であると言える．G-Linksは後者の問題に対して多領域の生物学リソースに関するRDFを高速生成することで解決を試みた．ユーザが必要とする大規模なRDFリソースを容易かつ高速に生成出来るだけでなくそのリソースを一意のURLで指定できるG-Linksは，Semantic Webが持つこれらの問題を肩代わりするプラットフォームとなりうる存在であると言える．

4.3 総括

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．第2章で多数のバイオインフォマティクスWebサービスに対して画一的なインタフェースを提供することで，その相互運用性の確保とそれによる効率的な連携の実現を行った．プロキシWebサービスとEMBOSS追加パッケージという2段階の標準化を行なっているため，CUIからでもプログラミングからでも他の解析ツールや解析Webサービス群とシームレスに連携を行い，複雑な解析ワークフローを構築・実行することが可能になる．

第3章では，Webサービスや生物学データベースまで含めた多領域生物学情報を効率的に統合し，そこからユーザが必要なデータセットを高速かつ自動的に取得するためのシステムの構築を

行った．本システムでは遺伝子情報を中心に全ての生物学情報の統合を行うというデータモデルを採用することで，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かつ効率的に探索・取得することを可能にしている．遺伝子や遺伝子セット，生物種を示すID，および塩基/アミノ酸配列という汎用的な入力に対応しており，これらの情報が含まれたURLにアクセスをするだけでユーザが対象とする遺伝子に関連する生物学情報を網羅的かつ高速に取得できる．複数のオプションを組み合わせることで，そのデータセットからユーザが必要とする情報だけを抽出する作業もURLの指定だけで実現可能である．また，本システムでは生物学データベースの情報だけではなく解析Webサービスによって出力される生物学情報のサポートも行なうことで，データベースだけにとどまらない全ての生物学Webリソースの統合を可能にしている．連携された各WebサービスはKBWSを経由することで画一化されたRESTインタフェースを通して利用されるため，KBWSの拡張を行うことで容易に新規Webサービスの追加を行うことが可能である．

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

本研究を行うにあたっては非常に多くの方にお世話になりました．まず初めに，学部時代に研究会に所属した当初から大学院にいたるまでアドバイザーを引き受けてくださり，多数の助言をくださった慶應義塾大学大学院政策・メディア研究科の荒川和晴特任講師に感謝を申し上げたいと思います．研究に関する細かい相談から私事まで多くの面でもお世話になり，私の研究生活を非常に楽しく有意義で価値のあるものにしていただきました．改めてここで，深く感謝の意を表明したいと思います．

河野暢明氏には研究会に所属する以前から公私にわたってお世話になりました．技術的な指導や研究自体の指針についてなどの幅広く深い議論やアドバイスは，研究自体の質だけではなく私自身の論理性も高めてくれました．また，自身が所属する研究グループメンバーの方々とのディスカッションは本研究を進めるにあたって非常に重要で，何事にも代え難いものでありました．アドバイジーとして共に研究生活を送ってくれた板谷英駿氏と宮原太陽氏や，同研究グループメンバーであった木戸信博氏をはじめとする，池上慶太氏，野崎慎氏，小川真菜氏，長谷部百合子氏，小峰栞氏，石黒宗氏，石野響子氏，橋本詩織氏，川本夏鈴氏，西脇友哉氏，吉田祐貴氏らに深く感謝申し上げます．

慶應義塾大学環境情報学部 内藤泰宏准教授と慶應義塾大学環境情報学部 佐野ひとみ専任講師には研究生活だけでなく，学業の面でもお世話になりました．また，同じ研究会のメンバーである小川隆氏，西野泰子氏，玉木聡志氏，新土優樹氏，大久保周子氏，真流玄武氏もこの場を借りて感謝申し上げます．

慶應義塾大学環境情報学部 萩野達也氏には本研究について情報工学的な視点から多くの議論をして頂き，情報工学的な視点から非常に有意義なアドバイスを多数頂きました．ここに感謝の意を評させて頂きます．

所属プロジェクト以外の友人にも多く助けていただきました．特にInteraction Design Projectの秋山博紀氏，白崎琢也氏，山本 伶氏には研究以外の様々な面でお世話になりました．本論文を作成する上で使用させていただいたLaTeXテンプレートの作者である田中佑樹氏にも心から感謝致します．

学部時代から6年間，様々な面から私の研究生活を支え，全力で背中を押してくれた両親に心から感謝致します．

最後になりましたが，本研究を行うにあたって多くの面からご支援をしてくださった慶應義塾大学環境情報学部 冨田勝教授にこの場をお借りして感謝の意を表させていただきます．

参考文献

Al-Masri, E. and Mahmoud, Q. H. (2008). Investigating web services on the world wide web. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, 795–804, New York, NY, USA. ACM.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**(3), 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**(17), 3389–3402.

Arakawa, K. and Tomita, M. (2006). G-language System as a platform for large-scale analysis of high-throughput omics data. *J. of Pest. Sci.*, **31**, 282–288.

Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y., and Tomita, M. (2003). G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics*, **19**(2), 305–306.

Arakawa, K., Suzuki, H., and Tomita, M. (2008). Computational Genome Analysis Using The G-language System. *Genes, Genomes, Genomics*, **2**, 1–13.

Arakawa, K., Oshita, K., and Tomita, M. (2009a). A web server for interactive and zoomable Chaos Game Representation images. *Source Code Biol. Med.*, **4**, 6.

Arakawa, K., Tamaki, S., Kono, N., Kido, N., Ikegami, K., Ogawa, R., and Tomita, M. (2009b). Genome Projector: zoomable genome map with multiple views. *BMC Bioinformatics*, **10**, 31.

Arakawa, K., Kido, N., Oshita, K., and Tomita, M. (2010). G-language genome analysis environment with REST and SOAP web service interfaces. *Nucleic Acids Res.*, **38**(Web Server issue), W700–705.

authors listed, N. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**(D1), D8–D20.

Bairoch, A. (2000). The enzyme database in 2000. *Nucleic Acids Res.*, **28**(1), 304–305.

Bairoch, A., Boeckmann, B., Ferro, S., and Gasteiger, E. (2004). Swiss-Prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**(1), 39–55.

Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, **41**(5), 706–716.

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.

Bussey, K. J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W. C., Zeeberg, B., Ajay, W., and Weinstein, J. N. (2003). MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.*, **4**(4), R27.

Carver, T. and Bleasby, A. (2003). The design of Jemboss: a graphical user interface to EMBOSS. *Bioinformatics*, **19**(14), 1837–1843.

Chatr-Aryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**(D1), D816–823.

Clark, T., Martin, S., and Liefeld, T. (2004). Globally distributed object identification for biological knowledgebases. *Brief. Bioinformatics*, **5**(1), 59–70.

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.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.*, **14**(6), 1188–1190.

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.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**(23), 4636–4641.

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.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**(5), 1792–1797.

Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S., Wiebe, V., Kitano, T., Monaco, A. P., and Paabo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**(6900), 869–872.

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.

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. Ph.D. thesis, University of California, Irvine. AAI9980887.

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., Bairoch, A., Schomburg, D., Tipton, K. F., and Apweiler, R. (2004). IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**(Database issue), D434–437.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garcia-Giron, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kahari, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S. M. (2013). Ensembl 2013. *Nucleic Acids Res.*, **41**(D1), 48–55.

Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., and Kanehisa, M. (1998). DBGET/LinkDB: an integrated database retrieval system. *Pac. Symp. Biocomput.*, 683–694.

Gordon, P. M. and Sensen, C. W. (2007). Seahawk: moving beyond HTML in Web-based bioinformatics analysis. *BMC Bioinformatics*, **8**, 208.

Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K. (2009). Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**(4), 465–473.

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.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S. L., Tacker, M., and Schuster, P. (1994). Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, **125**, 167–188.

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., and Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**(Web Server issue), W585–587.

Huang, d. a. W., Sherman, B. T., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2008). DAVID gene ID conversion tool. *Bioinformation*, **2**(10), 428–430.

Huang, d. a. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**(1), 44–57.

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.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**(Web Server issue), W729–732.

Imanishi, T. and Nakaoka, H. (2009). Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Res.*, **37**(Web Server issue), 17–22.

Jacso, P. (2004). Thoughts about federated searching. *Information Today*, **21**(9), 17–20.

Juty, N., Le Novere, N., and Laibe, C. (2012). Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.*, **40**(Database issue), D580–586.

Kall, L., Krogh, A., and Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**(5), 1027–1036.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**(Database issue), D109–114.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.

Katayama, T., Arakawa, K., Nakao, M., Ono, K., Aoki-Kinoshita, K. F., Yamamoto, Y., Yamaguchi, A., Kawashima, S., Chun, H. W., Aerts, J., Aranda, B., Barboza, L. H., Bonnal, R. J., Bruskiewich, R., Bryne, J. C., Fernandez, J. M., Funahashi, A., Gordon, P. M., Goto, N., Groscurth, A., Gutteridge, A., Holland, R., Kano, Y., Kawas, E. A., Kerhornou, A., Kibukawa, E., Kinjo, A. R., Kuhn, M., Lapp, H., Lehvaslaiho, H., Nakamura, H., Nakamura, Y., Nishizawa, T., Nobata, C., Noguchi, T., Oinn, T. M., Okamoto, S., Owen, S., Pafilis, E., Pocock, M., Prins, P., Ranzinger, R., Reisinger, F., Salwinski, L., Schreiber, M., Senger, M., Shigemoto, Y., Standley, D. M., Sugawara, H., Tashiro, T., Trelles, O., Vos, R. A., Wilkinson, M. D., York, W., Zmasek, C. M., Asai, K., and Takagi, T. (2010a). The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*. *J. Biomed. Semantics*, **1**(1), 8.

Katayama, T., Nakao, M., and Takagi, T. (2010b). TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services. *Nucleic Acids Res.*, **38**(Web Server issue), W706–711.

Katayama, T., Wilkinson, M. D., Vos, R., Kawashima, T., Kawashima, S., Nakao, M., Yamamoto, Y., Chun, H. W., Yamaguchi, A., Kawano, S., Aerts, J., Aoki-Kinoshita, K. F., Arakawa, K., Aranda, B., Bonnal, R. J., Fernandez, J. M., Fujisawa, T., Gordon, P. M., Goto, N., Haider, S., Harris, T., Hatakeyama, T., Ho, I., Itoh, M., Kasprzyk, A., Kido, N., Kim, Y. J., Kinjo, A. R., Konishi, F., Kovarskaya, Y., von Kuster, G., Labarga, A., Limviphuvadh, V., McCarthy, L., Nakamura, Y., Nam, Y., Nishida, K., Nishimura, K., Nishizawa, T., Ogishima, S., Oinn, T., Okamoto, S., Okuda, S., Ono, K., Oshita, K., Park, K. J., Putnam, N., Senger, M., Severin, J., Shigemoto, Y., Sugawara, H., Taylor, J., Trelles, O., Yamasaki, C., Yamashita, R., Satoh, N., and Takagi, T. (2011). The 2nd DBCLS BioHackathon: interoperable bioinformatics Web services for integrated applications. *J. Biomed. Semantics*, **2**, 4.

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.*, **537**, 39–64.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**(4), 656–664.

Kono, N., Arakawa, K., Ogawa, R., Kido, N., Oshita, K., Ikegami, K., Tamaki, S., and Tomita, M. (2009). Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS ONE*, **4**(11), e7710.

Labarga, A., Valentin, F., Anderson, M., and Lopez, R. (2007). Web services at the European bioinformatics institute. *Nucleic Acids Res.*, **35**(Web Server issue), 6–11.

Lassmann, T. and Sonnhammer, E. L. (2006). Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. *Nucleic Acids Res.*, **34**(Web Server issue), W596–599.

Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., and Wanner, B. L. (2005). Minimum information requested in the annotation of biochemical models (miriam). *Nat. Biotechnol.*, **23**(12), 1509–1515.

Lim, A. and Zhang, L. (1999). WebPHYLIP: a web interface to PHYLIP. *Bioinformatics*, **15**(12), 1068–1069.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**(5), 660–665.

Lord, P., Alper, P., Wroe, C., and Goble, C. (2005). Feta: A light-weight architecture for user oriented semantic service discovery. In A. G. Pérez and J. Euzenat, editors, *Proceedings of the European Semantic Web Conference 2005*, volume 3532 of *Lecture Notes in Computer Science*, 17–31. Springer-Verlag.

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**(5), 955–964.

Lukashin, A. V. and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**(4), 1107–1115.

Mackey, A. J., Haystead, T. A., and Pearson, W. R. (2002a). Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell Proteomics*, **1**(2), 139–147.

Mackey, A. J., Haystead, T. A., and Pearson, W. R. (2002b). Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell Proteomics*, **1**(2), 139–147.

Martin-Requena, V., Rios, J., Garcia, M., Ramirez, S., and Trelles, O. (2010). jORCA: easily integrating bioinformatics Web Services. *Bioinformatics*, **26**(4), 553–559.

McWilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., Miyar, T., and Lopez, R. (2009). Web services at the european bioinformatics institute-2009. *Nucleic Acids Res.*, **37**(Web Server issue), W6–10.

Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T., and Goble, C. (2010). Taverna, Reloaded. In M. Gertz and B. Ludäscher, editors, *Scientific and Statistical Database Management*, volume 6187 of *Lecture Notes in Computer Science*, chapter 33, 471–481. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Mudunuri, U., Che, A., Yi, M., and Stephens, R. M. (2009). bioDBnet: the biological database network. *Bioinformatics*, **25**(4), 555–556.

Nakai, K. and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, **11**(2), 95–110.

Nakamura, Y., Cochrane, G., and Karsch-Mizrachi, I. (2013). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **41**(D1), D21–D24.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**(1), 205–217.

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.

Object Management Group (1991). The common object request broker : architecture and specification.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A., and Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences: Research articles. *Concurr. Comput. : Pract. Exper.*, **18**(10), 1067–1100.

Oshita, K., Arakawa, K., and Tomita, M. (2011). KBWS: an EMBOSS associated package for accessing bioinformatics web services. *Source Code Biol. Med.*, **6**, 8.

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.

Ramirez, S., Munoz-Merida, A., Karlsson, J., Garcia, M., Perez-Pulido, A. J., Claros, M. G., and Trelles, O. (2010). MOWServ: a web client for integration of bioinformatic resources. *Nucleic Acids Res.*, **38**(Web Server issue), W671–676.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**(6), 276–277.

Rios, J., Karlsson, J., and Trelles, O. (2009). Magallanes: a web services discovery and automatic workflow composition tool. *BMC Bioinformatics*, **10**, 334.

Roure, D. D., Goble, C., and Stevens, R. (2009). The design and realisation of the myexperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, **25**, 561–567.

Sarachu, M. and Colet, M. (2005). wEMBOSS: a web interface for EMBOSS. *Bioinformatics*, **21**(4), 540–541.

Sato, K., Hamada, M., Asai, K., and Mituyama, T. (2009). CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**(Web Server issue), W277–280.

Saunders, B., Lyon, S., Day, M., Riley, B., Chenette, E., Subramaniam, S., and Vadivelu, I. (2008). The Molecule Pages database. *Nucleic Acids Res.*, **36**(Database issue), D700–706.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**(Database issue), D674–679.

Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996). Entrez: molecular biology database and retrieval system. *Meth. Enzymol.*, **266**, 141–162.

Senger, M., Rice, P., and Oinn, T. (2003). Soaplab - a unified sesame door to analysis tools. In *In UK e-Science All Hands Meeting*, 509–513.

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.

Shafer, K., Weibel, S. L., Jul, E., and Fausey, J. (1996). Introduction to persistent uniform resource locators. In *Proceedings of International Networking Conference INET'96*, BFC–1–BFC–9, Montréal, Canada.

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.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**(1), 22–28.

The Gene Ontology Consortium (2013). Gene Ontology Annotations and Resources. *Nucleic Acids Res.*, **41**(D1), D530–D535.

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.

van Engelen, R. A. and Gallivan, K. A. (2002). The gSOAP Toolkit for Web Services and Peer-to-Peer Computing Networks. In *Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid*, CCGRID '02, 128, Washington, DC, USA. IEEE Computer Society.

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.

Wilkinson, M. D., McCarthy, L., Vandervalk, B., Withers, D., Kawas, E., and Samadian, S. (2010). SADI, SHARE, and the in silico scientific method. *BMC Bioinformatics*, **11 Suppl 12**, S7.

Wilkinson, M. D., Vandervalk, B., and McCarthy, L. (2011). The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. *J. Biomed. Semantics*, **2**(1), 8.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W., and Su, A. I. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**(11), R130.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., and Brinkman, F. S. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**(13), 1608–1615.

Zhou, J. and Rudd, K. E. (2013). EcoGene 3.0. *Nucleic Acids Res.*, **41**(D1), D613–624.

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

2013年5月30日　　初版発行

著者　大下和希

監修　冨田勝

発行　慶應義塾大学 湘南藤沢学会
〒252-0816　神奈川県藤沢市遠藤5322
TEL:0466-49-3437

Printed in Japan　　印刷·製本　　ワキプリントピア

ISBN 978－4－87762－266－4
SFC-MT 2013–001

■本論文は修士論文において優秀と認められ、出版されたものです。