生物分野におけるWeb APIの適用

Web API for Biology

重元 康昌^{1*}, 桑名 良和², 權 娟大³, 菅原 秀明⁴

Yasumasa SHIGEMOTO, Yoshikazu KUWANA, Yeondae KWON, Hideaki SUGAWARA

*1 富士通株式会社

Fujitsu Limited

〒144-0035 東京都大田区南蒲田2-16-1 NOFテクノポートカマタセンタービル

E-mail: yshigemo@genes.nig.ac.jp

2 株式会社東海ソフトウェア

Tokai Software

〒410-0011 静岡県沼津市岡宮 1463-86

E-mail: ykuwana@genes.nig.ac.jp

3 国立遺伝学研究所

National Institute of Genetics

〒411-8540 静岡県三島市谷田 1111 生命情報・DDBJ 研究センター

E-mail: yekwon@lab.nig.ac.jp

4 国立遺伝学研究所

National Institute of Genetics

〒411-8540 静岡県三島市谷田 1111 生命情報・DDBJ 研究センター

E-mail: hsugawar@genes.nig.ac.jp

生物の研究現場ではさまざまな解析プログラムやデータベース、文献などの情報があり、それらを組み合わせることが多い。そこで、国立遺伝学研究所では120以上の機能についてWeb APIを開発し公開した。これによって、利用者は国立遺伝学研究所の大規模システムが手元にあるかのようにさまざまなサービスを自由に組み合わせて利用することができるようになるとともに、組み合わせのノウハウをワークフローとして共有することも可能になった。本稿ではこの生物分野におけるWeb APIの適用事例を紹介する。

There are various analytical programs, databases and bibliography in biology and they are often integrated and used. Therefore, we at National Institute of Genetics (NIG) developed Web APIs of over 120 functions and opened them to the public. As a result, users of NIG systems can combine several services as if the user has the large scale system. Furthermore, the know-how of the combination can be shared as a workflow. We introduce the application of Web API in biology.

キーワード: Web サービス, 相互運用性, ワークフロー, バイオインフォマティクス, ナビゲーション

Web service, interoperability, workflow, bioinformatics, workflow navigation

1 はじめに

生物学の研究者はある1つの研究・解析を行うためにさまざまなシステムを利用する。このときに最も時間がかかり、大変な作業の1つは、あるシステムの出力を別のシステムの入力とするために変換することである[1]。この変換は手作業によるコピー・ペーストで実施されることが多く作業が煩雑であるとともに、手順が残らないためノウハウが共有されない問題もある。そこで、これらの問題を解決するために国立遺伝学研究所 DNA Data Bank of Japan (DDBJ) では生物分野における Web APIサーバ (WABI) を公開した。また、複数のWeb APIを部品としてつなげたワークフローやある解析の結果から次に遷移可能なサービスを列挙するナビゲーションシステムも構築した。

2 生物分野におけるWeb API

2.1 Web APIとは

APIは Application Programming Interface の略で、プログラミング言語から利用可能な機能や関数の集合を表す。そして、Web APIはAPIの実体がインターネット上にあるAPIであり、それらのAPIを組み合わせることで新たなサービスを作るマッシュアップ技術も注目されている。

2.2 生物分野におけるWeb API

生物分野ではさまざまなデータベースや解析プログラム、文献などがあり、これらを統合して使うことが多い。そこで、DDBJでは、プログラミング言語から利用可能なSimple Object Access Protocol (SOAP[2]) や REpresentational State Transfer (REST) サーバを開発し公開した。このサーバでは、表1に示すようなデータベース検索・取得や解析など21サービスを公開している。

表1 国立遺伝学研究所で公開しているWeb APIサービス

サービス名(Web API数)	サービス概要
DDBJ (7), ARSA (4), GetEntry (44)	20以上の生物データベースにおけるキーワード検索やデータ取得
Blast (6), ClustalW (4), Fasta (5), VecScreen (4)	相同性検索や多重アライメントなどのDNA配列やアミノ酸配列を用いた解析
Gib (11), Gtop (3), GTPS (8), GIBV (8), GIBEnv (1), GIBIS (1), SPS (2)	DDBJオリジナルデータベース(微生物/ウィルスゲノム、環境由来配列、ゲノムのORF再評価、タンパク構造)
TxSearch (5), RefSeq (1), GO (3), Ensembl (4), OMIM (2), NCBI Genome Annotation (4)	DDBJ以外の研究所で開発されたデータベース

2.3 Web APIの利用方法

国立遺伝学研究所で公開しているWeb APIの1つであるRESTサービスの利用方法を説明する。RESTサービスは PerlやJavaなどのプログラミング言語で利用することが可能である。

国立遺伝学研究所で構築しているDNA

のデータベースDDBJには1億件を越すDNA配列が収録され、各エントリにはユニークIDであるアクセッション番号が割り振られている。Perl言語を用いてRESTサーバからアクセッション番号を指定して1件のDNA配列のエントリを取得する例を図1に紹介する。

```
use LWP::UserAgent;
$ua = new LWP::UserAgent;

# 接続情報の設定
$req = new HTTP::Request POST => 'http://xml.nig.ac.jp/rest/Invoke';
$req->content_type('application/x-www-form-urlencoded');

# パラメータの設定 (アクセッション番号の指定)
$req->content('service=GetEntry&method=getDDBJEntry&accession=AB000100');

# RESTサーバへ接続
$res = $ua->request($req);

# 結果の表示
print $res->content;
```

図1 RESTサービスの利用例

2.2.4 SOAPとREST

国立遺伝学研究所では、当初SOAPサービスを開発し公開していたが、全DNA配列であるゲノム情報やそれらの情報を用いた解析結果の情報が大きくなり、SOAPでは大規模なデータを扱う際にパフォーマンスの劣化が激しいことからRESTサービスも追加して公開した。例えば大腸菌ゲノムのエントリは約13MBあり、SOAPで取得する場合は19.4秒かかるがRESTで取得する場合、5.1秒しかかからない。さらに、RESTではさまざまな言語で同じように利用することができるとともにWebブラウザやtelnet, wget コマンドを使っても利用することができる。一方、SOAPも利用者のコーディング量が少なくすみ、拡

張性に優れているなどの利点があるため、両方継続して公開している。

3 ワークフロー

3.1. ワークフローとは

ワークフローは一連の手続きの流れのことである。国立遺伝学研究所で公開しているWeb APIを部品として組み合わせることでワークフローを簡単につくることができる。国立遺伝学研究所では以下に示す8つのワークフローを公開している。

- ホモロジーワークフロー

DNA配列を用いたヒトと他の生物種の比較・解析

- ・ヒトの染色体の遺伝子ワークフロー
染色体ごとの遺伝子数をカウントし
ダウン症などの疾患との関連探索
- ・DNA多型ワークフロー
人種により異なるDNA多型解析
- ・BLAST-ClustalWワークフロー
配列解析プログラムBLASTとClustalW
を組み合わせたワークフロー
- ・スプライシングワークフロー
スプライシング (遺伝子の構造) の観
点でヒトとマウスの遺伝子の比較解
析
- ・OMIMワークフロー
ヒトの疾患関連遺伝子と他の生物の
遺伝子との比較
- ・SNPワークフロー
ヒトの遺伝子とSNP (個人差がある部
位) の関係抽出

3.2 ホモロジーワークフロー

Web APIを組み合わせたワークフローについて説明するために、例としてホモロジーワークフローを取り上げる。

3.2.1 ワークフロー概要

DDBJにはヒトをはじめチンパンジーやマウスなどさまざまな生物種のDNAデータが登録されている[3]。このデータを利用することで複数の生物種における遺伝子の比較解析を行うことができる。そこで、ヒトの遺伝子と同様な遺伝子を持つ生物種とその類似度をひとまとめに検索することが可能なワークフローを構築した。図2にUMLを用いたワークフローのアクティビティ図を示し、表2にワークフローに ABO を入力した場合の結果例を示す。なお、ABO は血液型を決定する遺伝子名である。

図2 DNA配列解析ワークフロー

表2 ワークフローの結果例 (ヒトの遺伝子ABOの結果例)

生物種名	DDBJ ID (アクセッション番号)	DNA配列の 類似度
ゴリラ	AY138476	98.69
カニクイザル	AF100981	95.79
アカゲザル	AF071830	95.66
チンパンジー	AF071831	97.90
アヌビスヒヒ	AF019417	96.37

3.2.2 ワークフローの手順

ワークフローの手順を説明する。まず、

ARSAシステム[4]を用いて入力されたヒトの遺伝子名でDDBJデータベースを検索

する。次にその検索結果から最初のエントリのDNA配列をgetentryシステムによって取得する。そして、そのDNA配列を元に他の生物種のDNA配列と類似しているものを検索するBLASTシステムを実行し、その結果から生物種名の取得などを行い整理することで表2のような結果を取得することができる。ここで使われているARSAやgetentry、BLASTはWeb APIである。

4 ワークフローナビゲーションシステム

4.1 ワークフローナビゲーションシステムとは

Web APIを組み合わせることでホモロジーワークフローのようなさまざまなワークフローを作成することができる。しかし、Web APIは120以上もあるため、利用者がそれらをどのように組み合わせてよいか分からないこともある。また、さまざまな部品(Web API)を試行錯誤して組み合わせながらワークフローを構築するような利用方法も想定される。そこで、このような利用形態に対応するために、Web APIをWebブラウザで実行することができ、さらにその実行結果を用いて次に遷移可能な部品を列挙するワークフローナビゲーションシステムを開発した。

4.2 ワークフローナビゲーションシステムの構成

ワークフローナビゲーションシステムは、以下の3つの部品で構成される。

- メタ情報

各Web APIのカテゴリ情報やメソッド

名、パラメータ名などを定義する。

- 辞書

各Web APIの入出力情報やカテゴリ情報を標準化するための辞書。入出力情報を辞書で共有化することによって、あるWeb APIの出力の型と合致する入力情報を受け付けるWeb APIを検索することによって、次に遷移可能なWeb APIを列挙することが可能になる。

- Webインタフェース生成システム

メタ情報や辞書を参照することによってWeb APIをWebブラウザ上で実行することが可能なページを自動的に生成する。

4.3 ワークフローナビゲーションシステムの利用例

ワークフローナビゲーションシステムで複数のWeb APIを遷移する例として、1件のDDBJエントリを取得し、BLASTを用いて解析する手順を図3に説明する。図3-1はDDBJデータベースからDDBJアクセス番号を指定して1エントリを取得するページの例である。この画面はメタ情報や辞書を参照して自動的に生成される。適当なアクセス番号を入力し、ボタンをクリックすると、図3-2に示すように実行結果とその実行結果を用いて次に遷移可能なサービスが列挙される。そして、その列挙されたサービスのいずれかをクリックすると図3-3に示すように再びその実行画面が自動生成されるとともに、前の実行結果から抽出されて画面に入力される。

3-1

3-2

次に遷移可能なサービス

3-3

DNA配列でDDBJに対してBLAST実行

DDBJを参照データベースとしてBLAST (blastn) を実行します。クエリ配列としてDNA配列を指定することが可能です。この機能は計算が終了するのに時間がかかることがあります。

図3 ワークフローナビゲーションシステム

5 おわりに

ワークフローナビゲーションシステムを利用することにより、Web APIをただ単にプログラムから利用するだけでなく新たなワークフローを提案することもできるようになった。より複雑なワークフローにも対応できるように、今後もWeb APIを拡充し、ワークフローナビゲーションシステムの機能を拡張していきたい。

謝辞

本研究開発は、科学技術振興機構 (JST) BIRD 事業の支援で実施しています。

参考文献

- [1] Stein, L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119-120.
- [2] Sugawara, H. and Miyazaki, S. (2003) Biological SOAP servers and Web services provided by the public sequence data bank, *Nucleic Acids Res.*, **31**, 3836-3739.
- [3] Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. (2004) DDBJ in the stream of various biological data, *Nucleic Acids Res.*, **32**, D32-D34.
- [4] Sugawara, H., Ogawara, O., Okubo, K., Gojobori, T., and Tateno, Y. (2008) DDBJ with new system and face, *Nucleic Acids Res.*, **36**, D22-D24.