バイオ基幹情報資源の高準化と共用化

国立遺伝学研究所生命情報・DDBJ 研究センター
菅原　秀明

Enhancement of quality and usability of primary biological information resources

Hideaki Sugawara
Center for Information Biology and DDBJ, National Institute of Genetics

We have substantially enriched the biological information resources by Web API for Biology（WABI）, a metadatabase and method ontologies. WABI provides 124 APIs in 18 functional categories, 13 sample workflows, a workflow navigation system and CookBook. In addition, workflows have been applied to actual DDBJ services. The metadatabase improves the accessibility of primary biological information resources by accumulating access methods of 189 databases selected by the analysis of citations in biomedical literature. The method ontologies for wet experiments and bioinformatics are prepared for the refined retrieval of information on methods, protocols and workflows.

１．はじめに

現代のバイオ研究は、専門化かつ高度化する実験技術と情報技術によって膨大かつ多様なデータを産出する。このデータは、図１に示すように、そのままあるいは解析や解釈を加えられて、種々の文献データベースやさまざまな観点のファクトデータベースの形をとってバイオ情報資源として流通する。これらの多彩なバイオ情報資源を選択・統合利用することから新たなバイオ研究が展開し、バイオ情報資源も拡大再生産される。一方で、どのコンテンツが信頼できるのか、どのようにすればメソッドを使えるのか、多彩なバイオ情報資源をどのように組み合わせれば良いのか、は自明ではない。そこで、バイオ基幹情報資源の水準を高めて（高準化）、作り手とは異なる観点からでも情報資源を容易に利用可能とすること（共用化）を試みた。具体的には、

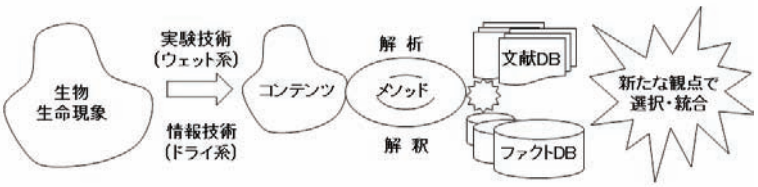

図１　バイオ研究における情報資源の生成と利用

- バイオ基幹情報資源の一つである日本 DNA データバンク（DDBJ）を主な対象とする Web サービスとワークフローに関する研究
- バイオ基幹情報資源のメタデータベースに関する研究

- 実験技術（ウエット系）と情報技術（ドライ系）のメソッドの発見と利用を促進するメソッド・オントロジーに関する研究

を進めた。

2. 研究開発の成果

2.1 Web サービスとワークフローに関する研究

2.1.1 Web API for Biology (WABI)

(1) Web サービスの高準化

DDBJ は、データ駆動型研究の展開を見越して 2002 年に、コンピュータプログラムから DDBJ のコンテンツやメソッドを活用できるように Web サービスを開始したが[1]、2006 年からの本研究課題において、Web API for Biology（WABI）としてその大幅な拡充と共用化を果たした[2]。

例えば、大量データへの対応やクライアントとの相互運用性の向上のために、当初採用した通信規約 SOAP によるサービスに、通信規約 REST によるサービスを加えた。この拡充によって、WABI においては、ゲノム配列も無理なく扱えるようになり、Java、Perl プログラミング言語に加え、Ruby、Python、C 言語も利用可能となった。

また、利用者からの要望に応えて、DDBJ が提供してきたキーワード検索や BLAST 検索も高速化した（DDBJ サービスの高準化）。その結果、WABI を利用して利用者がローカルに構築したワークフローが実用に耐えるようになり、結果的にバイオ分野におけるクラウドサービスを展開することになった。

この他にもきめ細かな改良を加え続けたこともあり、年間の利用者数（1 年間にアクセスしてきたユニークな IP 数）が 2006 年以後毎年増え続け、2010 年には 3,800 人を超えた。また、2010 年の Web API のメソッド利用回数は 837 万回を超えた。WABI に対する信頼が拡がってきている証左である。

(2) Web サービスの共用化

今では、さまざまなバイオ情報資源で Web API が用意されるようになったが、ワークショップなどで「どのような Web API があるのか分からない」と言われることが多い。そこで、Wiki サイト CookBook（http://wabi.ddbj.nig.ac.jp/CookBook_ jp/）を用意して、WABI の Web API に加えて以下の機関の Web API 群も登録させていただいた：

- DBCLS （http://togows.dbcls.jp/）
- G-language （http://www.g-language.org/）
- H-InvDB （http://www.jbic.or.jp/activity/i_db_pj/h-inv_db.html）
- PDBj（http://www.pdbj.org/index_j.html）
- EBI （http://www.ebi.ac.uk/Tools/webservices/）
- NCBI （http://eutils.ncbi.nlm.nih.gov/）

CookBook では、MediaWiki（http://www.mediawiki.org/）のテンプレート機能を用いて Web API の仕様を標準形式で構造化している。したがって、CookBook を使って、世界中に散在している多様なバイオ情報資源から、求める機能を持っている Web API を見つけ出すことができる。

また、このテンプレートには Web API の入出力型も定義されているので、ある Web API の出力型と別の Web API の入力型を照合して、それらをワークフローの部品として組み合わせることが可能か否かを判定できる。この機能は、次項で報告するワークフロー・ナビゲーションシステムで実現した。

2.1.2 ワークフローの例示からワークフロー・ナビゲーションシステムへ展開[2]

我々は、Web API を品揃えするとともに、CookBook によって他機関が提供する Web API の情報も提供してきた。さらに、複数の Web API を部品とするワークフローについても、ヒトの遺伝子頻度解析や遺伝子と疾患の相関分析などをモデルとするワークフローを構築・提供した。しかし、以下のような理由から研究現場で具体的なバイオの課題を解決するワークフローを開発することはなかなか難しいようである：

- 多様な部品があるため、それらの組み合わせ（ワークフロー）は膨大な数になる。
- その中で、試行錯誤を繰り返しながらワークフローを設計・開発する必要がある。

そこで、図 2 に示すワークフロー・ナビゲーションシステム／Workflow Navigation System（以下、WNS）を構築した。

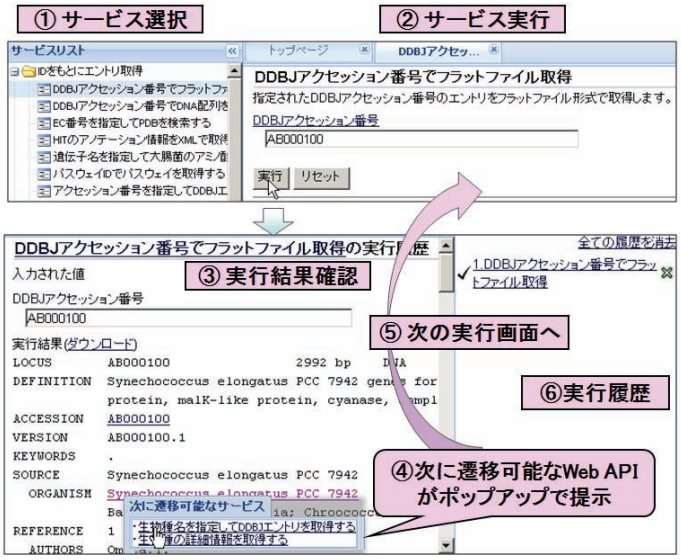

図 2　ワークフロー・ナビゲーションシステム画面例

図 2 の①から⑤を Web ブラウザ上で繰り返すことで、利用者は Web API を次々につなげていくことができる。この WNS は、CookBook（2.1.1（2）項参照）に登録された内容をもとに、Web API の選択画面や実行画面を自動生成している。したがって、WNS によって、Web ブラウザ上で多様な Web API を複数組み合わせたワークフローを試行することができる。また、WNS に利用者登録すれば、実行履歴をサーバに保存し、再利用することもできる（図 2 の⑥）。

2.1.3 ワークフローを恒常的サービスへ展開

(1) 微生物ゲノム情報の高準化（Gene Trek in Prokaryote Space(GTPS)）[3]

GTPS では DDBJ で練り上げたワークフローに従って、International Nucleotide Sequence Database （INSD, http://www.insdc.org/）から公開されている細菌とアーケアの完全ゲノム配列を、毎年１回その時点での最新参照データベースと照合し直して、再アノテーションを施し、ORF の確からしさをランク付けして、その結果をデータベースとして公開している。年間の一意な訪問者数と訪問件数が、2009 年の 4,897/46,299 から 2010 年の 21,509/154,631 へと、伸びてきている。

(2) 微生物ゲノム自動アノテーション（Microbial Genome Annotation Pipeline(MiGAP)）

微生物ゲノム配列を高速に自動解析するワークフローを設計して、MiGAPとして2009年6月に公開した。解析累積件数および解析塩基総数ともに着実に増加し、公開後18カ月間の解析累積件数が780件、解析塩基総数が1.5Gbases余り、予測したCDS総数が139万個余りに達した。

(3) DDBJ も新世代シーケンサ由来データに対応（DDBJ Sequence Read Archive（DRA））

DDBJ として、米国 National Center for Biotechnology Information（NCBI） と欧州 European Bioinformatics Institute（EBI）と協調して[4]、新世代シーケンサからの１次データを対象とする DDBJ Sequence Read Archive （DRA）の登録・管理・公開のワークフローを設計し DRA システムとして実装した [5]。このシステムは順調に稼働し、2010 年 12 月末までに 358 件 （データサイズ約 3.2 TB）の登録を捌いた。

2.2 メタデータベースに関する研究

ワークフローの展開に寄与することを目的に、Nucleic Acids Research の 2006 年データベース特集号[6]に収録されていたサイトについて調査と分析を行い、メタデータベースを構築・公開した（http://www.ps.noda.tus.ac.jp/biometadb/index.html）。

他のメタデータベースに対する第１の特徴は、データベースを利用するCGIやJava Appletを分析してその構造と出力形式もメタデータとして蓄積し再利用した点である。したがって、このメタデータベースによって、各バイオ情報資源の機能とその使い方を効率よく知ることができ、ひいては、新たなWeb APIとワークフロー開発の必要性も判断することができる。

第２の特徴は、PubMedに収録された論文のアブストラクトへの出現頻度を元に、バイオ基幹情報資源189件を選定して、これらについては特に詳細な分析を継続してメタデータベースを更新している点である。

第３の特徴は、バイオ情報資源の場合、インターネット上で見える記述と文献での記述の対応関係が曖昧であったり不明であったりする場合が多いため、その対応関係を整理した点である。

2.3 メソッド・オントロジーに関する研究

バイオ情報資源利用をメソッドの観点から有効利用できるように、ウェット系とドライ系それぞれのメソッドとワークフローについてオントロジーを中心に研究を行った。

2.3.1 ウェット系メソッド・オントロジー

- 実験手法名の収集と辞書の構築

  分子生物学プロトコールの洋書、生化学実験講座や各種ラボマニュアルなどの和書、国立情報学研究所のバイオポータルプロジェクトの用語辞書、ライフサイエンス統合データベースプロジェクト（以下、統合プロジェクト）の学術用語集などを対象として、実験手法の用語を収集した。また、実験書の目次の構造や、Protocol Online などのウェブサイトの構造を軽量なオントロジーと見立てて、上位や下位の用語の整理に利用した。最終的に収集した用語は統合プロジェクトの TogoDB システムを利用して公開した（http://togodb.dbcls.jp/）。TogoDB からは辞書の閲覧編集の他、用語の検索やダウンロードが可能である。

- 実験手法に関する検索サービスの開発

  前項の辞書やオントロジーを活用して、通常のサーチエンジンに比べて、精度よくメソッドを検索できるサービスを公開した（http://lifesciencedb.jp/dbsearch/bird/）。検索対象は、東大分子細胞生物学研究所や東大医生化学教室など、各研究機関や研究室によって維持公開されているラボマニュアル、蛋白質科学会アーカイブプロトコールなど学協会のプロトコール、ターゲットタンパク研究プログラムが手法の収集と標準化のために構築したデータベース等の国内のサイトの他、Nucleic Acids Research、Nature Protocol、Nature Methods などの実験手法に特化した論文誌、タカラやシグマなどメーカーが提供するサイトである。検索結果には、本研究課題のメタデータベース、ドライ系のメソッドのデータベース、ならびに WABI のワークフローを参照した結果も含まれる。

  この検索サービスは統合プロジェクトの横断検索システムを利用しているため、本研究課題が終了後も JST のバイオサイエンスデータベースセンター（http://biosciencedbc.jp/）のサービスとして継続される見込みである。一方、ウェブサイトや論文に対するデータベースやドライ系メソッドのマッピングは現在人手で行っているため、参照を最新のものに保つためには、本成果を発展させ自動化することが必要である。

2.3.2 ドライ系メソッド・オントロジー

ドライ系のメソッドを再利用可能な形式に記述するための辞書やオントロジーを構築し、これを利用してバイオロジーの文献情報からワークフロー情報を抽出してデータベース化し、ワークフロー検索システムを一般に提供することを目指した。

- 文献情報からのワークフロー抽出手法の開発

  手作業による試行を経て自動抽出の手法を開発し、自動抽出の後に手作業で修正、補完する方法によりワークフロー抽出をルーチン化したが、数百の論文の解析を現実的にするためには、自動抽出の精度をさらに上げる必要がある。

- 文献情報からのワークフローの収集

  前項のワークフロー抽出手法を使って、真核および原核生物のゲノム解析分野 390 論文と発現解析分野 280 論文からワークフローを抽出してデータベース化した。また、バイオ分野全体から文献を選定し、ツール名とデータベース名を抽出し、集計結果か

らバイオ分野全体で使われている解析ツールやデータベースの俯瞰マップを作成した。

- メソッド・オントロジーの構築

  初期に真核生物のゲノム解析分野の論文30報から手作業で抽出したワークフローをもとに、解析ツールやデータベース、入出力データ等のワークフロー構成要素を概念化・分類するためのオントロジーをRDF/RDFS形式にて構築した。その後、ワークフロー抽出対象論文の拡充に伴い、オントロジーを追加更新した。

- ワークフロー検索システムの開発

  キーワード、カテゴリー、関連ツール・ワークフロー推薦機能、WABIとメタデータベースへのリンク機能及びバイオ分野俯瞰マップ機能を備えたワークフロー検索システムを2010年度末までに公開する予定である。これによって文献に記述されたワークフローを、WABIの部品を使って手軽に再構築して実行することも可能になろう。

3. まとめ

Web API とワークフローの研究開発は、DBCLS の TogoWS や Biohackathon[7]、海外の Taverna（http://www.taverna.org.uk/）、myExperiment （http://www.myexperiment.org/）、BioMoby（http://www.biomoby.org/）、など国内外で盛んに行われている。WABI は、利用者が年々着実に増えていることに加えて、TogoWS、Taverna、myExperiment などでも登録・利用されていることから、これからますますバイオ基幹情報資源として国内外で重要な役割を果たしていくものと思われる。今後は、バイオ情報資源の高準化と共有化をさらに促進するために、セマンティック Web（http://www.w3schools.com/semweb/default.asp）や Liked Data（http://linkeddata.org/）の実装を拡充することも検討したい[8]。

バイオ・メタデータベースは、各バイオ情報資源の仕様を一つ一つ分析することなく直接利用することを可能としたために、バイオ情報資源の検索だけでなく、新たなワークフローの設計にも貢献する。一方で、各情報資源のアクセス方法は時とともに変更されるため、メタデータの維持更新を継続する必要がある。

メソッド・オントロジーは、メソッドの観点からバイオ情報資源に新たな観点のタグを付与し、実験のプロトコールや解析のワークフローの再利用を促す試みでもある。本研究課題でタグ付けされた実験プロトコールやワークフローは今後 WABI の機能向上に貢献していくが、そこで止まらずに、タグ付けをバイオ分野全般に広げていきたい。また、メソッドを主題としたジャーナル Nature Methods（http://www.nature.com/nmeth/）と Journal of Visualized Experiments（http://www.jove.com/）が、それぞれ 2004 年と 2006 年に刊行され、Nature Protocols（http://www.nature.com/nprot/index.html）も 2006 年に稼働し、また、YouTube にも実験手法に関する動画が多数アップされていることからも、メソッド・オントロジーを拡充して、メソッドの観点からのタグ付けを普及させたい。

4. 研究開発実施体制

代表研究者　菅原　秀明（国立遺伝学研究所生命情報・DDBJ 研究センター）

研究開発題目

(1)「Web サービスとワークフローに関する研究」

グループリーダー　菅原　秀明（国立遺伝学研究所生命情報・DDBJ 研究センター）

(2)「バイオ・メタデータベースに関する研究」

グループリーダー　宮崎　智（東京理科大学薬学部）

(3)「メソッド・オントロジーに関する研究」

グループリーダー　川本　祥子（ライフサイエンス統合データベースセンター）

5. 参考文献

[1] Sugawara H, Miyazaki S. Biological SOAP servers and web services provided by the public sequence data bank. Nucleic Acids Res. 2003 Jul 1;31(13):3836-9

[2] Kwon Y., Shigemoto Y., Kuwana Y., Sugawara H., (2009) “Web API for biology with a workflow navigation system,” *Nucleic Acids Res*., 37:W11-6

[3] Kosuge T., Abe T., Okido T., Tanaka N., Hirahata M., Maruyama Y., Tomiki A., Kurokawa M., Himeno R., Fukuchi S., Miyazaki S., Gojobori T., Tateno Y., Sugawara H., (2006) “Exploration and grading of possible genes in 183 bacterial strains by a common fine protocol lead to new genes: Gene Trek in Prokaryote Space (GTPS),”*DNA Res*., 13, 245-254.

[4] Shumway M., Cochrane G., Sugawara H.,(2010)“Archiving next generation sequencing data,” *Nucleic Acids Res*., 38: D870-871

[5] Kodama Y., Kaminuma E., Saruhashi S., Ikeo K., Sugawara H., Tateno Y., Nakamura Y.,(2010)“Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technologies,” *Adv Exp Med Biol*., 680:125-35.

[6] Galperin M. Y. (2006) “The Molecular Biology Database Collection: 2006 update” *Nucleic Acids Res*., 34: D3-D5

[7] Katayama T, *et al.* (2010) “The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium”*J Biomed Semantics.* Aug 21;1(1):8.

[8] 重元康昌、桑名良和、宮本 青、權 娟大、菅原秀明、“WABI から SABI への展開（Web サービスからセマンティック Web サービスへ）”、第 33 回日本分子生物学会年会、第 83 回日本生化学会大会、4P-1204、神戸、2010 年 12 月