

研究開発課題別中間評価結果

1. 研究開発課題名

バイオ基幹情報資源の高準化と共用化

2. 代表研究者名

菅原 秀明 （大学共同利用機関法人情報・システム研究機構 国立遺伝学研究所 特任教授）

3. 研究開発概要

バイオ分野の研究成果は長らく文献に蓄積されてきたが、今日では、文献データベースに加えて多様なファクトデータベースにも蓄積されている。また、データベースからデータ・情報・知識を抽出するツール群も Web サイトに蓄積されている。現代のバイオ研究には、種々の手法を参考にしながら、分散している Web サイトから有用なバイオ情報資源（データベースやツール）群を探し出し、それらを組み合わせて利用できる情報環境が必須である。本研究は、バイオ情報資源の水準を高め（高準化）かつバイオ情報資源の利用者の利便性を高め（共用化）て、バイオ分野の問題解決をもたらす情報環境の充実に貢献することを目的としている。

既存のバイオ情報資源、ウェット系の実験手法ならびにドライ系の解析手法を調査分析して、それらのコミュニティによる共用を可能とする記述方式を考案している。同時に、多様な情報資源をコンピュータプログラムから直接呼出し可能とする「Web サービス Application Program Interface」（Web API）を準備し、Web API そのものと共に Web API で構成されるワークフローも新たな情報資源として提供している。

4. 中間評価

4-1. 研究開発の進捗状況と今後の見込み

PubMed に収録されたアブストラクトに出現する頻度をもとに選定したバイオ基幹情報資源について、それらが提供するデータセット、ツール、利用法などを分析した結果を、バイオ・メタデータベースに蓄積・公開している。メタデータの 1 種として、利用者が Web ブラウザで検索や解析を実行した際に、Web ブラウザから Web サーバに送られる指示（CGI プログラム）を含んでいる。また、Java applet 利用ページの増加に伴い、CGI テンプレートパラメータの解析ツールを Java applet 対応に拡張している。

バイオ分野の実験手法とデータ解析手法を共用可能な形で記述するために、文献からウェット系の実験手法名を収集してオントロジーを試作している。文献を利用した実験手法名抽出の進捗は遅れている。ドライ系の解析手法については、ゲノム解読分野を対象を絞って、文献情報からワークフローを収集してデータベースに蓄積するとともにオントロジーを試作しており、予定どおりに進捗している。

DDBJ が提供するデータベースとツールを主な対象として Web API を拡充し、また、それらを部品とするワークフローの例も拡充するとともに、ワークフローの理解と生成を助けるナビゲーションシステム、Web API for Biology（WABI）を構築している。WABI にかかわるノウハウ共有のための Wiki サイトも構築している。

メソッド・オントロジーのように当初の目論見どおりには進捗していないものがある一

方、ナビゲーションシステムのように当初予定していなかった開発が進められているものもあり、概ね順調に進捗しているといえる。WABI の機能を拡張し、さらに使いやすく改良して一般の利用に公開している。今後は、特にユーザから見てそれを利用する意義をより明確にしていく必要があると考える。

4-2. 研究開発成果の現状と今後の見込み

web 公開状況 <http://www.jst-bird2.nig.ac.jp/>

バイオ基幹情報資源 189 件のメタデータをデータベースに収録して検索可能としている。多くのサイトが複数のデータセットやツールを提供しているため、メタデータベース登録のレコード件数は 2 万件に達している。バイオ基幹情報資源を記述する標準形式に到達したと考えられる。

実験書より収集した実験手法に関する専門用語、英語（約 500 用語）、及び日本語（約 500 用語）については用語集として平成 20 年度末までに公開が予定されている。また、ゲノム解読分野に対象を絞った文献情報からのワークフロー抽出作業に想定以上の負荷がかかり、自動抽出システムの構築を余儀なくされたが、結果的に、ワークフローの大量抽出への道が開け、30 件の論文からのワークフローを蓄積したデータベースは、平成 20 年度末までに公開予定である。

キーワード検索からアミノ酸配列の解析までを、22 種類のサービスにおいて 213 メソッドを開発・公開し、加えて、15 種類のワークフローを構築・公開している。また、Unified Modeling Language (UML) を使ってワークフロー編集のツールに依存しない標準形式でワークフローを記述することを試みている。

本課題で開発された Web サービス API はかなりユーザに利用されているようであるが、どのような研究にどれくらい効果があったかが明確ではない。また基盤整備に徹しているためか、このプロジェクトの研究成果が見えにくいように思われる。

4-3. 今後の研究開発にむけて

バイオ・メタデータベース、メソッド・オントロジーおよび Web サービスとワークフローのいわばモジュールはこれまでに概ね構想通り研究開発が進んでいる。しかし、これら 3 つのモジュールの間の連携については、ドライ系オントロジーが提案する機能の記述と Web サービスの機能の記述との対応付けなど、一層の努力が必要である。

セマンティック Web 技術の現状調査の結果を参考にしながら、3 つのモジュールの拡充と 3 つのモジュールの連携強化を基盤にしたボトムアップで、バイオ分野における問題解決環境の場の構築を目指している。この過程で、DDBJ のサービスに対する Web API のメソッドを改修・拡充し、ワークフローのサンプルも拡充することとしている。また、メソッド・オントロジーの事例も大幅に増加する見込みである。

研究開発は、必要な人材をほぼ揃えて進められており、今後の研究開発項目も、ばらつきはあるものの比較的妥当である。セマンティック Web の開発よりも、現在進めている開発間の連携強化に今後の方向を定めていることは評価できる。利用者のニーズを汲み取れるような体制にしていくことが望ましい。何が必要とされているかに関する客観的な検討も望まれる。

4-4. 国内外のデータベース高度化・標準化の動向・状況と本課題の位置づけ

バイオ・メタデータベースにおいては、文献アブストラクトへの出現頻度を手がかりに

選定した基幹情報資源について、利用可能な CGI の形式と実例を自動抽出してデータベース化したところに特徴がある。バイオ・メタデータベースのレコードを参照することで、利用者は各情報資源の仕様を一つ一つ分析することなく、コンピュータプログラムからの高度な利用を実現することができる。一方で、各情報資源の仕様は時とともに変更されるため、バイオ・メタデータベースの保守を継続することが必要である。

メソッドに関しては、従来の実験書に加えて、Nature Methods が刊行されたり、実験操作のビデオ投稿を受け付ける Journal of Visualized Experiments が PubMed の採録対象雑誌になるなど、ダイナミックな共用について関心が高まっている。また、メソッドのデータベース化についても、米国の Protein expression purification and crystallization DataBase (<http://pepcdb.pdb.org/>) や我が国のターゲットタンパク研究プログラムの例のように、取り組みが行われている。しかし、メソッド記述の標準とその運用については、未だ試行錯誤が続いており、本研究開発課題では、文献、Web API、ワークフロー、データベースなどに対してメソッドの観点からのタグ付を可能にすることを目指した研究開発となっている。

Web API とワークフローの研究開発は国外には先行事例があり、BioMoby や Taverna のようなプロジェクトにおいて記述の標準化が試みられている。また、国内でもライフサイエンス統合データベースプロジェクトによって、多様な Web サービスを統一的に扱えるプロキシサーバの提供も始まっている。それらを踏まえ、本研究開発課題においては、各グループの分担課題を貫く整合性を確保しながら、Web API の拡充、ワークフローの拡充ならびにコミュニティの拡充に注力することとしている。その上で、成功事例を具体的に示すことで実質的な国際標準の構築に貢献することを目指している。

上記のように国際的な動向に足並みを揃え、わが国が持つデータベースの高度化・標準化の研究開発が進められているが、国内外の他の Web サービス提供元との連携が図られているかどうかは明らかではない。作成したもののインパクトを判断する方策の検討も必要と思われる。

4-5. 研究開発成果の社会への貢献

バイオ・メタデータベースとメソッド・オントロジーによって、要素となるバイオ情報資源を発見し、インターネット上の Web サイトに散在する Web API を組み合わせることでワークフローを短時間で構築することができる。しかし、一般の利用における効果の程度が明確でない面があり、現在のまま進んだ場合、限られたユーザの利用しか期待できないのではないかと懸念される。今後の研究の方向性や期待される成果が明確ではないように思われる。

5. 総合的評価

多様なデータベースを使うための Application Program Interface の開発を行い、公表したことは評価される。メソッド・オントロジーを取り入れることにより、解析目的に適した新規なワークフローが提示されるような例が多く出てくるのか、予測が難しいように思われた。むしろ、ワークフローシステムのさらなる拡充や、ウェット系の研究者が遂行可能なデータ解析の種類と規模を拡大することが有効であろうと思われる。また、積極的な広報活動を行い、利用者のニーズをより詳細に反映した方策を取ることが望まれる。