A-01

メタデータを利用した生物情報解析ワークフローの作成支援手法の提案

A Proposal for a Method to Support Bioinformatics Workflow Composition using Metadata

福本 貴紀† 木戸 善之‡ 瀬尾 茂人† 竹中 要一† 松田 秀雄†
Takanori Fukumoto Yoshiyuki Kido Shigeto Seno Yoichi Takenaka Hideo Matsuda

1. はじめに

生物情報科学の分野では、研究に必要なデータが生物種や DNA やタンパク質の配列や構造といった種類ごとに別々のデータベースに格納されており、それら複数のデータベースの情報を統合して利用したいという要求がある。しかし、生物情報データベースは、地理的に分散して多数存在し（文献[1]によると 1230 個）、データ量も大きく頻繁に更新されるため、一箇所の Web サイトで物理的に統合することは困難であり、広域に分散したデータベース群を利用する手法の開発が求められている。

そこで、各データベースの Web サイトで提供される Web サービスを利用して、各研究者が自分の行いたい解析に応じてワークフローを構成することで、複数のデータベースから必要な情報を取得することが広く行われている。ここでワークフローとは、Web サービスの処理手順を定義することで Web サービスの連携を自動的に行うことができる技術である。

現在、アメリカの National Center for Biotechnology Information (NCBI)では、Entrez Utilities[2]と呼ばれる DNA 配列データベースや生物分類等に関するデータベースからデータを取得するための Web サービス群を提供している。ヨーロッパの European Bioinformatics Institute (EBI)では、DNA やタンパク質配列データベースからのデータ取得や、DNA やタンパク質配列の解析のための EBI Web Services[3]を提供している。日本の国立遺伝学研究所生命情報・DDBJ 研究センターでは、DNA やタンパク質配列、タンパク質構造等に関するデータベースからのデータ取得や、DNA やタンパク質配列の解析のための Web API for Biology (WABI) [4]を提供している。京都大学化学研究所バイオインフォマティクスセンターでは、パスウェイや薬等に関するデータベースからデータを取得するための KEGG API[5]を提供している。また、BioCatalogue[6]には世界中の生物情報解析 Web サービスに関する情報が集められており、検索により利用者が望む Web サービスを探すことができる。

また、GUI を利用した生物情報解析ワークフロー作成のためのツールとして Taverna[7]が存在する。また、既存の生物情報解析ワークフローを蓄積したレポジトリとして myExperiment[8]が存在する。

生物情報解析ワークフローの作成では、前述のように複数のデータベースの情報を統合的に利用できるようにしたいという要求がある。しかし、データベースが変わると、取得できるデータ形式も変わる。取得できるデータ形式が変わると、そのデータ形式に対応する入出力に変更する必要があるため、Web サービスも変更する必要がある。ワークフローで使われる Web サービスが変わると、入出力が変わるためデータを受け渡すことができる Web サービスも変わるため、ワークフロー内の構成も変わる。このことから、ワークフローで対応するデータベースを変更するとワークフロー内の Web サービスの構成を変更しなければならない。しかし、データベースを変更する場合にどの Web サービスに変更するか、どのように Web サービスを連携させるかの情報が無いため、利用者はどのようにワークフロー内の構成を変更すればよいか把握しにくいという問題があった。

本研究では、データベースと Web サービスを対応づけるためのメタデータを作成し、それを利用して利用者が既存ワークフローで使われるデータベースを変更した時に、解析処理自体は原則変更せずに既存ワークフローで使われる Web サービスを変更することで利用可能なデータベースを増やし、より多くの情報を得られるようなワークフロー作成の支援を実現する。先行研究として、tf・idf を利用した類似ワークフロー検索がある[9]。この研究でも、異なるデータベースを用いた同種の処理を行う既存ワークフローを検索可能である。しかし、変更したいデータベースの指定ができない。提案手法では、変更したいデータベースを指定することが可能である。

2. 提案手法

本章では、提案手法に関して説明する。まず提案手法で使用するメタデータについて説明し、その後、提案手法の概要を説明する。

2.1 メタデータ

メタデータの内容は、2 種類の表を用いて表現できる。1 番目はデータベースと機能名に対応する Web サービスを示した表である。Web サービスをデータベースに対応づける条件は、（1）そのデータベースからデータを取得している、または、（2）そのデータベースのデータを利用できる、または、（3）そのデータベースの検索に利用できるデータを取得できることである。機能名は、データ解析やデータ取得に関するものとする。よって、単純なデータ変換やデータ抽出のみを行うような Web サービスはこちらには登録されない。例としては表 1 のようになる。括弧内の数字は上記の当てはめる条件である。

†大阪大学大学院情報科学研究科、Graduate School of Information Science and Technology, Osaka University
‡大阪大学臨床医工学融合研究教育センター、 The Center for Advanced Medical Engineering and Informatics, Osaka University

表 1 ：機能名とデータベースによる Web サービス対応表

DB / 機能名	Ensembl	Uniprot
Sequence Retrieval	getMMusSequence (1) getHSapSequence (1) getRNorSequence (1)	getMMusProteinSequence (1) getHSapProteinSequence (1) getRNorProteinSequence (1)
Multiple Alignment	emma (2)	runClustaW2 (2)
Identifier Retrieval	hsapiens_gene_ensembl (1) hsapiens_gene_external (1)	hsapiens_gene_external (3)

この表から、Multiple Alignment を行う時は、遺伝子データベースである Ensembl[10]を利用するならば emma、タンパク質データベースである UniProt[11]を利用するならば runClustaw2 という Web サービスを利用すること等が分かる。hsapiens_gene_external は、利用データベースは Ensembl であるが UniProt でも利用できるデータを出力する。このように、同じ Web サービスが複数のデータベースに属する場合もある。

2 番目は、Web サービス同士の関係を示した表で、Web サービスを連携させるために必要なスクリプトや Web サービスを書き込む。例としては表 2 のようになる。

表 2 ：Web サービス同士の接続関係表

始点 / 終点	0	1	2	3	4	5	6	7	8	9
(0) getMMusSequence							△			
(1) getHSapSequence							△			
(2) getRNorSequence							△			
(3) getMMusProteinSequence								△		
(4) getHSapProteinSequence								△		
(5) getRNorProteinSequence								△		
(6) emma										
(7) runClustaW2										
(8) hsapiens_gene_ensembl	△	△	△							
(9) hsapiens_gene_external				△	△	△				

表 2 では、始点の出力データを終点の入力に受け渡すことができるかを示す。表中に存在しないが、○は直接受け渡すことができることを示す。△は間に他のスクリプトや Web サービスを挿入することで受け渡すことができることを示す。空白は受け渡し方がないことを示す。○の場合は、どの出力とどの入力が繋がるかの情報を付加する。△の場合は、それに加えて挿入されるスクリプトや Web サービスに関する情報も付加する。データ変換やデータ抽出のみを行うような Web サービスはここで登録される。

メタデータの情報は原則 BioCatalogue と myExperiment を参考にした。BioCatalogue では、Web サービスの機能的分類や利用データベースを知ることができる。表 1 の機能名に関しては分類を規定している Service Categories の表記から取っている。この分類では、同機能の場合でも対象が異なれば別の分類となる。例えば、Protein Multiple Alignment と Nucleotide Multiple Alignment は、それぞれタンパク質配列、塩基配列に対する Multiple Alignment の機能を持つことを示す。しかし、本研究では異なる種類のデータベースでも対応を取るため、対象が異なっても同じ分類になるようにする。よって、表 1 の機能名では Multiple Alignment に統一する。BioCatalogue だけでは分からない機能がある場合は、各 Web サービスに関するドキュメントも参考にする。myExperiment では、既存のワークフローから特定の Web サービスの連携方法を知ることができる。既存のワークフローで使われた連携方法は、実際に動作する可能性が高いと考えられる。既存のワークフローで使われている物とは異なる Web サービスの場合でも、入出力や機能が類似している場合は既存の連携方法から一部 Web サービス、スクリプトを入れ替えて連携可能とした。

なおメタデータの表現では、表の各項目が複数の値を取ったり、空値を取ったりすることが多いことを考慮して XML で記述している。

2.2 提案手法の概要

提案手法の入力と出力は以下のようになる。

- 入力: ワークフロー、ワークフロー内で利用されているデータベースから変更するデータベース名
- 出力: 入力のワークフローと同様の解析処理でデータベースを変更したワークフロー

まず、入力としたワークフロー内の Web サービスが、データベースと機能名に対応する Web サービスを示した表（表 1）のどの部分に属するかを見る。そして、その Web サービスが属する機能名と変更するデータベースに対応する Web サービスを表から見つける。これがその Web サービスを置き換える Web サービスとなる。これをワークフロー内の全ての Web サービスに対して行う。

次に、置き換える Web サービスの入出力をつなぎ合わせる。Web サービス同士の関係を示した表（表 2）からつなぎ方を判断し、ワークフローを再構築する。最終的に完成したワークフローを出力として返す。表 1 で置き換える Web サービスの候補が複数ある場合は、表 2 で受け渡し方の情報がある Web サービスを優先する。受け渡し方の情報がある Web サービスが複数ある場合は、挿入するスクリプトや Web サービスの少ない Web サービスを優先する。

3. 結果

メタデータを実際に作成し、それを用いてワークフローを変更する。

図 1 のワークフローのデータベースを変更した結果を図 2 に示す。図中では、四角のオブジェクトは Web サービスやスクリプトを示す。そして、矢印は入出力データの受け渡しを示す。また、終点が丸い線は制御関係を示す。始点の Web サービスやスクリプトの処理が終わってから終点の Web サービスやスクリプトの処理を始める。さら

に、一番下の点線で囲まれた“Workflow output ports”の部分はワークフローの出力を示す。これらは図 2 でも共通である。参照するメタデータは表 1 と表 2 である。

図 1 のワークフローでは、Ensembl という遺伝子データベースからヒトとマウスとラットで同じ機能を持つオーソロガスな遺伝子配列を取得し（hsapiens_gene_ensembl, getMMusSequence, getHSapSequence, getRNorSequence）、取得した遺伝子配列をマルチプルアライメントで比較する解析処理（emma）を行う。このワークフローで使用されるデータベースを UniProt というタンパク質データベースに変更した結果が図 2 である。

図 1 では、hsapiens_gene_ensembl, getMMusSequence, getHSapSequence, getRNorSequence, emma が解析やデータ取得に関する Web サービスである。これらを表 1 で照らし合わせた結果、図 2 ではそれぞれ hsapiens_gene_external, getMMusProteinSequence, getHSapProteinSequence, とが分かる。そして、表 2 でこれらの Web サービスの接続関係を参照し、接続できる場合はあらかじめ定義されたスクリプトや Web サービスによってデータを受け渡せるようにする。結果として、UniProt からヒトとマウスとラットで同じ機能を持つオーソロガスなタンパク質配列を取得し（hsapiens_gene_external, getMMusProteinSequence, getHSapProteinSequence, getRNorProteinSequence）、取得したタンパク質配列をマルチプルアライメントで比較する解析処理（runClustalW2）を行うことが可能となった。

4. おわりに

データベースと Web サービスの対応関係を定義したメタデータにより、データベースを変更する場合のワークフロー作成支援を行えることを示した。今後は、実際のシステムとして実装することを目指す。メタデータの格納や検索は Resource Namespace Service(RNS)[12]を使って実装する予定である。

参考文献

[1] Cochrane, G.R. and Galperin, M.Y.: The 2010 Nucleic Acids Research Database Issue and online Database Collection: a Community of Data Resources. *Nucleic Acids Research*, Vol.38, No.suppl_1, pp.D1–D4 (2010).

[2] National Center for Biotechnology Information: NCBI Entrez Utilities Web Service (online), available from 〈 http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/DOC/esoap_help.html〉 (accessed 2010-07-27).

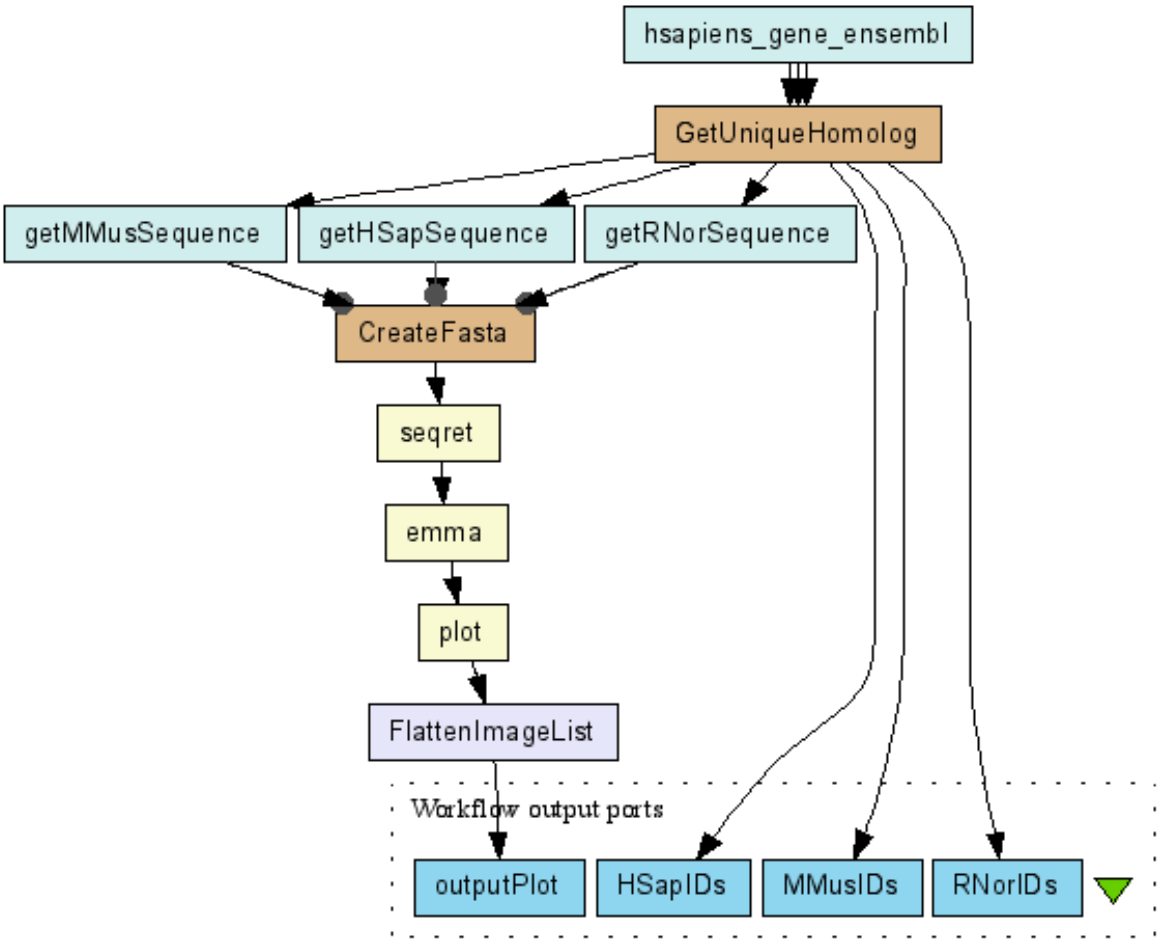

図 1 ：変更前のワークフロー

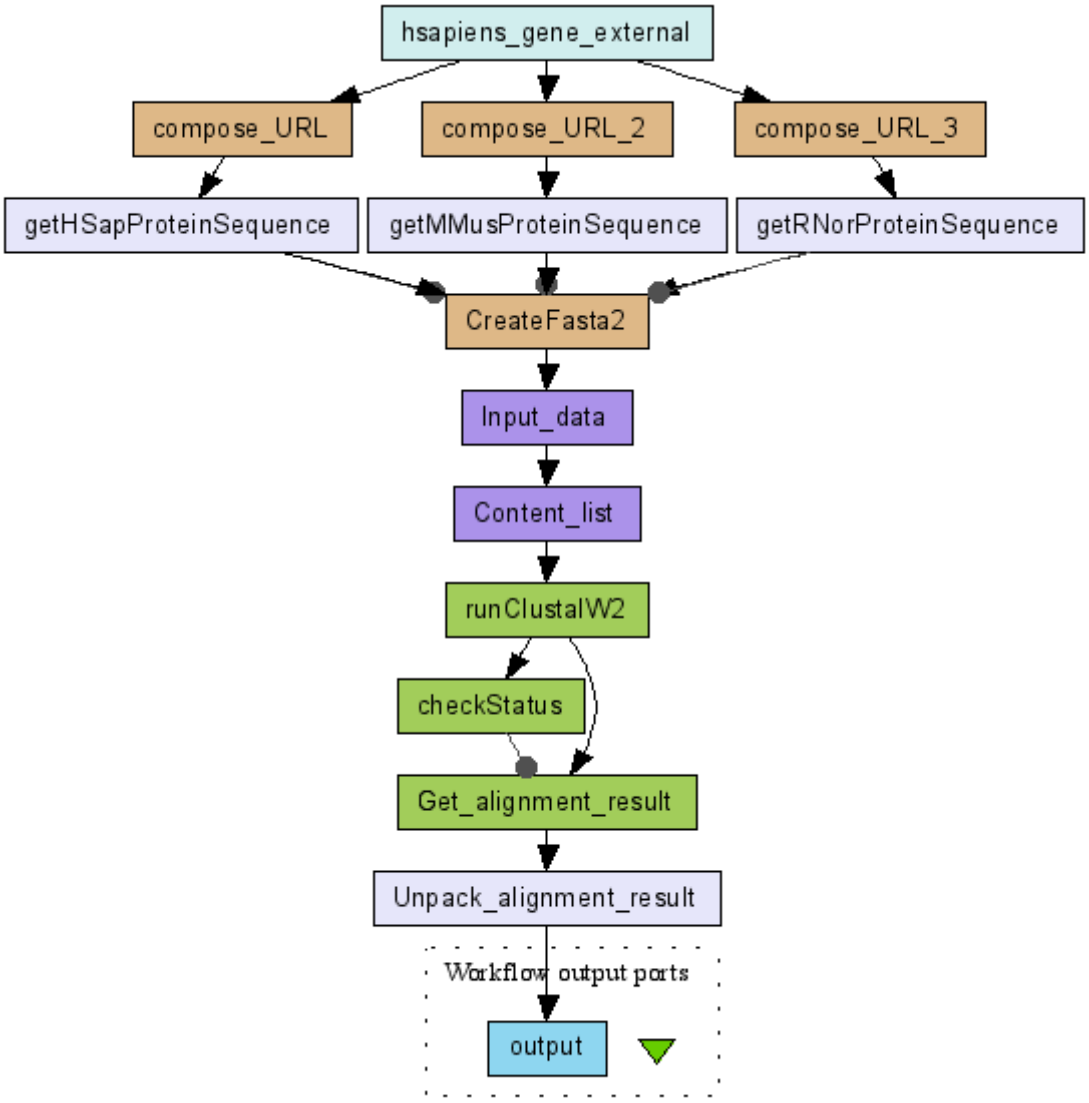

図２：変更後のワークフロー

[3] European Bioinformatics Institute: EBI Web Services at the EBI (online), available from〈http://www.ebi.ac.uk/Tools /webservices/〉 (accessed 2010-07-27).
[4] Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics: Web API for Biology, DNA Data Bank of Japan (online), available from 〈http://www.xml.nig.ac.jp/index.html〉 (accessed 2010-07-27).
[5] Bioinformatics Center, Institute for Chemical Research, Kyoto University: KEGG API, Kyoto Encyclopedia of Genes and Genomes (online), available from 〈http://www.genome.jp/kegg/soap/〉 (accessed 2010-07-27).
[6] Bhagat, J., Tanoh, F., Nzuobontane, E., et al.: BioCatalogue: A Universal Catalogue of Web Services for the Life Sciences. *Nucleic Acids Res*, Vol.38, pp.W689-W694 (2010).
[7] Oinn, T., Addis, M., Ferris, J., et al.: Taverna: a Tool for the Composition and Enactment of Bioinformatics Workflows. *Bioinformatics*, Vol.20, No.17, pp. 3045–3054 (2004).
[8] Goble, C.A., Bhagat, J., Aleksejevs, S., et al.: myExperiment: A Repository and Social Network for the Sharing of Bioinformatics Workflows. *Nucleic Acids Research*, Vol.38, No. suppl_2, pp.W677-W682 (2010).
[9] Seo, J., Seno, S., Takenaka, Y. and Matsuda, H.: Retrieving Functionally Similar Bioinformatics Workflows using TF-IDF Filtering, *IPSJ Transactions on Bioinformatics*, Vol.48, No.2, pp.20-29 (2007).
[10] Hubbard, T., Barker, D., Birney, E., et al.: The Ensembl Genome Database Project, *Nucleic Acids Research*, Vol.30, No.1, pp.38-41 (2002).
[11] Bairoch, A., Apweiler, R., Wu, C.H., et al.: The Universal Protein Resource (UniProt), *Nucleic Acids Research*, Vol.33, No.suppl_1, pp.D154-D159 (2005).
[12] Morgan, M., Grimshaw, A. and Tatebe, O.: RNS Specification 1.1, Open Grid Forum Proposed Recommendation (in public comment, online), available from 〈 http://www.ogf.org/Public_Comment_Docs/Documents/2010-03/RNS-spec-v11.pdf〉 (accessed 2010-07-26).