1D1-6

文献からのバイオサイエンス研究手法の収集・整理による研究支援セマンティックWebサービスの実現

Extraction and Structuring Research Protocols from Biological Articles to Implement Semantic Web Services for Biologists

荒木 次郎[*1] Jiro Araki　川本 祥子[*2] Shoko Kawamoto　藤山 秋佐夫[*4] Asao Fujiyama　菅原 秀明[*3] Hideaki Sugawara　大久保 公策[*3] Kousaku Okubo　武田 英明[*4] Hideaki Takeda

[*1](株) 三菱総合研究所 Mitsubishi Research Inst. Inc.　[*2]新領域融合研究センター Transdisciplinary Research Integration Center　[*3]国立遺伝学研究所 National Inst. of Genetics

[*4]国立情報学研究所 National Inst. of Informatics

Recently many bioinformatics tools and databases are developed and available as web services. Furthermore workflow building tools have been developed so that one can compose web services as workflow arbitrarily. But these tools assume that user can convert his research purpose into web services without trouble. This is the reason that biologists cannot use web services in bioinformatics. Therefore, we aim to develop a semantic web service system which can provide workflows for biological purposes given by biologists. For this goal we collect research methods from biological literatures and structure them as workflows. And then we propose a data model to relate workflow with web services to execute it.

1. 背景と目的

近年バイオサイエンス分野では、ゲノム解読やプロテオーム解析に代表されるような多くの網羅的解析プロジェクトが進められ、莫大なデータが生産され続けている。しかし、それらのデータはあくまでも生命の断片情報であり、これを情報科学手法を使って解析し、生命の再構築、知識発見を行うのがバイオインフォマティクスの役目である。

バイオインフォマティクスではその当初から、開発された解析ツールやデータベースがWebサービスとしてフリーで公開されるものが多く、それらを組合せて複雑な解析 (ワークフロー) を Web 上で行うことが事実上可能となっている。例えば、近年 myGrid プロジェクト [1] で開発されたワークフロー構築ツール Taverna[2] などを使えば、ワークフローの作成、実行、解析結果の参照までを 1 つの環境の中で行うことが可能である。

しかし、このようなワークフロー構築環境では、ユーザが 1) 個々の解析ツールのデータ処理機能を理解し、2) 解析するデータやデータベースのデータ形式、生物学的意味を理解した上で、さらに 3) どのように解析ツールやデータベースを組合せれば生物学的に意味のある解析を行うことができるかを理解している、という前提条件を要求している。この前提条件は一見当り前のようであるが、現実は難しく、データ形式・解析アルゴリズムと、生物学的な目的・データの意味、の両方を理解できる人材は非常に少ない。例えば、情報科学研究者やシステムエンジニアは個々のデータ形式やツールには詳しいが、実験で得られたデータの生物学的意味や解析目的が理解できないことが多い。逆に、生命科学研究者は解析の目的はあるものの、どのような解析ツールやデータベースを使い、それをどのように組合せれば目的が達成できるかが分らない。

このような状況を解決する一つの方法は、情報科学研究者と生命科学研究者がチームを組み、生命科学研究者が実現したい生物学的な解析目的を提示し、情報科学研究者がそれを具体的な解析ツールやデータベースの組合せに解釈し直すための情報共有環境を構築することである。しかしこのような環境が構築されているのは一部の大きな研究室だけで、多くの生物系研究室ではいまだに研究室内などの身近な人から聞いた情報をもとに解析ツールやデータベースを選択し、限られたデータ解析しか行えていないのが現状である。

そこで本研究では、生物学的な解析目的を実現するためのワークフローを、生命科学研究者自身によって検索・実行できるセマンティック Web サービス環境の構築を目指す。本稿では、生命科学研究者と情報科学研究者との連携によってこれまでに構築されてきた解析手法・研究手法を文献から網羅的に収集し、それを個々の Web サービスと対応付けるための整理方法を考案する。

連絡先: 荒木 次郎, (株) 三菱総合研究所 先端科学研究センター, 〒 100-8141 東京都千代田区大手町 2-3-6, jiro@mri.co.jp

2. 研究支援セマンティック Web サービス

我々が目指すバイオサイエンス分野の研究支援セマンティック Web サービスの対象ユーザは、「生命科学研究者」である。生命科学研究者は生命の解明を目標に、仮説を立ててウェット実験を実施し、得られた生命の断片情報であるゲノム配列や遺伝子発現情報などのデータに対し、バイオインフォマティクス解析によって生命の再構築、知識発見を行う。生命科学研究者は、生命科学ドメインの概念や言葉で表現される明確な解析目的をもっているものの、それを実現するためのバイオインフォマティクス解析ツールやデータベースの組合せを知らない。そのため、本研究では、研究目的を問合せキーとし、それを Web サービスとして実行するための最適なワークフローを答えとして返すサービスを実現する。また、生命科学研究者が実際に知りたいと考えられる表 1 のような検索目的にも対応できるように配慮する。例えば、実験で得られたゲノム配列中から遺伝子の場所を予測する目的に対し、それを実行するためのワークフローを提案してくれたり、過去の類似研究事例と結果を比較したい場合に、そこで用いられていたワークフローや文献自体を提示してくれたりなどの機能を実現する。

表 1: 本サービスで想定するユーザ検索目的

検索目的	問合せ	回答
解析目的を達成するための解析手法を知りたい	解析目的 (「遺伝子予測」)	ワークフロー
既往研究で用いられている解析手法を参考にしたい	研究分野 (「ゲノム解析」)	ワークフロー
同一処理機能をもつ最適な解析ツールを知りたい	処理機能 (「配列比較」)	解析ツール
どのような処理機能をもつ解析ツールであるかを知りたい	解析ツール (「BLAT」)	処理機能
特定のデータ形式を解析できるツールを知りたい	データ形式 (「ゲノム配列」)	解析ツール

このようなユーザ検索目的を満すことのできるセマンティック Web サービスを実現するために、次の 3 つから構成される「解析手法知識」を収集・整理し (図 1 参照)、検索に利用する。

- Web サービス
  情報科学研究者によって整備されてきた、世の中のバイオインフォマティクス解析 Web サービスを網羅的に収集し、その処理機能、入出力データ別に整理する。
- 解析目的/研究分野
  生命科学ドメインの概念や言葉で表現される、生命科学研究者の解析目的や研究分野を文献から収集し、タスクとして階層的に整理する。
- ワークフロー
  生命科学研究者と情報科学研究者との連携によってこれまでに構築されてきたワークフローを文献から収集し蓄積する。
  文献中にはワークフローの Web サービス構成とともに、その解析目的が記述されていることが多いため、文献からのワークフロー抽出によって生命科学研究者と情報科学研究者の知識を融合できることになる。

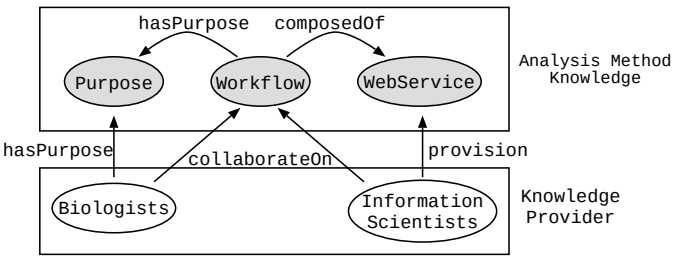


図 1: 解析手法知識の構成概要

3. Web サービスの収集

さまざまなリンク集 [3] [4][5] を参考にすると、現在世の中で公開されているバイオインフォマティクス解析 Web サービスの数は、解析ツールが 500〜1000、データベースが 1000〜1500 程度であると我々は推測している。これらのリンク集や文献などを参照しながら、可能な限り多くの解析ツールやデータベースの情報を収集する作業を現在進めている。この作業の中で提供サイトのアクセス情報とともに、解析ツールの場合はその処理機能を、データベースの場合は収録されているデータの種類を整理している。

解析ツールの処理機能の中には、遺伝子予測ツールのように生物学的目的 (「遺伝子を予測する」など) や入力データの生物学的意味 (「ゲノム」配列など) が特定されたツールもあるが、多くのツールは、配列比較ツールのようにデータ形式 (生物学的意味は含まれない) だけが特定された汎用ツールである。表 2 に、解析ツールの主な処理機能分類をまとめる。これから分かるように、バイオインフォマティクスの解析ツールの処理機能は、他のドメインにも共通する基本的なデータ処理機能である。ドメイン依存する部分としては、解析対象のデータ形式が単なる文字列や数値ではなく、具体例で示されているような、配列や立体構造データである点である。但し、ここには生物学的意味は含まれておらず、この点が生命科学研究者が解析目的から Web サービスにつなげにくい原因になっていると考えられる。

表 2: 解析ツール Web サービスの処理機能分類

処理機能分類				具体例
Data1	→ 変換	→ Data2		核酸配列からアミノ酸配列への翻訳
Data1, Data2	→ 比較	→ Similar Parts		配列比較、立体構造比較、発現比較
Data1, DB1	→ 検索	→ Data2 in DB1		配列 DB 検索
Data1... DataN	→ 分類	→ Group1... GroupM		複数配列の分類
Data1... DataN	→ ルール発見	→ Rule1... RuleM		複数配列からのパターン発見

一方、データベースの分類は、解析ツールの解析対象と同様にデータ形式 (配列、構造など) で分類されるとともに、そのデータが生物内のどの場所/物質/現象から計測されたかという生物学的意味で分類される (図 2 参照)。

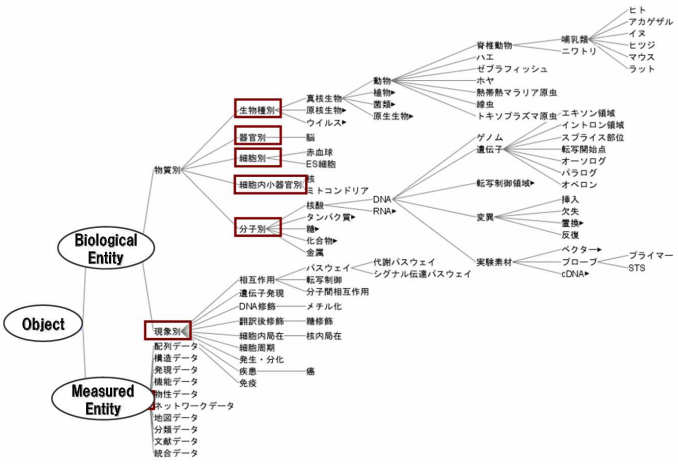


図 2: データ/データベースの分類 (抜粋)

では、解析ツールを利用する際に、どの時点で生物学的目的/意味が生れてくるのかを考える。解析ツールには、解析対象となる入力データと、検索ツールのように検索対象とするデータベースが指定される (図 3 参照)。入力データやデータベースが単に「配列データ」のようなデータ形式だけしか特定されていない場合には、生物学的意味は含まれず、配列比較という処理機能でしかない。しかし、入力データやデータベースに対し図 2 に示したようなデータの生物学的意味が付与されることによって、解析ツールに「ゲノム上から遺伝子を予測する」のような生物学的意味が生れる。このように、生物学的

意味が付与された入力データ、データベースを構成要素として含むワークフローの形になって初めて、解析ツールに生物学的意味が生れてくるのであり、ワークフロー情報を収集することは、セマンティック Web サービスを完結させる上で非常に重要である。

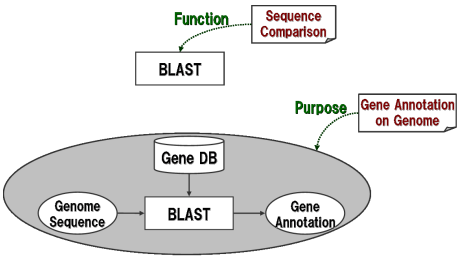


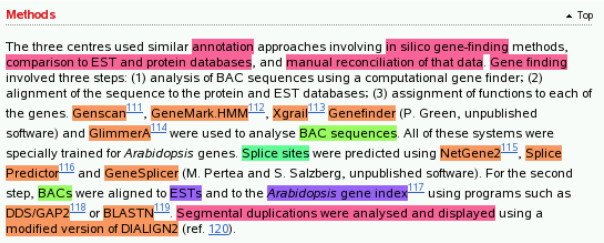
図 3: 解析ツールの処理機能と生物学的目的

4. 文献からの解析目的・ワークフローの収集

生命科学研究者と情報科学研究者との連携によってこれまでに構築されてきたワークフローを収集・整理することは、生命科学研究者が解析目的を入口として Web サービスを活用する上でのキーポイントとなる。

では有用なワークフローを収集する情報源として何が適切かを考えると、一つは従来から行われてきたように研究グループ内で収集、共有することである。しかしこれでは非常に限れた範囲の知識しか収集できない。次に Web2.0 のように Web 上で不特定多数から収集できるかというと、最初からそれを望むことは難しい。ワークフロー収集と再利用の有用性を皆が認識して初めてコミュニティの形成・活生化が可能と考えられる。このようなことから、現時点で多くの有用なワークフローを収集する術は、過去の文献から収集することと考える。生命科学の文献では、まず研究の背景や目的、研究手法の概要が述べられた後、結果に多くのページが割かれる。そして、用いられた実験手法やバイオインフォマティクス解析手法については別章を設けるか、分量が多い場合は付録として付けられることが多い。このように生命科学の文献から研究手法の記述を発見し読むべき範囲は限定されており明確である。図 4 は、2000 年に発表されたシロイヌナズナのゲノム解読文献 [6] 中での解析手法の記述である。Genscan などの解析ツール名 (オレンジ色)、Arabidopsis gene index などのデータベース名 (青色) が記述されている。また、解析対象となる入力データ (薄緑色) として BAC sequences、解析結果 (緑色) として Splice sites が記述されている。さらに、解析の目的 (紫色) として、Gene finding や annotation、それらをタスク分解した in silico gene-finding などが記述されている。この例から分かるように、入力・出力データは必ずしも明記されている訳ではなく、文脈から推測する必要がある。また解析ツールについては、この例以外の場合では具体的なツール名が書かれず、配列比較ツールのように処理機能名として記述されることもある。

このように、文献中に記述された解析ツール名、データベース名、解析対象 (入力データ)、解析結果 (出力データ)、解析目的、の 5 つの項目をマーキングし、文脈などをもとに記述されていない情報を補間した上で、ワークフローとしてデータ化する。ワークフローの記述方法には、OWL-S や WSMO、myGrid プロジェクトで使われている Scufl などがあるが、文献から抽出されるワークフローは必ずしも具体的な Web サービスまで記述されているとは限らず、またアノテータの作業の簡易性などを考慮すると、ここでは簡易な XML 形式で記録することとした。XML 形式のワークフローはさらフロー図として図示化できるようにした。図 5 は、図 4 で示したシロイヌナズナゲノム解読文献から抽出したワークフローである。解析ツールを長方形で示し、入出力データ及びデータベースを楕円形で示している。また大きな解析目的単位でブロックに区切っている。この例では、シロイヌナズナの解読されたゲノム (BAC) 配列に対して、遺伝子モデルをもとにした Ab initio な手法と、既知の配列データベースとの配列比較手法を複数種利用して遺伝子を予測し、その予測結果を組み合せることによって精度の良い遺伝子候補を求めている。


Methods ▲ Top

The three centres used similar annotation approaches involving in silico gene-finding methods, comparison to EST and protein databases, and manual reconciliation of that data. Gene finding involved three steps: (1) analysis of BAC sequences using a computational gene finder; (2) alignment of the sequence to the protein and EST databases; (3) assignment of functions to each of the genes. Genscan[111], GeneMark.HMM[112], Xgrail[113] Genefinder (P. Green, unpublished software) and GlimmerA[114] were used to analyse BAC sequences. All of these systems were specially trained for *Arabidopsis* genes. Splice sites were predicted using NetGene2[115], Splice Predictor[116] and GeneSplicer (M. Pertea and S. Salzberg, unpublished software). For the second step, BACs were aligned to ESTs and to the *Arabidopsis* gene index[117] using programs such as DDS/GAP2[118] or BLASTN[119]. Segmental duplications were analysed and displayed using a modified version of DIALIGN2 (ref. 120).

図 4: 文献中のワークフロー情報

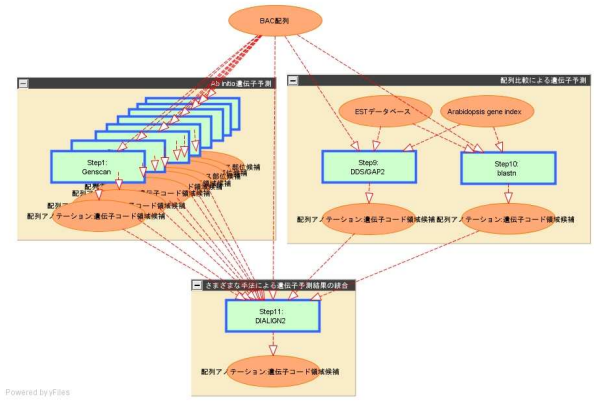


図 5: ワークフローの図示化

このような文献からのワークフロー抽出作業を、まずは「ゲノム解読」に関連する文献に限定し作業を進めた。またその際に、解析手法がより詳細に記述されている文献を優先的に選択するために、前章の Web サービスの収集で得られた解析ツール名リストを利用し、「ツール名の記述数が多い文献ほど、詳細に解析手法が記述されている」と仮定し、文献に対しツール名検索を行った。その結果、Nature の文献に多くのツール名が記述されていたことから、Nature を優先した。またこの検索結果から、世の中で評価の高い解析ツールを客観的に推定することができ、同じ処理機能をもつツールの中から推薦を行う際にこれを利用できる。

図 6 にさまざまなゲノム解読文献 [6] [7] [8] [9] [10] から抽出したワークフローの例を挙げる。

5. 解析手法知識の整理

生命科学ドメインの概念や言葉で表現される解析目的から具体的な Web サービスの実行までをつなげるために、収集した Web サービスとワークフロー、解析目的情報を統合するためのデータ構造を検討する。

文献から抽出されたワークフローはそれだけも利用価値は高いが、既存のワークフローの分解、結合、同類処理機能の

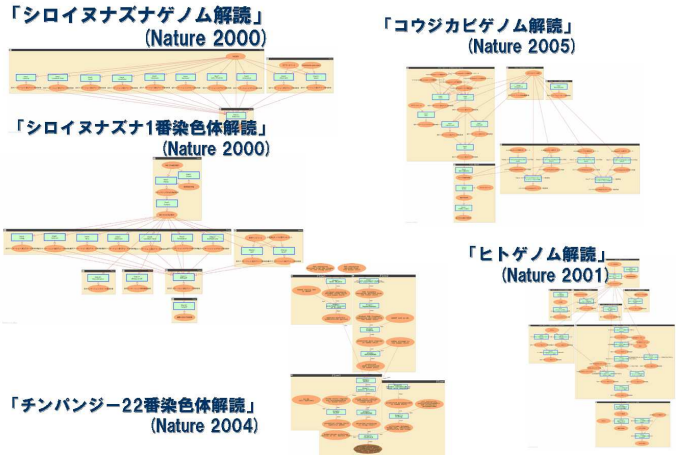


図 6: 文献から抽出したゲノム解読ワークフロー例

Web サービスへの代替などができれば、ワークフローの総和以上の効果が現れる。また、ワークフローの利用だけでなく、個々の研究者が把握できていないバイオサイエンス分野の研究手法の全体像やその推移も把握できる可能性がある。

図 1 で示した解析手法知識の構成概要に対し、これまでの検討結果を加えてより詳細化したものが図 7 である。ワークフロー (Workflow) には生物学的な解析目的 (Purpose) があり、解析対象データ (Data) と、解析ツール (Tool) 及びデータベース (Database) の Web サービス (WebService) から構成されている。解析ツール (Tool) は処理機能 (Function) を持ち、処理対象のデータ形式 (Measured Entity) と処理行動 (Function Action) によって表現される。一方、ワークフロー (Workflow) の解析目的 (Purpose) の対象は、生物学的な対象物 (Biological Entity) で表現される。解析対象データ (Data) は、少なくともデータ形式 (Measured Entity) が付与されており、さらにそのデータの出所 (derivedFrom) である生物学的対象 (Biological Entity) が付与されることによって初めて、それを入力とする解析ツール (Tool) に生物学的な目的 (Purpose) が生れる。

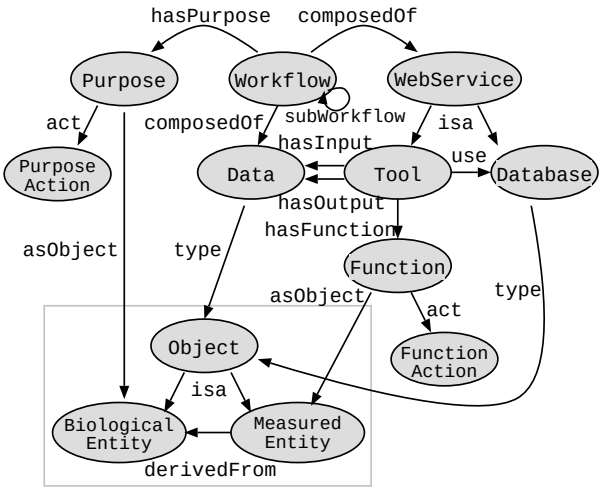


図 7: 解析手法知識の構成詳細

収集したワークフローと Web サービス情報を本データ構造に従って仮想的に整理し、データ構造の妥当性を検証した。まず、本データ構造を規定することによって、ワークフローと Web サービス情報を別々に収集してもその連携を取ることができるを確認した。生命科学ドメインの概念や言葉で表現される解析目的から具体的な Web サービスまで到り着くことができ、さらに意味のある解析目的単位でワークフローの分解、結合、Web サービスの代替も可能である。また、さまざまなワークフローを収集し、その共通部分を結合することによって、「フルスペック」のワークフローを合成することが原理的に可能である。例えば、図 6 に示したワークフロー群を結合することで、「配列アセンブリング」「ゲノムアノテーション」「機能予測」「ゲノム比較」を行う、「ゲノム解読」のフルスペックワークフローが得られる。但し、この合成過程を自動化し意味のある組合せを実現するためには、解析精度などのより高次の解析手法知識の構造化が必要である。

6. まとめ

生命科学ドメインの概念や言葉で表現される解析目的と具体的な Web サービスを結び付けるために、文献からのワークフロー収集を行い、その整理手法を提案した。収集したワークフローに対し本整理手法を適用することで、解析目的をベースとした従来にはないセマンティック Web サービスの実現の形を示せた。今後は、この整理手法を実装したシステムを開発するとともに、ワークフロー収集範囲を拡げ、バイオサイエンス分野全体の研究手法を網羅することを目指していきたい。

参考文献

[1] myGrid: http://www.mygrid.org.uk/

[2] Taverna: http://taverna.sourceforge.net/

[3] NAR Web Server Issue: http://www.oxfordjournals.org/nar/webserver/cap/

[4] JST ゲノム解析ツールリンク集: http://www-btls.jst.go.jp/Links/

[5] Bio Database Showcase: http://www.ps.noda.tus.ac.jp/biometadb/

[6] Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature, Vol.408, No.6814, pp.796-815, 2000.

[7] Theologis A et.al, Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana, Nature, Vol.408, No.6814, pp.816-20, 2000.

[8] Watanabe H et.al, DNA sequence and comparative analysis of chimpanzee chromosome 22, Nature, Vol.429, No.6990, pp.382-8, 2004.

[9] Machida M et.al, Genome sequencing and analysis of Aspergillus oryzae, Nature, Vol.438, No.7071, pp.1157-61, 2005.

[10] Lander ES et.al, Initial sequencing and analysis of the human genome, Nature, Vol.409, No.6822, pp.860-921, 2001.