3F2-4

バイオサイエンス論文からの解析手法知識自動抽出アルゴリズムの開発

An Automatic Extraction Method of Workflow Knowledge from Biological Papers

荒木 次郎*1 Jiro Araki　藤山 秋佐夫*2 Asao Fujiyama　菅原 秀明*3 Hideaki Fugawara　武田 英明*2 Hideaki Takeda

*1(株) 三菱総合研究所 Mitsubishi Research Inst. Inc.　*2国立情報学研究所 National Inst. of Informatics　*3国立遺伝学研究所 National Inst. of Genetics

We research how to collect and reuse workflows (= analysis protocols) in order to develop a semantic web service system for bioinformatics. We could collect only thirty workflows from biological papers by hand last year. Therefore, we develop an automatic extraction method of workflows. Because in the workflow extraction from papers it is most difficult problem to extract relation between workflow components, we use both dependency analysis in NLP and case-frames for expressing workflows.

1. はじめに

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るため、情報科学手法を使って断片知識の統合、大規模データからの知識発見を行って初めて、生命の理解につながる。

例えば、2001 年に国際ヒトゲノムコンソーシアムによって雑誌 Nature 上にヒトゲノム解読の成果が発表された [Lander 01]。このプロジェクトでは、30 億文字にも及ぶヒトゲノム配列データが各国の研究機関の分担によって解読された。解読されたゲノム配列データは断片化されているために、まずはそれを繋ぎ合せるための情報解析が行われ、次にゲノム配列全体の中から高々数パーセントにも満たない遺伝子に当る部分を見つけ出すための解析が行われた。最後にそれぞれの遺伝子がどのような機能をもつタンパク質の設計図であるかを予測するための解析が行われた。これらのバイオインフォマティクス解析も各国の研究機関によって分担して行われ、その解析手順と結果が 60 ページに及ぶ論文の中にまとめられている。我々は論文中の記述からこのような解析手順をワークフローの形で抜き出す作業を行い、昨年度の本学会で報告した [荒木 07a]。図 1 は、このヒトゲノム解読論文から手作業で抜き出したワークフローである。図内の長方形が解析ツール名を表し、その前後に繋がる楕円形が解析ツールの入力データ、出力データを表している。また、解析ツールに左右から繋がる楕円形はデータベース検索に用いられたデータベースを表す。図から分かるように、1 生物種のゲノム配列を解析するだけでもツールとデータベースの非常に複雑な組合せが必要であり、このワークフローは各国の研究者が持ちよった知識の集積と言える。2001 年のヒトゲノム解読以降、現在までに数百種のゲノム解読が完了し、論文が公開されている。それぞれの論文では、過去の論文の解析手法を参考にしつつも、その生物種や研究目的に特化して手法をアレンジしている。また、実験手法や解析ツール、計算機処理の高性能化に伴って、例えば複数の生物種の解析を同時に行うメタゲノム解析など、より高度な解析が行われるようになってきている。このように、より高度な解析が必要となってきている現在、解析を実行するためのツールやデータベースの整備とともに、解析手法知識の整理が重要視されている。

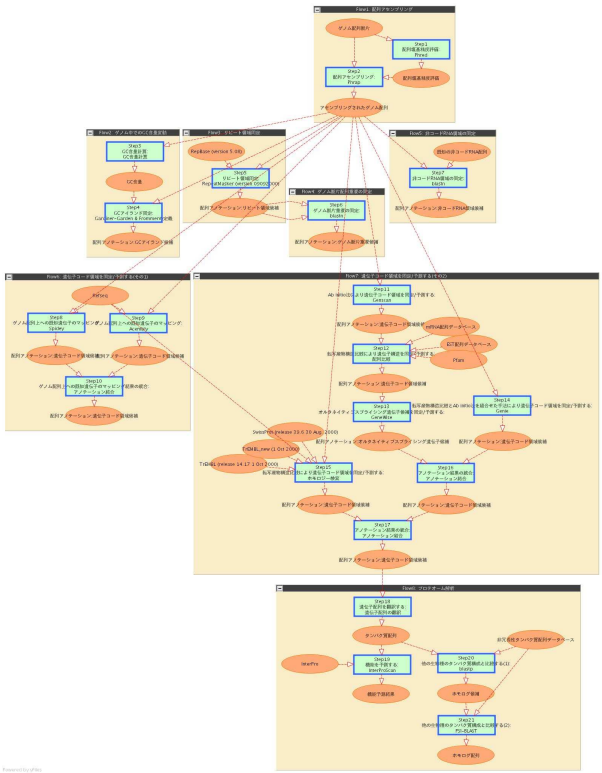

図 1: ヒトゲノム解読論文から手作業で抽出したワークフロー

バイオインフォマティクス分野では、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を Web 上で行うことが可能となっている。例えば、近年英国の e-Science プロジェクトの一つ myGrid プロジェクト [myGrid] で開発された Taverna ワークフロー構築ツール [Taverna] 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どのようなツール・データベースをどのように組合せればよいのかを知っていることを前提としてお

連絡先: 荒木 次郎, (株) 三菱総合研究所 科学･安全政策研究本部, 〒 100-8141 東京都千代田区大手町 2-3-6, jiro@mri.co.jp

り、不慣れな生物学者が使うにはハードルが高い。そのため、myGrid プロジェクトでは不特定多数のユーザがワークフローを投稿、共有するためのサイトを立ち上げ、解析手法知識の集積を行おうとしている [myExperiment]。myGrid のこの試みは一般向けサービスの Yahoo! pipes と同じ方向性をもっているが [Yahoo! Pipes]、プロジェクト関係者以外の投稿になかなか広がっていないのが実状である。

そこで我々は、ゲノム解読研究だけでも既に数百件が出版されている論文に注目し、論文からのワークフロー抽出を始めた。しかし、論文を 1 件ずつ読み手作業で抽出する場合、精度は確実なものの収集数に限界があることから、論文からワークフローを自動抽出する手法を開発した。本論文で提案する手法では、1) 構成要素抽出のためのの固有名辞書、2) 係り受け関係抽出のための構文解析、3) 係り受け関係をもとにワークフロー上での構成要素の役割を規定するための格フレーム辞書、4) 構成要素がワークフロー上の役割を満す概念であるかを規定するワークフローオントロジー、を組み合せる。

第 2 章では、提案するワークフローの自動抽出手法を説明する。第 3 章では、開発した自動抽出手法を使って実際に自動抽出した実験結果についてまとめる。

2. 論文からのワークフロー自動抽出手法

本章では、我々が提案するワークフロー自動抽出手法について説明する。まず最初にワークフローが何から構成されているかを定義し、論文テキストからワークフローを抽出する上で課題となる点を明らかにする。その次に提案する手法の全体の流れを説明する。

2.1 ワークフローの定義

我々が抽出対象とするワークフローはバイオインフォマティクス解析のためのワークフローである。ワークフローは図 2 に示すような 4 種類の要素から構成されており、要素間はデータの入出力関係によって結ばれている。即ち、「解析ツール」に対してその解析対象となる「入力データ」と、解析結果となる「出力データ」が存在する。またバイオインフォマティクスの一部の解析ツールには、ツールとは独立して存在する任意の「データベース」を利用して解析するものもある。あるツールの出力データは、さらに別のツールの入力データとなることで、複数のツールから成るワークフローが構成される。出力結果に応じて次の解析ツールが異ったり、ある条件を満すまで解析を繰り返すなど、実際の解析においてははさまざまな制御が入り得るが、本研究ではそれらの制御ノードは考慮しない。

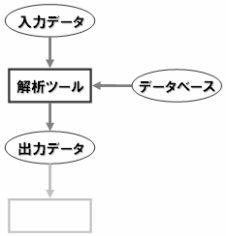

図 2: ワークフローの構成要素

2.2 ワークフロー抽出の課題と開発手法の要件

論文中に記述されているワークフロー事例を紹介し、論文テキスト中からワークフローを抽出する上での課題について議論する。図 3 は、シロイヌナズナという植物のゲノム解読論文 [The Arabidopsis Genome Initiative 00] 中の研究方法を記述したセクションの全文である。ワークフローを構成する解析ツール、データベース、入力データ、出力データの記述を色分けして表している。例えば、3 文目では、”Genscan”、”GeneMark.HMM”などが解析ツールを表し、”BAC Sequences”が”analyse”対象の入力データであることを示す。但し、この文では出力データに当る記述が明記されていない。このように論文中のワークフローの記述では一方の記述が省略されることが多い。この例の場合、これより前の文に記述された解析の目的”gene-finding”をもとに、出力は”gene”であると解釈する必要がある。また場合によっては解析目的すら記述されていない場合もあり、その場合は解析ツールの機能を前提として入出力データの一方を補完する必要がある。

以上のような記述の特長をもつワークフロー記述からワークフローの構成要素を自動抽出するためには、”Genscan”などの固有名を辞書化しておくことが一つの重要な手段である。次に、ツールと入出力データ間のような要素間関係を抽出するためには、係り受け関係や照応関係が手がかりとなることから構文解析、照応解析を活用することが必要である。ただし、構文解析の係り受け関係だけでは、例えば述語-目的語関係や修飾語-被修飾語関係のように、語の構文上の役割しか規定しておらず、ワークフロー上の意味的役割は規定できない。また、図 2 に示したような、解析ツールを中心としたワークフロー構成要素間の関係を抽出するためには、オーバーラップする係り受け関係の組み合せを見る必要がある。そのため、係り受け関係をもとにワークフロー上での構成要素の役割を規定するための辞書が必要である。さらに、構成要素がワークフロー上の役割を満す概念を本当に意味しているかを規定する概念辞書も抽出精度を上げる上で重要である。

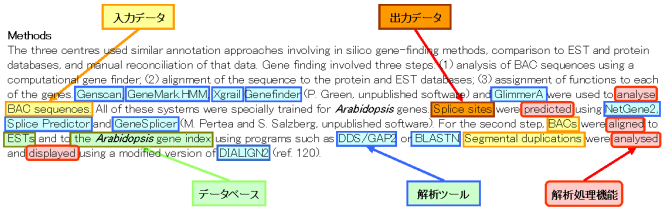

図 3: 論文中のワークフロー記述例

2.3 提案する手法の流れ

前節のワークフロー抽出手法の要件をもとに、我々は図 4 に示すような自動抽出手法を提案する。但し、最初から全ての要件を全て満すことは非常に困難であるので、本論文では次のような制限を設けた手法を提案し、その達成点と限界を明らかにした上で、今後の課題検討につなげることとする。

- 文間の記述を関係付ける照応解析は行わない
- 文に記述されていない記述 (入出力データの一方など) を前提知識として補完することは行わない

a. **構文解析による係り受け関係の抽出**
まず最初に構文解析ツールを使って論文テキストから係り受け関係を抽出する。本研究ではバイオサイエンスの論文を解析対象とするので、バイオメディカル分野のコーパス GENIA で学習された Enju HPSG パーサ [Enju 07] を利用した。構文解析の結果は、ワークフローの構成要素間の関係抽出に利用するとともに、次の固有名辞書による構成要素抽出にも利用する。

b. **固有名辞書によるワークフロー構成要素の抽出**
ワークフローを構成する解析ツール、データベース、入

力データ、出力データの記述を論文テキストから抽出するために、これらの固有名辞書を整備し利用する。抽出する固有名はワークフローにおける概念に相当するので、後述のワークフローオントロジーと対応付けて整備する。

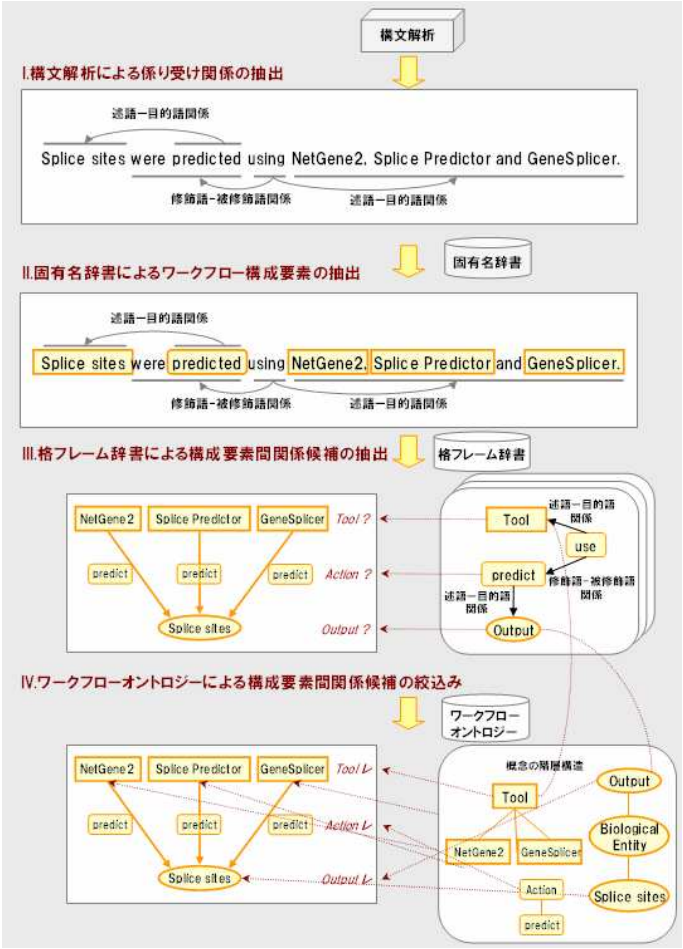

図 4: 提案するワークフロー自動抽出手法

現在、解析ツール名、データベース名いずれも約 400 種を登録している。一方、入出力データは標準的なデータ種を数十種類登録している。しかし、入出力データは解析ツールやデータベース名に比べデータ名の表記は多様であることから、辞書とのマッチングだけに頼らず、マッチしなかった記述についても後述の格フレーム辞書とのマッチングにおいて候補となり得る場合は残すものとする。

なお、固有名辞書には、例えば”protein database”、”EST database”のような複合語も登録されている。しかし論文中の記述では”protein and EST database”のように 2 つが併記される場合もあるので、辞書とのマッチングの際に隣接する語だけではなく、係り受け関係でつながる語 (この例の場合、”protein”が”database”に係る) も隣合う語として扱うことで、離れた語から構成される複合語の抽出も可能とする。

c. **格フレーム辞書による構成要素間関係候補の抽出**
格フレームは、言語学者の C. フィルモアによって提唱された文法理論である。動詞に対して、その動作主や対象、道具などの格の制約を記述する。現在、UC バークレーのグループが FrameNet と呼ばれる汎用の格フレームデータベースの整備を行っている [FrameNet]。また、同グループはフレームやその格のオントロジー化による、自然文記述の自動推論を研究している [Scheffczyk 06]。我々はこの考えを利用して、ワークフローに特化した格フレーム辞書を構築し、既に構築済みのワークフローオントロジー (図 6 参照) と連携させることで、構文解析の結果から意味的な関係抽出を行う。

格フレーム辞書は、図 5 に示すような XML 形式で記述する。記述は、フレームごとに 1) 述語情報、2) 格情報、3) 述語-項構造パターン、の 3 種類の情報から構成されている。述語、格それぞれには、ワークフローオントロジー上の意味クラスの制約 (Semantic Type) とワークフロー上の役割 (Workflow Role) が付与されている。述語-項構造パターンは、このようなフレームの意味構造と構文上の構造とを関係付けるためのものである。この例は、”Predicting”(予測する) の格フレームを記述したものである。述語”Predicting”の意味クラスはワークフローオントロジーの Action クラス下の”Predict”に相当する。”Predicting”の格には 2 種類あり、対象格”Object”の意味クラスはオントロジーの”Object”に相当し、ワークフロー上の役割は出力データ (Output Data) になる。道具格”Tool”のクラスは”Tool”であり、ワークフロー上の役割も解析ツール (Tool) となる。このような”Predicting”の意味構造が構文上どのように記述されているかを規定するのが述語-項構造パターンで、述語”Predicting”の目的項が対象格に相当し、”Predicting”の修飾語”Use”の目的項が道具格に当ることが記述されている。

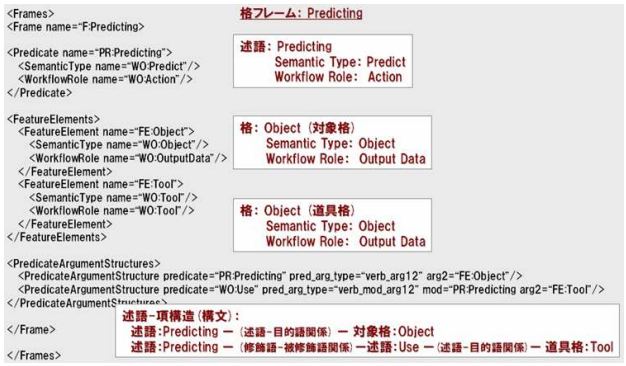

図 5: 格フレーム辞書の例

d. **ワークフローオントロジーによる構成要素間関係候補の絞り込み**
格フレーム辞書は構文的制約によるワークフロー構成要素間の関係候補の抽出を行ったが、各構成要素が本当にその要素となり得るかは概念クラス上の位置関係を調べる必要がある。例えば、格フレームの構文上の制約から「”NetGene2”を使って”Splice Sites”を予測する」という関係候補が抽出されたが、”NetGene2”が”Tool”概念クラスに属するか、”Splice Sites”は”Output”概念クラスに属するかは概念階層構造を定義したオントロジーを参照する必要がある。我々は、ワークフローの構成要素を意味関係を定義したワークフローオントロジーを構築し、昨年の全国大会で報告した。図 6 の上図は、ワークフローの構成要素間の意味関係を定義したものである。下図は構成要素の意味属性を定義付けるための生物学ドメインオントロジーの主要概念関係を表す。この図には示していな

いが、前述したように解析ツール (Tool)、データベース (Database)、入出力データ (Data) の下には数十〜数百種類のサブクラスの概念がぶら下っており、図 4 の概要図で示すように、”NetGene2”は”Tool”概念クラスに、”Splice Sites”は”Output”概念クラスに属することが定義付けられている。このように構成要素の意味定義を参照することで構成要素間関係候補の絞り込みを行うことができる。

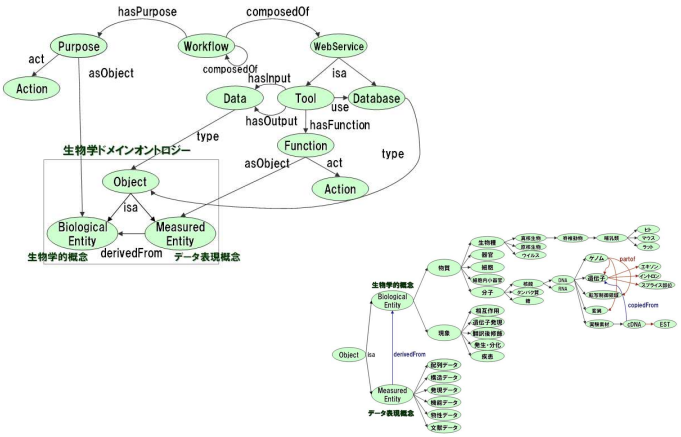

図 6: ワークフローオントロジーの概要

3. ワークフロー自動抽出の評価実験

提案した手法の精度評価を目的として、昨年手作業で論文から収集整理したワークフローデータ [荒木 07a] を利用して、一部のワークフローをもとに各フレーム辞書を構築し、残りのワークフローに対し比較検証を行った。図 7 にワークフロー自動抽出例を示す。図の上側が自動抽出結果を、下側が手作業による正解ワークフローを示す。正解ワークフローと比較すると、構成要素については殆どを自動抽出することができた。一方、構成要素間関係については、前述したように、本手法では照応解析や未記述の入出力データ補完は行わないために、ワークフローが文単位で断片化されている。

自動抽出した他の結果も検証した結果、次のようなことが言える。

- 各フレームの登録数を増やせば抽出精度はより向上する傾向にある
- 照応解析や未記述の入出力データ補完を行わない限りは構成要素間関係抽出に限界がある
- 構文解析による係り受け解析自体に間違いが含まれることがあり、それを利用する限りは精度に限界がある

4. まとめ

本論文では、構文解析・固有名辞書・格フレーム辞書・オントロジーを組み合せることで論文テキストからワークフローデータを自動抽出する手法を開発した。但し、本論文では、照応解析や未記述の入出力データ補完を行わないとの制限を設けたため、ワークフローが断片的にならざるを得なかった。今後の課題として取り組みたい。

なお、本手法は、本論文ではバイオサイエンスの論文からバイオインフォマティクス解析のワークフローを抽出することに適用したが、方法論自体はさまざまな分野に適用可能と考えられる。例えば、筆者が [荒木 07b] で提案した「学術分野動向把握のための論文タイトルの構造化」にも適用可能であり、これまで主に行われていた用語共起性だけをもとにしていた知識抽出の限界を超えることができると考えられる。

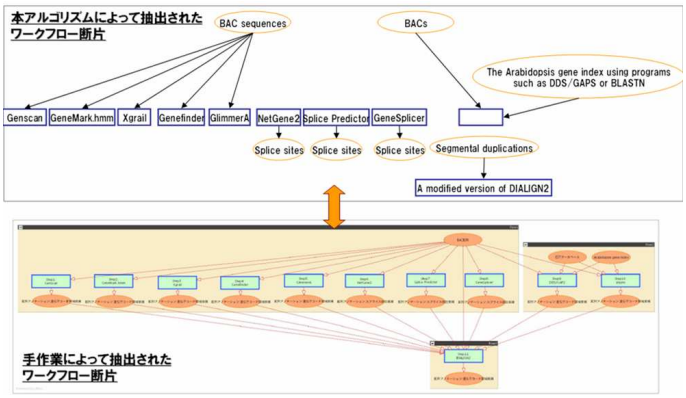

図 7: 論文からのワークフロー抽出例

謝辞

本研究はの一部は、科学技術振興機構バイオインフォマティクス推進事業 (Bird) 「バイオ基幹情報資源の高準化と共用化」の支援を受けている。ここに記して感謝する。

参考文献

[荒木 07a] 荒木、川本、藤山、菅原、大久保、武田: 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, 2007 年度人工知能学会全国大会, 2007.

[荒木 07b] 荒木: 学術分野動向把握のためのオントロジー構築, 第 16 回セマンティックウェブとオントロジー研究会, 2007.

[Enju 07] Tsujii Laboratory: Enju - A practical HPSG parser. http://www-tsujii.is.s.u-tokyo.ac.jp/enju/, 2007.

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

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

[myExperiment] http://www.myexperiment.org/

[Yahoo! Pipes] http://pipes.yahoo.com/pipes/

[FrameNet] http://framenet.icsi.berkeley.edu/

[Scheffczyk 06] J.Scheffczyk, C.F.Baker, and S.Narayanan: Ontology-based reasoning about lexical resources. In Proceedings of the OntoLex Workshop at the 5th International Conference on Lexical Resources and Evaluation (LREC), 2006.

[Lander 01] Lander ES et.al, Initial sequencing and analysis of the human genome, Nature, Vol.409, No.6822, pp.860-921, 2001.

[The Arabidopsis Genome Initiative 00] Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature, Vol.408, No.6814, pp.796-815, 2000.