2ZB-8

バイオインフォマティクス解析における Web サービス統合利用のためのメタサービスの提案

A proposal of Meta Service for Integration of Web Services in Bioinformatics Analysis

小野圭亮　瀬尾茂人　竹中要一　松田秀雄

大阪大学大学院情報科学研究科バイオ情報工学専攻

1. はじめに

バイオインフォマティクス分野では，遺伝子やタンパク質などの生物学情報の増加に伴い，各種のデータを解析するためのツールが提供されるようになった．これらのツールは単体で利用されるだけではなく，複数用いることで統合的解析（ワークフロー）を行うことも多い．

本研究では，Web サービスを利用したワークフローを想定し，類似サービス群の統合利用，サービス連携のための，オントロジを利用したメタサービスを提案する．

2. バイオインフォマティクス解析におけるワークフローとその問題点

ワークフローにおけるツール利用には，類似ツールを複数実行させるものと，ツールを連結させるものとが存在する．前者は，対象とするデータに対し，複数の解析ツールを実行し，その結果からコンセンサスをとるものがある．後者は，ある解析ツールの結果を別の解析ツールに渡すことで複雑な解析が可能になる．特に，バイオインフォマティクスの解析では，複数のデータベースやツールが必要となる．各研究機関のウェブサイトにアクセスすることでブラウザ上からのデータベースやツール利用は可能であるが，これらを目的に応じて，ワークフローとして自動化するにはコストがかかる．このためネットワーク上に存在するツールをプログラムから呼び出し利用することができる Web サービス技術が注目されており，これをワークフローに組み込むことで解析の効率化が期待されている．

現状の Web サービス利用における仕様定義 WSDL(Web Service Description Language)には，サービスの内容，入出力の意味，どのようにサービスを組み合わせるのかなどのワークフロー構成のための情報が不足している．そこで，セマンティック Web 技術によるオントロジを利用することで Web サービスを補完しようとする試みが行われている. 例えば，FETA[1]では，RDF 記述を用いて, サービスやデータのクラス分類をドメインオントロジとして記述し，これを利用してバイオインフォマティクス Web サービスに対して，メタデータを付与している．しかし，現状のオントロジは，単体のサービス発見には有用ではあるが，サービスの統合や連携には適したものではない．例として，図１の配列解析のワークフロー作成を考える．このワークフローは，類似配列の検索を行う相同性検索の Web サービスを複数実行し，出力結果の比較や，マルチプルアライメントサービスへの連携を行う．このワークフローで用いられている，Blast, Fasta などの相同性検索 Web サービスを，サービスに付与されたメタデータ，”Pairwise aligning”などを基に発見することはできる．しかし，それぞれの相同性検索サービスの入出力フォーマットは異なる．このため，入力設定をサービス毎に行う必要がある．また出力結果の比較において，関連データを参照するときや，出力結果を次のマルチプルアラインメントサービスの入力に繋げるためには，必要なデータを抽出し整合をとるための Shim[2]と呼ばれるプログラムを個別に作成しなければならない．このようなフォーマットの違い，それに伴うデータ整合の煩雑さがワークフローの作成，実行の問題となっている．

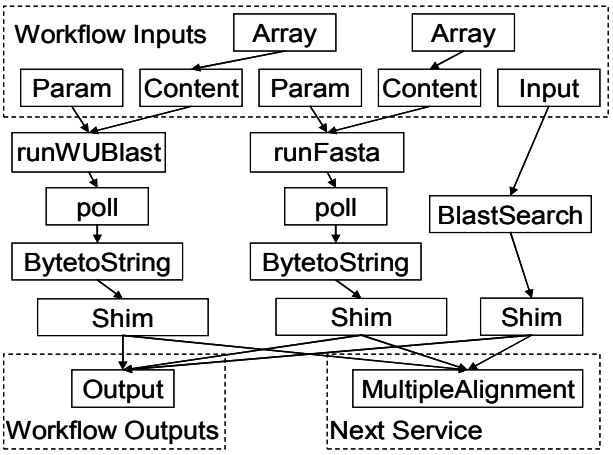

図１：ワークフロー

3. メタサービスの提案

本研究では，類似サービス群の統合利用を容易にするために，サービスの入出力定義に着目する．類似サービス群は，それぞれ異なる手法で解析を行うが，目的は共通している．そのため各サービスの入出力も共通するデータを持っていると考えられる．前述した相同性検索のサービス群を見ても，各サービスは入力にデータベース名, 出力にヒット配列の概要など共通したデータを持っている．このようなデータを統合して扱うことができれば，これまでサービス毎に行っていた作業を統一した方法で行えるようになり，ワークフローにおけるデータ整合の煩雑さを解消できると考えた．そこで本研究では，入出力データの関連を記述したインターフェースオントロジを作成し，これを利用する統合インターフェースを提供するメタサービスを提案する．

3.1. メタサービス概要

メタサービス概要（図 2）について説明する．類似サービス群の入出力定義の XML から，単体のデータ表現を行う要素を全て抽出し，各要素間のマッピングを行い対応関係を得る．本研究での対応関係を，ある XML の 1

要素が別の XML の 1 要素と同種のデータを持つという関係とする．対応関係をインターフェースオントロジとして記述する．統合インターフェースは，作成したインターフェースオントロジを基に，入力用のテンプレートを作成し，クライアントに提供する．統合インターフェースは，個別サービスにアクセスするためのラッパーへ，クライアントがテンプレートに入力したデータを適切に分配する．ラッパーは個別サービスにアクセスし，出力 XML を統合インターフェースへ返す．統合インターフェースは，各出力 XML から対応するデータを抽出し，結果をまとめた XML を生成しクライアントに返す．

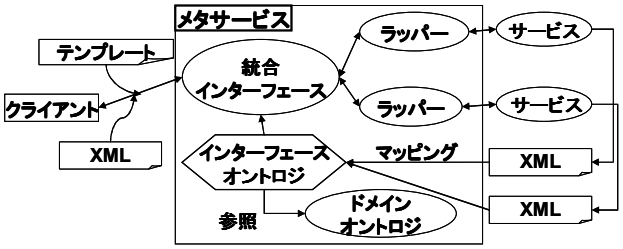

図 2：メタサービス概要

3.2. データマッピングとインターフェースオントロジの記述

インターフェースオントロジ作成時のマッピングを人手で行うのは，コストがかかる．そこで本研究では，XML 要素の名前とデータタイプの比較を行うアルゴリズムを用い，マッピングの自動化を試みる．マッピングに際し，以下の点を考慮する．(1)入力データのマッピングには，各サービスの WSDL の Input 記述，出力データには，各サービスが出力する XML のスキーマ記述を利用する．(2)データタイプの比較には，W3C 勧告の XML Schema[3]のデフォルトデータ型を用いる．(3)要素の名前の類似度の計算には，Levenshtein Distance[4]を用いる．(4)あるスキーマの 1 要素と別スキーマの全要素を比較する．データタイプが一致し，名前の類似度がもっとも高い要素を，対応関係を持つ要素とする．

マッピングにより，類似サービス群の入出力 XML の対応関係を得ることができる．この情報を基に W3C 勧告のオントロジ記述言語 OWL を用い，インターフェースオントロジの記述を行う．記述の方針を以下に示す．(1)クラス”MetaService”を定義する．(2)“MetaService”は”input”，”output”クラスを持つ．データは”owl:DatatypeProperty”で記述される．個別のスキーマへのリファレンスを”owl:equivalentProperty”で記述する．(3)myGrid プロジェクト[5]で利用されるドメインオントロジから，扱う類似サービス群のクラス表現を入手し，”Meta Service”，”input”，”output”クラスの内部に”owl:equivalentClassof”で記述する．メタサービスは，このクラス表現をメタデータに持つサービス群を取り扱うということを意味する．

4. プロトタイプ開発と成果

上述したシステムを相同性検索のサービス群に対し実装を行った．利用する Web サービスは DDBJ が提供する BLAST，EBI が提供する FASTA,WU-BLAST の３つとする．また，メタサービス利用のクライアントシステムには Taverna[6]を利用した．Taverna は，バイオインフォマティクス解析の Web サービスを利用したワークフロー構成のための GUI ツールである．メタサービスは Web サービスとして実装されているので，Taverna からの利用が可能となる．

統合インターフェース作成の結果を示す．各サービスの入出力のスキーマの要素数はそれぞれ，BLAST が（Input 4, Output 45），FASTA, WU-BLAST が(5, 62)である．入力要素に関しては，サービス実行に必須な要素のみとした．また FASTA, WU-BLAST の出力スキーマは共通したものを用いている．これらを直接利用した場合，扱う入出力要素の総数は，(14, 107)である．人手で要素間のマッピングを行い，統合インターフェースを作成した場合は，(6, 80)，本研究でのマッピングアルゴリズムを用いた場合は，(7, 92)であった．

Taverna からの通常の Web サービス利用とメタサービス利用を比較する．通常利用と比較し，統合インターフェースからサービス群の入力を統合したテンプレートを与えることで，これまで個別のサービス毎に設定していた要素をサービス群単位で一括して設定できるようになった． また，実行結果から必要なデータ要素を指定することで，類似サービス群の出力データから関連する情報を取得できることを確認した．これにより，例えば，図１のワークフローで相同性検索の次のマルチプルアラインメントサービスに必要な配列データを取得することができ，これまで個別のサービスごとに行っていた出力データの整合のための Shim 作成も統一した方法で行えるようになり，サービス連携が実現できる．

サービスの通常利用とメタサービス利用の並列実行の際の実行速度について比較する．例としてアクセッション番号”Q9NRA8”のエントリ配列を入力として 5 回実行した際の平均時間を計測した．結果は通常の利用が約 45 秒，メタサービス利用の場合は，約 49 秒であった．メタサービスを利用してもワークフロー全体の実行時間に与える影響は少ないと考える．

5. まとめ

本研究では Web サービス統合利用のためのメタサービスの提案を行った．メタサービスを利用することで，類似サービス群の統合利用，サービス連携が可能になった．

今後の課題として，より精度の高いデータマッピングアルゴリズムの検討や，データ仲介機能の追加などが挙げられる．

参考文献

[1] P. Lord, P.Alper, C.Wroe, and C.Goble.
Feta: A lightweight architecture for user oriented semantic service discovery. In The Semantic Web: Research and Applications: Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece: 17-31(2005)
[2]U.Radetzki,U.Leser,S.C.Schulze-Rauschenbach, J.Zimmermann,J.Lussem,T.Bode and A.B.Cremers. Adapters, shims, and glue-service interoperability for n silico experiments, Bioinformatics, 22:1137-1143 (2006).
[3]The World Wide Web Consortium: http://www.w3.org/
[4]V.I,Levenshtein: Binary Codes Capable of Correcting Deletions, Insertions and Reversals, Soviet Physics, Doklady, Vol.10, No.8:707-710(1966)
[5]myGrid: http://www.mygrid.org.uk/
[6]T.Oinn, M.Addis, J.Ferris, D.Marvin, M.Greenwood, T.Carver, M.R.Pocock, A.Wipat and P.Li .
Taverna: a tool for the composition and enactment of bioinformatics workflows,Bioinformatics, 20: 3045-3054(2004)