Workflow 기반의 생명정보 분석 자동화 환경 구축에 관한 연구

Bioworks – A scientific workflow platform for problem solving in biological domain

한영만, 이상주
한국과학기술정보연구원

Youngmahn Hahn, Sang-Joo Lee
Korea Institute of Science and Technology Information

요약

Workflow 형태로 수행되는 BT 분야에서의 생명정보 분석과정을 효과적으로 모델링하고 자동화하기 위한 통합 Bio-Workflow 시스템(Biowoks)을 개발하였다. 사용자는 Bioworks 시스템을 통하여 복잡한 생명정보 분석과정에 대한 Workflow 모델을 손쉽게 구성할 수 있으며, 이를 실행하여 단계별 중간 결과물을 생성할 수 있다. 또한 각각의 중간 결과물에 대한 가시화 및 검증 모듈을 플러그 인 형태로 제공함으로써 보다 손쉽게 분석 업무를 수행할 수 있다. 작성된 생명정보 분석 Workflow를 XML 형태로 생성하여 웹 서비스를 통해 공유함으로써 연구자 간의 협업 연구를 통한 시너지 효과를 극대화 할 수 있다.

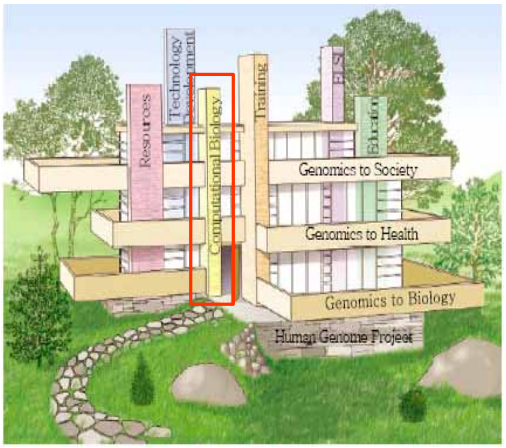
Abstract

Bioworks is designed to make, visualize, automate and execute a model of biological analysis processes as a workflow in biotechnology field. It provides for constructing a workflow model of complex biological analysis processes more easily and reporting the analysis results of each step. Users can perform their analysis process by using a visuality and validation module provided as a plug-in program. It supports sharing their workflow in XML format, to which conversion is supplied by Bioworks, with Web Services to improve the efficiency of their study.

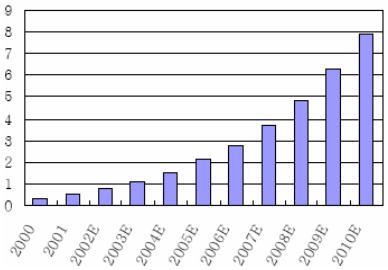
I. 서 론

1. 필요성

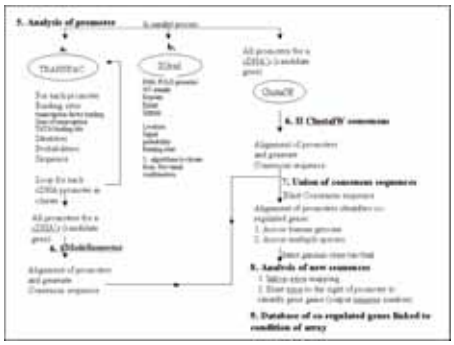
미국의 국립인간게놈연구소(NHGRI)의 소장이었던 Francis Collins가 제시한 향후 Genomics의 발전방향에 대한 삽화(<그림 1>)에서 보는 바와 같이 생명정보학 분야는 게놈 유전체 학에서의 하나의 큰 기둥으로서의 역할을 담당하며 다른 연구 분야와의 유기적 협력관계를 유지할 것으로 예상된다. 또한 <그림 2>에서 예시한 것처럼 생명 정보 연구 분야의 폭발적인 시장 수요 증가가 현실화 되고 있는 시점에 있다[1]. 따라서 유전자 구조 및 기능, 진화상 관계 등 생명정보 분야에서의 중요한 문제들에 대해 매일 발견되는 새로운 지식을 종합적으로 신속하게 분석하는 것이 매우 중요하다. 일반적인 생물정보 관련 연구는 <그림 3>에서 예시한 바와 같이 여러 생물 정보 데이터베이스를 검색하여 다양한 정보를 축출하고 이에 대한 다양한 분석도구의 적용 및 결과물 분석 등의 여러 단계의 단위 분석 업무의 Workflow 형태로 수행된다. 따라서 보다 효과적이고 체계적인 생명정보 연구를 위해서는 Workflow기반의 생명정보 시스템 개발이 중요하다. 유럽의 유수의 5개 대학과 EBI(European Bioinformatics Institute)를 포함한 관련 그룹에서는 생명정보 연구 분야와 관련한 eScience 실현을 위한 myGrid 프로젝트를 수행하고 있으며 이 프로젝트의 핵심 목표는 Web Service를 통한 생명정보 데이터 및 애플리케이션의 통합을 위한 다양한 툴 및 Workflow 개발에 있다[2]. 또한, 생명정보 연구를 위한 Workflow 플랫폼이 개발 및 상용화되고 있으며 특히 대표적인 IT 솔루션 제공업체인 IBM은 BioWBI(Bioinformatics Workflow Interface)[3]의 개발 및 서비스를 통해 생명정보 데이터의 통합 및 생명정보 연구를 위한 Workflow 플랫폼을 제공하고 있다.



▶▶ 그림 1. Francis Collins가 제시한 향후 Genomics의 발전 방향



▶▶ 그림 2. 전 세계 생물정보학 시장규모($US Billon), "Bioinformatics Opportunities," Digital Vector Market Report, 2004.



▶▶ 그림 3. Biological Workflow for Promoter Identification

II. 본 론

1. 생명정보 분석 Workflow의 특징

Workflow라는 단어가 'Work + Flow'로 구성되어 있는 것처럼 Workflow 시스템 구현에 있어 중심 관점은 대상 시스템이 처리하는 일의 영역에 따라 달라지며 특히 생명공학 분야에서의 Workflow는 가설과 검증, 그리고 중간 단계에서의 결과물의 분석 등이 중요하기 때문에 일반적인 Business Workflow와는 다른 특징들을 갖고 있다. 이를 <표 1>에 정리하였다.

[표 1] Business Workflow와 Scientific Workflow 차이점

구 분	Business Workflow	Scientific Workflow
중심대상	Transaction Process & state	Problem solving Knowledge
개방성	파트너 간 폐쇄적	개방적 커뮤니티를 통한 과학적 데이터 교류
검증과 오류처리	상호 간의 협의된 정책에 의해 일관된 방식	Learning from mistakes 중간 결과물에 대한 검증이 중요
재사용	중요하지 않음	재사용이 중요
유연성	Workflow는 좀처럼 변경되지 않음	Based on experiments Rapid & flexible
데이터 유형	Small amounts of data 정형화된 타입	Large amounts of data 다양하고 비정형화된 데이터 타입과 포맷
흐름유형	Data와 Control 흐름이 분리되어 진행 Process 흐름에서의 객체 상태가 중요	Data와 Control 흐름이 통합 되어 진행 Process 흐름에서의 데이터 값이 중요

2. Bioworks 시스템의 중요 요구사항

앞서 제시한 생명정보 분석 Workflow의 특징을 고려하면 Bioworks 시스템의 핵심 요구사항을 도출해 낼 수 있다. 첫째, 생명정보 분석 Workflow는 Business 분야의 Workflow와는 달리 데이터 흐름 중심으로 진행된다. 따라서 Workflow를 구성하는 컴포넌트는 데이터를 처리하는 Process, 데이터 흐름에 대한 Link, 그리고 단계별 입출력을 저장하는 입출력 채널로 구성된다. 둘째, 생명정보 분석 Workflow의 데이터 흐름에서는 대부분 데이터 변환을 수반한다. 따라서 각각의 Process와 Link 객체는 내부적으로 데이터 변환을 수행할 수 있도록 구성되어져야 한다. 셋째, Workflow의 각 실행 단계에서의 중간 결과물과 결과물간의 연관성을 추적할 수 있어야 한다. 넷째, 다양한 생물정보 관련 서비스들과 연계를 쉽고 확장성 있게 구성할 수 있어야 한다. 다섯째, 연구자들이 단계별 실행 결과물에 대해 보다 효율적으로 분석하고 검증할 수 있도록 해당 결과물에 대한 다양한 가시화 도구를 플러그 인 형태로 제공해야 한다.

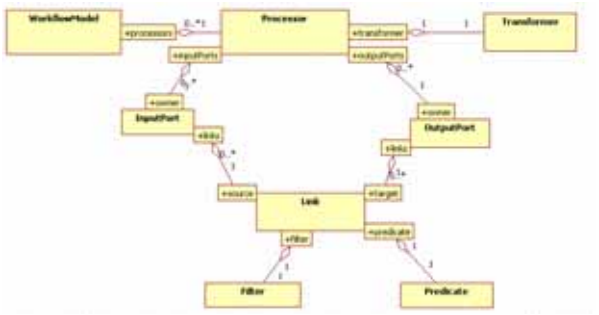
3 생명정보 분석 Workflow 실행 순서

<그림 4>은 하나의 생명정보 Workflow의 실행 개념도이다. 하나의 Workflow는 Process를 Vertex로 하고 Process간 Link를 Edge로 하는 'Directed Graph'로서 표현될 수 있다. 하나의 초기 실행 Process는 최상위 Vertex에 해당하는 Process가 되며 각각의 Process는 Link에 의해 연결되어 부모 Process와 자식 Process를 갖게 된다. 하나의 Process는 상위 Process들이 모두 종료되어 그것의 Output이 해당 Process의 Input으로 적합하게 설정되었을 때 비동기적으로 실행된다. Process 간의 Link에 의한 데이터 전달이 일어나는 시점에서 특정 전이 조건에 대한 검사와 데이터 변환이 이루어진다.

▶▶ 그림 4. Workflow 실행 개념도

4. Bioworks 시스템 핵심 컴포넌트

앞서 제시한 실행 개념도를 바탕으로 하여 Bioworks 시스템을 구성하는 핵심 컴포넌트 모델에 대한 클래스 다이어그램을 <그림 5>에 도시하였다.



▶▶ 그림 5. Bio-Workflow 핵심 클래스 다이어그램

5. 생명정보 분석 Workflow에 대한 XML 스키마

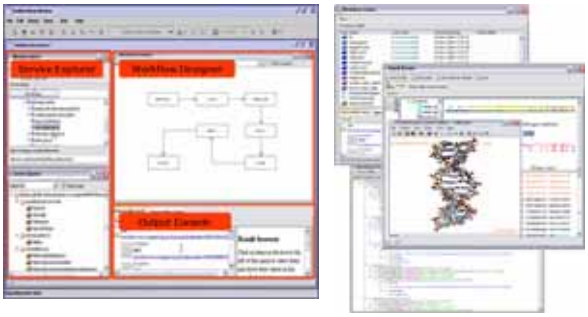
하나의 생명정보 분석 Workflow는 구조화된 'Directed Graph'로서 표현되어 질 수 있다. 따라서 Workflow 모델은 트리 구조로 구조화 된 객체 집합을 적합하게 기술할 수 있는 XML을 통하여 재정의 될 수 있다. 앞서 우리는 생명정보 분석 Workflow는 일반적인 Business Workflow와는 중심 관점 및 Process 처리 방식 등이 매우 다름을 분석하였고 이에 따른 생물 정보 분석 Workflow에서의 핵심 요구 사항 및 추상화된 객체 모델을 도출하였다. 이를 기반으로 하여 XML 스키마를 <그림 6>에 도시하였다.

```
<?xml version="1.0" encoding="UTF-8"?>
<s:schema elementFormDefault="qualified" xml:lang="EN"
          targetNamespace="http://bioworks.org/workflow"
          xmlns="http://bioworks.org/workflow"
          xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="example"/>
  <s:workflow id="test_001" name="Test">
    <s:description>Test workflow</s:description>
    <s:input-source id="in1" name="seq_id">
      <value>A0001</value>
    </s:input-source>
    <s:output-sink id="out1" name="align_result">
      <value>/user/blast_result.res</value>
    </s:output-sink>
    <s:processor id="p1" name="get_seq_by_id">
      <s:transformer>
        <s:ref-id>service_op_002</s:ref-id>
      </s:transformer>
      <s:input-port index="0" name="seq_id"/>
      <s:output-port index="0" name="out_seq"/>
    </s:processor>
    <s:processor id="p2" name="blast_align">
      <s:transformer>
        <s:ref-id>service_op_005</s:ref-id>
      </s:transformer>
      <s:input-port index="0" name="seq"/>
      <s:output-port index="0" name="out_seq"/>
    </s:processor>
    <s:link source="in1" target="p1:in_0"/>
    <s:link source="p1:out_0" target="p2:in_0">
      <s:predicate class="org.bioworks.predicate.NotNullOrEmpty"/>
      <s:filter s:ref-id="filter_002"/>
    </s:link>
    <s:link source="p2:out_0" target="out1"/>
  </s:workflow>
</s:schema>
```

▶▶ 그림 6. Bio-Workflow XML 스키마

6. Bioworks 사용자 인터페이스

앞서 제시한 Bioworks 시스템의 요구사항을 충족하기 위한 전체 사용자 인터페이스를 <그림 7>에 도시하였다. Bioworks 사용자 인터페이스는 사용자가 손쉽게 생명정보 분석과정에 대한 Workflow 손쉽게 작성할 수 있도록 사용자 중심의 시각화된 유저인터페이스를 제공하고 있다. 각각의 세부 컴포넌트에 대해 살펴보면, 분산된 생명정보 서비스를 키워드/카테고리 별로 검색하는 기능을 제공하는 Service Explorer, Drag & Drop 방식으로 손쉽게 Workflow를 작성할 수 있도록 하는 Workflow Editor, Workflow 실행 현황 검증을 위한 Output Console, 그리고 단계별 결과물의 가시화 및 분석 기능을 제공하는 Result Browser로 구성되어 있다.



▶▶ 그림 7. Bioworks 시스템 사용자 인터페이스

III. 결 론

바이오 분야는 최근 급격히 세계시장이 증가하고 있는 분야이며, 향후 국가 전략 산업으로 육성되고 있는 분야이다. 특히 유전체 연구 등 최근 급속히 진행되는 연구는 대량의 자료를 분석하고 이를 활용하여 다양한 응용연구가 이루어지고 있는 실정이다. 이에 발맞추기 위해서는 생명과학분야의 다양한 분석 프로그램을 이용해야 한다. 하지만 국내외의 생명과학연구 분야는 이러한 대규모 분석 처리과정에 어려움을 겪고 있는 실정이다. BioWorks는 연구자들이 어려워하는 복잡한 분석 과정은 자동화하고 이를 공동연구자들과 같이 공유하면서 연구의 효율성을 극대화 할 수 있는 도구로써 향후 국가 생명과학 연구 발전에 기여할 수 있는 도구가 될 것으로 예상된다.

▌참 고 문 헌▐

[1] Goldfarb, Debra, Bio-IT Infrastructure Market Forecast 2001-2006, IDC, 2002.

[2] Tom Oinn, et al., Taverna: A tool for the composition and enactment of bioinformatics workflows, Bioinformatics, 20, 17, pp. 3045-3054, 2004.

[3] Life Sciences Practice Team, BioWBI and WEE: Tools for Bioinformatics Analysis Workflows, IBM Business Consulting Services-AIS, 2004.