바이오인포매틱스 도구 통합을 위한 워크플로우 기반의 멀티에이전트 시스템

손봉기*○ 이건명* 황경순* 김영창**
*충북대학교 전기전자 및 컴퓨터공학부/AITrc, **충북대학교 생명과학부
dobest○@aicore.chungbuk.ac.kr

A Workflow-Based Multiagent System for Integrating Bioinformatics Tools

Bongki Sohn*○ Keonmyung Lee* Kyungsoon Hwang* Youngchang Kim**
*School of Electronic and Computer Engineering, Chungbuk National University/AITrc
**School of Life Science, Chungbuk National University

요 약

이 논문에서는 여러 가지 도구를 논리적인 순서로 사용함으로써 이루어지는 작업을 워크플로우로 보고, 이러한 관점에서 바이오인포매틱스 도구를 통합하는 새로운 멀티에이전트 시스템을 제안한다. 제안한 시스템은 기존의 도구를 랩퍼 에이전트로 구현하고, 에이전트간의 통신은 XML 형식의 메시지로 이루어진다. 수신 에이전트는 송신 에이전트가 전송하는 정보를 명시적으로 알리지 않고도 메시지로부터 필요한 정보를 추출할 수 있다. 제안한 시스템의 이러한 특징은 바이오인포매틱스 도구와 데이터베이스의 통합을 용이하게 한다. 또한, 제안한 시스템에서는 워크플로우를 여러 가지 제어 구조를 이용하여 정의할 수 있으며, 워크플로우 진행을 모니터링할 수 있는 기능을 제공한다. 제안한 시스템의 가용성을 보이기 위해 박테리아 *Sphingomonas Chungbukensis DJ77* 의 유전자 주해(gene annotation) 작업에 제안한 시스템을 적용하여 구현하고 있다.

1. 서 론

shot-gun 방법이나 효율적인 assemble 소프트웨어 같은 여러 가지 고속의 서열화 기술의 발달로, 많은 종의 유전체 정보가 밝혀지고 있다. 유전체 프로젝트에서, 긴 DNA 서열은 2kb 정도의 여러 개의 겹치는 단편으로 분할되고, 단편에 대한 DNA 서열이 결정되어 긴 컨티그(contig)로 조립된다. 이러한 컨티그들로 유전체 지도가 제작된다[1]. 유전체 지도에서, 유전자를 포함하고 있을 것 같은 ORFs(Open Reading Frame)을 찾고, 이에 대한 기능을 예측하여 유전자 기능과 같은 관련있는 정보를 유전체 데이터베이스에 대해 주해 작업을 한다.

많은 바이오인포매틱스 도구와 데이터베이스가 유전체 프로젝트에서의 여러 가지 작업을 지원하기위해 개발되고 있다. 유전체 분석 대상종(*organism*)과 수행할 작업 및 연구의 관심사항에 따라 연구자는 선택적으로 도구를 사용할 수 있다. 적절한 도구를 선택하기 위해서는 연구자가 각 도구의 기능과 작동 방법을 이해해야 한다. 연구자가 도구와 데이터베이스에 대해 충분히 이해하지 못하면, 주어진 작업을 처리하기 위해 적절한 도구를 선택하는 것이 쉽지 않다. 대부분의 바이오인포매틱스 도구와 데이터베이스는 웹을 통해 제공되며, 일부는 자유롭게 다운로드하여 설치할 수 있도록 공개된다. 그러나 각각의 도구에 대한 입력과 출력 데이터 형식 및 작동 방법을 습득하는 것은 어려운 일이다.

이러한 어려움을 줄이기위해, 이 논문에서는 지능적인 바이오인포매틱스 도구 통합을 제공하는 워크플로우 기반의 멀티에이전트 시스템을 제안한다. 대체로 생물학적 데이터 처리는 여러 도구를 선행관계에 의해 실행함으로써 이루어지는데, 작업에 대한 도구 실행의 논리적인 흐름을 워크플로우로 볼 수 있다[2,3]. 제안한 시스템에서는 워크플로우를 관리하는 방법과 같이 바이오인포매틱스 도구를 통합한다. 기존의 도구는 랩퍼 에이전트를 개발하여 쉽게 멀티에이전트 시스템에 추가한다[4]. 에이전트간 통신은 XML 형식의 메시지로 이루어진다. 송신하는 에이전트가 전송하는 정보에 대해 명시적으로 알리지 않아도 수신 에이전트는 필요한 정보를 추출할 수 있어 에이전트 통합이 유연하다. 여러 가지 통합 방법이 제안되었지만 대부분은 새로운 도구의 추가가 용이하지 않은 정적인 방법이다[5-9].

이 논문의 구성은 다음과 같다: 2장에서 바이오인포매틱스 도구 통합을 위한 워크플로우 기반의 멀티에이전트 시스템을 설명한다. 3장에서는 제안한 시스템의 구현에 대해 알아보고, 4장에서 결론을 맺는다.

2. 바이오인포매틱스 도구 통합을 위한 워크플로우 기반의 멀티에이전트 시스템

2.1 워크플로우 기반의 멀티에이전트 시스템 구조

그림 1은 워크플로우 관점에서 바이오인포매틱스 도구를 통합하기 위한 제안한 멀티에이전트 시스템의 구조를 보인 것이다. 사용자 인터페이스 에이전트를 통해 사용자는 도구를 접근하고 제어하며, 중간 처리 결과를 검색한다. 디렉토리 에이전트(Directory Agent)는 도구와 데이터베이스가 어디서 어떻게 사용되는 지와 어떤 서비스를 제공하는 지에 대한 정보를 관리한다. 계획 에이전트(Planning Agent)는 사용자의 요청을 받아들이고, 주어진 작업의 처리 방법에 대하여 적절한 워크플로우 정의를 제안한다. 워크플로우 관리 에이전트(Workflow Management Agent)는 모든 워크플로우 인스턴스 에이전트를 조정하고 모니터링한다. 각 워크플로우에 대해 워크플로우 인스턴스 에이전트를 생성하고, 사용자의 요청에 의해 실행중인 워크플로우를 종료한다. 또한 사용자의 요청을 해당 워크플로우 인스턴스 에이전트로 전달하며, 향후 의사결정에 도움이 될 수 있는 워크플로우와 관련 에이전트에 대한 통계적 데이터를 수집한다. 워크플로우 인스턴스 에이전트(Workflow Instance Agent)는 초기화된 워크플로우를 관리한다. 즉, 워크플로우 정의에 따라 적절한 타스크 에이전트를 활성화시키고, 타스크 에이전트간의 통신을 조정한다. 또한 중간 결과에 대한 검색 서비스를 제공하며 현재 워크플로우의 인자를 갱신한다. 타스크 에이전트에는 랩퍼 에이전트와 특수 목적 에이전트(Special Purpose Agent)가 있다. 랩퍼 에이전트는 기존의 도구이나 어플리케이션이 다른 에이전트와 통신하고 특정 작업을 수행할 수 있는 자율적인 개체처럼 보이게 하는 에이전트이다. 또한 랩퍼 에이전트는 사용자가 도구의 데이터 형식이나 옵션의 의미를 이해하지 않고도 도구를 사용할 수 있게 한다. 특수 목적 에이전트는 적절한 도구가 존재하지 않는 특정 데이터 처리를 위해 개발된 에이전트이다. 이러한 타스크 에이전트를 워크플로우 정의에 따라 실행으로써 바이오인포매틱스 작업을 수행한다.

본 연구는 첨단정보기술 연구센터(AITrc)를 통해서 과학재단의 지원에 의한 것입니다.

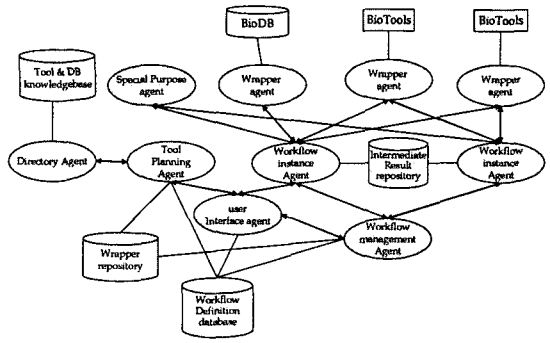

그림 1. 워크플로우 기반의 멀티에이전트 시스템 구조

제안한 시스템에는 여러 가지 레포지트리가 있다. Tools/DB는 워크플로우에 사용될 타스크 에이전트에 대한 정보를 포함하며, 디렉토리 에이전트에 의해 사용된다. 랩퍼 레포지트리(Wrapper Repository)는 랩퍼와 다른 타스크 에이전트의 코드를 관리한다. 타스크 에이전트가 필요할 때, 이 코드가 인스턴스화되어 실행된다. 워크플로우 정의(Workflow Definition) 데이터베이스는 미리 정의된 워크플로우 정의를 저장한다. 중간 결과 레포지트리(Intermediate Result Repository)는 타스크 에이전트 사이의 공유 메모리 역할을 한다. 즉, 타스크 에이전트의 처리 결과는 중간 결과 레포지트리에 저장되고, 사용자에 의해 중간 결과가 검색될 수 있고, 메일박스처럼 메시지를 교환하는데도 사용된다.

2.2 구성 에이전트 구조

시스템을 구성하는 에이전트는 처리 가능한 메시지 종류, 가능한 상태, 구성 모듈 등이 시스템에서의 역할에 따라 각각 다른 구조를 가진다.

타스크 에이전트

타스크 에이전트는 워크플로우 인스턴스 에이전트에 의해 활성화되고, 수신 메시지를 처리하여 결과를 반환한다. 또한 도구나 작업에 대한 기본 인자값을 가지는데, 워크플로우 인스턴스 에이전트로부터의 메시지에 의해 대체될 수 있다.

타스크 에이전트는 워크플로우 인스턴스 에이전트로부터의 새로운 작업 할당과 작업 처리 종료에 대한 제어 메시지와 작업 처리 상태에 대한 질의 메시지를 처리할 수 있다.

타스크 에이전트는 *Finished with Success, Finished with Failure, Timer-Out, Terminated, Processing, Failed with no proper input information, Idle*의 상태에 속할 수 있다.

타스크 에이전트는 다음과 같은 모듈로 구성된다. 통신 모듈(communication module)은 메시지를 송수신하고, 메시지를 파싱하여 필요한 정보를 추출한다. 작업 제어 모듈(task control module)은 특정 종류의 메시지에 해당하는 이벤트 처리를 제어한다. 작업 제어 모듈은 추론 엔진과 규칙을 저장하는 규칙 베이스(rule base), 에이전트의 상태 정보를 관리하기 위한 객체 베이스(object base)를 사용하여 지능적인 처리를 한다. 랩핑 모듈(wrapping module)은 랩핑된 도구에 요청을 전송하고 도구로부터 결과를 수신한다. 특정 목적 에이전트는 처리 작업을 수행하는 여러 개의 작업 처리 모듈을 가질 수 있다.

워크플로우 인스턴스 에이전트

워크플로우 인스턴스 에이전트는 워크플로우 인스턴스를 관리한다. 워크플로우 관리 에이전트로부터 워크플로우 정보를 받아 워크플로우 내부 자료 구조와 제어 구조를 생성한다. 즉, 워크플로우 정보를 포함하는 XML 메시지를 파싱하여 워크플로우를 추적하고 타스크 에이전트에게 작업을 할당하는데 사용되는 방향성 그래프를 생성하고, 규칙 베이스로 제어 구조 정보를 표현한다. 규칙 베이스는 어떤 액션이 취해질 것인가를 결정하는데 사용되고, 작업 상태와 인자값은 객체 베이스에서 관리된다. 워크플로우 정보에 따라 워크플로우 인스턴스 에이전트는 타스크 에이전트를 초기화하고, 타스크 에이전트 ID, 현재 상태, 시작 시간, 타임아웃 플래그, 타임아웃값, 관련된 규칙의 ID와 생성된 결과에 대한 포인터를 포함하는 객체를 생성하여 초기화된 타스크 에이전트에 전송한다. 타스크 에이전트에 전송되는 정보는 성공적으로 종료된 이전 타스크 에이전트들의 결과와 사용자에 의해 입력된 인자들이다. 워크플로우 인스턴스 에이전트는 성공적으로 종료된 타스크 에이전트들의 결과를 합병함으로써 결과를 생성한다.

워크플로우 인스턴스 에이전트는 새로운 워크플로우를 할당하고, 진행중인 워크플로우를 종료하고, 워크플로우 인스턴스 에이전트를 비활성화하는 제어 메시지와 워크플로우 상태나 구성 작업에 대한 질의 메시지를 처리한다.

워크플로우 인스턴스 에이전트는 *Finished with Success, Finished with Failure, Timer-out, Terminated, Processing, Idle*의 상태에 속할 수 있다.

워크플로우 인스턴스 에이전트는 다음의 구성요소로 구성된다. 통신 모듈은 메시지를 송수신하며, 수신 메시지를 파싱하여 메시지에 대한 해당 이벤트를 생성하여 프로세스 제어 모듈(process control module)에 전송한다. 프로세스 제어 모듈은 워크플로우 정의에 따라 타스크 에이전트를 초기화하고, 워크플로우 진행과 상태 정보를 관리하는 객체 베이스, 워크플로우의 제어 흐름에 관한 규칙을 저장하는 규칙 베이스와 추론 엔진을 모니터링 한다.

워크플로우 관리 에이전트

워크플로우 관리 에이전트는 워크플로우 인스턴스 에이전트를 생성하고, 워크플로우 정보를 전송하고 에이전트를 비활성화시키기도 한다.

워크플로우 관리 에이전트는 새로운 워크플로우 인스턴스를 생성하고 진행중인 워크플로우를 종료하거나 워크플로우 인스턴스 에이전트를 비활성화하는 제어 메시지와 특정 워크플로우의 상태에 대한 질의 메시지를 처리할 수 있다.

워크플로우 관리 에이전트는 통신모듈, 관리 제어 모듈(management control module), 객체 베이스, 규칙 베이스, 추론 엔진을 포함한다.

계획 에이전트

계획 에이전트는 워크플로우 계획을 생성한다. 계획 에이전트는 바이오인포매틱스 작업에 대해 미리 정의된 워크플로우 정의를 관리하고, 영역 전문가 지식을 포함하는 규칙 기반 시스템(rule base system)을 사용해서 선택된 워크플로우 정의에 포함된 작업에 대해 후보 도구를 추천한다. 또한 사용자가 기존의 워크플로우 정의와 제어구조를 수정하고 편집할 수 있게 한다. 계획 에이전트는 통신모듈, 객체 베이스와 규칙 베이스에 의한 추론 엔진, 이벤트 지향적인 관리 제어 모듈, 워크플로우 편집을 위한 GUI로 구성된다.

디렉토리 에이전트

디렉토리 에이전트는 사용가능한 도구에 대한 정보를 등록하고 도구에 대한 질의에 응답한다. 새로운 도구는 도구 정보를 디렉토리 에이전트에 등록하고, 도구에 대한 랩퍼에이전트를 랩퍼 레포지트리에 등록함으로써 멀티에이전트 시스템에 통합된다.

사용자 인터페이스 에이전트

사용자 인터페이스 에이전트를 통해 사용자는 새로운 워크플로우에 대한 정의를 초기화하고, 워크플로우 인스턴스를 활성화.종료.비활성화시킨다. 또한, 워크플로우 인스턴스의 진행을 모니터링하고 중간 결과를 검색하며, 진행중인 워크플로운 인스턴스를 수정한다. 이러한 기능은 다른 에이전트와의 메시지 교환을 통해 가능하다.

2.3 XML 기반의 워크플로우 정의 및 통신

도구와 데이터베이스 사이의 데이터 형식 불일치 문제를 해결하기 위해 제안한 시스템에서는 XML 기반의 데이터 표현을 적용한다. 워크플로우 정의와 에이전트간 메시지는 XML 형식으로 표현된다. 워크플로우는 *sequence, fork, join, choice, merge, fork with choice(s), choice with fork(s), join with merge(s), merge with join(s)*와 같은 다양한 제어구조로 정의될 수 있다. 예를 들어, 다음 XML 기술은 *fork*

with choice 제어 구조를 나타낸 것이다. 작업 t_i 종료 후에 세 개의 쓰레드가 생성되고, t_{i2}와 t_{i3}중 하나가 조건에 따라 선택되어 처리된다.

<fork>
<base> <tid> t_i </tid> </base>
<outgoing nthreads = 3>
<tid> t_{i1} </tid>
<choice nested = true>
<flow><condition>$cond_{i2}$ </condition><tid>t_{i2}</tid></flow>
<flow><condition>$cond_{i3}$ </condition><tid>t_{i3}</tid></flow>
</choice></thread>
<thread><tid>t_{i4}</tid></thread>
</outgoing>
</fork>

생물학적 정보에 대한 데이터 개체의 태그명은 NCBI 시스템에서 사용되는 태그로 하고, 그 이외의 태그명은 직접 정의한다. 각 메시지는 응답 메시지를 수신 에이전트가 알 수 있도록 유일한 ID를 가진다. 타스크 에이전트가 메시지 전송을 통해 서로 통신하지만, 어떤 정보가 어떤 수신 에이전트에 전송되어야한다는 것을 명시적으로 표현하지 않는다. 타스크 에이전트는 메시지에서 어떤 정보를 추출할 것인지를 알고 있기 때문이다. 이러한 방법은 특정 타스크 에이전트 이후에 어떤 에이전트가 호출될 지에 대해 고려하지 않고 랩퍼 에이전트를 개발할 수 있게 한다. 따라서 기존의 도구와 새로운 도구가 시스템에 쉽게 추가될 수 있다.

3. 구현

에이전트는 다른 작업을 진행 중이더라도 수신되는 메시지에 대해 응답해야 한다. 따라서, 에이전트는 다중 쓰레드의 자바 프로세스로 구현될 수 있다. 즉, 통신 모듈을 위한 쓰레드와 이벤트 처리를 위한 쓰레드가 필요하다. 에이전트간 통신은 자바 RMI(Remote Method Invocation)를 통해 이루어진다. RMI에 의해 수신 에이전트에게 새로운 메시지를 고지하면 수신 에이전트는 중간 결과 레포지트리에서 메시지를 가져와 파싱하여 작업을 수행한다. 메시지의 유일한 ID는 메시지 검색과 대응하는 메시지에 대해 응답할 때 사용된다.

제안한 시스템의 가용성을 보이기 위해, 박테리아 *Sphingomonas Chungbukensis DJ77*의 유전자 주해 작업을 위한 바이인포매틱스 도구 통합 시스템을 개발하고 있다[10]. 워크플로우는 워크플로우 편집기를 통해 정의된다. 그림 2는 워크플로우 편집기로 편집한 주해 작업 워크플로우를 나타낸 것으로, 사각형은 타스크 에이전트, 타원은 각 타스크 에이전트의 입력 형식을 의미한다.

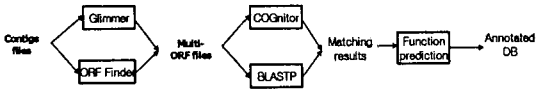

그림 2. 주해 작업 워크플로우

정의된 워크플로우의 실행은 워크플로우의 선택과 필요한 최초 입력값이 주어짐으로써 시작되며, 워크플로우 진행에 사용자의 처리작업이 필요한 경우도 있다. 그림 3은 시스템의 통합 인터페이스와 주해 작업 워크플로우의 최초 입력 및 중간 결과를 나타낸 것이다. 중간 결과는 사용자의 기능 예측(function prediction)에 필요한 요약된 정보로써, 이 정보에 대한 사용자의 부가적인 처리작업을 거쳐 주해작업이 종료된다.

4. 결론

이 논문에서는 바이오인포매틱스 도구 통합을 위한 워크플로우 기반의 멀티에이전트 시스템을 제안하였다. 제안한 시스템에서는 다양하고 이질적인 바이오인포매틱스 도구가 필요하고, 많은 양의 정보를 반복적으로 처리해야하는 생물학적 데이터 처리 작업을 워크플로우로 정의하고 실행한다. 워크플로우에 의한 작업 수행은 도구의 적절한 선택과 사용법에 대한 충분한 이해없이도 주어진 작업을 처리할 수 있게 한다. 또한 제안한 시스템은 개방된 멀티에이전트 구조를 채택하기 때문에, 새로운 도구를 통합하는데 매우 유연하다[11]. 현재 제안한 시스템 구조에 기반하여 박테리아 유전자 주해작업을 위한 시스템을 개발 중에 있다.

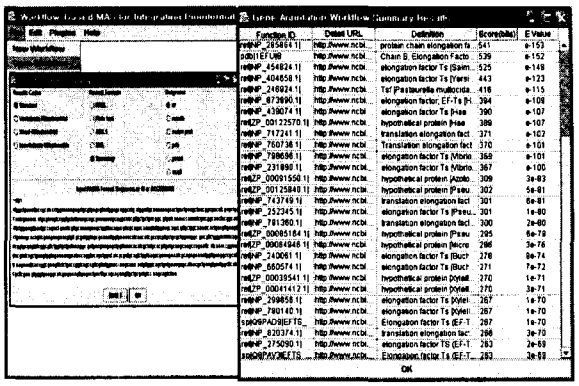

그림 3. 통합 인터페이스 및 주해 워크플로우 진행 결과

참고 문헌

[1] L. Hunter, Molecular Biology for Computer Scientists, *Artificial Intelligence and Molecular Biology*(L. Hunter, eds.), AAAI Press, 1993.

[2] H. Schuschel, M. Weske, Integrated Workflow Planning and Coordination. *Proc. of the 14th International Conference on Database and Expert Systems Applications*, 2003.

[3] F. Wan, S. K. Rustogi, J. Xing, M. P. Singh, Multiagent Workflow Management, *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 8, pp.105-117, 1999.

[4] L. Chen, H. M. Jamil, On using remote user-defined functions as wrappers for biological database interoperability, *international Journal of Cooperative Information Systems*, Vol.12, No.2, pp.161-195, 2003.

[5] K. -H, Cheung, P. Miller, A. Sherman, S. Strtmann, M. Schultz, et al., Graphically-Enabled Integration of Bioinformatics Tools Allowing Parallel Execution, *Proc. of the 2000 AMIA Annual Symposium*, Nov. 2000.

[6] S. Moller, U. Leser, W. Fleischmann, R. Apweiler, EDITtoTrEMBL: A Distributed Approach to High-Quality Automated Protein Sequence Annotation, *Proc. of the German Conference on Bioinformatics*, 1998.

[7] D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, H. -W. Mwes, Functional and Structural Genomics Using PEDANT, *Bioinformatics*, Vol.17, No.1, pp.44-57, 2001.

[8] K. Bryson, M. Luck, M. Joy, D. T. Jones, Agent Interaction for Bioinformatics Data Management, *Applied Artificial Intelligence*, Vol.15, No.10, pp.917-947, 2001.

[9] P. G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, R. Stevens, TAMBIS-Transparent Access to Multiple Bioinformatics Information Sources, *Proc. of the Sixth International Conference on Intelligent Systems for Molecular Biology*, ISMB98, Montreal, 1998.

[10] S. -J. Kim, J. Chun, K. S. Bae, Y.-C. Kim, Polyphasic assignment of an aromatic-degrading Pseudomonas sp., strain DJ77, in the genus Shingomonas as Sphingomonas chungbukensis sp. nov., *international Journal of Systematic and Evolutionary Microbiology*, Vol.50, pp.1641-1647, 2000.

[11] G. Weiss, *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence(eds)*, The MIT Press, 1999.