ISBN 978-89-6211-486-7 93500

오믹스 정보 분석 통합시스템(Bioworks)을 활용한 생명정보 분석 시나리오 구현

한 영 만(hans@kisti.re.kr), 이 상 주(lsj@kisti.re.kr)

한국과학기술정보연구원
Korea Institute of Science and Technology Information

목 차

1. 개요 ·····1
2. Bioworks 시스템 소개 ·····3
2.1. Bioworks 개요 ·····3
2.2. 슈퍼컴퓨팅 기반의 시스템 아키텍처 ·····4
2.3. Bioworks 클라이언트 프로그램 ·····4
2.4. Biworks 서버 측 워크플로우 실행엔진 ·····5
2.5. Bioworks 시스템의 주요 기능 및 특징 ·····6
3. Bioworks 시스템 사용법 ·····7
3.1. Java Web Start를 이용한 Bioworks 클라이언트 설치 ·····7
3.2. Bioworks 사용자 계정 등록 ·····7
3.3. Bioworks 시스템 로그인 및 사용자 작업 환경 ·····8
3.4. 사용자 카테고리 등록 및 수정 ·····9
3.5. 스크립트 도구 등록 및 수정 ·····10
3.6. 생명정보 자료 등록 및 수정 ·····11
3.7. 워크플로우 생성 및 편집 ·····13
3.8. Zoom In/Out, 자동 레이아웃, 프린트 ·····13
3.9. 워크플로우 노드 입력값 설정 ·····13
3.10. 입출력 데이터 변환 링크 스크립트 작성 ·····15
3.11. 워크플로우 실행 및 현황 모니터링 ·····16
3.12. 워크플로우 실행 결과물 저장 ·····16
3.13. 워크플로우 실행 시 결과물 자동 저장 ·····17
3.14. 사용자 자료 공유 설정 ·····18
3.15. 공유 사용자 자료 복사 ·····18
4. Bioworks 시스템을 활용한 생명정보 분석 시나리오 구현 ·····20
4.1. Design a primer set for the conserved region ·····20
4.2. Prokaryotic gene prediction ·····21
4.3. Human start codon prediction using a decision tree ·····21

4.4. Random mutation and alignment ····· 23
4.5. Comparison of MSA programs ····· 24
4.6. Searching polyketide synthase genes against a genome ····· 24
4.7. Protein profile ····· 26
4.8. Protein information ····· 26
4.9. Domain search using HMMER ····· 27
4.10. Protein motif discovery ····· 28
4.11. Get drugs related to a pathway ····· 29
4.12. Get KEGG pathway from DNA sequence ····· 30
4.13. Workflow 관련 PubMed 검색 ····· 31
5. 맺음말 ····· 31

그림 차례

그림 1. 일반적인 생명정보 분석과정(Promoter Identification) ·····1
그림 2. 생명정보 분석 연구를 위한 Bioworks 활용 흐름 ·····3
그림 3. 슈퍼컴퓨팅을 연계한 Bioworks 클라이언트-서버 아키텍처 ·····4
그림 4. Bioworks 클라이언트 UI 컴포넌트 구성 ·····5
그림 5. Bioworks 서버 엔진 실행 개념도 ·····5
그림 6. Bioworks 시스템 주요 기능 구성도 ·····6
그림 7. 사용자 계정 등록 화면 ·····8
그림 8. 사용자 작업 환경 ·····9
그림 9 사용자 카테고리 등록/수정 화면 ·····9
그림 10. 스크립트 도구 초기 설정 화면 ·····10
그림 11. 스크립트 도구 입출력 설정 화면 ·····10
그림 12. 스크립트 소스 설정 화면 ·····11
그림 13. 신규 생명정보 자료 생성 화면 ·····12
그림 14. 생명정보 자료 편집 화면 ·····12
그림 15. 신규 워크플로우 생성 화면 ·····14
그림 16 워크플로우 편집 화면 ·····14
그림 17 Zoom In/Out, 자동 Layout, Print 기능 예시 ·····15
그림 18. 워크플로우 실행 노드 입력값 설정 화면 ·····15
그림 19. 입출력 데이터 변환 링크 스크립트 작성 화면 ·····16
그림 20. 워크플로우 실행 및 현황 모니터링 ·····17
그림 21. 워크플로우 실행 결과물 저장 화면 ·····17

그림 22. 실행 결과 값 자동 저장 선택 화면 ·····18
그림 23. 사용자 자료 공유 설정 화면 ·····19
그림 24. 공유 사용자 자료 검색 및 복사 화면 ·····19

1. 개요

생명과학분야는 생물학, 분자생물학, 미생물학, 생화학, 유전학, 의학, 및 약학 등과 같은 다양한 학문들을 포함하고 있으며, 최근 들어 기존의 생명과학 연구에 전산학적인 기법들을 응용한 다양한 오믹스(-omics) 데이터를 대상으로 하는 유전체학, 단백체학, 대사체학 및 피지옴과 같은 생명정보 연구영역도 생명과학 분야에서 중요한 위치를 차지하기 시작하고 있다. 기존의 생명과학 연구는 주로 개개의 연구실 단위로 수행되었던 반면에 최근에는 인터넷 기술의 발달로 인하여 전 세계적으로 생산된 생물학 정보들이 공공의 데이터베이스를 통하여 수집 및 가공되고 있으며, 또한 다양한 종류의 생명정보 분석 도구들도 웹상에서 서비스 및 배포되고 있다.

그러나 이러한 생명정보 분석 기술의 발전에도 불구하고, 생명과학 연구자들이 자신의 연구에 적합한 기술을 활용하기에는 몇 가지 문제점이 있다. 첫째, 생명정보 데이터 및 도구의 이질성이다. 현재까지 국내는 물론 전 세계 생명 과학 분야 기업이나 연구 기관들은 연구 결과로부터 얻어진 생물 정보 데이터를 각기 다른 독자적인 포맷으로 저장되고 배포되어 왔으며 생명 과학 연구에 필요한 분석 도구들도 역시 각자의 프로그래밍 언어와 개발 환경을 기반으로 개발된 것이 현실이다. 이러한 까닭에 실로 다양하고 이질적인 생명정보 분석도구들이 각기 다른 입출력 포맷과 사용자 인터페이스를 갖고 분산되어 있어, 연구자들은 자신의 연구 환경에 맞는 분석도구를 선택하고 그것을 활용하는 데 있어 많은 어려움을 겪게 된다. <그림 1>에서 보는 바와 같이 일반적인 생명정보 분석 과정은 여러 생물 정보 데이터베이스를 검색하여 다양한 정보를 축출하고 이에 대한 다양한 분석도구의 적용 및 결과물 분석 등의 여러 단계의 단위 분석 도구의 입출력 연계로 이루어지기 때문에, 입출력 포맷이 상이한 분석 도구 간의 입출력 연계를 원활히 수행할 수 있는 통합 연구 환경이 절실히 요구된다.

둘째, 대용량•대규모 분석 환경이 미흡하다는 것이다. '인간 게놈 프로젝트'가 성공적으로 진행된 이래로 생명정보 데이터는 폭발적으로 증가해 왔으며 그러한 대용량 데이터를 누가 더 빠르고 정확하게 분석하여 생물학적으로 의미 있는 정보를 유추해내는 것이 가장 중요한 일이다. 게다가 최근 선진국들을 중심으로 빠르게 발달하고 있는 첨단 생명과학 연구는 기존의 소규모 연구방식에서 벗어나서 점차적으로 대용량의 생물학 정보들을 대상으로 복잡한 계산과정을 요구하기 때문에 고도의 컴퓨팅 인프라를 필요로 하게 되었다. 특히, 암과 같은 다인자 유발 질병의 경우, 기존의 단편적인 생명과학 연구방법으로는 한계가 있으므로 세포내의 전체적인 요인들에 대한 통합적인 분석법이 요구되며, 이를 위해 현재까지 밝혀진 방대한 데이터를 통합하고 이를 대규모 컴퓨팅자원을 활용하여 분석할 수 있는 슈퍼컴퓨팅 인프라 환경이 필요하다. 그러나 이와 같은 조건에 부합하는 전산자원들은 하드웨어 자

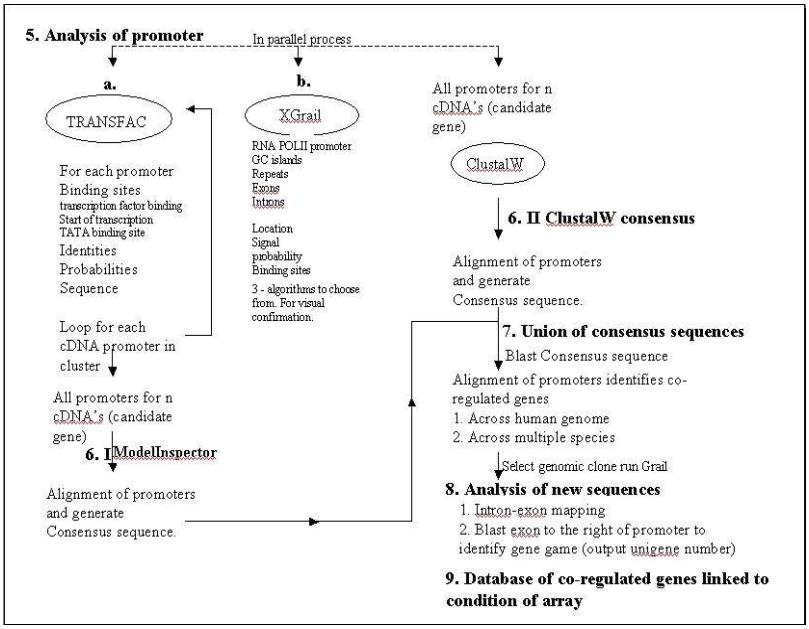


그림 1. 일반적인 생명정보 분석과정(Promoter Identification)

체의 단가가 매우 높을 뿐 아니라, 이를 효과적으로 관리하고 활용할 수 있는 전문 인력을 요구하기 때문에 개별적인 연구실 단위로는 제대로 된 연구기반을 갖추는 것이 현실적으로 불가능한 실정이다. 따라서 이러한 문제점들을 해결하기 위하여 도입된 것이 e-Science 환경구축을 통한 생명과학 분야에서의 첨단 전산자원의 공동 활용 방안이다. 이러한 문제점들을 해결하기 위하여 한국과학기술정보연구원(KISTI)에서는 생명과학 연구자들이 보다 손쉽게 자신의 연구에 필요한 생물정보 분석 도구들을 효과적으로 활용할 수 있도록 하기 위한 슈퍼컴퓨팅 인프라 기반의 오믹스 정보 통합시스템(Bioworks 시스템)을 개발하였다.

본 연구보고서에서는 Bioworks 시스템에 대한 주요 특징 및 기능을 소개하고 Bioworks 시스템을 활용하여 생명정보 연구 분야에서 주로 활용될 수 있는 시나리오(워크플로우) 구현 방법에 대해 제시하고자 한다.

2. Bioworks 시스템 소개

2.1. Bioworks 개요

Bioworks 시스템은 워크플로우 형태로 수행되는 다양한 생명정보 분석 과정을 효과적으로 모델링하고 자동화하기 위해 고안된 시스템이다. Bioworks 시스템은 다양한 조립 부품(생명정보 분석도구)을 사용하여 자신이 원하는 워크플로우를 손쉽게 만들 수 있도록 하는 편리한 사용자 인터페이스를 제공하고, 완성된 워크플로우를 슈퍼컴퓨팅 기반에서 빠

르고 정확하게 실행할 수 있도록 하는 최적의 실행 환경을 제공한다. 사용자는 Bioworks 시스템에서 제공하는 부품을 서로 연결하여 어떠한 일련의 작업을 수행하도록 하기까지 그냥 그림 그리듯이 각각의 부품을 컴퓨터 화면상에 마우스로 끌어다 놓고 서로 연결하여(끼워 맞추어) - 만약 맞지 않는다면 두 부품을 연결해주는 새로운 이음쇠를 직접 만들어 계속해서 연결하면서 - 자신이 원하는 시나리오를 만들고 실행하기만 하면 되는 것이다. 각각의 중간 결과물에 대한 가시화 및 분석 모듈을 플러그인 형태로 제공함으로써 보다 손쉽게 분석업무를 수행할 수 있다. 또한 작성된 워크플로우에 접근 수준을 설정하여 다른 사용자에게 게시하고 공유할 수 있는 기능을 제공함으로써 연구자 간의 협업 연구를 통한 시너지 효과를 극대화 할 수 있다. <그림 2>는 생명정보 분석 연구에 있어 Bioworks 시스템의 활용 업무 흐름을 보여준다.

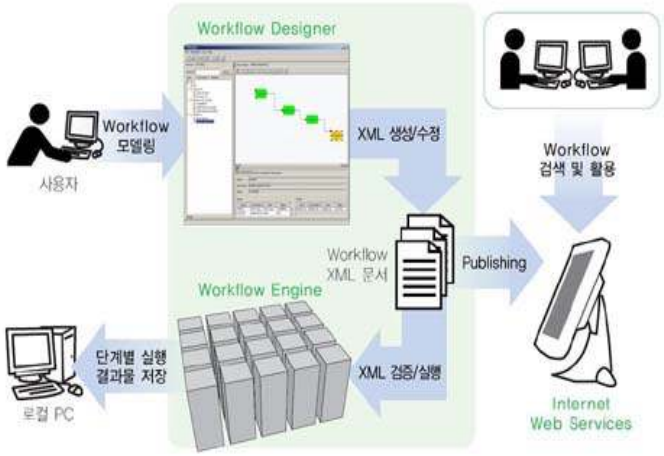


그림 2. 생명정보 분석 연구를 위한 Bioworks 활용 흐름

2.2. 슈퍼컴퓨팅 기반의 시스템 아키텍처

Bioworks 시스템은 <그림 3>과 같이 웹서비시스 방식의 클라이언트-서버 아키텍처로 구성된다. 사용자는 시각화된 유저인터페이스인 Bioworks 클라이언트 프로그램을 통하여 자신이 원하는 생명정보 분석 시나리오를 하나의 워크플로우 형태로 손쉽게 작성할 수 있다. 작성된 워크플로우는 Process와 Link로 구성된 XML 형태로 Bioworks 서버 측으로 전달되고 워크플로우에 포함된 각 실행 명령들은 배치 큐 서버에 배치 실행 작업으로 등록된다. 등록된 워크플로우 배치 작업은 워크플로우 엔진을 통하여 슈퍼컴퓨팅 서버 상에서 빠르고 정확하게 실행되고, 단계별 결과물은 사용자 데이터베이스에 저장된다. 사용자는 클라이언트 프로그램을 통하여 실시간으로 워크플로우 실행 현황과 단계별 결과물을 확인하고 분석할 수 있다.

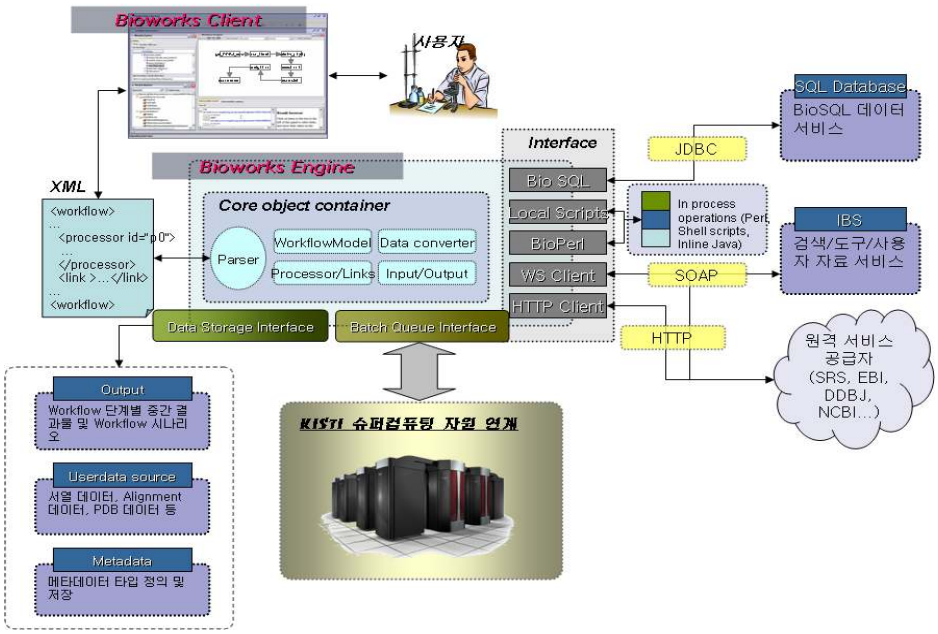


그림 3. 슈퍼컴퓨팅을 연계한 Bioworks 클라이언트-서버 아키텍처

2.3. Bioworks 클라이언트 프로그램

Bioworks 클라이언트 프로그램은 사용자가 손쉽게 생명정보 분석과정에 대한 워크플로우를 작성할 수 있도록 사용자 중심의 시각화된 인터페이스를 제공하고 있다. 특히 Java Web Start 기술을 도입하여 사용자가 별도의 프로그램 설치 및 버전 업그레이드 과정 없이 웹을 통해 자동으로 프로그램을 설치하여 사용할 수 있도록 구현되었다. 또한, Bioworks 클라이언트 프로그램은 Internalization(I18N) 지원을 통해 다양한 언어 환경의 사용자 맞춤형 사용자 인터페이스를 제공한다. Bioworks 클라이언트 프로그램은 <그림 4>에서 보여 지는 바와 같이 작업공간 탐색창, 사용자 자료 편집 창, 메시지 콘솔 창, 결과물 분석 창 등으로 구성된다.

2.4. Bioworks 서버 측 워크플로우 실행엔진

<그림 5>은 하나의 생명정보 워크플로우의 실행 개념도이다. 하나의 워크플로우는 Process를 정점으로 하고 Process간 Link를 간선으로 하는 '방향성 그래프(Directed Graph)'로서 표현될 수 있다. 하나의 초기 실행 Process는 최상위 정점에 해당하는 Process가 되며 각각의 Process는 Link에 의해 연결되어 부모 Process와 자식 Process를 갖게 된다. 하나의 Process는 상위 Process들이 모두 종료되어 그것의 출력 값이 자식 Process의 입력 값으로 적합하게 설정되었을 때 병렬적으로 실행된다. 각각의 Process 실행 시에 사용자가 지정한 단위 분석 프로그램들이 호출되며 배치 큐 인터페이스를 통하여 슈퍼컴퓨팅 기반에서 실행된다. Process 간의 Link에 의한 데이터 전달이 일어나는 시점에서 특정 전이 조건에 대한 검사와 사용자가 정의한 스크립트 코드에 따른 데이터 변환이 이루어 질 수 있다.

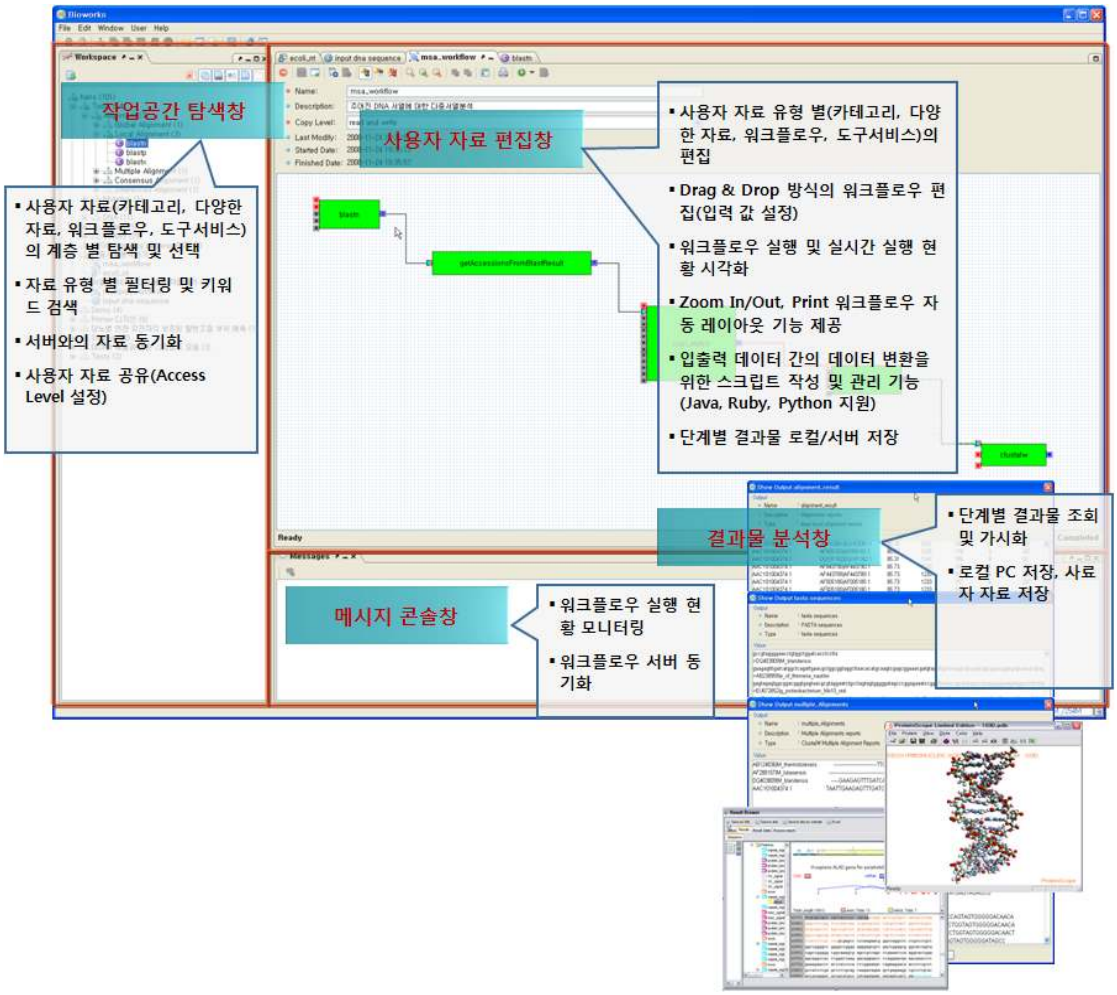


그림 4. Bioworks 클라이언트 UI 컴포넌트 구성

2.5. Bioworks 시스템의 주요 기능 및 특징

Bioworks 시스템은 <그림 6>에서 보여 지는 바와 같이 크게 실행서비스 관리, 사용자 자료 관리 및 공유, 워크플로우 편집/실행 관리, 결과물 저장 및 조회 등의 기능으로 구성된다. Bioworks 시스템의 주요 특징 및 장점을 살펴보면 다음과 같다.

- XML 기반 도구 Configurations: 다양한 생명정보 도구 서비스를 XML 기반으로 설정할 수 있도록 하여 손쉽게 신규 도구 서비스의 추가 가능
- 동적 스크립트 처리 및 보안 실행: 동적 스크립트에 대한 보안 실행 기능을 지원함으로써, 사용자 정의 도구 서비스 및 서비스 노드 간 링크에서의 입출력 데이터 변환 스크립트(Java, Python, Ruby 등)를 직접 입력하여 적용 가능
- 온톨로지 기반 생명정보 타입: 다양하고 이질적인 생명정보 데이터 타입에 대한 온톨로지 기반 계층적 스키마(OWL 정의)를 통하여 효과적인 생명정보 메타 데이터 통합

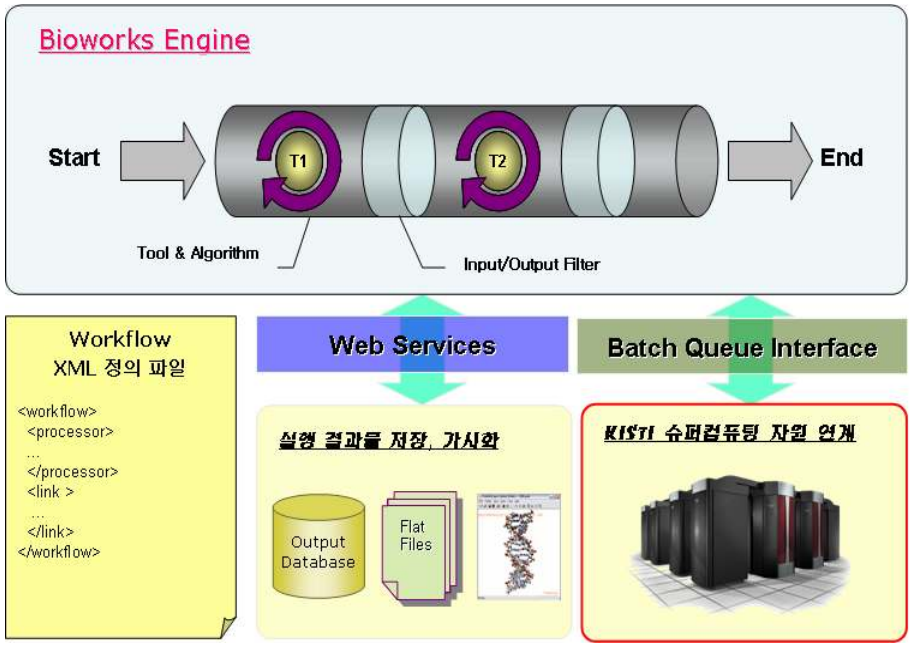


그림 5. Bioworks 서버 엔진 실행 개념도

그림 6. Bioworks 시스템 주요 기능 구성도

- 사용자 자료 공유: Access Level 적용을 통한 사용자 자료(도구 서비스, 생명정보 데이터, 워크플로우)의 실시간 공유를 지원하여 사용자 간의 워크플로우 기반 최적의 협력 연구 환경 제공
- 웹 기반 원클릭 설치 및 업그레이드: Java WebStart 기술을 도입하여 별도의 프로그램 다운로드 및 설치 없이, Bioworks 웹 포털 사이트 상에서 원클릭 자

동 프로그램 설치 및 업그레이드 가능

- Java GUI 기반 사용자 인터페이스: 계층적 사용자 자료 관리를 위한 Workspace Explorer, Drag & Drop 방식의 Visual Workflow Designer, 워크플로우 실행 현황 및 결과물 관리를 위한 Result Browser 등 직관적이고 사용이 편리한 사용자 인터페이스 제공

3. Bioworks 시스템 사용법

3.1. Java Web Start를 이용한 Bioworks 클라이언트 설치

앞서 말한 바와 같이 Bioworks 시스템은 Java WebStart 기술을 도입하여 웹페이지를 통한 원클릭 설치 및 업그레이드가 가능하다. Java Web Start는 Java 기술 기반 응용프로그램을 위한 새로운 배포 기술로 사용자가 웹에서 바로 응용프로그램을 시작하고 관리할 수 있도록 컴퓨터와 인터넷 사이를 이어주는 다리 역할을 하며 웹페이지에서의 클릭 한 번으로 응용프로그램을 쉽게 작동하고 복잡한 설치나 업그레이드 절차 없이도 항상 최신 버전을 실행할 수 있도록 해준다.

Bioworks 클라이언트 프로그램을 설치하기 위해서는 Bioworks 베타 서비스 포털 사이트 (http://bioworks.kisti.re.kr)에 접속하여 메인페이지에서 Bioworks 클라이언트 실행 링크를 클릭하여 설치한다. 이때 주의할 점은 Bioworks 클라이언트 프로그램이 JAVA 프로그램으로 개발되었으므로 사용자 PC에 Java 프로그램을 구동하기 위한 JRE(Java Runtime Environment)가 설치되어 있어야 한다. JRE의 설치방법은 Java Sun 사이트(http://java.sun.com)를 참조한다.

3.2. Bioworks 사용자 계정 등록

Java WebStart를 통하여 Bioworks 클라이언트 프로그램의 설치가 완료되면 초기 실행화면(로그인 화면)이 나타난다. Bioworks 시스템을 사용하기 위해서는 먼저 사용자 계정을 등록해야 한다. 사용자 등록을 통해서 사용자 자료 및 워크플로우 관리를 위한 서버 측 리소스가 할당되고 기본 사용자 작업 환경이 설정된다. 사용자 계정 신청 작업은 다음의 순서대로 진행한다.

- “User>Register…” 메뉴를 선택하여 사용자 등록 창 팝업(<그림 7>)
- 사용자 등록 창에서 사용자 정보 입력하고 “Save” 버튼을 누르면 등록 완료
- “Create Default Services”를 선택하면 Bioworks에서 제공되는 기본 생명정보 도구(현재 208개)와 기본 사용자 작업 공간에 자동 복사되어 사용 가능

3.3. Bioworks 시스템 로그인 및 사용자 작업 환경

Bioworks 클라이언트 프로그램을 실행하면 사용자 로그인 화면이 나타난다. 로그인 화면에서 계정과 패스워드를 입력하면 <그림 8>과 같은 사용자 작업 환경이 보여 진다. 사용자 작업 환경은 카테고리, 생명정보 자료, 분석도구, 워크플로우

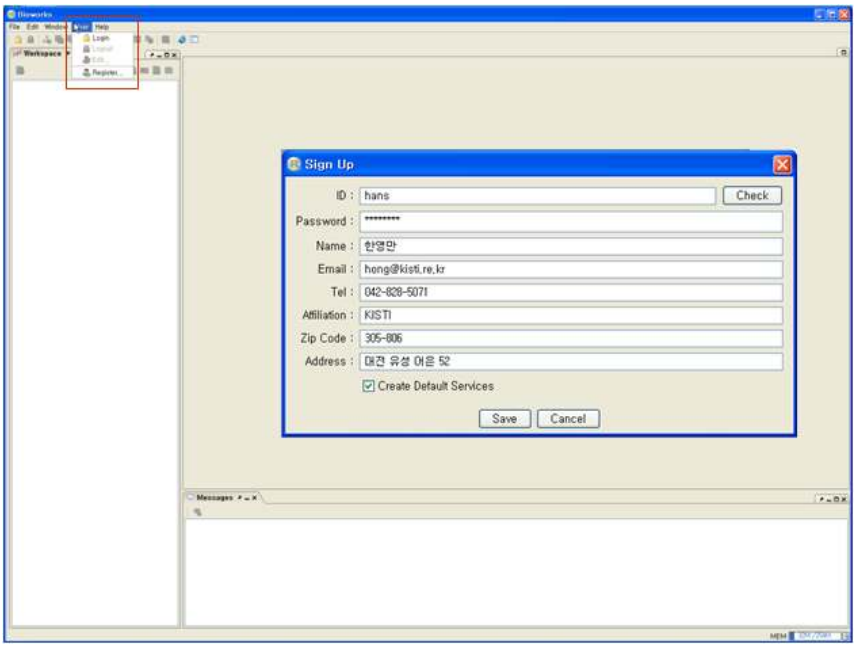


그림 7. 사용자 계정 등록 화면

등을 포함하는 사용자 자료의 트리 구조 - 사용자 아이디를 최상위 루트로 하는 - 계층적 탐색을 위한 작업 공간 탐색 창, 사용자 자료 편집 및 워크플로우 실행과 실시간 실행현황 가시화를 위한 사용자 자료 편집창, 그리고 워크플로우 실행 현황 모니터링 및 각종 로그를 표시하는 메지지 콘솔창으로 구성된다.

3.4. 사용자 카테고리 등록 및 수정

신규 카테고리를 등록하기 위해서는 사용자 작업공간 탐색창에서 상위 카테고리를 선택 한 후 오른쪽 마우스 버튼을 클릭, 팝업메뉴에서 “New>Category"를 선택하면 <그림 9>의 신규카테고리 등록 창을 팝업된다. 신규 카테고리 등록 창에서 카테고리 이름을 입력한 후 “Save" 버튼을 클릭하면 입력한 이름의 카테고리가 선택 상위 카테고리 하위에 생성된다.

3.5. 스크립트 도구 등록 및 수정

신규 스크립트 도구를 등록하고 수정하기 위한 작업 절차는 아래와 같다.

- 사용자 작업공간 탐색창에서 상위 카테고리를 선택한 후 오른쪽 마우스 버튼을 클릭, 팝업메뉴에서 “New>Service>ScriptService" 항목을 선택하면 <그림 10>의 스크립트 도구 등록 초기 화면이 팝업된다.

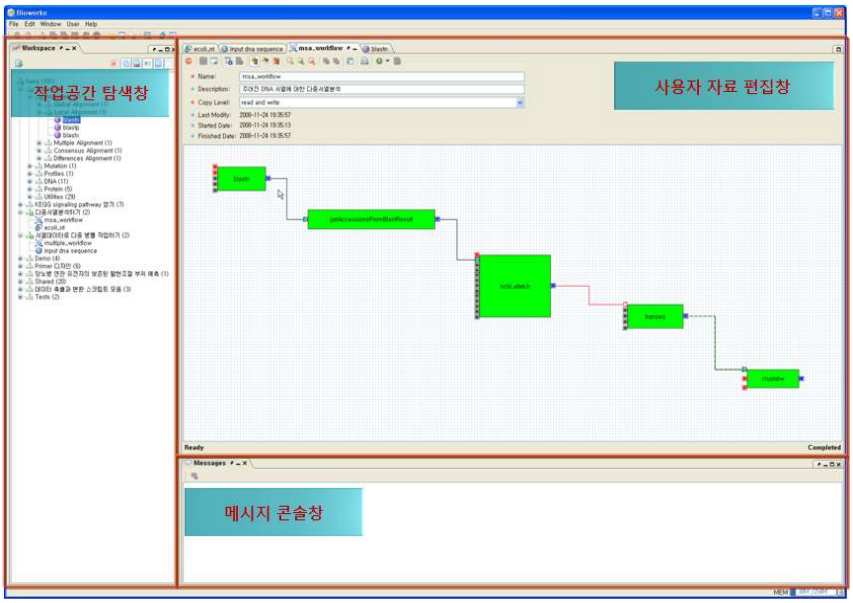


그림 8. 사용자 작업 환경

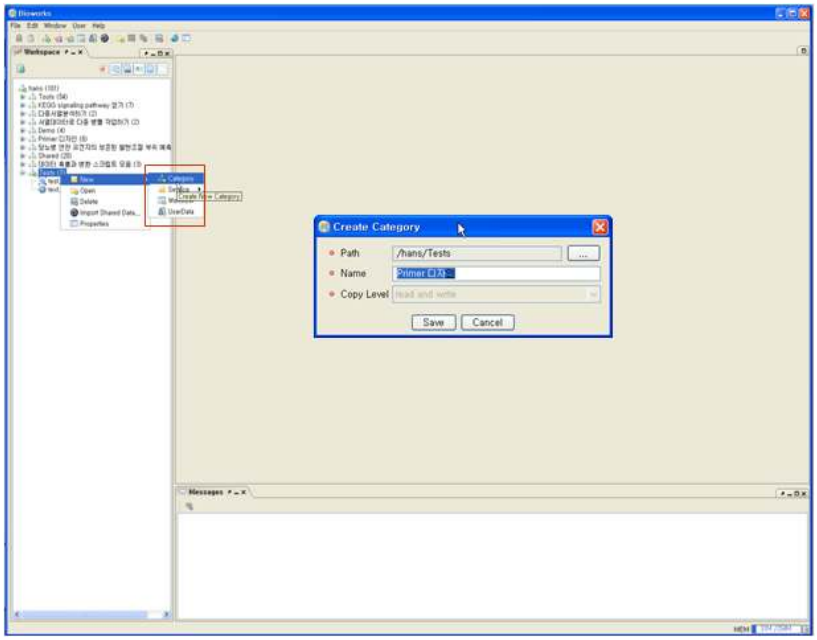


그림 9 사용자 카테고리 등록/수정 화면

- 스크립트 도구 등록 초기 화면에서 도구 이름과 공유레벨(Copy Level)을 설정한 후 “Save" 버튼을 클릭하면 신규 스크립트가 생성되고 사용자 자료 편집창에 스크립트 도구 편집 화면이 탭으로 표시된다.

- 스크립트 도구 편집 탭의 Input 테이블 영역에서 오른쪽 마우스 버튼을 클릭, 팝업메뉴에서 “Add..." 메뉴를 선택하면 입력 설정 팝업창에서 각 필드 값과 데이터 타입을 설정 한 후 ”Apply" 버튼을 클릭하면 해당 스크립트 도구의 신규 Input 설정이 추가된다<그림 11>.
- 해당 스크립트 도구에 대한 Output 설정도 스크립트 도구 편집 탭의 Output 테이블 영역을 클릭한 후 Input 설정과 같은 방법으로 추가한다.

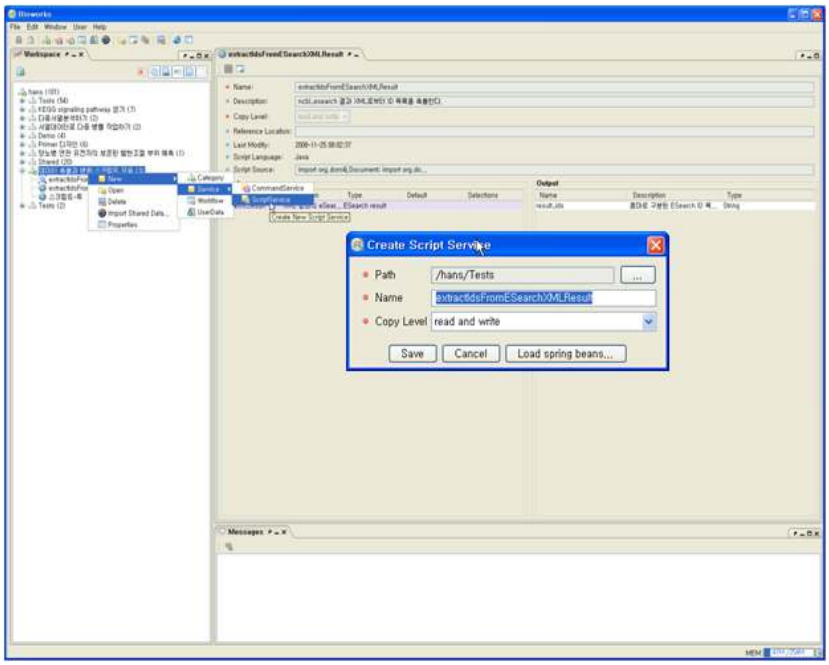


그림 10. 스크립트 도구 초기 설정 화면

- 스크립트 도구 편집 탭의 Script Source 입력 필드를 더블 클릭하면 스크립트 소스 설정 창이 <그림 12>과 같이 팝업된다.
- 스크립트 소스 설정창에서 Language(Java, Python, Ruby)를 선택하고 스크립트 소스를 편집한다. 이때 입력 스크립트가 정상적으로 동작하는 지 확인하기 위해서는 “Test" 버튼을 눌러 테스트 실행 창을 팝업하고 여기에서 입력값을 설정하여 테스트를 실행한다.

3.6. 생명정보 자료 등록 및 수정

다양한 양식의 텍스트 기반의 생명정보 데이터를 등록하기 위한 작업 절차는 아래와 같다.

- 먼저 사용자 작업 공간 탐색창에서 특정 상위 카테고리 선택한 후 오른쪽 마우스 버튼을 클릭, 팝업 메뉴에서 “New>UserData" 메뉴를 선택하여 <그림 13>와 같은 신규 생명정보 자료 생성 창을 팝업한다.

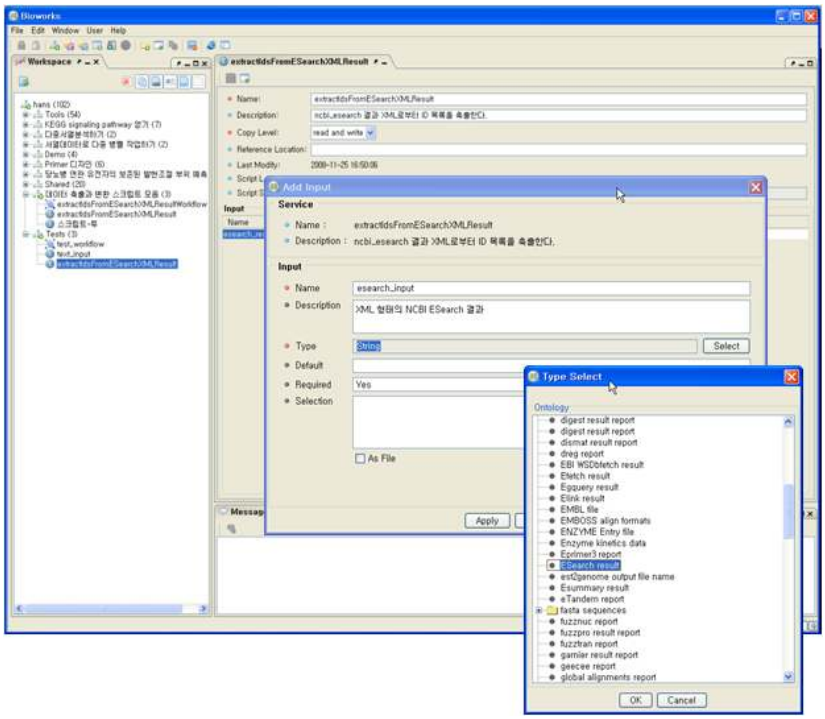


그림 11. 스크립트 도구 입출력 설정 화면

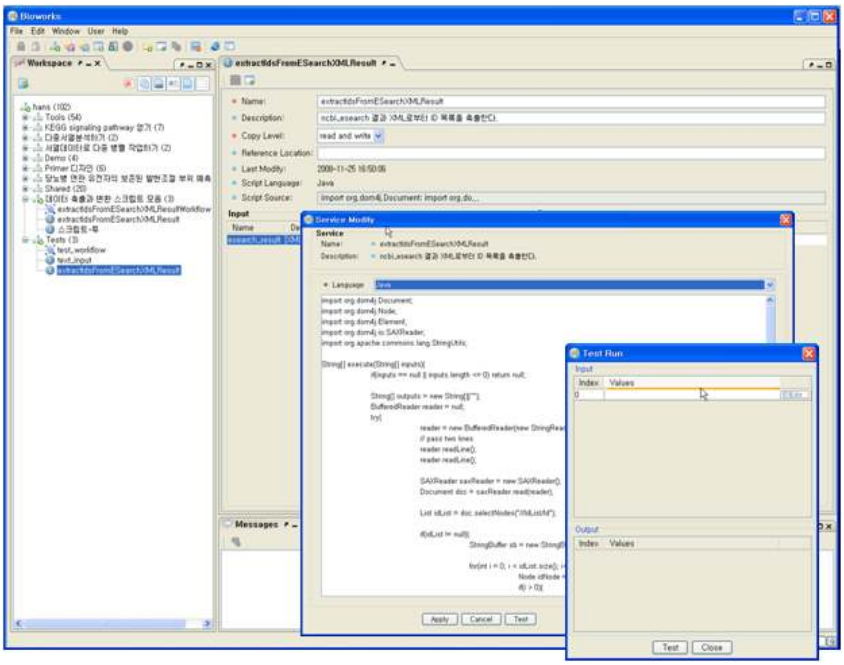


그림 12. 스크립트 소스 설정 화면

- 신규 생명정보 자료 생성창에서 이름과 공유레벨을 설정한 후 “Save" 버튼을 클릭하면 입력 생명정보 자료가 서버에 저장되고 <그림 14>와 같은 해당 자료 편집 탭 창이 나타난다.
- 신규 자료 편집 탭에서 각 입력 필드 값을 입력하고 데이터 유형을 설정한다. 이 때 “From URL", "From File" 버튼을 클릭하여 URL 또는 로컬 파일을 불

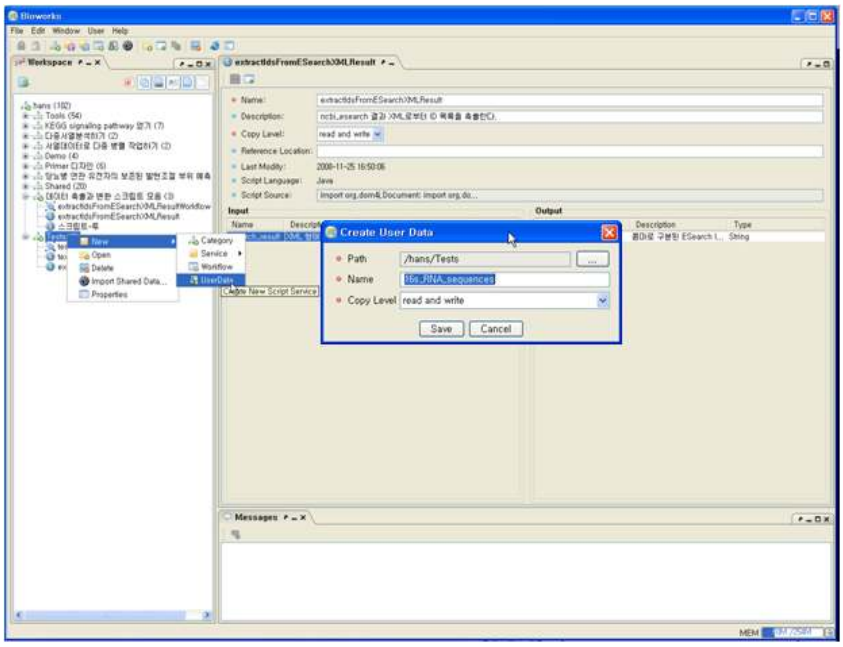


그림 13. 신규 생명정보 자료 생성 화면

러와 Source 값을 설정할 수 있다.

3.7. 워크플로우 생성 및 편집

신규 워크플로우를 생성하고 편집하기 위한 작업 절차는 아래와 같다.

- 먼저 사용자 작업공간 탐색 트리에서 특정 카테고리를 선택한 후 오른쪽 마우스 버튼을 클릭, 팝업 메뉴에서 “New>Workflow" 메뉴를 선택하여 <그림 15>과 같은 신규 워크플로우 생성 창을 팝업한다.
- 워크플로우 생성 창에서 이름과 공유레벨을 설정한 후 “Save" 버튼을 클릭하면 입력 워크플로우가 서버에 저장되며 해당 워크플로우 편집 탭 창이 <그림 16>과 같이 보여진다.
- 워크플로우 편집 창에서 각 입력 필드를 입력하고 편집 캔버스 위에 작업 공간 트리에서 선택한 도구 서비스를 Drag & Drop 하여 워크플로우 편집을 수행하고 저장 버튼을 눌러 현재 편집 중이 워크플로우를 저장한다.

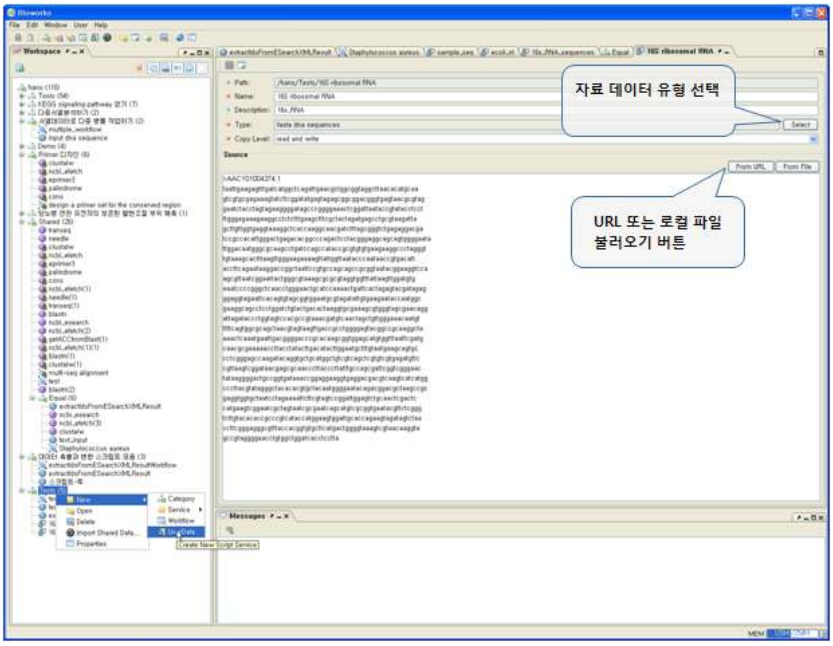


그림 14. 생명정보 자료 편집 화면

3.8. Zoom In/ Out, 자동 레이아웃, 프린트

<그림 17>에서 보여 지는 바와 같이 워크플로우 편집 탭 상단의 “Zoom In/Out", ”Layout", 그리고 ”Print" 툴바 버튼을 눌러 편집 중인 워크플로우 캔버스의 뷰 스케일 조정, 워크플로우 노드들의 최적 레이아웃 자동 설정, 워크플로우 캔버스 프린트 등의 작업을 각각 수행할 수 있다.

3.9. 워크플로우 노드 입력값 설정

워크플로우 편집 캔버스에서 대상 노드를 더블 클릭하면 <그림 18>와 같이 선택 실행 노드의 입력 값 설정 화면이 팝업된다. 노드 입력 값 설정 창에서 특정 입력 필드를 선택한 후 “Edit" 버튼을 클릭하면 해당 입력 필드에 대한 값 설정 창이 되며, 여기에 직접 입력하거나 로컬 파일, URL, 사용자 자료를 불러와 값을 설정할 수 있다.

3.10. 입출력 데이터 변환 링크 스크립트 작성

두개의 워크플로우 도구 실행 노드 간의 입출력 링크를 만들 때 간혹 입출력 데이터 포맷이 달라 링크가 연결 되지 않을 경우가 발생한다. 이럴 경우 입출력 데이터 변환 링크 스크립트를 작성하여 워크플로우 실행 시 동적으로 데이터 변환이 수행되도록 할 수 있다. 입출력 데이터 변환 링크 스크립트 작성을 위한 작업 순서는 아래와 같다.

- 입출력 데이터 변환을 수행할 대상 링크를 더블 클릭하여 입출력 간 데이터

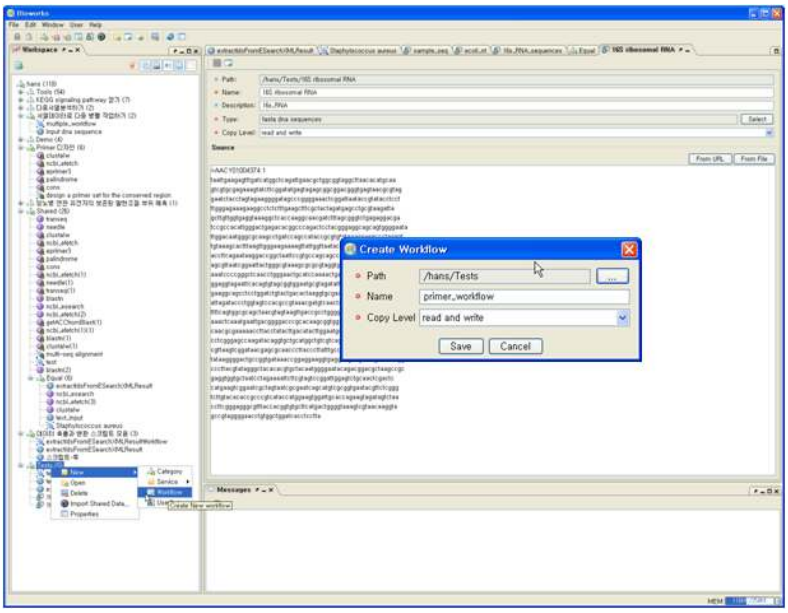


그림 15. 신규 워크플로우 생성 화면

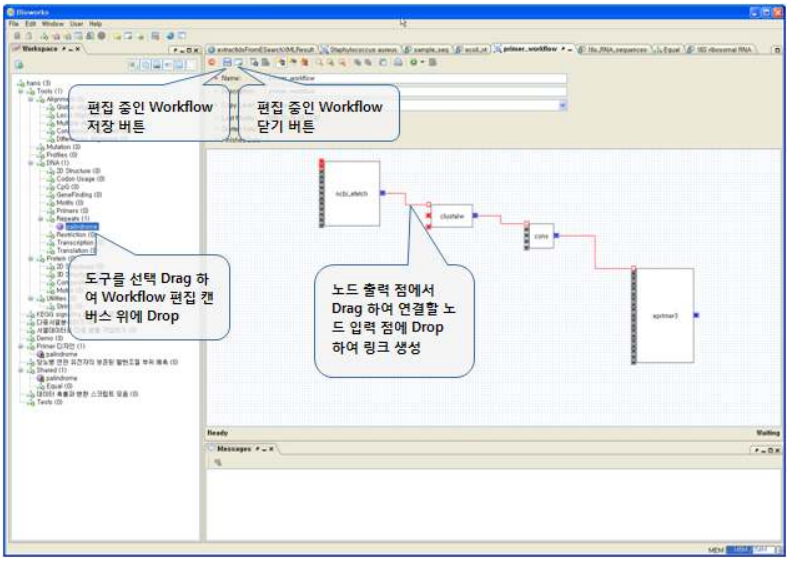


그림 16 워크플로우 편집 화면

변환 스크립트 작성 창(<그림 19>)을 팝업한다.

- 데이터 변환 스크립트 작성 창에서 Language(Java, Python, Ruby)를 선택하고 입력 스크립트를 입력한다.
- 작성 중인 스크립트 소스 코드를 테스트 해보기 위해서는 “Test" 버튼을 클릭

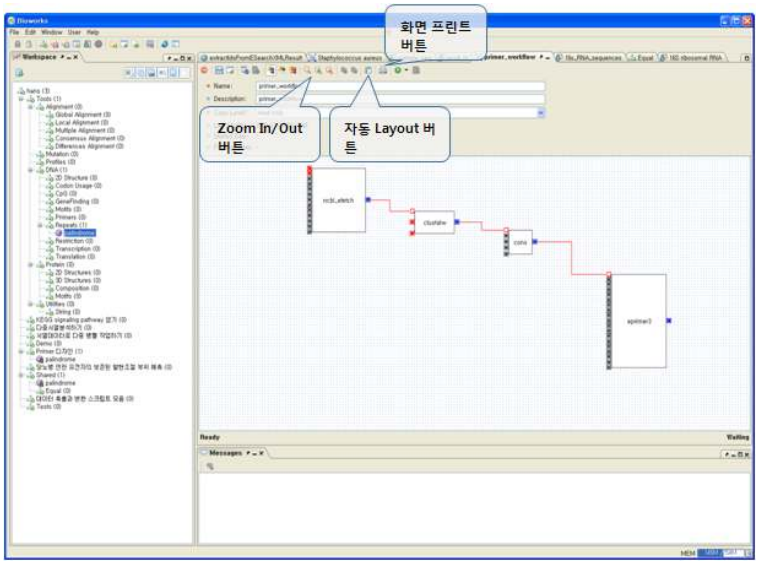


그림 17 Zoom In/Out, 자동 Layout, Print 기능 예시

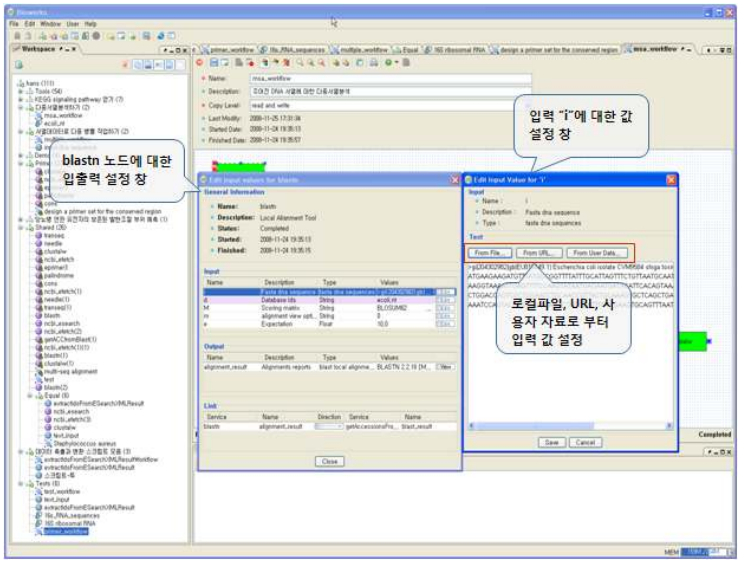


그림 18. 워크플로우 실행 노드 입력값 설정 화면

하여 테스트 실행 창을 팝업하고 여기에서 임시 입력 값을 설정하여 테스트를 수행한다.

- 스크립트 소스 작성이 완료되면 “Save" 버튼을 클릭하여 작업을 종료한다.

3.11. 워크플로우 실행 및 현황 모니터링

작성된 워크플로우를 실행하기 위해서는 워크플로우 편집 탭에서 “Run" 툴바 버

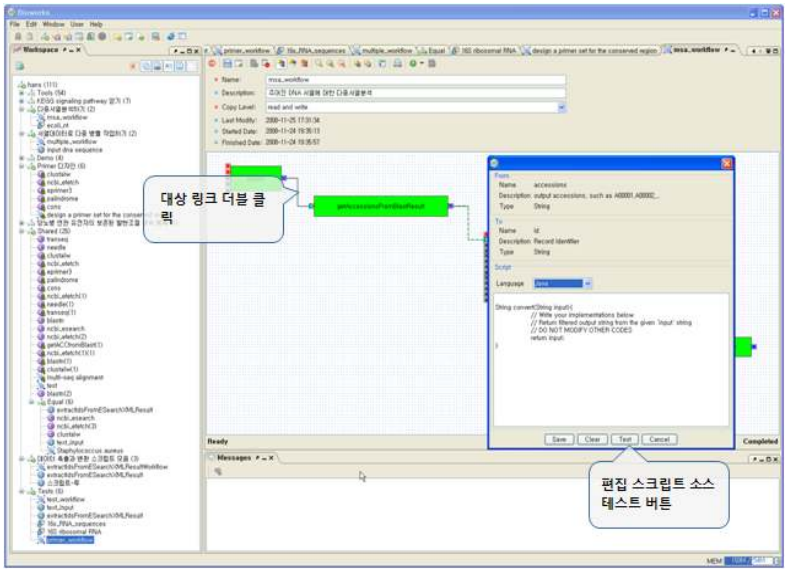


그림 19. 입출력 데이터 변환 링크 스크립트 작성 화면

튼을 클릭한다. 해당 워크플로우의 실행 현황은 <그림 20>에서 보여 지는 바와 같이 각 워크플로우 실행 노드의 배경색으로 표시된다(대기: 흰색, 실행중: 오렌지색, 완료: 초록색, 실패: 빨간색). 각 실행 노드를 더블 클릭하면 해당 노드의 실행 상태, 시작/종료 일시를 알 수 있다.

3.12. 워크플로우 실행 결과물 저장

실행이 완료된 워크플로우의 각 실행 노드의 출력 포트를 더블 클릭하면 <그림 21>와 같은 출력 값 조회창이 팝업된다. 여기에서 출력 값을 확인하고 “Save As File..." 또는 "Save As UserData..."를 클릭하여 로컬 PC 또는 사용자 자료로 출력 값을 저장할 수 있다.

3.13. 워크플로우 실행 시 결과물 자동 저장

사용자는 실행 완료된 워크플로우의 각 중간 결과물을 일일이 저장할 필요 없이 결과물 자동 저장 기능을 이용하여 일괄적으로 특정 사용자 카테고리에 자동 저장할 수 있다. 작업 순서는 아래와 같다.

- 워크플로우 편집 탭에서 “Run Configuration..." 툴바 버튼을 클릭하여 <그림 22>과 같이 결과 값 저장 선택창을 팝업한다.
- 결과 값 저장 선택창에서 결과 값을 자동 저장할 사용자 카테고리를 선택한다.
- 결과 값 저장 선택창에서 “Run" 버튼을 눌러 워크플로우를 실행한다(실행 완료 후 선택 결과 값은 선택 카테고리에 자동 저장 됨).

3.14. 사용자 자료 공유 설정

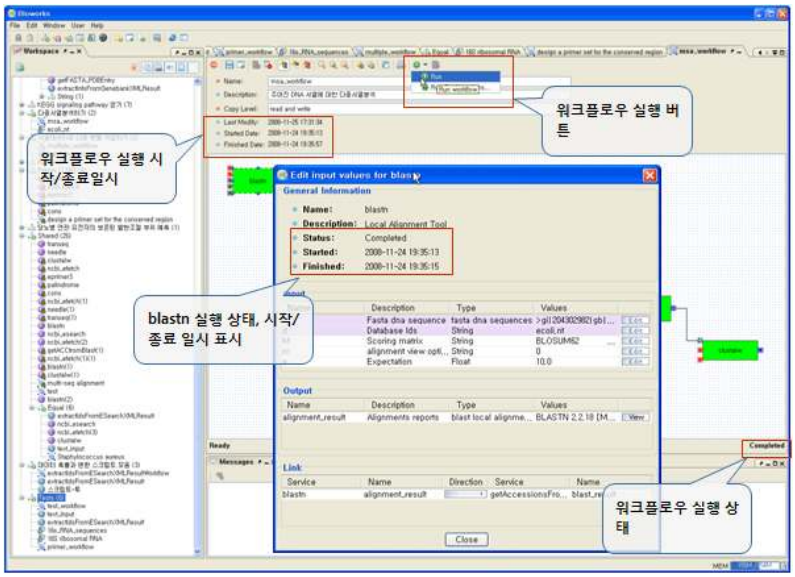


그림 20. 워크플로우 실행 및 현황 모니터링

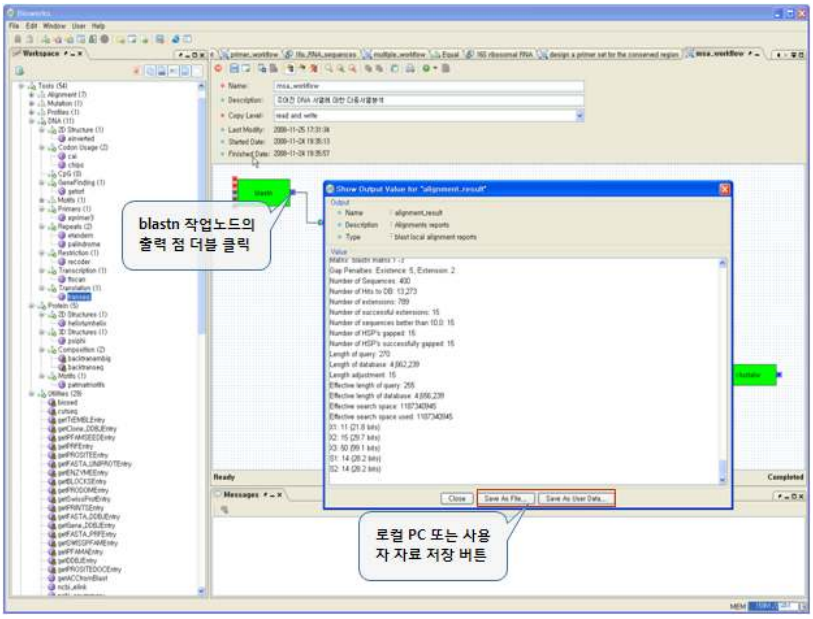


그림 21. 워크플로우 실행 결과물 저장 화면

Bioworks 시스템은 클라이언트-서버 환경으로 각 사용자 작업 환경과 자료는 서버에 저장된다. 이러한 서버 환경을 기반으로 각 사용자는 <그림 23>에서 보이는 바와 같이 모든 사용자 자료(생명정보 데이터, 도구 서비스, 워크플로우)에 대한 Access Level 설정을 통하여 다른 사용자와의 공유가 가능하다. 특정 사용자 자료에 대한 공유 설정은 None, Read only, Read and Write로 할 수 있다.

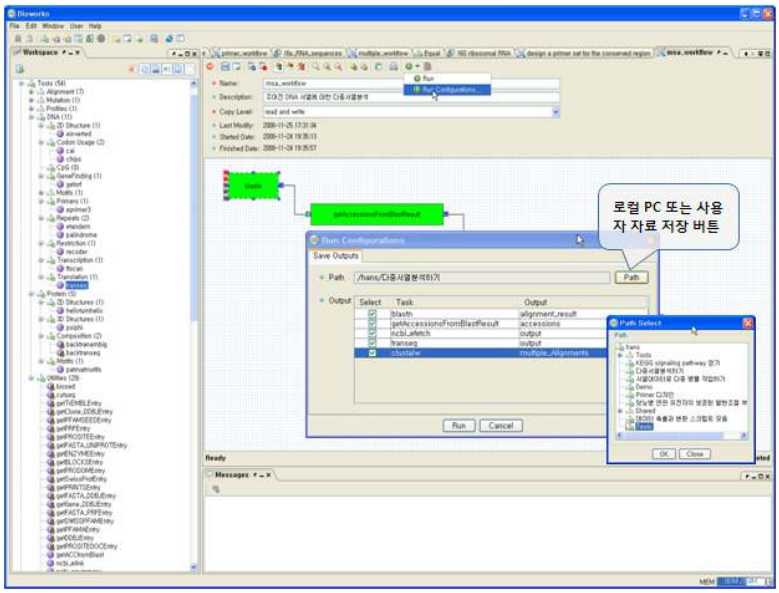


그림 22. 실행 결과 값 자동 저장 선택 화면

3.15. 공유 사용자 자료 복사

Bioworks 클라이언트 프로그램을 통하여 사용자는 다른 사용자들이 공유해 놓은 사용자 자료를 조회하고 특정 공유 자료를 복사하여 자신의 자료로 활용할 수 있다. 공유 사용자 자료 복사 기능을 이용하기 위한 작업 순서는 아래와 같다.

- 특정 카테고리 선택 후 오른 쪽 마우스 버튼 클릭, 팝업 메뉴에서 “Import Shared Data..." 메뉴를 선택하여 <그림 24>와 같이 공유 자료 검색창을 팝업한다.
- 공유 자료 검색창에서 사용자 자료 이름, 사용자 이름, 키워드, 그리고 자료유형 등의 검색 필터를 입력하고 “Search" 버튼을 눌러 조건에 맞는 공유 사용자 자료를 검색한다.
- 검색 결과 목록에서 복사할 사용자 자료를 선택하고 “OK" 버튼을 클릭하여 앞서 선택한 상위카테고리로 복사한다.

4. Bioworks 시스템을 활용한 생명정보 분석 시나리오 구현

4.1. Design a primer set for the conserved region

NCBI_efetch를 이용하여 "XM_001075341, XM_001136125, XM_001080119"에 해당하는 염기 서열들을 가져온 후 clustalw를 이용하여 다중서열을 실행하고 그 결과로부터 cons를 이용하여 공통서열을 추출한다. 그 공통서열로부터 eprimer3를 이용

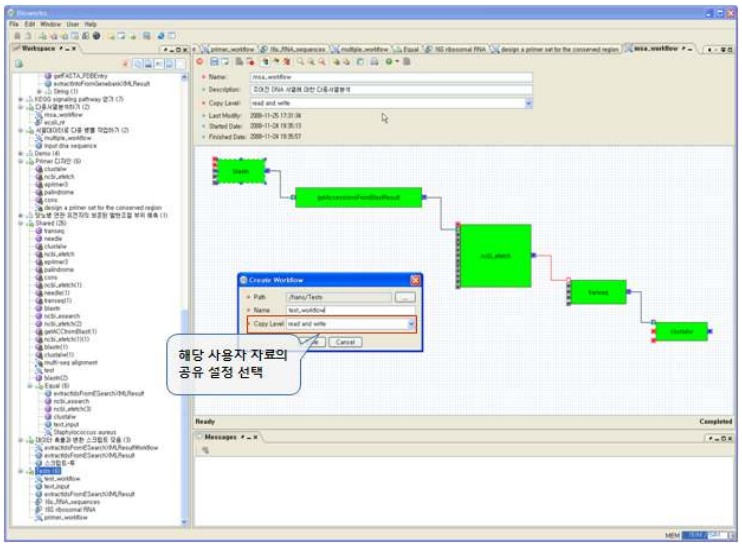


그림 23. 사용자 자료 공유 설정 화면

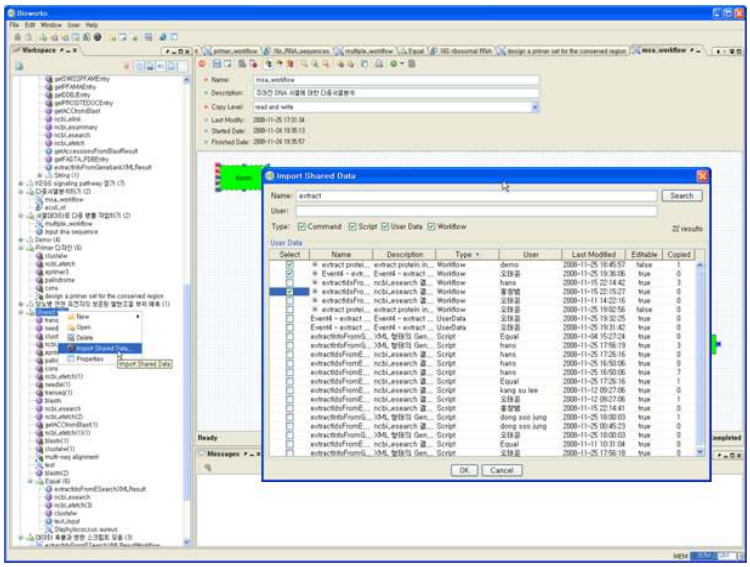
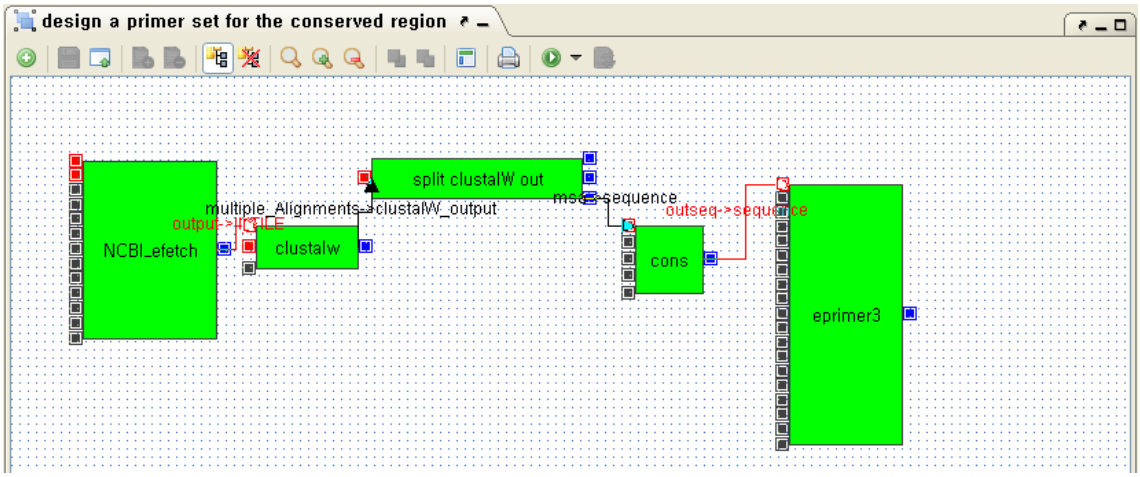


그림 24. 공유 사용자 자료 검색 및 복사 화면

하여 primer를 디자인한다.



1) NCBI_efetch : NCBI로부터 ID에 해당하는 데이터를 가져올 때 사용한다.

- db : sequences
- id : XM_001075341,XM_001136125,XM_001080119
- remax : 3
- retmode : text
- rettype : fasta

2) clustalw : 다중서열을 입력받아서 정렬한다.

- options : all defaults

3) split clustalW out : clustalw의 출력을 포맷에 맞게 나누어서 출력한다.

4) cons : cons는 다중서열정렬 결과를 입력받아서 공통 서열을 찾는다.

- options : all defaults

5) eprimer3 : 입력된 DNA 서열로 부터 PCR 반응에 필요한 primer를 추출한다.

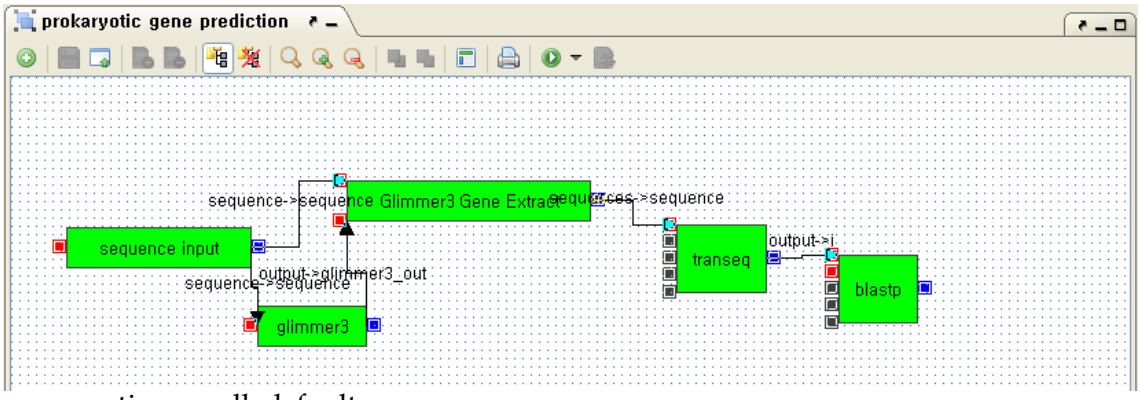
- options : all defaults

4.2. Prokaryotic gene prediction

원핵생물 유전체(genome) 서열로부터 유전자(gene)들의 위치를 예측하고, blastp를 이용하여 예측된 유전자들의 기능을 유추한다.

1) glimmer3[1] : 원핵생물 유전체 서열로부터 유전자들의 위치를 예측한다.

2) Glimmer3 Gene Extract : 유전체서열과 glimmer3의 결과로부터 유전자들의 서열을 추출한다.



- options : all defaults

3) transeq : DNA 서열을 단백질 서열로 번역한다.

- options : all defaults

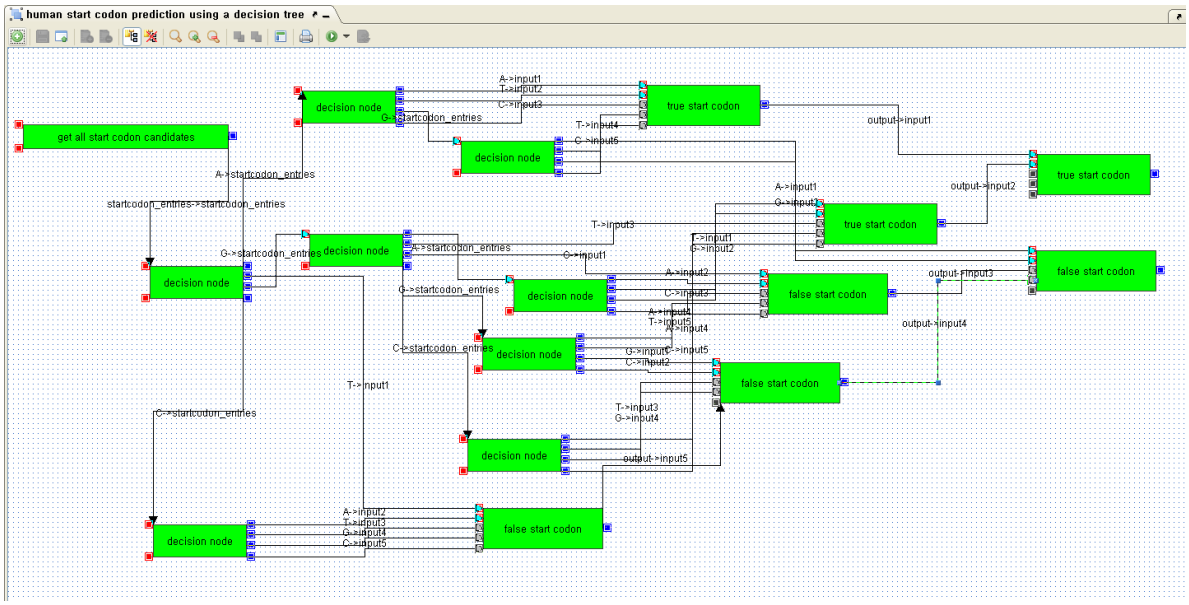
4) blastp[2] : 서버에 미리 구축되어있는 단백질서열 데이베이스로 부터 입력된 단백질서열과 유사한 서열 부분을 찾아낸다.

- d : nr

- e : 0.01

4.3. Human start codon prediction using a decision tree

human의 start codon로부터 decision tree 알고리즘을 이용하여 구축된 tree를 workflow에 적용시켰다



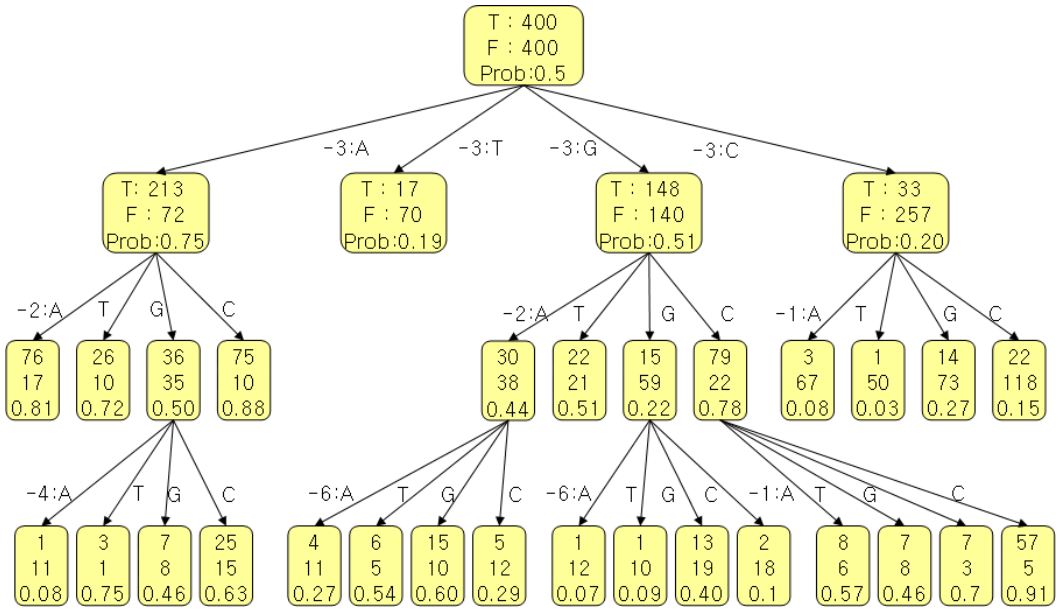
1) get all start codon candidates : 입력 유전체 서열로부터 ATG를 포함하는 부위를 모두 추출한다.

- lengthOfUpstream : 10 (ATG와 같이 추출될 ATG 이전서열의 길이)

2) decision node : 해당 위치의 염기 구성에 따라 네 그룹(A, T, G, C)으로 나누어 출력한다.

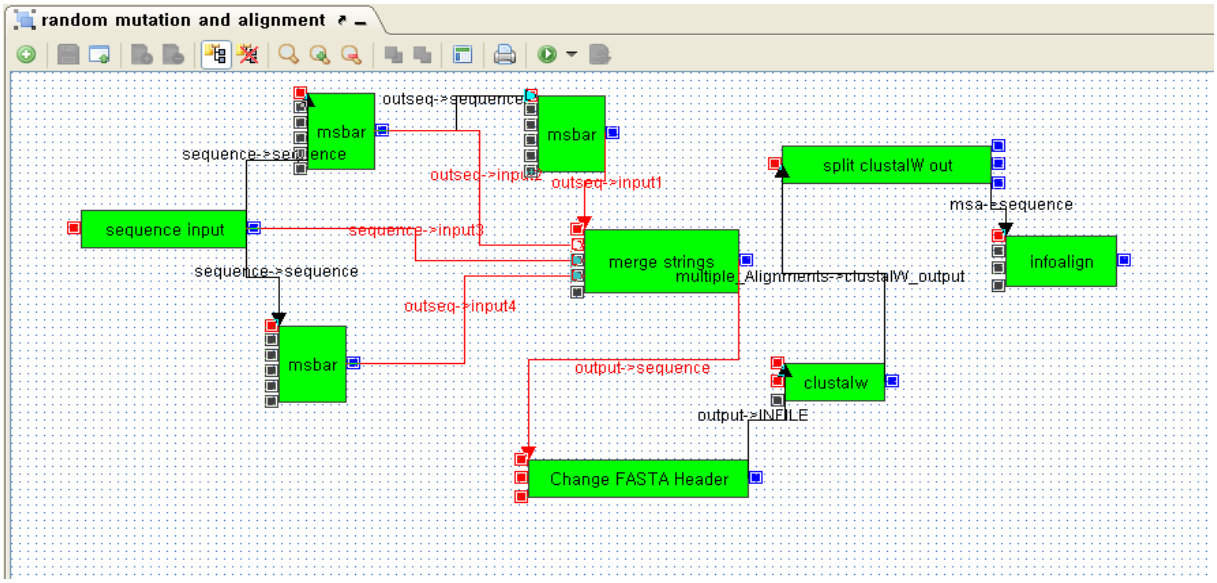
3) true start codon : decision tree의 결과에 따라 true start codon으로 판단된 단편서열들

4) false start codon : decision tree의 결과에 따라 true start codon이 아닌 것으로 판단된 단편서열들



4.4. Random mutation and alignment

DNA 서열을 무작위로 돌연변이 시킨 후 원본 서열과 다중서열 정렬을 한다.



1) msbar : 주어진 염기서열이나 단백질서열의 임의의 위치에 돌연변이를 가한다.

- count : 10
- point : 1
- block :1
- codon : 1

2) merge strings : 여러 입력 문자열을 하나의 문자열로 합친다.

3) Change FASTA Header : 서열의 FASTA header를 바꾼다. 여러 FASTA 서열이 같은 header을 가지고 있을 때 사용된다.

4) clustalw[3] : 다중서열을 입력받아서 정렬한다.

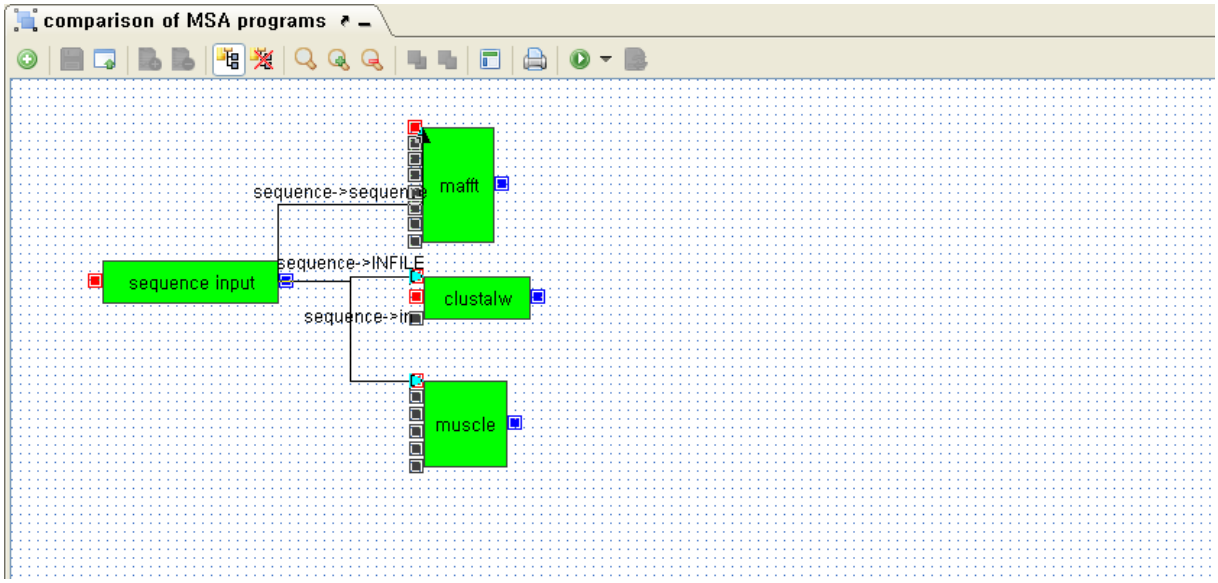
- options : all defaults

5) split clustalW out : clustalw의 출력을 포맷에 맞게 나누어서 출력한다.

6) inforalign : 다중서열정렬정보를 통계적으로 보여준다.

4.5. Comparison of MSA programs

mafft[4], clustalw, muscle[5] 세 개의 다중서열정렬 프로그램의 결과를 비교한다.

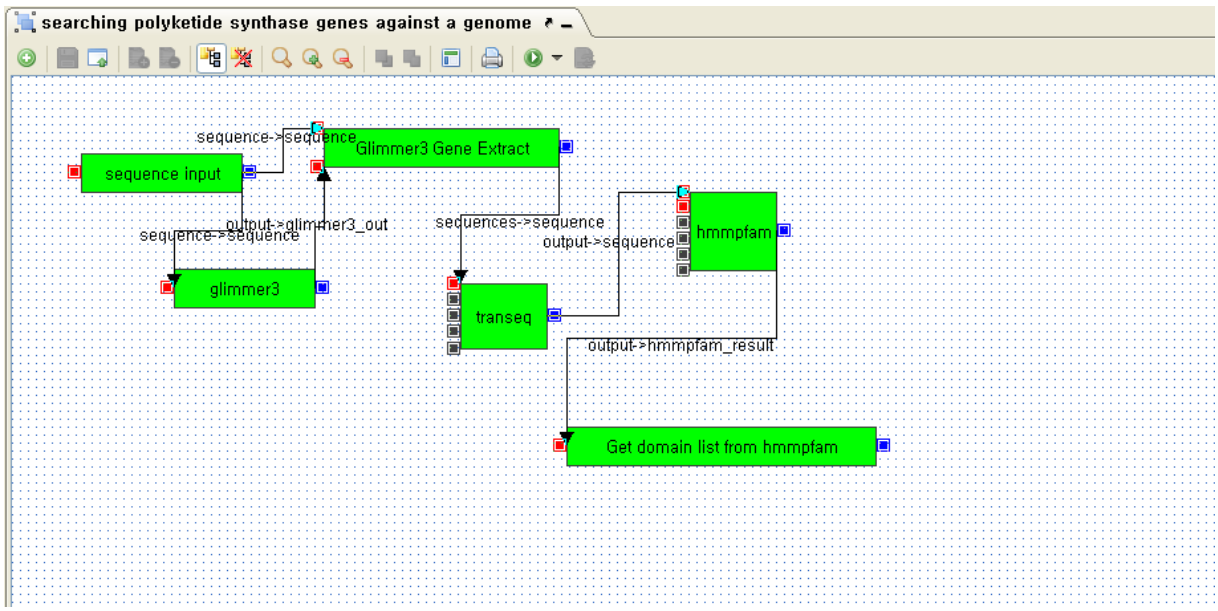


1) mafft, clustalw, muscle : 다중서열을 입력받아서 정렬한다.

- options : all defaults

4.6. Searching polyketide synthase genes against a genome

유전체 서열로부터 glimmer3를 이용하여 유전자를 예측한 후 그 유전자들로부터 polyketide의 도메인을 HMMER[6] 프로그램 중 하나인 hmmpfam을 이용하여 검색한다.



1) sequence input : Acaryochloris marina MBIC11017 plasmid pREB1

2) glimmer3 : 원핵생물 유전체 서열로부터 유전자들의 위치를 예측한다.

3) Glimmer3 Gene Extract : 유전체서열과 glimmer3의 결과로부터 유전자들의 서열을 추출한다.

- options : all defaults

4) transeq : DNA 서열을 단백질 서열로 번역한다.

- options : all defaults

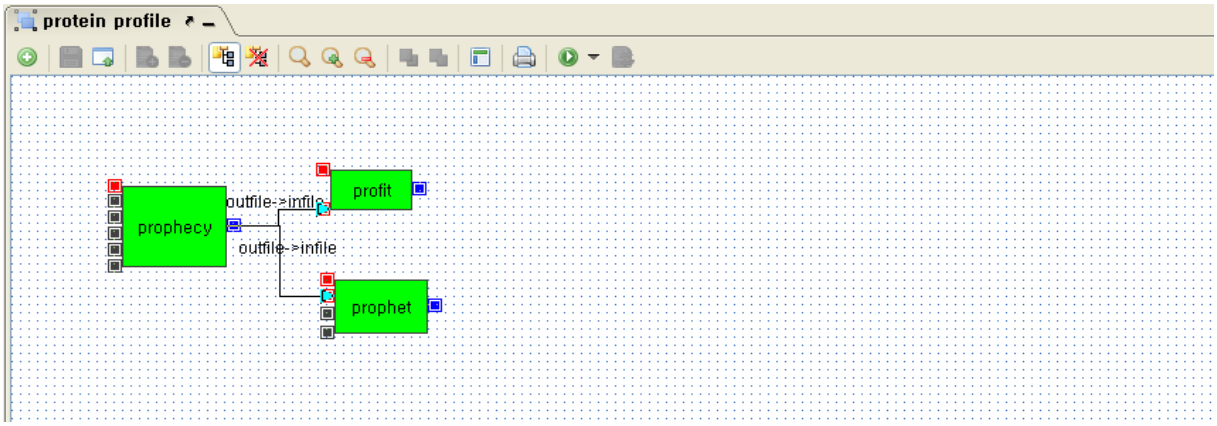
5) hmmpfam : 여러 서열로부터 hmmbuild에 의해 만들어진 HMM profile을 이용하여 도메인을 검색한다.

- E : 0.1

6) Get domain list from hmmpfam : hmmpfam 결과로부터 도메인 목록을 가져온다.

4.7. Protein profile

다중서열에 대한 프로파일을 만들고, 프로파일을 이용하여 입력서열로부터 프로파일에 해당하는 위치를 검색한다.



1) prophecy : 다중서열이 정렬된 서열로부터 프로파일을 만든다.

- options : all defaults

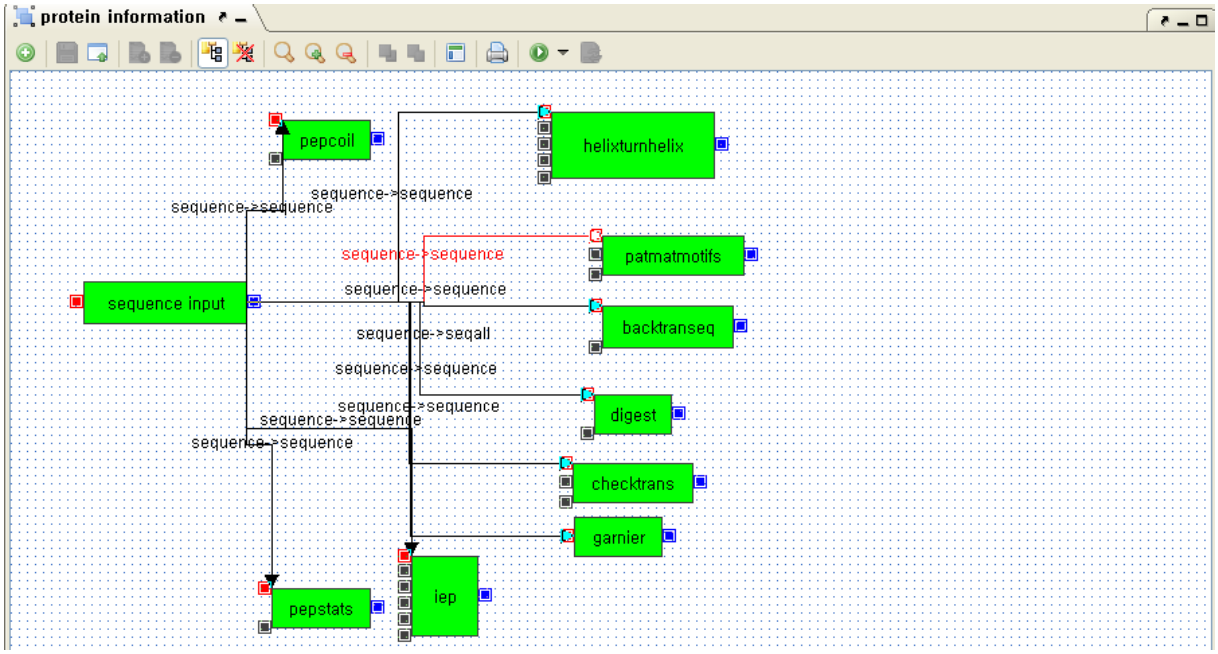
2) profit : 입력서열로부터 프로파일을 이용하여 해당위치를 찾아서 위치를 표시한다.

3) prophet : 입력서열로부터 프로파일을 이용하여 해당위치를 찾아서 상세 정보를 표시한다.

- options : all defaults

4.8. Protein information

단백질의 정보를 보여주는 도구들을 이용하여 단백질 서열의 모든 분석 정보를 얻는다.



1) pepcoil : 단백질 구조에서 코일 부분을 검색한다.

2) helixturnhelix : 단백질 서열로부터 DNA 단편이 바인딩 하는 부분을 검색한다.

3) patmatmotifs : 단백질 서열로부터 PROSITE의 motif를 검색한다.

4) backtranseq : 단백질 서열을 생성하는 DNA 서열을 예측한다.

5) digest : 단백질 분해 효소의 작용부분을 예측한다.

6) checktrans : STOP codon과 ORF에 대한 통계분석을 한다.

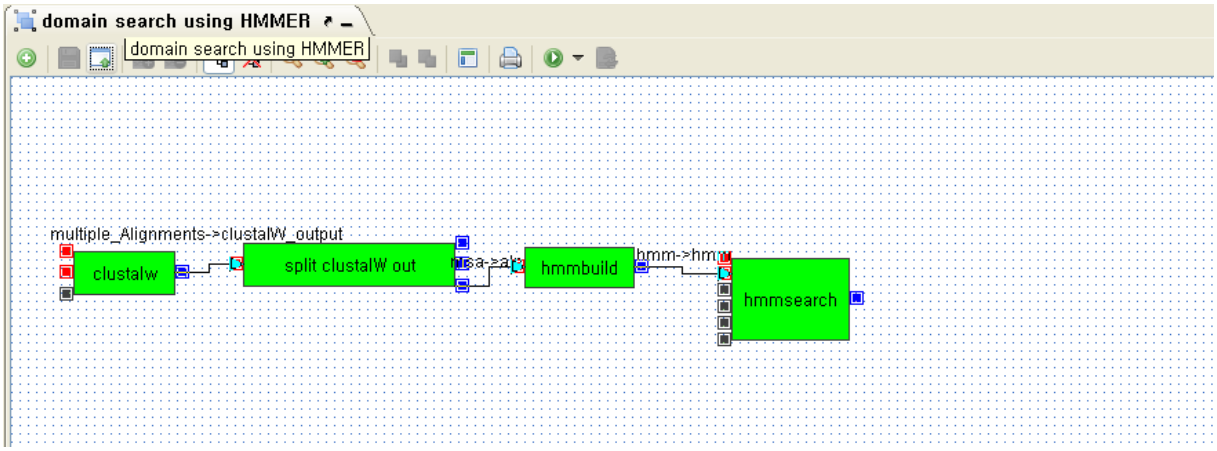
7) garnier : GOR 방법을 이용하여 단백질의 이차구조를 예측한다.

8) iep : 단백질의 isoelectric point를 계산한다.

9) pepstats : 단백질 서열에 대한 통계적 속성을 계산한다.

4.9. Domain search using HMMER

clustalW와 HMMER 패키지의 도구들을 이용하여 단백질 도매인을 검색한다.



1) clustalw : 다중서열을 입력받아서 정렬한다.

- options : all defaults

2) split clustalW out : clustalw의 출력을 포맷에 맞게 나누어서 출력한다.

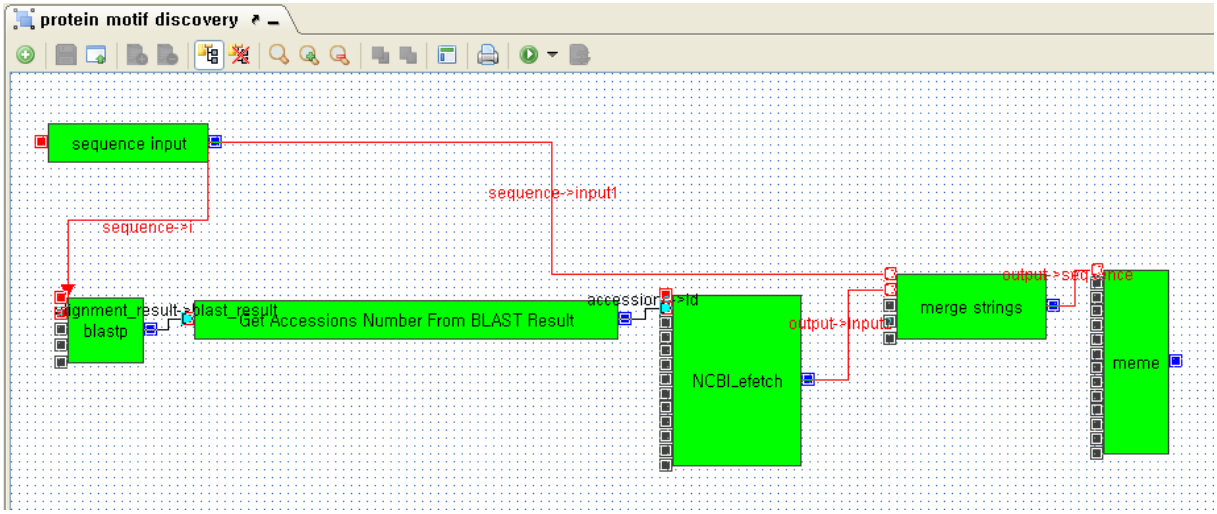
3) hmmbuild : 다중서열의 정렬 결과로부터 HMM 프로파일을 작성한다.

4) hmmsearch : 여러 서열로부터 hmmbuild에 의해 만들어진 HMM profile을 이용하여 도메인을 검색한다.

- E : 0.01

4.10. Protein motif discovery

입력 단백질과 유사한 단백질들을 blastp과 NCBI_efetch를 이용하여 추출하고 그들로부터 motif를 검색한다.



1) blastp : 프로그램은 서버에 미리 구축되어있는 단백질서열 데이베이스로 부터 입력된 단백질서열과 유사한 서열 부분을 찾아낸다.

- d : nr
- e : 0.01

2) Get Accessions Number From BLAST Result : BLAST 결과물에서 Accession numer를 가져온다.

3) NCBI_efetch : 해당 ID의 데이터를 가져온다.

- remax :10
- retmode : text
- rettype : fasta

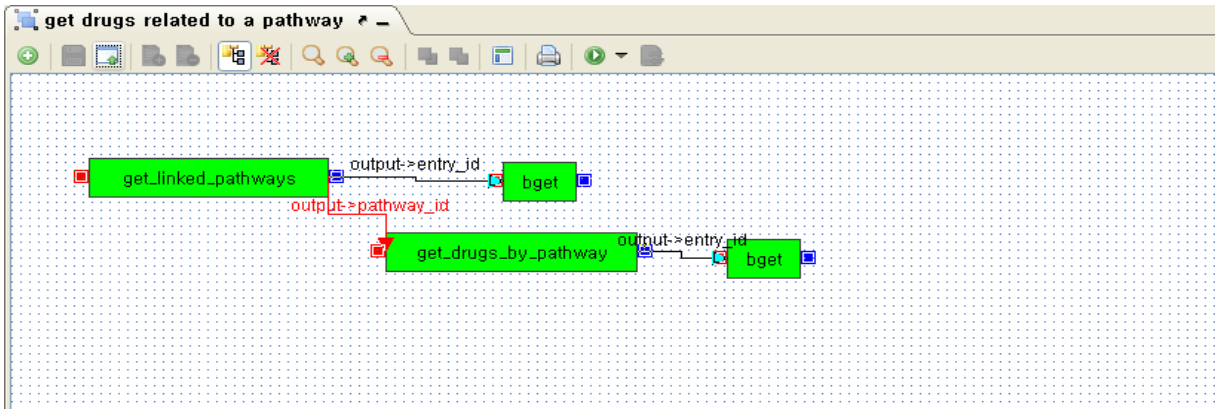
4) merge strings : 여러 문자열을 하나의 문자열로 통합한다. 여기서는 여러 단백질 서열을 하나의 출력으로 통합한다.

5) meme : 다중서열로부터 공통된 부분을 찾아 통계적인 분석을 하고 표시한다.

- options : all defaults

4.11. Get drugs related to a pathway

특정 pathway와 관련된 drug들을 찾는다.



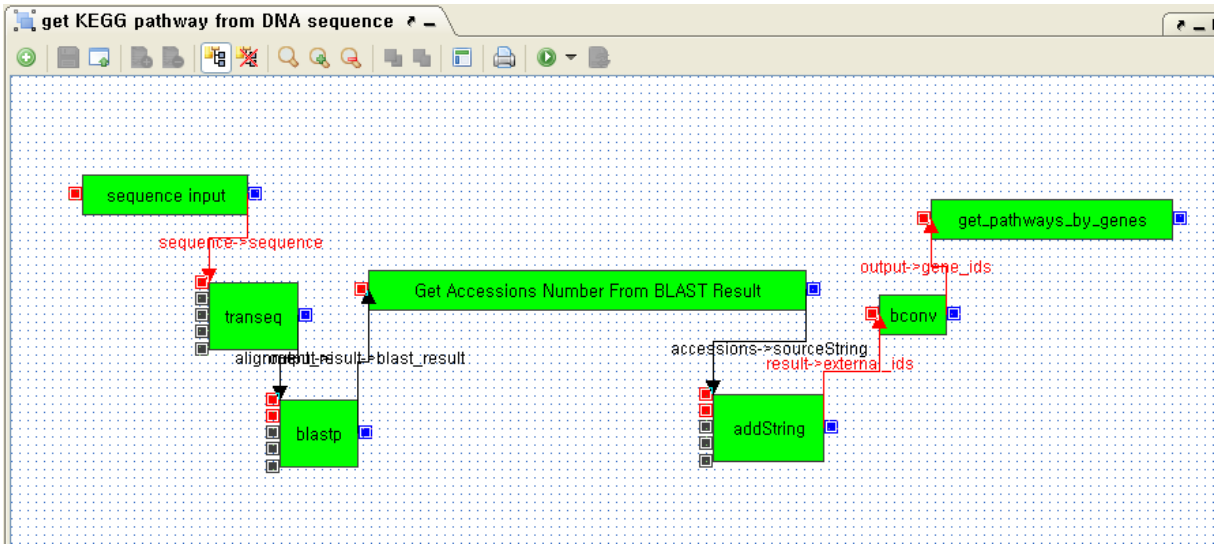
1) get_linked_pathways : 주어진 pathway와 연결된 pathway들을 출력한다.

2) get_drugs_by_pathway : 주어진 pathway와 관련된 drug들을 출력한다.

3) bget : 해당 엔터리의 상세 정보를 보여준다.

4.12. Get KEGG pathway from DNA sequence

입력 염기서열을 단백질로 번역한 후 blastp를 이용하여 swissprot 데이터베이스로부터 유사단백질들을 찾고, 그 단백질들과 관련된 pathway를 검색한다.



1) transeq : DNA 서열을 단백질 서열로 번역한다.

- options : all defaults

2) blastp : 서버에 미리 구축되어있는 단백질서열 데이베이스로 부터 입력된 단백질서열과 유사한 서열 부분을 찾아낸다.

- d : swissprot
- m : 0
- e : 0.5

3) Get Accessions Number From BLAST Result : BLAST 결과물에서 Accession number를 가져온다.

4) addString : 입력 문자열로부터 구분자를 이용하여 엔트리 목록으로 변환 후 각 엔트리에 접두사 또는 접미사를 붙인다.

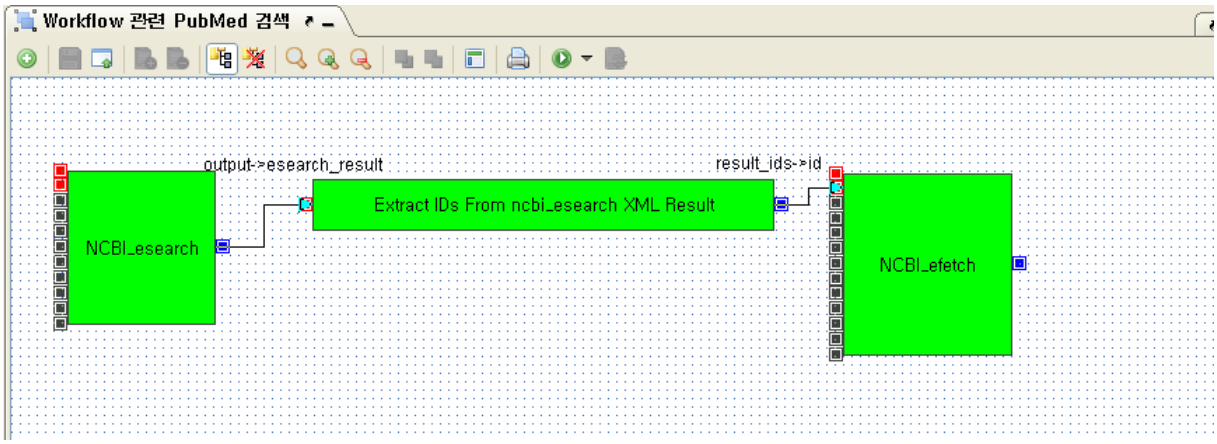
- appendString : uniprot:
- inputSeparator : ,
- direction : prefix
- outputSeparator : SPACE

3) bconv : 외부 데이터베이스 ID를 KEGG 데이터베이스 ID로 변환한다.

4) get_pathways_by_genes : 유전자 ID로부터 pathway ID를 가져온다.

4.13. Workflow 관련 PubMed 검색

Workflow 관련된 논문을 NCBI PubMed로부터 검색한다.



1) NCBI_esearch : NCBI의 특정 DB를 선택해서 검색하면 결과에 해당하는 ID를 XML 형태로 제공한다.

- db : pubmed
- query : workflow

2) Extract IDs From ncbi_esearch XML Result : ncbi_esearch 결과 XML로부터 ID 목록을 축출한다.

3) NCBI_efetch : ID에 해당하는 데이터를 NCBI 데이터베이스로부터 가져올 때 사용.

- db : pubmed
- retmode : text
- rettype : abstract

5. 맺음말

생명정보학 분야는 최근 급격히 세계시장이 증가하고 있는 분야이며, 향후 국가 전략 산업으로 육성되고 있는 분야이다. 특히 생명정보 분석 결과는 유전체학, 전사체학, 단백질체학, 대사체학, 약리유전체학 등 분자생물학의 모든 분야에서의 질문들에 대한 잠재적 의미 있는 답을 줄 수 있다. 따라서 유전자 구조 및 기능, 진화상 관계 등 생명정보 분야에서의 중요한 문제들에 대해 발견되는 새로운 지식을 종합적으로 신속하게 분석하여 생명과학 연구에 활용하는 것은 매우 중요한 일이다. 그

러나 이러한 생명정보 분석 기술의 발전에도 불구하고, 생명과학 연구자들이 자신의 연구에 활용하기에는 데이터 및 도구의 이질성, 일관된 사용 환경의 부재, 대용량•대규모 분석 환경 미흡, 그리고 IT 기술이 부족한 생명공학 연구자들의 생명정보 분석 시나리오 구현의 어려움 등의 문제점이 있다. 이러한 문제점들을 해결하기 위하여 KISTI에서는 생명과학 연구자들이 보다 손쉽게 자신의 연구에 필요한 생명정보 분석 도구들을 효과적으로 활용할 수 있도록 하기 위한 슈퍼컴퓨팅 인프라 기반의 Bioworks 시스템 서비스 제공하고 있다. 본 연구보고서에서는 Bioworks 시스템의 주요 특징 및 활용 방법에 대해 설명하였고, 실제 생명정보 응용연구에 활용될 수 있는 Bioworks 시스템을 활용한 주요 생명정보 분석 시나리오 구현 방법을 제시하였다. Bioworks 시스템은 연구자들이 어려워하는 복잡한 생명정보 분석과정을 효과적으로 자동화하고 이를 공동연구자들과 같이 공유하면서 연구의 효율성을 극대화 할 수 있는 시스템으로써 향후 국가 생명과학 연구 발전에 크게 기여할 수 있을 것이라 예상된다.

참고문헌

1. Arthur L. Delcher, et al., *Improved microbial gene identification with GLIMMER*, Nucleic Acids Research, 27(23)4636-4641, 1999.

2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, *Basic local alignment search tool*, J. Mol. Biol., 215(3):403-410, 1990.

3. Chenna R, et al., *Multiple sequence alignment with the Clustal series of programs,* Nucleic Acids Research, 31(13):3497-3500, 2003.

4. Katoh K, Misawa K, Kuma K, Miyata T., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform,* Nucleic Acids Research, 30(14):3059-3066, 2002.

5. Edgar RC, *MUSCLE: multiple sequence alignment with high accuracy and high throughput,* Nucleic Acids Research, 32(5):1792-97. 2004.

6. Durbin, Richard; Sean R. Eddy, Anders Krogh, Graeme Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,* Cambridge University Press, 1998