단백질 구조예측을 위한 에이전트 시티 네트워크 기반의 다중 에이전트 시스템

김현식[0] 남덕우 진 훈 김인철
경기대학교 전자계산학과
(advance7[0], culyhair76, jinun, kic)@kyonggi.ac.kr

An Agentcities Network-Based Multiagent System for Supporting Protein Structure Predictions

Hyun-Sik Kim[0] Duek-Woo Nam Hoon Jin In-Cheol Kim
Dept. of Computer Science, Kyonggi University

요 약

휴먼지놈프로젝트이후 컴퓨터를 이용한 연구는 점차로 활발하게 진행되고 있는데 그 중 단백질의 기능예측과 관련하여 보다 많은 연구가 이루어지고 있다. 단백질의 기능예측을 위해서는 3차원 구조정보가 많이 이용된다. 3차원 구조를 형성하는 것은 주로 아미노산 서열이나 1차, 2차 구조 정보가 보다 구체적인 단백질 구조예측을 위해 이용되고 있다. 전세계적으로 다량의 단백질 구조정보 및 예측을 위한 방법들이 소개되고 있지만 각 자원들마다 저장, 관리 형식이 다를 뿐만 아니라, 정보를 이용하는 방법도 어렵다. 본 논문에서는 다양하게 존재하는 단백질 구조 데이터베이스 자원들을 에이전트화여 통합성과 재사용성을 지향하였고, 에이전트시티 네트워크에 연결함으로써 개방성과 확장성, 분산성을 높이도록 하였다.

1. 서론

1990년대 들어 시작된 휴먼지놈프로젝트를 토대로 90년대 중반 이후 컴퓨터를 이용하여 밝혀진 서열정보를 이용하여 단백질의 구조 예측을 시도하는 수많은 연구들과 성과를 얻게 되었다. 현재까지 약 330여 종류의 단백질 관련 데이터베이스들과 다수의 소프트웨어 도구들이 개발되어 있다. 단백질의 기능을 제대로 예측하기 위해서는 단백질의 3차원 구조를 알 필요가 있다. 단백질의 3차원 구조는 20개의 아미노산들로 이루어진 서열정보에 의해 주로 결정되지만 정확한 구조예측을 위해서는 이 외에도 서열을 이루는 원자 간 힘, 수소결합력, 친수성과 소수성, 전하 간 상호작용 등 다양한 정보를 필요로 한다. 또한 이와 같이 많은 정보들을 잘 이용한다고 해도, 폴딩(folding) 구조는 생물학적 환경과 그 변화에 따라 달라질 수 있기 때문에 구조예측에 어려움이 많다. 구조예측기술이 많은 발전을 이루었음에도 불구하고 단백질 구조를 예측하는 과정은 대단히 복잡하기 때문에, 아직까지도 단백질의 구조예측을 위한 정확한 프로토콜은 정해져 있지 않은 상황이고[3], 그 동안은 주로 단백질 구조에 대해 부분적인 연구가 이루어졌을 뿐이다. 최근 들어 이와 같이 존재하는 여러 자원들을 연결하여 정보를 공유하고자 하는 연구들이 진행되었고 결과적으로 기존의 단백질 데이터베이스들을 서비스하는 유명 사이트들이 복합적인 형태의 정보를 제공하는 형태로 변모하게 되었다. 이를 통해 단백질의 구조 예측기술이 발전하였지만 여전히 해결해야 할 문제가 많다. 첫째, 각 자원들은 구조예측작업을 위한 연계작업을 고려하여 개발되지 않았다. 각각의 자원들은 개인별, 기관별로 개발되어 사용되고 있으며, 서로 간 데이터의 연결이나 통합을 전제로 개발된 것들이 아니기에 데이터 양식 및 용어체계가 다르고 이용방법 면에서도 차이가 난다[1]. 둘째, 단백질 구조예측을 위해 기 존재하는 다양한 자원들을 효율적으로 연결하지 못한다. 셋째, 자원들은 세계 도처에 흩어져 존재하며 이러한 자원 정보를 판단하고 분석하여 최종 예측작업을 수행하는 전문가들 역시 흩어져 있다. 넷째, 현재 이용이 가능한 연결된 형태의 자원들을 이용한다고 하더라도 완전한 형태의 단백질 구조예측작업을 수행하기 어렵다. 이와 같은 면을 고려할 때 모듈성과 분해성, 그리고 협업성과 재사용성 및 확장성을 특징으로 하는 멀티 에이전트 시스템을 이용하는 것이 합리적인 선택이 될 수 있다[4]. 현재까지 알려진 이러한 연구들로는 지놈 분석을 위해 다수의 데이터베이스 자원들을 에이전트화한 GeneWeaver, 의학 정보 검색을 위해 개발된 MeLiSA, 생물과학 연구를 지원하는 목적의 BioAGent, 단백질 구조예측을 위한 에이전트들 간의 행위를 주로 다룬 MACoP 등이 있다. 본 논문에서는 개방형 다중에이전트 플랫폼인 JADE를 이용하여 전세계적인 분산 네트워크를 지원하는 에이전트시티 네트워크에 연결하도록 하였다. 그렇게 함으로써 플랫폼 간의 이질성을 극복하고 에이전트 간 협업체제를 통한 통합성과 타 시스템에서의 이용이 가능하도록 재사용성을 높이도록 하였다. 또한 세계 어느 곳에서라도 접근하여 이용이 가능하도록 개방성과 분산성, 확장성을 높인 에이전트 기반의 단백질 구조예측 지원 시스템에 대하여 기술하고자 한다.

2. 관련 연구

2.1 에이전트 시티 네트워크

에이전트시티는 FIPA 표준을 따르면서 다양하고 이질적인 에이전트들 간의 서비스를 지원할 수 있는 온라인 형태의 전세계적인, 협력작업을 지원하는 개방형 네트워크 시스템이다[5]. 에이전트시티 환경을 채용하는 것은 이질적 에이전트 플랫폼들과의 서비스를 융합하기 위해 ACL메시지를 통한 상호 교환과 공유에 기반 하는 적응성이 강하고 지능적 에이전트의 메커니즘을 제공하여 융통성을 높이기 위해서이다. 현재까지 97개의 플랫폼이 등록되어 있으며 플랫폼 디렉토리, 에이전트 디렉토리, 서비스 디렉토리 서비스를 통해 이미 개발된 다양한 종류의 에이전트 네트워크 서비스를 이용할 수 있다. 현재 에이전트시티에 접근 가능한 대표적인 에이전트 플랫폼들로는 April, Corntec, FIPA-

OS, JADE, LEAP, ZEUS 가 있다.

2.2 단백질 구조 예측

지금까지의 단백질 구조는 2차원 구조 예측을 위한 1D에서의 예측 연구, 2차원 구조에서도 사이드 체인 및 서열을 이루는 잔기, 또는 서열 가닥들 간의 관계 및 성질을 분석하는 2D에서의 예측 연구, 그리고 3차원 구조를 예측하기 위한 연구들로 이루어져 왔다. 3차원 구조에 관한 연구에서, 두 개의 염기서열이 서로 유사하다고 인정될 때 3차 구조는 다수의 경우 같게 나타나지만 다를 수도 있으며 3차 구조가 유사하더라도 염기서열 상으로는 전혀 다른 원연체일 수도 있다. 이는 서열 정보가 단백질의 구조를 결정하는 기본 골격이 되는 것은 분명하지만 서열정보가 아닌 부가적인 여러 정보(예. 분자론적 동적 방법)를 포함하여 분석과정을 거친 후에야 단백질의 3차 구조를 예측할 수 있다는 것을 의미한다.

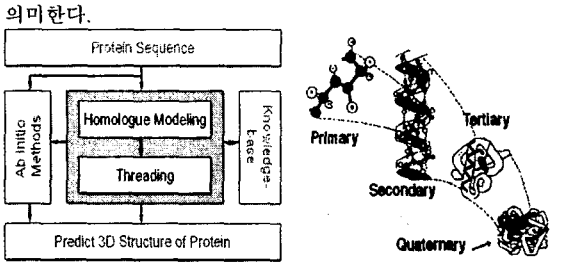

그림 1 단백질 구조예측을 위한 방법

단백질의 3차 구조예측을 위한 방법은 아래와 같이 단백질 구조 예측 기술 평가단(CASP)이 분류하는 바에 따라 상동성 모델링(Homologue Modeling), 스레딩(Threading), 순이론 예측(Ab Initio Methods)로 나뉠 수 있다[2]. 상동성 모델링 방법은 3차 구조의 경우가 단순 서열보다 진화를 거치는 과정에서도 더 잘 보존된다는 점에 착안하여 질의서열과 상동성이 있으면서 3차 구조를 가진 서열을 기반으로 하여 질의 서열의 3차 구조를 예측하는 방법이다. 스레딩 방법은 상동성 모델링 방법을 위한 상동성이 25%미만일 경우의 서열에 대하여 정렬상의 대상 서열을 선정하여 3차 구조를 찾은 후 서열과의 일치성을 평가하여 예측하는 방법이다. 이에 반해 순이론 예측방법은 단순 서열정보에 의거하여 물리화학적인 특성들을 컴퓨터를 이용하여 측정하고 분석하여 3차 구조를 예측하는 방법이다. CASP초기에는 별로 주목 받지 못했지만 CASP4에서 발표된 Rosetta 알고리즘의 개발 이후로 점차로 많이 이용되고 있다.

3. 시스템 설계

우리는 다양하게 존재하는 단백질 관련 데이터베이스들을 에이전트화하여 이들을 플랫폼의 특성, 위치에 상관없이 통합 운용하여 단백질에 관한 다양한 연구를 돕는 시스템(APSS, Agent-based Proteomics Support System)을 개발하고자 한다. 그리고 본 논문에서는 이에 앞서 단백질 구조예측을 지원하는 기능을 갖는 시스템을 구현하고 이에 관해 서술하고자 한다. APSS는 인터넷 상에 다양하게 각 처에 존재하는 단백질 자원들을 에이전트화하여, 조정 에이전트를 통해 대신 정보를 요청하고 수신할 수 있도록 하며 중간처리단계에서 발생되는 다른 자원들로의 접근, 처리결과의 변환 및 소프트웨어 에이전트의 사용유무 등은 에이전트들 간의 상호작용을 통해 이루어질 수 있도록 하였다.

3.1 시스템 구성

[그림 2]과 같이 에이전트 네트워크에 접속되는 플랫폼들은 최소한 AMS, RMA, DF 에이전트를 포함하며 UI 에이전트를 통해서 사용자는 예측 과정과 관련되는 모든 작업을 수행할 수 있다. 수행되는 모든 작업은 조정 역할을 수행하는 조정에이전트가 담당한다. 우측의 플랫폼2는 연결된 다른 플랫폼을 예로 든 것이다. 자세히 보면 플랫폼2에는 2차 구조 예측 서버 및 3차원 모델 예측 서버를 위한 에이전트가 존재한다. 그리고 이들 에이전트들과 플랫폼1이 점선으로 연결된 것처럼 하나의 시스템을 구성하며 이는 전세계에 존재하는 수많은 생물정보 자원들을 대신하는 에이전트는 하나의 플랫폼 내에만 존재해야 할 필요가 없으며, 에이전트시티 네트워크 서비스에 등록만 되어있으면 다른 시스템에서 이미 개발되었거나, 앞으로 개발될 에이전트들까지 쉽게 연결하여 하나의 시스템으로 포함될 수 있음을 의미한다.

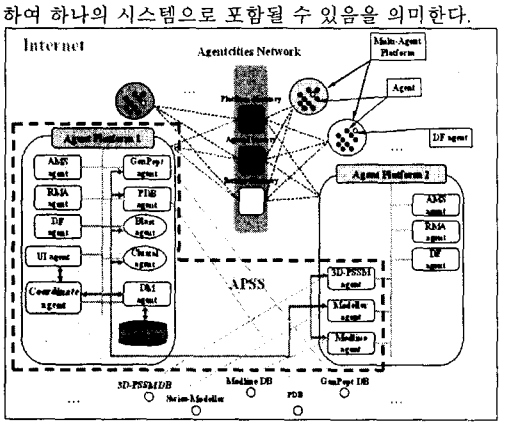

그림 2 시스템 구조도

시스템을 구성하는 에이전트들을 역할 및 기능에 따라 분류하면 아래와 같다.

① 자원 에이전트들 : GenPept, 3D-PSSM, PDB, Medline, SwissProt, MODELLER
② 분석 에이전트들 : Blast, Clustalw
③ 관리조정 에이전트들 : Coordinate(조정), DM, AMS, DF
④ 사용자 인터페이스 에이전트: UI

자원 에이전트는 실질적으로 단백질 구조예측을 위해 이용해야 할 단백질 구조 데이터베이스에 접속하여 질의를 하고 정보를 찾는 에이전트로서 일반 사용자들이 해당 데이터베이스에 웹 브라우저를 이용, 접속하여 사용하는 과정을 대신한다. 분석 에이전트는 자원 에이전트로부터 얻어진 정보들을 계산하고 분석하는 과정을 수행하며, 관리조정 에이전트는 에이전트 간 동작을 제어하고 조정하며 필요에 따라 저장기능까지도 수행하는 에이전트이다.

3.2 단위 에이전트

PDB 에이전트는 APSS를 이루는 주된 역할을 수행하는 에이전트이다. PDB 에이전트는 현재까지 밝혀진 모든 단백질 구조정보를 포함하는 PDB 데이터베이스(http://www.rcsb.org/pdb/)에 접속하여 질의 및 응답 과정을 수행한다. PDB 데이터베이스는 기본적으로 PDB ID 와 키워드 및 맞춤검색을 지원한다. 질의를 통한 응답결과로는 요약 정보, 구조 파일 보기, 구조 파일 받기, 유사 구조 파일 정보, 및 1차 서열 파일 등이다. PDB 에이전트는 UI 에이전트를 이용하여 사용자가 직접 PDB 데이터베이스 정보를 사용할 수 있도록 하기 위해 질의 종류와 요구 정보, 2가지로 구분하였으며 아래와 같다.

표 1 PDB 에이전트 질의 정보

질의 종류	요구 정보
PDB ID	요약 정보
	구조 파일
	서열
	링크 정보
서열	유사 단백질의 PDB ID 들
	유사 단백질의 구조 파일들
	유사 서열들
	유사 단백질의 링크 정보
키워드	유사 단백질의 PDB ID 들
	유사 단백질의 구조 파일들
	유사 서열들
	유사 단백질의 링크 정보

PDB ID 로 검색 시에는 직접 해당 단백질 구조 정보를 찾아 돌려주지만 서열과 키워드 검색 시에는 상동성을 갖는 단백질 구조들을 찾은 후 PDB ID 리스트를 모으게 된다. 그 후 다시 PDB ID를 이용하여 질의를 수행하여 해당 정보들을 찾게 된다. 이때 PDB 에이전트로 취합된 정보들은 자바 오브젝트로 생성되어 최종적으로 UI 에이전트에게 전달된다.

3.3 에이전트 동작

먼저 GenPept에이전트는 NCBI의 GenPept 데이터베이스에 접속하여 질의서열과 blasting한 결과를 통해 상동성이 존재하는 유전자의 Annotation 정보 또는 서열을 구하는 작업을 수행한다(②). 이를 위해 Blast 에이전트와 정렬과정 수행을 위한 Clustalw에이전트가 이용된다. Medline에이전트는 유전자와 관련된 문헌정보를 사용자에게 제공하여 구조예측을 돕는 작업을 수행한다. MODELLER는 SWISS-Modeller서버를 대신하는 에이전트이다. [그림 3]은 처리과정을 예로 나타낸 것이다.

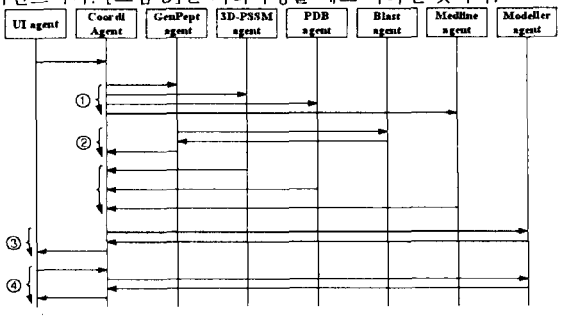

그림 3 시퀀스 다이어그램

미지의 서열에 대한 3차 구조를 예측하기 위해서는 이미 얻어진 서열들과 유사한 모형(template) 구조를 찾고(③) 서열과 모형구조와의 맞춤작업(fitting) 후 이를 서버로 전달한다. 최종결과를 수신하면 사용자에게 전달한다(④). 조정에이전트는 UI에이전트로부터 질의된 내용을 처리하여 데이터베이스 에이전트들에 차례로 전달하고 필요에 따라 소프트웨어 에이전트를 호출하여 작업을 지시하기도 한다. 처리된 결과들은 다시 조정에이전트에게 전달되며 적절한 처리과정을 거쳐 UI에이전트에게 전달된다(①).

4. 구현

현재까지 구현된 시스템의 구현환경은 다음과 같다[그림 4].

- H/W : Intel Pentium CPU 4 1.6Ghz, 512M Memory
- S/W : Windows XP professional OS, SDK 1.4.1, JADE 2.61

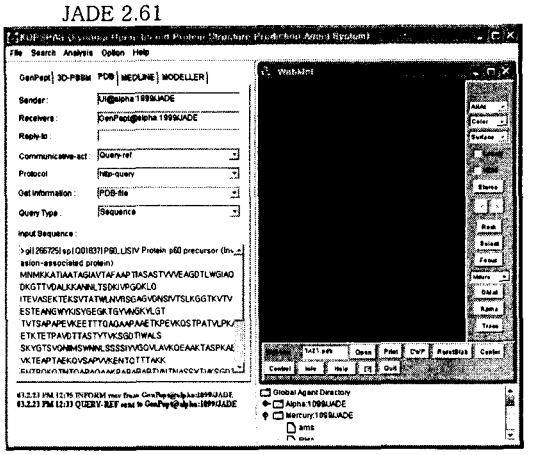

그림 4 통합 사용자 인터페이스

UI는 크게 다섯 부분으로 분류된다. 상단 메뉴 부분과 탭 메뉴로 구성된 각 에이전트 제어 부분, 에이전트 간 메시지 송수신 상태 확인을 위한 부분, 그리고 작업 요청에 대한 에이전트로부터의 정보를 수신하여 결과를 보여주는 결과창 부분, 마지막으로 현재 디렉토리 서비스에 등록되어 동작중인 에이전트의 목록 및 상태를 트리 형태로 보여주는 부분으로 구성되어 있다.

5. 결론

지금까지 본 논문에서 개발하고 있는 시스템의 개발 목적 및 방법, 설계 그리고 구현결과에 대하여 기술하였다. APSS는 재사용성을 높일 수 있을 뿐만 아니라 다양하게 존재하는 단백질 자원들을 관리하고 사용자 측면의 편리성을 제공하기 위하여 에이전트화하였다. 또한 이들을 에이전트시티 네트워크에 연결함으로써 시스템 측면에서의 개방성 및 분산성과 확장성을 높였다. 향후 계획으로는 시스템의 완성도를 높이는 것이며 나아가 APSS를 단백질체학 전 영역에 걸쳐서 지원이 가능한 시스템이 되도록 하는 것이다.

참고 문헌

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

[2] Cynthia Gibas, Per Jambeck., *Developing Bioinformatics Computer Skills*, O'Reilly, 2001.

[3] Rost, B., "Protein structure prediction in 1D, 2D, and 3D", *The Encyclopaedia of Computational Chemistry* (eds. PvR Schleyer, NL Allinger, T Clark, J Gasteiger, PA Kollman, HF Schaefer III and PR Schreiner), No.3, pp.2242-2255, 1998.

[4] Sycara, Katia, "*Multiagent Systems*", *AI Magazine* Vol.19, No.2, pp79-92, 1998.

[5] Willmott, S.N., Dale, J., Burg, B., Charlton, C. and O'brien, P., "*Agentcities: A Worldwide Open Agent Network*", *Agentlink News*, Vol.8, pp.13-15, Nov. 2001.