Virtual Screening을 위한 e-Science 프로젝트 동향

The Trend of e-Science Projects for Virtual Screening

김남규, 안선일, 이세훈, 이준학, 황순욱
한국과학기술정보연구원 e-Science 사업단

KIM Nam Gyu, AHN Sun Il, LEE Sehoon, LEE June H., HWANG Soon Wook
Korea Institute of Science and Technology Information

요약

e-Science는 이전의 연구수행방법을 최신의 정보통신기술을 적용하여 연구생산성을 향상시키는 새로운 패러다임이다. 이러한 e-Science를 적용하는 연구 분야 중에서도 바이오 분야는 바이오인포매틱스 학문이 만들어질 정도로 정보통신 기술을 많이 사용하는 분야이다. 이 바이오 분야에 적용되는 IT기반의 기술 중에서 대용량의 컴퓨팅 자원을 바탕으로 한 e-Science가 적용되는 분야가 신약 개발 과정 중 하나인 Virtual Screening이다. 이 Virtual Screening을 수행하는 e-Science 프로젝트 중 WISDOM과 Avian Flu에 대해 알아보고 국내에서 진행되고 있는 프로젝트의 진행 상황을 살펴본다. 그리고 앞으로 국내에서 KISTI와 전남대가 계획하고 있는 Virtual Screening에 대한 공동 연구에 대한 향후 방향을 설정해본다.

Abstract

e-Science is a new paradigm to increase the research efficiency by applying the state-of-the-art information technology to the classical research methodology. Among the research fields influenced by e-Science, Biology and Bioinformatics are the fields that are using IT very actively. And virtual screening, a procedure for the drug discovery, is one of the bioinformatics applications which requires a large amount of computing resources. In this paper, WISDOM, which is a e-Science project to design a new drug for Malaria and Avian flu, and related projects are introduced and the joint research between KISTI and Chun-nam university planned for the virtual screening research is explained.

I. 서 론

e-Science는 이전의 전통적인 연구수행방법을 정보통신기술을 적용하여 연구생산성을 향상시키는 새로움 패러다임이다. 이러한 e-Science에 대해 미국과 영국, 유럽을 비롯하여 많은 나라에서 연구 및 구축활동을 벌이고 있으며 항공, 바이오, 기상, 물리 등 다양한 연구 분야에 적용되고 있다. 이 중 바이오 분야는 바이오인포메틱스와 같은 새로운 학문을 통해 정보통신기술이 많이 적용되고 있는 연구 분야로써 대용량 계산 및 데이터를 많이 활용되고 있고 필요로 하는 분야이다. 이렇게 바이오 분야가 정보통신기술을 바탕으로 급속도로 발전하고 있는 분야이므로 자연스럽게 e-Science를 기반으로 한 프로젝트들이 많이 생겨나고 있다. 이 중 대용량 계산 자원 및 데이터를 이용하여 신약개발 시간을 단축시키고자하는 Virtual Screening 기법이 있다[1][2]

본 논문에서는 이러한 Virtual Screening을 위한 해외의 e-Science 프로젝트에 대해 기술하고 이러한 해외 프로젝트와 공동 연구를 수행하고 있는 KISTI에서 진행하고 계획하는 Virtual Screening에 대한 프로젝트에 대해서 기술한다.

관련연구에서는 Virtual Screening에 대한 간략한 기술하게 소개한다. 그 뒤 본론에서는 Virtual Screening을 위한 유럽의 EGEE 프로젝트 중 하나인 WISDOM 프로젝트와 이 프로젝트의 스핀 오프 프로젝트인 대만의 Avian Flu 프로젝트에 대해 설명한다[3][4][5]. 그리고 한국에서 진행되고 있는 Virtual Screening을 위한 작업들로써 KISIT의 EGEE 인프라를 기반으로 한 시스템의 아키텍쳐 및 운영에 대해서 기술하고 전남대 김도만 교수님 팀의 Virtual Screening을 결과의 동물 실험에 대해 기술한 결론 및 향우연구에서 앞선 국외 프로젝트 결과 및 앞으로 이러한 프로젝트와 연계하여 한국에서 KISTI와 전남대 김도만 교수님 팀의 Virtual Screening을 위한 프로젝트 계획에 대해 설명한다.

II. 관련연구

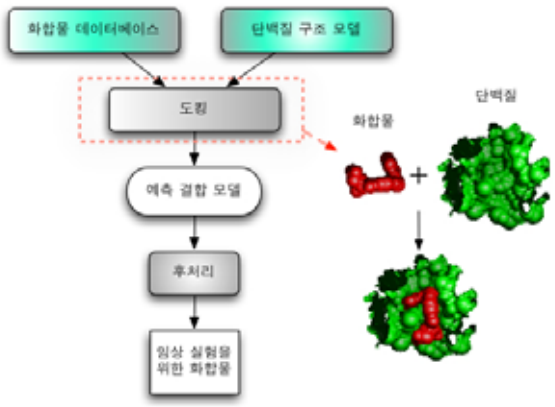
1. Virtual Screening

제약 산업에서 신약개발에 필요한 프로세스는 초기에 단백질과 결합하는 화합물을 추출하여 동물 실험과 임상 실험을

거쳐 약의 효능 및 부가작용에 대해 입증하고 정부의 허가를 받아 제품을 생산하기까지의 과정을 포괄한다.

신약 개발에서 가장 초기에 시행되어지는 작업은 병을 일으키거나 병을 일으키는데 관련된 단백질을 억제하는 물질을 찾는 과정이다. 예전에는 소수의 경험 많은 전문가들이 수백만 개의 화합물 중에서 경험을 통해 축적한 지식과 여러 가지 기법을 통해 해당 단백질과 관련이 없을 것 같은 화합물을 제거하고 수 십,수 만개의 화합물을 가지고 실제 화합물을 사거나 합성하여 단백질과 실제 결합시키는 실험을 수행하는 것이다. 이것을 Screening이라고 한다. 이 작업은 전체 신약개발 프로세스의 10%정도만 차지하고 있으나 총 기간 및 비용을 고려한다면 매우 큰 작업이다.

이러한 단순한 작업의 반복을 컴퓨터상의 시뮬레이션을 통해 수행하는 것이 Virtual Screening이다. Virtual Screening은 수백만 개의 화합물 데이터와 목표로 하는 단백질의 구조를 수치화 시켜 컴퓨터상에서 Docking 시뮬레이션을 통해 어떠한 화합물이 단백질과 제일 잘 결합하는 지를 찾는 작업이다.[1][2]. 이를 위해서 데이터베이스에 탑재된 수백만 개의 화합물과 단백질간의 결합을 계산하기 위한 대용량의 데이터베이스와 저장소 그리고 계산자원이 필요하다.[6] 이를 흐름도로 나타낸 것이 그림[1]이다.



▶▶ 그림 1. Virtual Screening 흐름도

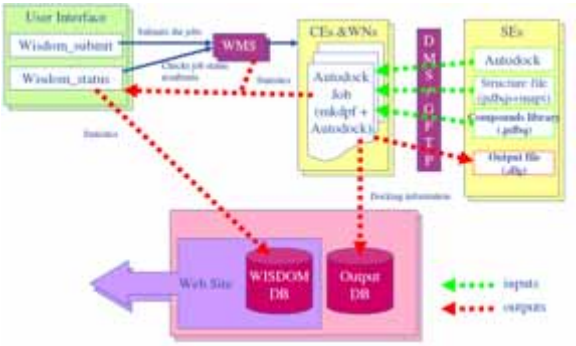
III. Virtual Screening 프로젝트

1. WISDOM

유럽의 EGEE 프로젝트 중 하나인 WISDOM 프로젝트는 초기 말라리아 신약 개발을 위한 Virtual Screening을 수행하는 것으로 시작한 프로젝트이다.[4] 말라리아는 계속해서 변종이 나오고 있으나 보통 말라리아 발병이 아프리카의 저개발 국가에서 자주 발생되므로 제약회사들이 신약개발에 소극적이여서 많은 아프리카 주민들이 고통을 받고 있다. 따라서 Virtual Screening을 통해 비용줄이고 시간을 단축시키므로써 빠른시간안에 적응 비용으로 말라리아에 대한 신약개발을 위한 후보물질을 발굴할 수 있다. WISDOM은 EGEE 그리드 인프라위에서 분자결합 응용에 대한 대규모 구축의 첫 사례이다. 이러한 대규모 구축을 위해 유럽은 물론 아시와의 많은 그룹들이 이 프로젝트에 참여했으며 대부분의 그룹은 Biomed 팀이 주축이 되었다.

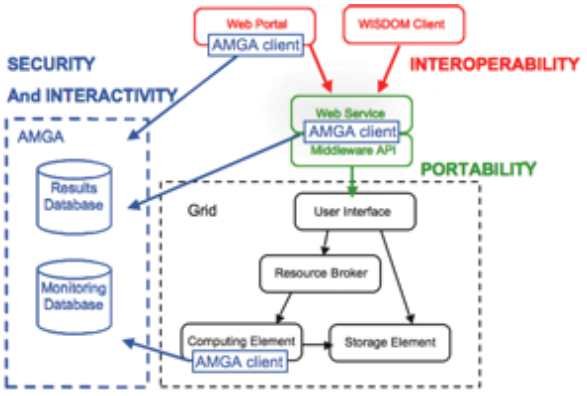
WISDOM-I은 4200만개의 작업을 생성하여 2005년 8월부터 2005년 9월까지 약 6주간 데이터 챌린지를 통해 약 80년 cpu 시간을 사용했다. 이 데이터 챌린지 결과 인간의 헤모글로빈의 초기 분열에 관련된 말라리아 원충의 아스파라산 단백질 효소에 대응하는 화합물을 1000개 찾았다[4].

아래 그림[2]은 WISDOM 프로젝트에 사용된 Virtual Screening의 흐름도이다. gLite-UI 노드에서 Wisdom_submit 스크립트로 WMS를 통해 Docking에 대한 Job을 실행시키면 gLite-CE를 통해 gLite-WN에서 SE에 있는 Auodock과 화합물, 단백질 구조를 가져와서 Docking을 하고 그 결과를 분석하여 데이터 베이스에 저장하고 결과 파일은 SE에 올려놓는다. 후에 데이터 베이스에 저장된 결과를 취합하고 후처리를 통해 결과가 좋은 화합물을 골라낸다.



▶▶ 그림 2. WISDOM 흐름도

현재 WISDOM은 데이터 챌린지를 통해서 EGEE 인프라에서 발견된 문제 및 기타 다양한 문제점을 보완하여 작업 실행 아키텍쳐를 변경하여 새로 WISDOM-II를 개발하고 있으면 이 시스템을 통해 올해 연말경에 새로운 화합물과 단백질을 가지고 데이터 챌린지를 시도하려 하고 있다[7]. 이 WISDOM-II의 아키텍쳐는 그림[3]과 같다. 가장 크게 변한 것은 데이터 베이스로 GSI기반의 gLite 공식 메타 데이터 카타로그인 AMGA를 사용한다는 것과 분산 처리 효과를 크게 하기 위해서 WISDOM의 작업 실행을 웹 서비스로 제공하는 것이다.



▶▶ 그림 3. New WISDOM 흐름도

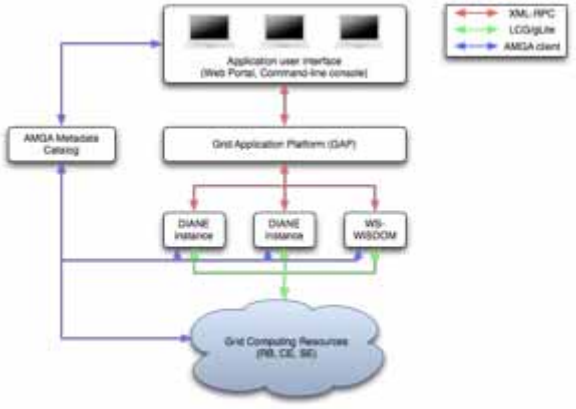
2. Avian Flu

1에서 언급했듯이 WISDOM은 많으 그룹들이 참여하여 수행한 프로젝트로서 그 중에 타이완의 ASGC도 참여하였다. WISDOM-I의 데이터 챌린지 후 ASGC는 WISDOM을 바탕으로 조류 독감의 백신을 위한 신약후보물질을 찾기 위한 Avian Flu 프로젝트를 만들었다[5]. Avian Flu는 현재 아시아에서 계속해서 발생되고 있는 조류 독감이 계속 변종을 만들어내고 그 중 H5N1은 인간에게 전염되는 특성을 가진 변종 바이러스에 대항 할 수 있는 신약후보물질을 찾기 위한 Virtual Screening 프로젝트이다.

Avian Flu에서는 WISDOM-I의 데이터 챌린지 결과를 분석한 뒤 WISDOM-II와 같이 작업을 분산처리하기 위하여 DIANE 프레임워크를 도입하였다. 그림 [4]는 Aviane Flu의 개략적인 구조도를 나타낸 것이다. DIANE은 분산 작업 처리를 위한 GANGA를 포함하여 더 효율적인 그리드 분산 작업을 처리한다. 그리고 각 DIANE Workers를 이용하여 작업을 처리한 뒤 그 결과를 DIANE Master Process가 취합할 수 있게한다. 이를 통해 많은 작업을 한 곳에서 통제할 수 있고 작업 실행에 대한 정확한 정보도 더 쉽게 얻을 수 있었다.

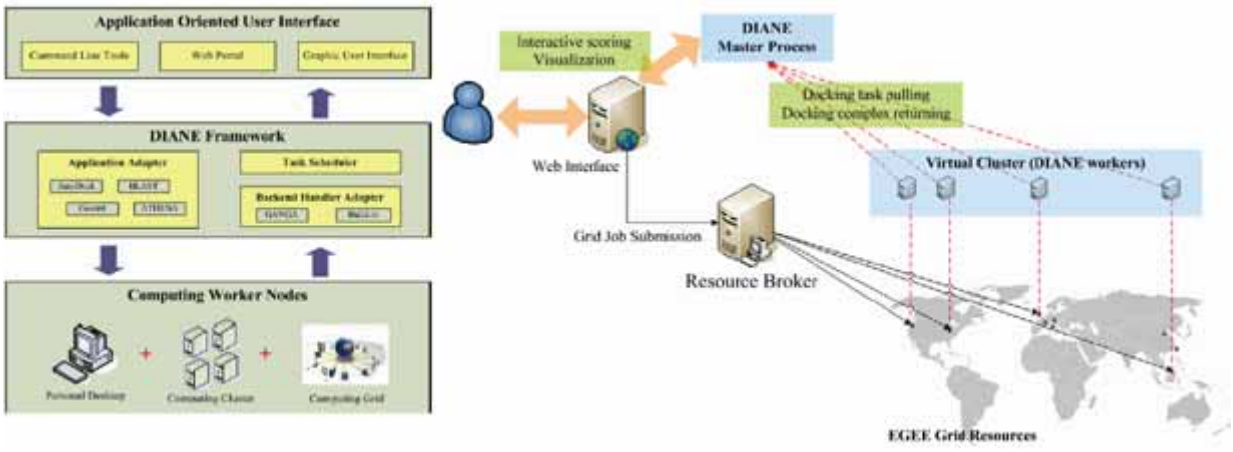
Avian Flu는 2006년 4월과 5월 사이에 30만개의 화합물을 H5N1 단백질 구조체와 Docking하는 데이터 챌린지를 수행하여 총 308,585개 작업을 한달 동안 수행하였다. 이를 통해 조류 독감 바이러스가 증식하는데 필수적인 효소를 분해하는 화합물을 찾을 수 있었다[5].

이 Avian Flu 프로젝트에서도 초기 데이터 챌린지 시 문제점을 발견하였는데 DIANE Master Process가 하나여서 많은 DIANE Workers가 작업을 받아오거나 결과를 전송하려고 한꺼번에 접속할 경우 300개 이상 연결이 불가능 하였다. 이유는 대부분의 리눅스에서 TCP/IP연결을 300개로 제한하여 발생하는 문제였다. 타이완의 ASGC팀은 확장성을 위해서 다수의 DIANE Master process를 유지하고 관리하기 위하여 새로운 아키텍처를 개발하였다[8]. 아키텍처의 개략적인 그림은 그림 [5]와 같다.



▶▶ 그림 5. New Avian Flu 구조도

새로운 Avian Flu 아키텍처는 WISDOM-II와 마찬가지로 AMGA를 데이터베이스로 사용한다. 그리고 여러개의 DIANE Master Process 또는 WISDOM 웹서비스를 이용하고 관리할 수 있도록 GAP(Grid Application Platform)이라



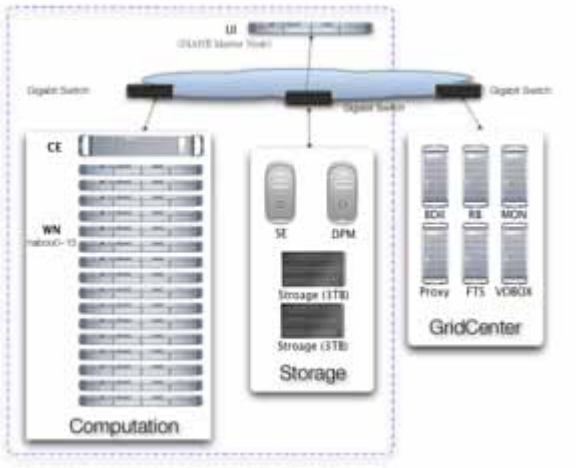
▶▶ 그림 4. Avian Flu 구조도

는 계층을 하나 더 만들어서 Avian Flu의 첫번째 데이터 챌린지에서 나타난 문제점을 해결하였다. Avian Flu의 두번째 데이터 챌린지는 2007년 9월부터 시작하여 10월초순경에 Phase I를 끝냈으며 10월 중순부터 Phase II를 시작하였다. 이 Avian Flu 데이터 챌린지는 EGEE 07 컨퍼런스에서 Best Demo 상을 받으므로써 내외적으로 훌륭하게 작업을 수행한 것으로 평가받았다.

3. 한국에서의 Virtual Screening 진행 현황

3.1 KISIT의 인프라 구축

KISTI는 작년부터 CERN연구소와 고에너지 물리분야의 응용분야 중 하나인 Alice를 위하여 EGEE 인프라를 구축하여 서비스하여 왔다. 이를 바탕으로 WISDOM과 Avian Flu와 협력하여 거대 스케일의 Virtual Screen을 위한 인프라를 추가 구축하였다. 그림[6]은 KISTI에 구축되어진 EGEE 인프라이다.



▶▶ 그림 6. KISTI EGEE 인프라 구조도

GridCenter는 이미 Alice를 위한 gLite 컴포넌트들을 설치하여 운영하고 있다. 정보서비스를 하는 gLite-BDII, 작업 브로커인 gLite-RB, 모니터링을 담당하는 gLite-MON, 프락시를 관리하는 gLite-Proxy, 파일 전송 컴포넌트인 gLite-FTS, 그리고 VO에 대한 관리를 위해 gLite-VOBOX가 설치되어 있다.

이러한 기본적인 EGEE 인프라에 더하여 대용량 데이터 처리를 위한 스토리지 컴포넌트인 gLite-SE, gLiet-DPM을 설치하고 6TB의 스토리지를 할당하였다. 또한 Virtual Screening을 위한 계산 자원용으로 gLite-WN를 15대 구축하고 WN를 관리하는 gLite-CE를 구축하였다. 마지막으로 Avian Flue에 Site Deployment로 참여하여 DIANE Master를 설치 운용하기 위한 gLite-UI를 구축하였다. 이러한 EGEE 인프라를 구축하여 KISTI는 2007년 9월달부터 시작된 Avian Flu 2차 데이터 챌린지에 참여하여 전체 작업 양의 1%의 작업을 할당 받아 처리하였다. 또한 이번 데이터 챌린지에 사용된 4개의 DIANE Master 노드 중 KISTI에 설치된 하나의 DIANE Master 노드가 전체 Avian Flu 작업의 25%를 EGEE 인프라에 제출하고 그 결과를 관리하였다.

3.2 전남대 팀의 실험실 실험

전남대학교 생명과학기술학부의 김도만 교수님은 WISDOM 프로젝트에서 나온 결과물, 즉 단백질과 잘 결합할 수 있는 화합물 중 최상위의 화합물들을 실제로 실험을 하는 유일한 생명과학자이다. 즉 계산 결과로 나온 화합물들을 보유하고 있는 회사에서 사들여서 단백질과 결합을 하여 실제 그 화합물이 단백질과 결합을 잘 하는지 실험을 하는 책임을 가지고 있다. 이러한 책임을 지고 있는 전남대 팀은 앞선 WISDOM 데이터 챌린지 결과는 물론 Avian Flu의 데이터 챌린지 결과를 바탕으로 결합 실험을 시행 중이다.

또한 전남대 팀은 연구실 내부에서 당뇨병을 일으키는 기전에 관계되는 단백질을 연구하고 있다. 이러한 단백질을 적절하게 억제하는 화합물을 찾기 위해서 전남태 님은 WISDOM과 Avian Flu의 Virtual Screening 팀과 협력하고 있다.

IV. 국내 Virtual Screenig 프로젝트 계획

현재 KISTI는 Virtual Screening을 위한 인프라를 구축 운영 중에 있으며 계속 확충하고 있다. 또한 WISDOM과 Avian Flu와 국제 공동 연구를 통해서 대규모 자원을 이용하고 관리하는 기술을 축적 중에 있다. 또한 한국-프랑스간 LIA협정을 통하여 프랑스의 WISOM 기술에 대한 공동 개발을 추진 중에 있다. 전남대는 현재 WISDOM과 Avian Flu에 대한 화합물과 단백질에 대한 실험실 실험을 프랑스의 지원 아래 진행하고 있으며 이를 통해 화합물과 단백질 결합에 대한 노하우를 축적 중에 있다

이와 같이 국내에서 신약후보물질발굴에 관한 초기 두가지 단개를 모두 수행할 수 있는 기관이 있으므로 국내에서도 독자적으로 신약후보물질발굴 프로젝트를 진행할 수 있는 기반이 마련되었다.

이에 따라 KISTI와 전남대는 향후 MOU를 체결하여 당뇨병 발병 기전에 관여하는 단백질에 대한 화합물을 찾기 위해서 Virtual Screening을 EGEE 인프라를 통해 수행하고 그 결과물을 전남대에서 실험할 것이다. 이를 통해 국내에 Virtual Screening에 대한 인식을 높이고 실제 신약개발을 위

한 여러 개발 주체들이 e-Science를 이용한 인프라 구축 및 개발에 참여하는 계기를 될 것으로 확신한다.

▌참 고 문 헌▐

[1] P.D. Lyne, Structure-based virtual screening: an overview, Drug Discov, Today 7, pp. 1047 - .1055, 2002

[2] M. Congreve et al., Structural biology and drug discovery, Drug Discov. Today 10, pp. 895 - 907, 2005

[3] F. Gagliardi, et al., Building an infrastructure for scientific grid computing: status and goals of the EGEE project. PhilosophicalTransactions: mathematical, Physical and Engineering Sciences 363, pp. 1729-1742 and <http://www.eu-egee.org/>, 2005

[4] Jacq, N., Salzemann, J., Legré, Y., Reichstadt, M., Jacq, F., Zimmermann, M., Maaß, A., Sridhar, M., Kasam, V. Schwichtenberg, H., Hofmann, M., Breton, V., Demonstration of In Silico Docking at a Large Scale on Grid Infrastructure, Studies in Health Technology and Informatics, 120, pp. 155-157, 2006

[5]H. C. Lee, et al, "Grid-enabled High-throughput in silico Screening against Influenza A Neuraminidases", IEEE Transaction on NanoBioscience, Vol. 5 Issue 4, pp. 288-2295, 2006.

[6] A. Chien et al., Grid technologies empowering drug discovery, Drug Discov. Today 7, pp. 176 - .180, 2002

[7] Jacq, N., Breton, V., Chen, H.-Y., Ho, L.-Y., Hofmann, M., Lee, H.-C., Legré, Y., Lin, S.C., Maaß, A., Medernach, E., Merelli, I., Milanesi, L., Rastelli, G., Reichstadt, M., Salzemann, J., Schwichtenberg, H., Sridhar, M., Kasam, V., Wu, Y.-T., Zimmermann, M., Virtual Screening on Large Scale Grids, Parallel Computing, 33, pp. 289-301, 2007

[8] N. Jacq, et al, "Virtual Screening on Large Scale Grids", Parallel Computing, Vol. 33, pp. 289-301, 2007