PreSPI: 단백질 상호작용 예측 서비스 시스템
(PreSPI: Protein-Protein Interaction Prediction Service System)

한 동 수 † 김 홍 숙 †† 장 우 혁 ††† 이 성 독 ††††
(Dong-Soo Han) (Hong-Soog Kim) (Woo-Hyuk Jang) (Sung-Doke Lee)

요 약 계산을 통한 단백질 상호작용 예측 기법의 중요성이 제기되면서 많은 단백질 상호 작용 예측 기법이 제안되고 있다. 하지만 이러한 기법들이 일반 사용자가 손쉽게 사용할 수 있는 서비스 형태로 제공되고 있는 경우는 드물다. 본 논문에서는 현재까지 알려진 단백질 상호작용 예측 기법 중 예측 기법의 완성도가 높고 상대적으로 예측 정확도가 높은 것으로 알려진 도메인 조합 기반 단백질 상호 작용 예측 기법을 이용하여 서비스 시스템으로 설계하고 구현하였다. 효모(Yeast)의 단백질 집합에 대하여 학습한 후, 학습된 단백질 집합과 공통된 도메인을 가지지만 학습 집합에 존재하지 않는 단백질 쌍들에 예측 기법을 적용하여 매우 높은 77%의 민감도(sensitivity)와 95%의 특이도(specificity)를 보였다. 더불어 DIP CORE, HMS-PCI, TAP 데이타의 테스트를 통해서 이 기법의 안정성을 확인하였다. 시스템의 기능들은 핵심 기능, 부가 기능 그리고 일반 서비스 기능으로 분류하였다. 시스템 설계의 주요 목표인 성능, 개방성 그리고 확장성에 따라, 개별 서비스들은 병렬화, 웹 서비스 표준 준수 및 계층화된 구조화를 지원하도록 구현하였다. 본 논문에서는 몇 가지 대표적인 사용자 인터페이스와 상세한 사용 지침도 소개한다.

키워드 : PreSPI, 단백질-단백질 상호작용 예측, 도메인 조합, 웹 서비스, AP 행렬, primary interaction probability

Abstract With the recognition of the importance of computational approach for protein-protein interaction prediction, many techniques have been developed to computationally predict protein-protein interactions. However, few techniques are actually implemented and announced in service form for general users to readily access and use the techniques. In this paper, we design and implement a protein interaction prediction service system based on the domain combination based protein-protein interaction prediction technique, which is known to show superior accuracy to other conventional computational protein-protein interaction prediction methods. In the prediction accuracy test of the method, high sensitivity(77%) and specificity(95%) are achieved for test protein pairs containing common domains with learning sets of proteins in a Yeast. The stability of the method is also manifested through the testing over DIP CORE, HMS-PCI, and TAP data. Performance, openness and flexibility are the major design goals and they are achieved by adopting parallel execution techniques, web Services standards, and layered architecture respectively. In this paper, several representative user interfaces of the system are also introduced with comprehensive usage guides.

Key words : PreSPI, protein-protein interaction prediction, domain combination, web services, AP matrix, primary interaction probability

1. 서 론

계산을 통한 단백질 상호작용 예측 기법의 효용성 및 중요성에 관한 인식이 확산되면서 단백질 상호작용을 계산적으로 예측하는 다양한 기법이 제안되었다[1-4]. 별도의 처리를 하지 않은 단백질 서열만을 가지고 직접 단백질-단백질 상호작용에 영향을 끼치는 부분 서열을 찾아 분석하는 것이 한 가지 접근 방법[5]이며, 단백질의 구조나 물리·화학적 특성을 분석함으로써 단백질 상호작용을 예측하는 방법도 알려져 있다[6]. 도메인 기

† 종신회원 : 한국정보통신대학교 공학부 교수
dshan@icu.ac.kr
†† 정 회 원 : 한국전자통신연구원 이동통신연구단 연구원
kimkk@etri.re.kr
††† 학생회원 : 한국정보통신대학교 공학부
torajim@icu.ac.kr
†††† 정 회 원 : 한국정보통신대학교 공학부 교수
sdlee@icu.ac.kr
논문접수 : 2005년 1월 18일
심사완료 : 2005년 9월 8일

반 단백질-단백질 상호작용 예측도 또 하나의 접근 방법이 될 수 있으며, 현재 여러 연구진들에 의하여 활발히 연구되어지고 있다[1,4,7].

그러나 지금까지의 대부분의 연구는 주로 예측 방법 고안에 머무르고 있으며, 극히 일부만이 일반 사용자를 위한 제한된 서비스만을 제공하고 있다. 그 이유는 계산을 통한 단백질 상호작용 예측 연구가 비교적 초기 단계라는 점과 제안된 예측 기법의 정확도가 서비스를 하기에는 만족스럽지 못하다는 점에 있다.

최근 Han을 중심으로 한 연구그룹에서 제안한 도메인 조합 기반(domain combination based) 단백질 상호작용 예측 기법은 그 완성도 면에서나 예측 정확도 측면에서 일반 사용자(특히, 생물학자)에게 서비스할 정도의 우수성을 보여주고 있다[8-10]. 이 예측 기법은 기존의 도메인 기반(domain based) 단백질 상호작용 예측 기법 보다 세련되고 신뢰성이 있다. 본 논문에서는 도메인 조합 기반 예측 기법의 우수성을 확인하기 위하여 이전에 실험과는 다른 데이타와 조건하에서 예측 기법의 정확도를 다시 측정하였다. 효모(Yeast)의 단백질 집합에 대하여 학습한 후, 학습된 단백질 집합과 공통된 도메인을 가지지만 학습 집합에 존재하지 않는 단백질 쌍들에 예측 기법을 적용하여 매우 높은 77%의 민감도(sensitivity)와 95%의 특이도(specificity)를 보였다. 이와 함께 DIP CORE, HMS-PCI, TAP 데이타를 이용한 평가를 통해서 이 기법의 안정성을 확인하였다. 비록 테스트 결과는 효모 생물에 한정되었지만 이와 같은 정확도와 안정성을 갖는 예측 시스템은 생물학자에게 유익한 정보를 제공할 수 있다.

본 논문에서는 이 기법을 이용하여 PreSPI(Prediction System for Protein Interaction)로 명명된 단백질 상호작용 예측 서비스 시스템을 설계하고 구현한다. PreSPI의 구현을 위하여 먼저 서비스 기능들을 나열하고, 설계 목표를 정했다. 그리고 서비스 기능들의 효과적인 적용과 설계 목적을 달성하기 위해서 PreSPI의 소프트웨어 구조가 고안되었다.

PreSPI의 주요 기능은 단백질 간 상호작용 가능성을 예측하는 것이지만, 이외에도 단백질 또는 단백질 상호작용의 연구자들을 위한 다양한 보조 기능들도 지원되어야 한다. PreSPI가 제공하여 주는 기능은 크게 핵심 기능 및 부가 기능 그리고 일반 서비스 등으로 분류된다. 본 논문에서 PreSPI의 상세한 설명 및 각 카테고리 내 기능들의 사용법을 서술하였다.

핵심 기능은 논문[8]에서 소개한 도메인 조합 기반 단백질 상호 작용 예측 기법을 서비스 형태로 만들어 제공하는 기능으로 입력 단백질 쌍에 대한 상호작용 예측, 상호 작용 확률 값 분포 표시, 복수의 단백질 쌍에 대한 카테고리 결정 및 상호작용 가능성 순위부여 기능 등을 포함한다. 부가 기능은 핵심 기능으로부터 파생되는 기능으로 도메인 조합 출현 확률 행렬 상에서 높은 값을 갖는 도메인 조합 쌍 검색 및 본 논문에서 소개한 기법을 이용하여 예측한 단백질 상호작용 데이타에 기반한 다양한 단백질 상호작용 네트워크 구성 그리고 예측 시스템 정확도 제시 기능 등을 포함한다. 일반 서비스 기능은 단백질 상호작용에 관하여 연구하는 연구자에게 도움이 되는 일반적인 기능을 모은 것으로 인터넷상에서 주어진 단백질 및 도메인 데이타에 대한 수집 및 구성 수행 등을 포함한다. 핵심 기능 및 부가 기능이 도메인 조합 기반 단백질 상호 작용 예측 기법에 준하여 제공하는 기능인 반면에 일반 서비스 기능은 이들 기법과는 직접적인 연관이 없이 제공될 수 있다는 점에서 구분된다.

일반적으로 단백질 상호작용에 관한 계산적 시도는 다양한 포맷으로 지속적으로 갱신되는 분산된 데이타를 기반으로 대량의 계산이 필요하다. 또한 PreSPI가 제공하여 주는 서비스는 그것이 곧바로 생물학자가 요구하는 최종적인 답을 제공하기 보다는 그러한 정보를 이용하여 다양한 시도를 할 수 있도록 하는 단초를 제공하여 주는 것이다. 따라서 다른 응용 프로그램이나 외부의 시스템이 예측 시스템의 서비스에 손쉽게 접근할 수 있는 수단을 제공하여 주는 것이 필요하다.

위와 같은 요구 사항을 반영하여 본 논문에서 소개하는 PreSPI는 시스템의 성능과 개방성 그리고 확장성을 목표로 설계되고 구현되었다. PreSPI의 성능은 일부 서비스를 병렬 처리함으로써 확보하였고, 시스템의 개방성은 웹 서비스 표준 기술[11-13]을 도입하고, PreSPI가 제공하는 기능을 웹 서비스 API(application program interface) 형태로 제공함으로써 외부의 응용 프로그램이나 시스템이 손쉽게 접근할 수 있도록 지원하고 있다. PreSPI의 확장성 및 유연성을 위해서는 시스템을 데이타 모듈과 서비스 모듈로 명확히 구분하는 계층적 구조를 사용하여 새롭게 갱신되는 데이타를 주기적으로 시스템에 반영할 수 있도록 구성하였다.

본 논문은 구성은 다음과 같다. 먼저, 2장에서는 본 논문에서 설계 및 구현 대상으로 삼고 있는 도메인 조합 기반 단백질 상호작용 예측 기법에 관해서 간략히 소개한다. 3장에서는 PreSPI 시스템의 기능을 소개하고 시스템의 구성에 관해서 기술한다. 4장에서는 PreSPI 시스템이 제공하는 사용자 인터페이스를 중심으로 PreSPI가 제공하는 기능과 사용에 관해서 소개하고, 마지막으로 5장에서 결론을 내린다.

2. 도메인 조합 기반 단백질 상호작용 예측 기법

본 장에서는 본 논문에서 대상으로 삼고 있는 도메인 조합 기반 단백질 상호작용 예측 기법에 관하여 간략히 소개한다. 상세한 도메인 조합 기반 단백질 상호작용 예측 기법은 참고문헌[8-10]에 소개되어 있다.

2.1 제안 배경

도메인 조합 기반 단백질 상호작용 예측 기법은 그동안 활발하게 연구되어진 도메인에 기반한 단백질-단백질 상호작용 예측 기법의 발전된 형태로 볼 수 있다[1,4,7]. 도메인 기반 예측 기법의 결점은 도메인 조합 기반 단백질 상호작용 예측 기법에 의해서 해결된다. 도메인 기반 예측 기법들은 단백질-단백질 상호작용 데이타로부터 도메인-도메인 상호작용 정보를 추측하고, 이를 토대로 단백질의 상호작용을 예측하는 것이 일반적이다. 하지만 도메인에 기반한 대부분의 기존 연구들은 계산의 편의상, 단백질의 상호작용이 독립적으로 발생하는 단일 도메인 쌍(single domain pair)의 결합에 의해 유발된다고 가정하고 있다. 그 결과 기존의 도메인에 기반한 단백질 상호작용 예측 기법의 정확도가 높지 않았다.

기존의 도메인에 기반한 단백질 상호작용 예측 기법이 낮은 예측 정확도를 보이는 것은 많은 이유가 있을 수 있겠지만 위에서 언급한 단백질의 상호작용이 독립적으로 발생하는 단일 도메인 쌍의 결합에 의해 유발된다는 가정에 오류가 있는 것으로 추정된다. 즉 단일 도메인 쌍 보다는 복수의 도메인들이 합동으로 단백질 상호작용에 영향을 미친다고 가정하는 것이 적절할 것으로 판단된다. 이러한 문제점을 극복하기 위하여 도메인 조합 기반 단백질 상호작용 예측 기법에서는 도메인 조합(domain combination)과 도메인 조합 쌍(domain combinations pair 또는 dc-pair)의 개념을 도입한다. 도메인 조합이란 용어는 하나의 도메인 집합에서 생성 가능한 도메인 부분 집합을 의미한다. 도메인 조합에 기반한 단백질-단백질 상호작용은 복수의 도메인 쌍이나 도메인 조합 간 상호작용의 결과로 인식하며, dc-pair를 단백질 상호작용의 기본 단위로 해석한다. 도메인 조합 기반 해결 방법과 도메인 기반 해결 방법의 대조는 참고문헌[8]에 잘 설명되어 있다.

2.2 도메인 조합 기반 단백질 상호작용 예측 기법

도메인 조합 기반 단백질 상호작용 예측 기법에서는 상호작용이 있는 것으로 알려진 단백질 쌍 집합에 있는 도메인 조합 쌍의 출현 확률 (appearance probability: AP) 행렬과 임의로 짝지어진 상호작용이 없는 것으로 가정된 단백질 쌍 집합에 있는 도메인 조합 쌍 정보의 출현 확률 행렬을 작성한다. 그 후 각각의 출현 확률 행렬을 기반으로 각 단백질 쌍 0과 1 사이의 실수 값으로 대응시키는 확률 함수를 고안한다. 다음으로, 상호작용이 있는 것으로 알려진 단백질 쌍에 고안된 함수를 적용하여 얻어지는 값의 분포와 상호작용이 없는 것으로 가정된 단백질 쌍에 고안된 함수를 적용하여 얻어지는 값의 분포를 각각 생성한다. 본 논문에서 이 실수를 PIP (primary interaction probability) 값이라 부른다. 모든 상호작용이 있는 단백질 및 상호작용이 없는 단백질 쌍 집합에 있는 모든 단백질 쌍에 대해서 PIP 값을 구하는 함수가 적용된다. 이 결과로 얻은 두 PIP 분포를 이용하여 미지의 새로운 단백질 쌍이 주어지면 PIP 값을 얻고, 얻어진 PIP 값이 어느 PIP 분포에 속하게 되는지 판단하여 그 상호작용 가능성을 결정한다[8-10].

2.3 검증 결과

논문[8-10]에 따르면, 도메인 조합 기반 단백질 상호작용 예측 기법은 기존의 도메인 기반 예측 기법에 비하여 현저히 향상된 예측 정확도를 보여준다. 그러나 이들 기법에 대한 기존의 검증은 실험에 사용된 단백질 쌍과 구성된 AP 행렬 간에 공유 도메인이 없을 때 그 응용에 있어서 무의미하다는 점에서 착오가 있었다. 다시 말하면, 단백질 쌍과 AP 행렬 간에 겹쳐지는 부분이 없을 때 도메인 또는 도메인 조합 기반 단백질 상호작용 예측 기법은 적용되지 않았어야 한다. 따라서 본 기법의 검증에서는 단백질 쌍의 실험 집합 내에서 AP 행렬과 겹쳐지지 않는 도메인을 포함한 모든 단백질 쌍을 제거하였다. 준비한 상호작용 및 상호작용이 없는 단백질 쌍의 학습 집합에 대하여 사전 검증을 통하여 엄밀하게 위와 같은 방법을 사용하였다. 검증을 위하여 우리는 두 가지의 단백질 쌍 집합을 이용하였다. 하나는 DIP 데이타베이스 (http://dip.doe-mbi.ucla.edu) [16]에서 구할 수 있는 상호작용 단백질 쌍의 집합으로 효모를 기반으로 한 15,174개의 상호작용 쌍이 사용되었다. 그러나 현재 도메인 정보가 알려진 단백질 쌍만이 사용 가능 하기 때문에 7,500 개의 상호작용 단백질 쌍이 검증에서 사용되었다. 단백질에 대한 도메인 정보는 PDB(http://www.ebi.ac.uk/proteome/) [14,15]에서 추출하였다.

반면 상호작용이 없다고 추정되는 단백질 쌍은 효모 유기체의 도메인 정보를 가진 임의의 쌍에서 인위적으로 생성되었다. 아직까지 상호작용이 없는 단백질 쌍에 대하여 공표된 정보가 없다. 효모는 개략 6,000개의 단백질이 있는 것으로 알려져 있다. 그 가운데 2,700개의 단백질이 도메인 정보를 가지고 있으며, 단백질 쌍의 상호작용이 없는 집합의 생성에 사용 될 수 있었다. 2,700개의 단백질을 임의로 분할하여 12,700개의 단백질 쌍이 생성되었다. 상호작용이 없는 단백질 쌍이 준비되었을

때 도메인 정보가 알려진 단백질 쌍 집단에서 상호작용이 있는 것으로 보고된 단백질 쌍 집단을 제거하는 방식을 통하여 임의로 생성되었고, 상호작용이 없는 단백질 쌍의 같은 개수가 초기 계산에 사용되었다.

기법의 예측 정확도를 평가하기 위하여 우선적으로 상호작용 및 상호작용이 없는 집합을 준비한 후, 집합의 각각을 학습 및 검증 집단으로 나누었다. 데이타 중에서 학습 집단으로 80%를 사용하였고, 20%를 검증 집단으로 사용하였다.

한편, 자연계에는 상호작용이 있는 단백질 쌍보다 상호작용이 없는 단백질 쌍이 더 많이 존재하는 것이 일반적이기 때문에 본 논문에서는 상호작용이 없는 단백질 쌍 집합의 크기를 점점 증가시켜 가면서 재평가를 시도하였다. 참고로 AP 행렬 내에서 겹치는 도메인이 없는 단백질 쌍은 측정하기 위한 평가 데이타에 포함시키지 않았다.

표 1 테스트 그룹의 민감도와 특이도

	Ratio	1.0	2.0	5.0	10.0
I	Sensitivity	96.77	92.96	85.98	78.73
	Specificity	73.20	83.62	91.03	95.00
II	Sensitivity	69.70	76.74	61.19	31.15
	Specificity	62.16	64.58	76.36	81.67
Total	Sensitivity	95.93	92.27	85.08	76.95
	Specificity	73.07	83.32	90.73	94.65

Note I: 일치 PIP 값을 갖는 단백질 쌍의 경우
II: 비 일치 PIP 값을 갖는 단백질 쌍의 경우

표 1은 단백질 쌍의 상호작용이 있는 집합과 상호작용이 없는 집합의 비율에 의한 각 테스트 그룹의 민감도(sensitivity)와 특이도(specificity)를 나타낸다. 각각의 그룹은 PIP 값 분포에 일치하는 값의 존재 여부에 따라 두 가지 하위 그룹으로 나뉘어 진다. 표 1에서 살펴보면 PIP 분포 내 일치하지 않는 PIP 값을 가지는 테스트 그룹(II)은 보통의 민감도와 특이도를 보여주는 반면에 일치하는 PIP 값을 가지는 테스트 그룹(I)은 상당히 높은 정확도를 보여준다. 실험에서 AP 행렬 상의 공통된 도메인을 가지는 단백질 쌍들은 PIP 값 역시 일치하는 경향을 보이고 있으며, 공통된 도메인을 가지는 단백질 쌍들 중 5% 미만의 쌍들만이 일치하지 않는 PIP 값을 가진다.

전반적으로 상호작용이 없는 단백질 쌍 집합의 상대적 비율이 증가함에 따라 상호작용 예측 정확도는 점차 개선된다. 상호작용이 없는 단백질 쌍 집합의 크기가 있는 집합 크기보다 10배인 경우, AP 행렬에 공통되는 도메인이 있는 테스트 단백질 쌍 들은 77%의 민감도와 95%의 특이도를 나타내고 있다.

또한 본 기법이 DIP 이외의 다른 데이타 집합에서도 안정된 예측 정확도를 제공하는지 확인하기 위해서 DIP CORE[16], HMS-PCI[17], TAP[18] 데이타를 이용하여 기법의 예측 정확도를 측정하였다. 표 2는 각 테스트 그룹의 민감도와 특이도를 나타낸다.

표 2 DIP, DIP CORE, HMS-PCI, TAP를 이용한 실험의 민감도와 특이도

	Ratio	1.0	2.0	5.0	10.0
DIP	Sensitivity	96.77	92.96	85.98	78.73
	Specificity	73.20	83.62	91.02	95.00
DIP CORE	Sensitivity	97.89	97.19	95.40	90.50
	Specificity	70.23	90.77	89.76	95.21
HMS-PCI	Sensitivity	94.64	96.98	95.71	93.08
	Specificity	62.50	72.92	91.96	93.91
TAP	Sensitivity	92.70	97.23	98.30	97.66
	Specificity	86.67	79.43	97.70	98.60

Note: 일치 PIP 값을 갖는 단백질 쌍의 경우

표 2에서 보면 데이타 집합에 관계없이 상당히 안정되고 높은 예측 정확도를 얻었다. 비율이 10배일 경우 DIP 데이타를 이용한 정확도는 다른 데이타의 정확도보다 낮다. 이것은 DIP 데이타가 다른 데이타 소스보다 더 많이 오류 데이타를 갖고 있음을 간접적으로 나타낸다. 이와 반면, TAP 데이타를 이용한 예측 정확도는 거의 완벽함을 알 수 있다. 이와 같은 결과에 의해서 도메인 조합 기반 단백질-단백질 상호작용 예측 서비스 시스템은 AP 행렬에서 겹치는 도메인을 갖는 단백질 쌍에 대하여 상당히 신뢰성 있는 계산적 예측 단백질-단백질 상호작용 정보를 제공하는 것으로 결론 내릴 수 있다.

3. PreSPI

도메인 조합 기반 단백질 상호작용 기법이 일반 사용자 또는 생물학자가 사용하기 위해서는 이들이 쉽게 접근할 수 있도록 서비스 시스템을 만들 필요가 있다. 3장에서는 2장에서 설명한 기법의 적용에 의해서 구성된 단백질 상호작용 예측 서비스 시스템을 소개하고, 단백질 및 단백질 상호작용에 관한 유익한 내용들을 소개한다.

3.1 PreSPI 기능

서론에서 소개한 바와 같이 PreSPI가 제공하는 기능은 크게 핵심 기능, 부가 기능 그리고 일반 서비스 등으로 분류된다. 핵심 기능은 단지 PreSPI에서 서비스 형태로 만들어 제공하는 기능이다. 이 기능은 단일 단백질 쌍에 대한 상호작용 예측, PIP 분포 표시, 복수의 단백질 쌍에 대한 카테고리 결정 그리고 상호작용 가능성 서열 결정 기능 등을 포함한다. 부가 기능은 핵심 기능

으로부터 파생되는 기능으로 AP 행렬 상에서 높은 값을 갖는 dc-pair 검색, 본 논문에서 소개한 기법을 이용하여 예측한 단백질 상호작용 데이타에 기반한 다양한 단백질 상호작용 네트워크 구성 및 예측 시스템 정확도 제시 기능 등을 포함한다. 일반 서비스 기능은 단백질 상호작용에 관하여 연구하는 연구자에게 도움이 되는 일반적인 기능을 모은 것으로 주어진 단백질에 대한 도메인 정보 검색 기능 및 DIP_ID, SWISSPROT_ID, PIRID 등의 Accession_ID를 상호 변환해 주는 기능 등을 포함한다. 핵심 기능 및 부가 기능이 논문[9,10]에서 소개한 기법을 구현하는 PreSPI 만이 제공할 수 있는 독특한 기능이다. 반면에 일반 서비스 기능은 본 논문에서 소개한 기법과는 직접적인 연관이 없이 단백질 상호작용에 관하여 연구하는 연구자에게 도움이 되는 일반적인 기능을 모은 것으로 인터넷상에서 주어진 단백질 및 도메인 데이타에 대한 수집 및 구성 수행 등을 포함한다. 표 3은 PreSPI가 제공하는 서비스 기능들을 위의 분류에 따라서 정리한 내용이다. 이 기능들은 PreSPI 웹사이트(http://prespi.icu.ac.kr)에서 자세하게 표현되어 있다.

3.2 PreSPI 구조

일반적으로 계산을 통한 단백질 상호작용 예측 시스템은 다양한 포맷으로 지속적으로 갱신되는 분산된 데이타를 기반으로 대량의 계산을 통하여 예측을 시도한다. 따라서 많은 경우 예측에 많은 시간이 소요되는 것이 보통이다. PreSPI는 예측을 위해서 종별로 약 10억 개 이상의 요소를 갖는 AP 행렬을 생성하고, 다시 학습에 사용된 단백질 쌍을 입력으로 한 PIP 분포를 얻는다. 이와 같이 서비스를 위한 준비가 완료된 뒤에는 3.1절에서 소개한 다양한 기능을 제공할 수 있는데 서비스에 소요되는 시간은 제공하는 기능에 따라서 수초에서 수십 시간이 소요되기도 한다.

또한 PreSPI가 제공하여 주는 서비스는 그것이 곧바로 생물학자가 요구하는 최종적인 답을 제공하기 보다

표 3 PreSPI의 서비스 기능

분류	세부적 기능	설명
핵심 기능 (Core Functions)	PIP 값 분포 표시	이 기능은 상호작용이 있는 것으로 알려진 단백질 쌍과 임의로 조합된 단백질 쌍의 PIP 값 분포를 대비하여 보여주는 기능으로 각 집단의 색을 달리하여 그 분포를 보여준다.
	단일 단백질 쌍의 상호작용가능성 예측	이 기능은 도메인 정보가 알려진 두 단백질의 상호작용 가능성을 도메인 조합 정보에 기반하여 예측해 주는 기능이다. 입력 단백질 쌍에 대한 PIP 값을 계산하여 PIP 분포 상에 나타내고, 시스템의 상호작용 가능성 판단 결과를 보여준다. 입력 단백질의 구별은 DIP_ID, SWISSPROT_ID, PIR_ID 등을 사용하는 것이 가능하다.
	복수 단백질 쌍의 상호작용 가능성 예측 및 서열 결정	이 기능은 복수의 단백질 쌍에 대해서 이들의 상호작용 가능성을 예측함과 동시에 어느 단백질 쌍이 더 상호작용을 일으킬 가능성이 높은지를 예측하는 기능이다. 입력된 복수의 단백질 쌍에 대해서 PIP 값들을 계산하여 PIP 분포 상에 나타내고 이들의 상호작용 가능성의 서열을 결정하여 보여준다. 이 과정에서 단백질 쌍들은 논문[9,10] 에서 제시한 방식으로 분류한 뒤 그 서열을 결정한다.
	상호작용 단백질 검색	이 기능은 주어진 하나의 단백질과 상호작용 할 가능성이 있는 단백질 들을 찾고 이것들을 그 서열에 의하여 상호작용 확률과 함께 리스트 형태로 출력한다.
부가 기능 (Subsidiary Functions)	도메인 조합 생성	이 기능은 주어진 단백질 쌍으로부터 생성 가능한 도메인 조합 쌍을 생성하여 보여준다.
	dc-pair 검색	이 기능은 상호작용이 있는 것으로 알려진 단백질 쌍 집합의 AP 행렬 상에 나타나는 dc-pair 중 큰 값을 갖는 300개의 dc-pair 리스트를 보여주고, 특정 dc-pair에 대해서는 AP 행렬 상의 해당 값을 찾아준다. 또한 특정 도메인 조합을 포함하면서 특정 값 이상을 갖는 dc-pair에 대해서도 리스트 형태로 출력하는 것이 가능하다. 생물학자는 이것을 통해 단백질 상호작용에 영향을 미치는 주요 도메인 또는 도메인 조합 쌍에 대한 정보를 얻을 수 있다.
	단백질 상호작용 네트워크 구성	이 기능은 PreSPI에 의해 상호작용이 있는 것으로 예측된 단백질 쌍을 기반으로 구성된 단백질 상호작용 네트워크를 보여준다. 사용자가 입력하는 확률 이상의 상호작용 가능성이 있는 단백질 쌍만을 대상으로 네트워크를 구성하는 것도 가능하며 주어진 쌍(two or more) 이상의 상호작용이 있는 것으로 추정되는 단백질만을 대상으로 네트워크를 구성하는 것도 가능하다.
	예측 시스템 정확도 표시	이 기능은 도메인 조합에 기반한 단백질 상호 작용 예측 기법의 정확도를 효모 단백질을 대상으로 검증하였을 때 결과를 표시하는 기능으로 상호작용이 있는 것으로 알려진 단백질 쌍 집단과 임의로 만들어진 단백질 쌍 집단의 비율 변화에 따른 정확도 변화를 보여준다.
일반 서비스 기능 (General Service Functions)	도메인 정보 검색	이 기능은 주어진 특정 단백질이 가지고 있는 도메인 정보를 검색하여 그 결과를 출력한다.
	단백질 정보 검색	이 기능은 주어진 특정 도메인을 포함하고 있는 단백질을 검색한 결과를 리스트 형태로 출력한다.
	Accession_ID 상호변환	이 기능은 DIP_ID, SWISSPROT_ID, PIR_ID를 상호 변환한 결과를 출력한다.

는 그러한 정보를 이용하여 다양한 시도를 할 수 있도록 하는 단초를 제공하여 주는 역할을 수행할 수 있어야 한다. 따라서 다른 응용 프로그램이나 외부 시스템이 본 논문에서 제안한 예측 시스템의 서비스에 손쉽게 접근할 수 있는 수단을 제공하도록 설계되어야 한다.

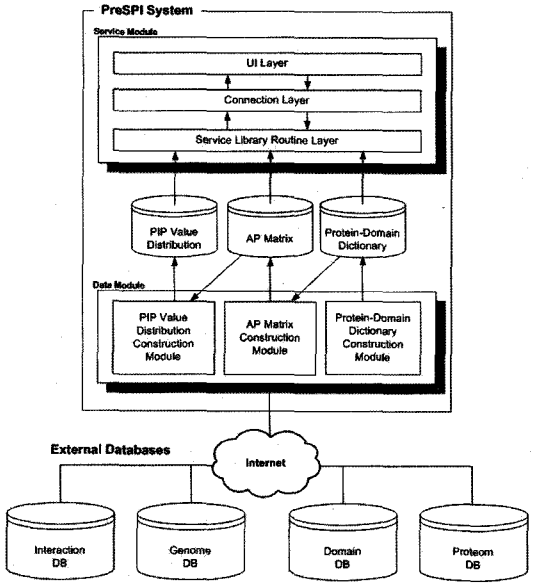

그림 1 PreSPI의 소프트웨어 구조

PreSPI는 위와 같은 요구 사항을 반영하여 시스템의 성능 및 확장성 그리고 개방성을 달성하는 것을 목표로 설계하고 구현하였다. 시스템의 성능은 일부 서비스 기능을 병렬 처리함으로써 확보하였고, 개방성은 웹 서비스 표준 기술을 도입하고 시스템이 제공하는 기능을 웹 서비스화 함으로써 외부의 응용 프로그램이나 시스템이 손쉽게 접근할 수 있도록 하였다. 시스템의 확장성을 위해서는 시스템을 데이타 모듈과 서비스 모듈로 명확히 구분하고 새롭게 갱신되는 데이타를 주기적으로 시스템에 반영할 수 있도록 구성하였다. 그림 1은 이러한 목표를 달성하기 위하여 설계된 PreSPI의 소프트웨어 구조를 보여주고 있다.

PreSPI는 크게 분류하면 서비스 준비를 위해서 관련 데이타베이스를 생성하는 데이타 모듈(Data Module)과 준비된 데이타베이스에 기반하여 사용자로부터 서비스 요청을 받아서 서비스하는 서비스 모듈(Service Module)로 구성되어 있다. 데이타 모듈은 다시 인터넷상에 산재해 있는 각종 단백질 및 도메인 관련 데이타를 모아서 단백질-도메인 사전을 구축하는 단백질-도메인 사전 구축 모듈(Protein-Domain Dictionary Construction Module), 상호작용이 있는 단백질 쌍 집합과 임의의 단백질 쌍 집합으로부터 AP 행렬을 생성하여 데이타베이스에 저장하는 AP 행렬 생성 모듈(AP Matrix Construction Module) 그리고 PIP 함수를 각 AP 행렬의 원소에 적용하여 PIP 값 분포를 얻어 데이타베이스에 저장하는 PIP 값 분포 생성 모듈(PIP Value Distribution Construction Module)을 포함한다. 이와 같이 세 가지 데이타베이스가 데이타 모듈에 의해서 생성이 완료되면 PreSPI는 사용자로부터 각종 서비스 요구를 받아서 처리할 초기 준비가 1차 완료된 상태로 볼 수 있다.

서비스 모듈은 데이타 모듈에 의해서 생성된 세 개의 데이타베이스를 기반으로 표 3 에서 제시된 기능 등을 제공하는 데 필요한 루틴 등을 준비하여 사용자의 서비스 요청에 대응한다. 서비스 모듈은 크게 사용자로부터 서비스 요청을 받고 그 서비스 결과를 보여주는 UI 층(UI Layer), 사용자 서비스 요청을 받은 후 해석하여 그것에 해당하는 서비스 루틴을 찾고 구동하는 연결 층(Connection Layer) 그리고 해당 서비스 루틴 들을 포함하고 있는 서비스 라이브러리 루틴 층(Service Library Routine Layer)으로 구성된다. PreSPI는 서비스 모듈과 데이타 모듈을 분리함으로써 인터넷상에 새로운 데이타가 추가되는 경우에 유연하게 대처할 수 있고 사용자의 새로운 서비스 요청에도 비교적 용이하게 대처할 수 장점을 가진다는 점이다. 즉 새로운 서비스 요청에 대해서는 서비스 라이브러리에 해당 루틴을 추가하고 관련 루틴을 UI 및 연결 층에 추가함으로써 층 구조의 특징을 유지시키는 것이 가능하다. 또한 인터넷상의 새로운 데이타의 추가는 서비스 모듈에는 영향을 미치지 않고 데이타 모듈의 해당 부분만을 확장함으로써 수용 가능하다.

한편 PreSPI는 많은 기능을 웹 서비스 API 형태로 제공함으로써 응용 프로그램이나 외부 시스템에 의하여 손쉽게 접근할 수 있는 개방성을 지원한다. 표 4는 PreSPI에 의하여 외부에 제공되는 API를 정리한 것이다. 그 외에 WSDL(Web Service Definition Language)은 본 시스템이 제공하고 있는 웹 사이트(http://prespi.icu.ac.kr)에서 자세히 살펴 볼 수 있다.

4. 구현

본 장에서는 본 논문에서 구현된 PreSPI 시스템의 대표적인 사용자 인터페이스 및 사용법에 관하여 간략히 소개한다. PreSPI 시스템의 각 페이지는 페이지 표현과 사용 설명서 형태로 구성하여 사용자가 페이지 표현을 통하여 페이지의 의미를 파악하고 사용 설명서를 참고하여 PreSPI가 제공하는 기능을 사용할 수 있도록 하였다.

PreSPI의 구현을 위하여 Python, Java 그리고 웹 서비스 기술을 사용하였다. PreSPI 시스템의 구현에 있어

표 4 PreSPI에 의하여 외부에 제공되는 API

이름	파라미터	설명
getDomain()	In: String protein, int kind	kind 타입의 단백질에 대하여 해당 도메인을 반환한다. (0: DIP, 1: SWISSPROT, 2: PIR, 3: GI, -1: 도메인 정보를 가지고 있지 않음)
	Out: String	
checkProPair()	In: String proteinA, String proteinB	단백질A와 단백질B의 PreSPI에서 계산가능 여부를 알려준다. (0: 도메인 정보를 가지고 있음, 1: 도메인 정보를 가지고 있으며 실험적으로 상호작용 함으로 증명된 쌍임, -1: 도메인 정보를 가지고 있지 않음)
	Out: int	
transID()	In: String protein, int from, int to	from 타입의 단백질을 to 형태의 ID로 변환한다. (0: DIP, 1: SWISSPROT, 2: PIR, 3: GI, ID가 없을 경우 -1 리턴)
	Out: String	
getDomainCombination()	In: String protein, int kind	kind 타입의 단백질에 대하여 해당 도메인의 멱집합(DC)을 반환한다. (0: DIP, 1: SWISSPROT, 2: PIR, 3: GI, -1: 도메인 정보를 가지고 있지 않음
	Out: String	
getDomainCombinationPairs()	In: String proteinA, String proteinB, int aKind, int bKind	aKind 타입의 단백질A와 bKind 타입의 단백질B에 대하여 DC-쌍을 반환한다. (0: DIP, 1: SWISSPROT, 2: PIR, 3: GI, 도메인 정보가 없을 경우 size가 0인 Vector 반환)
	Out: java.util.Vector	
getProteinsForGivenDomains()	In: String[] domains, int-andOr	도메인들을 입력받아, andOr조건으로 검색하여 포함하고 있는 단백질을 되돌린다. (0: or, 1:and)
	Out: java.util.Vector	
getPIPValue()	In: String proteinA, String proteinB	단백질A와 단백질B를 입력받아 PIP 값을 되돌린다. ID는 DIP 형태를 지원한다.
	Out: String	
getInterProb()	In: String pipValue	PIP 값을 입력받아 상호작용 확률을 되돌린다.
	Out: String	
getInterProb()	In: String proteinA, String proteinB	단백질A와 단백질B를 입력받아 상호작용 확률을 되돌린다. (-1: 도메인정보를 가지고 있지 않음)
	Out: String	
getIsInPIPDist()	In: String pipValue	pipValue를 입력받아 PIP 분포에 존재하는지 여부를 되돌린다.
	Out: boolean	
getIsInPIPDist()	In: String proteinA, String proteinB	단백질A와 단백질B를 입력받아 PIP 분포에 존재하는지 여부를 되돌린다.
	Out: boolean	
getIsProved()	In: String proteinA, String proteinB	단백질A와 단백질B를 입력받아 실험적으로 상호작용이 증명되었는지 여부를 되돌린다.
	Out: boolean	
getInteractingPairs()	In:	PreSPI에서 사용하고 있는 상호작용 쌍의 DIP ID와 도메인들을 되돌린다.
	Out: java.util.Vector	
getAllProteins()	In:	PreSPI에서 사용하고 있는 모든 단백질의 DIP ID를 되돌린다.
	Out: java.util.Vector	

서 데이타 모듈 부분은 주로 Python 2.2.2를 사용하였으며, MySQL 3.23.5를 데이타베이스 관리 시스템으로 사용하였다. 사용자 인터페이스는 웹으로 제공하기 위하여 Java, Java 애플릿, JSP를 사용하여 구현하였으며, Jython으로 Java와 Python 모듈을 연결하였다. 또한 Jakarta-tomcat-4.1.24를 웹 서버로 사용하였으며, Java에서 MySQL 데이타베이스 접근을 위한 JDBC 드라이버로는 Mysql-connector-java-3.0.8-stable을 사용하였다.

PreSPI 웹 서비스 API는 웹 서버로 Apache-tomcat 내 Jakarta-tomcat을 사용하고, 컨테이너로 Axis-1_1을 사용하여 구축하였다. PreSPI 기능을 웹 페이지로 접근하길 원하는 일반 사용자는 PreSPI 서비스 웹 사이트를 방문하여 사용할 수 있다.

4.1 PIP 분포 시각화

PIP 분포 시각화 기능은 상호작용이 있는 것으로 알려진 단백질 쌍과 임의로 조합된 단백질 쌍의 PIP 값 분포를 대비하여 보여주는 기능으로 각 집단의 색을 달리하여 분포를 보여준다.

그림 2의 스냅 사진은 'Regular-Interval PIP distribution'을 사용한 두개의 집단의 PIP 값 분포를 보여주고 있다. 사용자는 이 분포를 대비해 봄으로써 두개의 집단이 PIP 값을 매개로 잘 분리될 수 있는지에 대하여

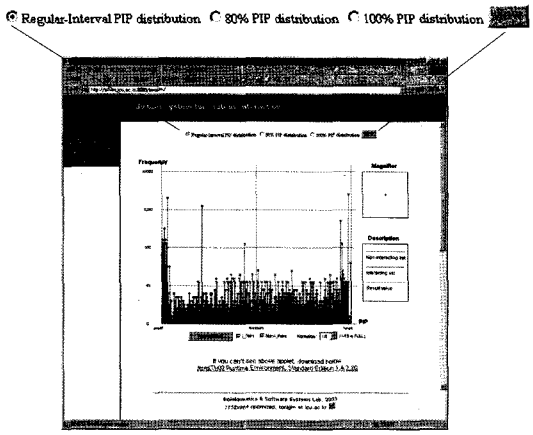

그림 2 일정 간격을 사용한 두개의 단백질 집단의 PIP 값 분포

직관적으로 판단할 수 있다. 일정 간격을 사용한 PIP 값 분포에서는 수평선 라인의 0과 1사이에 나타나는 서로 다른 PIP 값의 간격을 PIP 값에 관계없이 그 상대적 크기에 따라 동일한 간격으로 배열한다. 따라서 약 10,000여개의 서로 다른 PIP 값이 존재하는 경우에는 각 PIP 값의 간격은 1/10,000으로 일정하게 결정된다. 이와 같은 분포를 통해서 PIP 값이 상호작용이 있는 것으로 알려진 단백질 쌍 집단과 그렇지 않은 집단을 잘 분리해 주고 있는 지를 더 잘 표현할 수 있게 된다.

그림에서 오른쪽에 집중되어 있는 선(붉은 선)은 상호작용이 있는 것으로 알려진 단백질 쌍 집단의 PIP 값 분포를 나타내고, 왼쪽에 집중되어 있는 선(파란 선)은 임의로 짝지어진 즉 상호작용이 없는 것으로 알려진 단백질 쌍 집단의 PIP 값 분포를 나타낸다. 그림에서 보여 지는 바와 같이 두개의 집단은 PIP 값을 매개로해서 대체로 잘 분리되고 있음을 알 수 있다. 이것은 PIP 값이 두 개의 집단을 분리해 주는 분리자로서의 역할을 잘 수행하고 있음을 의미한다.

PIP 분포 시각화에서 수평축의 PIP 값을 일정 간격을 사용하지 않고 PIP 절대 값을 사용하기 위해서는 그림 상단의 '80% PIP distribution' 과 '100% PIP distribution' 버튼을 클릭하면 된다. '80% PIP distribution'은 학습 집단으로 상호작용이 있는 것으로 알려진 단백질 쌍 집합에서 80%를 사용한 경우의 PIP 분포이고, '100% PIP distribution'을 선택하면 100%를 사용한 경우의 PIP 분포를 보여준다.

4.2 단일 단백질 쌍 상호작용 예측

PreSPI의 단일 단백질 쌍 상호작용 예측 페이지에서 사용자는 도메인 정보가 알려진 두 단백질의 상호작용 가능성을 도메인 조합 정보에 기반하여 예측된 결과를 얻을 수 있다. 이 페이지에서는 입력 단백질 쌍에 대한 PIP 값을 계산하여 PIP 분포 상에 나타내고 시스템의 상호작용 가능성 판단 결과를 보여준다. 예측 결과는 각 단백질이 가지고 있는 도메인에 대한 정보, PIP 값, 계산된 PIP 값이 PIP 분포 상에 존재하는 지의 여부 그리고 실험적으로 상호작용이 있는지 확인된 것인지에 관한 정보 및 상호작용 확률 값으로 표시된다. 계산된 PIP 값이 PIP 분포 상에 존재하는 지의 여부는 계산된 PIP 값이 PIP 분포 상에 존재하는 경우 예측된 상호작용 확률 값의 신뢰도가 더 높다는 점에서 참고가 된다. 입력 단백질은 DIP_ID 외에도 SWISSPROT_ID, PIR_ID등을 사용하는 것이 가능하다.

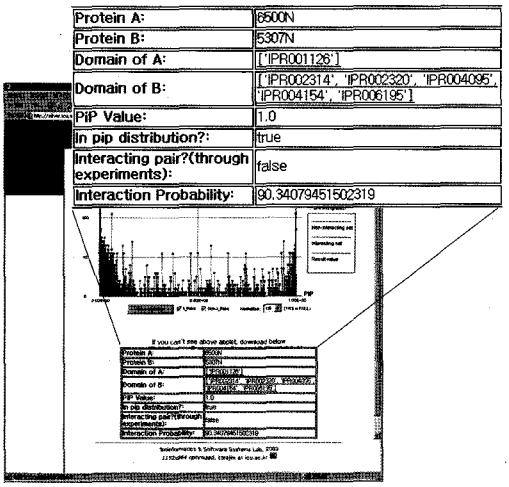

그림 3 6500N과 5307N의 상호작용 예측 결과

그림 3은 단일 단백질 쌍 <6500N(GID Number), 5307N>을 입력으로 하여 단일 단백질 쌍 상호작용 예측을 시행한 결과를 보여주고 있다. 그 결과 단백질 6500N은 하나의 도메인(IPR001126)을 갖고 단백질 5307N은 다섯 개의 도메인(IPR002314, IPR002320, IPR004095, IPR004154, IPR006195)을 갖는 것을 보여주고 있다. 한편 이 단백질 쌍의 PIP 값은 1.0 이고, In-pipdistribution 필드가 true 이어서 이 값은 PIP 분포 상에 존재하는 값임을 알 수 있다. 또한 이 단백질 쌍의 상호작용은 실험을 통하여 밝혀져 있지는 않은 상태이고, PreSPI를 통해서 계산적으로 예측한 이들의 상호작용 확률은 90.34%이어서 비교적 높은 상호작용 확률을 가지고 있는 것으로 나타나 있다. 이 값들은 비록 절대적으로 신뢰 가능한 것은 아니지만 하나의 단백질 쌍에 대한 상세하고 유용한 정보를 모아서 보여주고 있다.

4.3 복수 단백질 쌍 상호작용 예측

PreSPI의 복수 단백질 쌍 상호작용 예측 페이지에서

도 사용자는 복수의 단백질 쌍에 대해서 이들의 상호작용 가능성을 예측함과 동시에 어느 단백질 쌍이 더 상호작용을 일으킬 가능성이 높은지를 예측 할 수 있다. 복수 단백질 쌍 상호작용 예측 페이지에서 제공하는 기능을 이용하여 사용자는 많은 단백질 쌍에 대해서 하나씩 그 상호작용 가능성을 예측하고 비교하는 수고를 획기적으로 줄일 수 있다. 사용자는 이 페이지에서 복수의 입력 단백질 쌍을 테이블 형태로 주어지는 입력 필드에 입력하거나 미리 정해진 포맷으로 저장된 파일을 통하여 많은 단백질 쌍을 손쉽게 입력하고 그 결과를 받아볼 수 있다.

PreSPI는 복수의 단백질 쌍에 대한 예측을 단일 단백질 쌍 상호작용 예측과 동일하게 하나씩 실시하고 그 결과를 테이블에 나타내게 된다. 사용자는 예측 결과를 필드를 지정하여 상호작용 확률 또는 PIP 값의 순서로 손쉽게 입력 단백질 쌍을 나열하는 것도 가능하다.

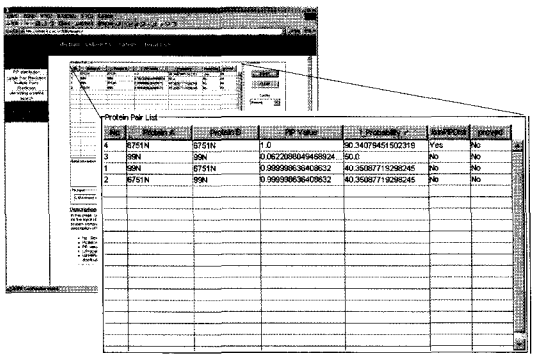

그림 4 단백질 쌍에 대한 상호작용 예측 결과

그림 4는 네 개의 단백질 쌍에 대해서 상호작용 가능성을 예측하고 그 결과를 상호작용 확률 값으로 정리하여 보여주고 있다. 마찬가지로 PreSPI가 제공하여 주는 이러한 상대적 순서가 절대적인 것은 아니지만 생물학자 들은 이러한 결과를 이용하여 복수의 단백질 쌍에서 어느 단백질 쌍의 상호작용 가능성이 상대적으로 높은지를 판단할 수 있는 단서를 찾을 수 있다.

4.4 상호작용 단백질 검색

상호작용 단백질 검색 페이지에서 사용자는 하나의 입력 단백질에 대해서 이것과 상호 작용할 가능성이 있는 단백질 들을 찾고 이것들을 그 서열에 의하여 상호작용 확률과 함께 리스트 형태로 출력하는 것이 가능하다. PreSPI는 입력으로 주어진 단백질과 나머지 단백질(효모의 경우에는 대략 2,700여개)과 쌍을 형성하여 각각의 단백질 쌍에 대한 상호작용 확률을 예측하고 그 결과를 정리하여 보여준다. 이 작업은 주어진 단백질과 쌍을 이루는 모든 단백질 쌍에 대해서 예측이 되어야 하는 만큼 많은 시간이 소요되는 것이 보통이다. 현재의 시스템에서는 하나의 단백질 쌍에 대한 상호작용 예측에는 약 5-8초가 소요되는 만큼 2,700여 쌍(효모의 경우)에 이르는 단백질 쌍에 대해서 모두 예측하는 데에는 약 6-7시간이 걸리게 된다. 그럼에도 불구하고 생물학자들은 PreSPI가 제공하는 이 기능을 통해서 관심이 있는 몇 개의 단백질과 상호작용을 일으킬 가능성이 있는 단백질에 관한 유용한 정보를 제공받을 수 있다.

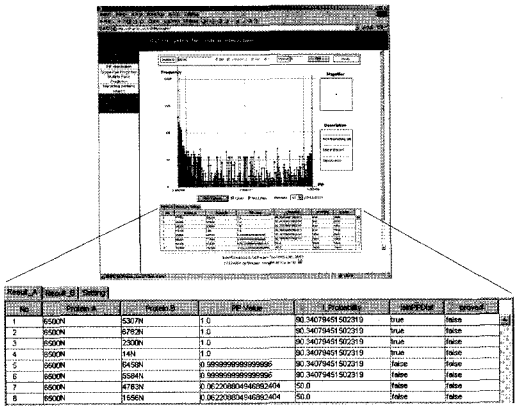

그림 5 상호작용 가능성이 있는 단백질 검출(6500N)

그림 5는 단백질 6500N에 대해서 이것과 상호 작용할 확률이 높으면서 PIP 값이 큰 8개의 단백질 쌍을 보여주는 화면이다. 화면은 약 30여분이 경과했을 때 약 2,800여 쌍의 대상 단백질 쌍 중에서 약 200여 쌍의 단백질의 상호작용 확률이 계산되었고, 그 중에서 8개의 단백질 쌍을 보여주고 있다. 8개를 제외한 나머지 190여 개의 단백질 쌍에 대한 예측 결과는 화면의 'Result B' 버튼을 누르면 볼 수 있다.

이 밖에도 PreSPI에는 빈번하게 도메인 및 도메인 조합에 관한 정보를 제공하는 페이지 등 많은 유용한 기능을 제공하고 있지만 지면 관계상 그 설명을 생략하기로 한다. 기타 기능에 대한 정보는 앞서 소개한 사이트를 방문하여 확인할 수 있다.

5. 결 론

본 논문에서는 도메인 조합에 기반 한 단백질 상호작용 예측 서비스 시스템인 PreSPI 시스템의 성능, 개방성 그리고 확장성 달성을 목표로 설계하고 구현하였다.

PreSPI 시스템을 구현한 결과, 예상대로 서비스 제공에 많은 시간이 소요됨을 확인할 수 있었으며 클러스터 시스템 상에서 시스템의 병렬화를 통하여 비교적 손쉽

게 시스템 성능을 향상 시킬 수 있음도 확인하였다.

구현된 PreSPI 시스템이 제공하는 기능은 사용자들이 필요로 하는 최종적인 답을 주기보다는 연구자들이 원하는 답을 찾는 과정에서 필요로 하는 유용한 정보를 제공한다고 볼 수 있다. 그런 점에서 PreSPI가 제공하는 기능은 다른 응용 프로그램 또는 외부 시스템과 손쉽게 상호 연결되는 것이 필요하다. 이를 위해서 PreSPI 시스템은 독립적인 서비스 또는 시스템 간의 상호 연결에 적절한 수단을 제공하는 웹 서비스 표준 기술을 활용하였고, 그 일환으로 PreSPI 기능을 웹 서비스 API 형태로 제공함으로써 개방성을 지원할 수 있는 것을 확인하였다.

한편 PreSPI는 인터넷상에 새롭게 공표되는 도메인 및 단백질 상호작용에 관한 데이타를 지속적으로 갱신할 필요가 있다. 이를 위해서 PreSPI는 데이타 모듈과 서비스 모듈을 분리하여 구성하였으며 이러한 구성이 새롭게 갱신되는 데이타에 대해서 손쉽게 확장할 수 있는 구조임도 확인하였다. 본 논문에서 고안한 PreSPI 구조는 비록 도메인 조합 기반 단백질-단백질 상호 작용 예측 기법의 구현을 위하여 설계되었지만 또 다른 단백질 상호 작용 예측 기법을 구현하는 경우에도 참조 모델로 활용될 수 있을 것으로 기대한다.

PreSPI가 위에서 언급한 바와 같이 효율적인 구조 위에서 단백질 및 단백질 상호 작용과 관련된 유용한 서비스를 제공하고 있지만 그 기능 측면에서 아직도 부족한 점이 많다. 무엇보다도 주로 효모 단백질을 중심으로 제공되는 서비스를 다른 종(예를 들면 *C.elegance*, *Drosophila*, *E. coli.*, Mouse 그리고 Human)으로 확장하는 것이 필요하며, 사용자가 단백질과 관련된 풍부하고 다양한 형태의 정보를 손쉽게 접근할 수 있는 기능이 지속적으로 보강되어야 한다.

참 고 문 헌

[1] M. Deng, S. Metah, F. Sun and T. Chen, Inferring Domain-Domain Interactions from Protein-Protein Interactions. *Genome Research*, 12, 1540-1548, 2002.

[2] A. J. Enright, I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402, 86-90, 1999.

[3] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg, Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285, 751-753, 1999.

[4] S. Ng, Z. Zhang and S. Tan, Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19, 923-929, 2003.

[5] A. J. Enright and C. A. Ouzounis, Chapter 33: Protein-Protein Interactions-A Molecular Cloning Manual, Cold Spring Harbor Laboratory Press, Cold spring Harbor, NY, 2002.

[6] J. R. Bock. and D. A. Gough, Prediction of protein-protein interaction from primary structure. *Bioinformatics*, 17, 455-460, 2001.

[7] J. Wojcik and V. Schächter, Protein-Protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17 Suppl., S296-S305, 2001.

[8] D. S. Han, H. S. Kim, J. M. Seo, and W. H. Jang, A Domain Combination Based Probabilistic Framework for Protein-Protein Interaction Prediction, *Genome Informatics*, No. 14, 250-259, 2003.

[9] D. S. Han, H. S. Kim, W. H. Jang, and S. D. Lee, Domain Combination Based Protein-Protein Interaction Possibility Ranking Method, *Proc. of 4th IEEE Sym. on Bioinfo. and Bioeng.*, 434-441, May, 2004.

[10] 한동수, 서정민, 김홍숙, 장우혁, 도메인 조합 기반 단백질-단백질 상호작용 확률 예측 틀, 정보과학회논문지:컴퓨팅의 실제, Vol. 10, No. 4, 299-308, August, 2004.

[11] W3C Web Services Architecture Working Group, Web Services Architecture, *World Wide Web Consortium*, http://www.w3.org/TR/ws-arch/, Feb., 2004.

[12] W3C Web Services Architecture Working Group, Web Services Architecture Requirements, *World Wide Web Consortium, http://www.w3.org/TR/wsa-reqs/*, Feb., 2004.

[13] W3C Web Services Architecture Working Group, Web Services Architecture Usage Scenarios, *World Wide Web Consortium*, http://www.w3.org/TR/ws-arch-scenarios/, Feb., 2004.

[14] R. Apweiler, *et al.*, The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29, 37-40, 2001.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.*, 28, 235-242, 2000.

[16] C.M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, Protein Interactions: Two methods for assessment of the reliability of high throughput observations, *Mol. Cell. Proteomics*, 1, 349-356, 2002.

[17] Y. Ho, *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415, 180-183, 2002.

[18] A. Gavin, *et al.*, Functional organization of the yeast proteom by systematic analysis of protein complexes, *Nature.*, 415, 141-147, 2002.

[19] S. Ng, Z. Zhang and S. Tan, InterDom: A database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.*, 31, 251-254, 2003.
[20] I. Xenarios, E. Fernandez, L. Salwinski, X. J. Duan, M. J. Thompson, E. M. Marcotte and D. Eisenberg, DIP: The Database of Inter acting Proteins: 2001 update. *Nucleic Acids Res.*, 29, 239-241, 2001.

한 동 수
정보과학회논문지 : 컴퓨팅의 실제
제 11 권 제 5 호 참조

김 홍 숙
정보과학회논문지 : 컴퓨팅의 실제
제 11 권 제 5 호 참조

장 우 혁
정보과학회논문지 : 컴퓨팅의 실제
제 11 권 제 5 호 참조

이 성 독
정보과학회논문지 : 컴퓨팅의 실제
제 11 권 제 5 호 참조