The 3rd Annual Conference of the Korean Society for Bioinformatics (KSBI2004)

PreSPI: Design and Implementation of Protein-Protein Interaction Prediction Service System

PreSPI:단백질 상호작용 예측 서비스 시스템 설계 및 구현

Dong-soo Han[1*], Hong-Soog Kim[1], Woo-hyuk Jang[1], Sung-Doke Lee[1]

[1] Bioinformatics and Software Systems Laboratory, School of Engineering, Information and Communications University, Daejeon 305 -714, Korea

*To whom correspondence should be addressed. E -mail: dshan@icu.ac.kr

Abstract

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 PreSPI(Prediction System for Protein Interaction)라는 서비스 시스템으로 설계하고 구현하였다. 구현된 시스템이 제공하는 기능은 크게 도메인 조합 기반 단백질 상호 작용 예측 기법을 서비스 형태로 만들어 제공하는 기능으로 입력 단백질 쌍에 대한 상호작용 예측이 중심이 된 핵심기능과, 핵심 기능으로부터 파생되는 기능인 부가 기능, 그리고 주어진 단백질에 대한 도메인 정보 검색 기능과 같이 단백질 상호작용에 관하여 연구하는 연구자에게 도움이 되는 일반적인 기능으로 구성되어 있다. 계산을 통해 단백질 상호 작용을 예측하는 시스템은 대규모 계산이 요구되는 경우가 많아 좋은 성능을 갖추는 것이 중요하다. 본 논문에서 구현된 PreSPI 시스템은 서비스에 따라 적절히 그 처리를 병렬화 함으로써 시스템의 성능 향상을 도모하였고, PreSPI 가 제공하는 기능을 웹 서비스 API 로 Deploy 하여 시스템의 개방성을 지원하고 있다. 또한 인터넷 환경에서 변화되는 단백질 상호 작용 및 도메인에 관한 정보를 유연하게 반영할 수 있도록 시스템을 계층 구조로 설계하였다. 본 논문에서는 PreSPI 가 제공하는 몇 가지 대표적인 서비스에 관하여 사용자 인터페이스를 중심으로 상술함으로써 초기 PreSPI 사용자가 PreSPI 가 제공하는 서비스를 이해하고 사용하는 데에도 도움이 되도록 하였다.

서론

계산을 통한 단백질 상호작용 예측 기법의 효용성 및 중요성에 관한 인식이 확산되면서 단백질상호작용을 계산적으로 예측하는 다양한 기법이 제안되고 있다[5,6,7,8]. 가공하지 않은 단백질 서열로부터 직접 단백질-단백질 상호작용에 영향을 끼치는 요소들을 발견하고 분석하는 것이 한 가지 접근 방법이며[9], 단백질의 구조나 물리화학적 특성을 분석함으로써 단백질 상호작용을 예측하는 방법도 알려져 있다[10]. 도메인에 기반한 단백질-단백질 상호작용 예측도 또 하나의 접근 방법이 될 수 있으며, 현재 여러 연구진들에 의하여 활발히 연구되어지고 있다[5,8,11].

하지만 지금까지의 대부분의 연구는 예측 방법 고안에 머무르고 있어 일반 생물학자들이 사용할 수 있는 서비스 형태로는 제공되지 않고 있다. 그 이유는 계산을 통한 단백질 상호작용 예측 연구가 비교적 초기 단계라는 점과 제안된 예측 기법의 정확도가 서비스를 하기에는 다소 떨어지는 점이 그 원인이 있는 것으로 판단된다.

최근 Han[Han03]을 중심으로 한 연구그룹에서 제안한 도메인 조합 기반 단백질 상호작용 예측 기법은 그 완성도 면에서나 예측 정확도 측면에서 일반에 서비스할 정도의 우수성을 보여주고 있다. 본 논문에서는 Han 이 제안한 기법을 중심으로 단백질 상호작용 예측 서비스 시스템을 설계하고 구현한다.

본 논문에서 설계하고 구현하는 예측 서비스 시스템이 제공하여 주는 기능은 크게 핵심 기능 및 부가 기능 그리고 일반 서비스 등으로 분류된다. 핵심기능은 논문 [Han03]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 서열 결정 기능 등을 포함한다. 부가 기능은 핵심 기능으로부터 파생되는 기능으로 도메인 조합 출현 확률 배열 상에서 높은 값을 갖는 도메인 조합 쌍 검색 및 본 논문에서 소개한 기법을 이용하여 예측한 단백질 상호작용 데이터에 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_ID 를 상호 변환해 주는 기능 등을 포함한다. 핵심 기능 및 부가 기능이 논문 [Han03]에서 소개한 도메인 조합 기반 단백질 상호 작용 예측 기법에 준하여 제공하는 기능인 반면에 일반 서비스 기능은 논문 [Han03]에서 소개한 기법과는 직접적인 연관이 없이 제공될 수 있다는 점에서 핵심 및 부가 기능과는 구분된다.

일반적으로 계산을 통한 단백질 상호작용 예측 시스템은 다양한 포맷으로 지속적으로 경신되는 분산된 데이터를 기반으로 대량의 계산을 통하여 예측을 시도한다. 또한 단백질 상호작용 예측 시스템이 제공하여 주는 서비스는 그것이 곧바로 생물학자가 요구하는 최종적인 답을 제공하기 보다는 그러한 정보를 이용하여 다양한 시도를 할 수 있도록 하는 단초를 제공하여 주는 측면이 강하다. 따라서 다른 응용 프로그램이나 외부의 시스템이 예측 시스템이 제공하여주는 서비스에 손쉽게 접근할 수 있는 수단을 제공하여 주는 것이 필요하다.

위와 같은 요구 사항을 반영하여 본 논문에서 소개하는 단백질 상호작용 예측 시스템은 시스템의 성능과 확장성 그리고 개방성을 목표로 설계되고 구현되었다. 시스템의 성능은 일부 서비스를 병렬 처리함으로써 확보하였고, 시스템의 개방성은 웹 서비스 표준 기술을 도입하고 시스템이 제공하는 기능을 웹 서비스 API 형태로 제공함으로써 외부의 응용 프로그램이나 시스템이 손쉽게 접근할 수

있도록 지원하고 있다. 시스템의 확장성 및 유연성을 위해서는 시스템을 데이터 모듈과 서비스 모듈로 명확히 구분하는 계층적 구조를 사용하여 새롭게 갱신되는 데이터를 주기적으로 시스템에 반영할 수 있도록 구성하였다.

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

도메인 조합 기반 단백질 상호작용 예측 기법

본 장에서는 본 논문에서 대상으로 삼고 있는 도메인 조합 기밥 단백질 상호작용 예측 기법에 관하여 간략히 소개한다. 상세한 도메인 조합 기반 단백질 상호작용 예측 기법은 참고문헌 [Han03]에 소개되어 있다.

제안 배경

도메인 조합 기반 단백질 상호작용 예측 기법은 그 동안 활발하게 연구되어진 도메인에 기반한 단백질-단백질 상호작용 예측 기법의 발전된 형태로 볼 수 있다[5,8,11]. 대부분의 도메인 기반 단백질-단백질 상호작용 예측 모델들은 단백질-단백질 상호작용이 도메인-도메인 상호작용의 결과물이라는 추측에서 출발한다. 이 방법들은 단백질-단백질 상호작용 데이터로부터 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 예측 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 쌍(single domain pair)의 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 간의 상호작용의 결과로 인식하며, *dc-pair* 를 단백질 상호작용의 기본 단위로 해석한다.

그림 1(b) 에서는 각각 3 개와 2 개의 도메인을 갖는 단백질이 상호작용하는 경우,

잠재적인 상호작용 *dc-pair* 를 보여주고 있다. 또한, 기존 방법에서 사용했던 도메인 쌍 기반 방법과 도메인 조합 쌍 기반 접근 방법의 차이점을 보여준다.

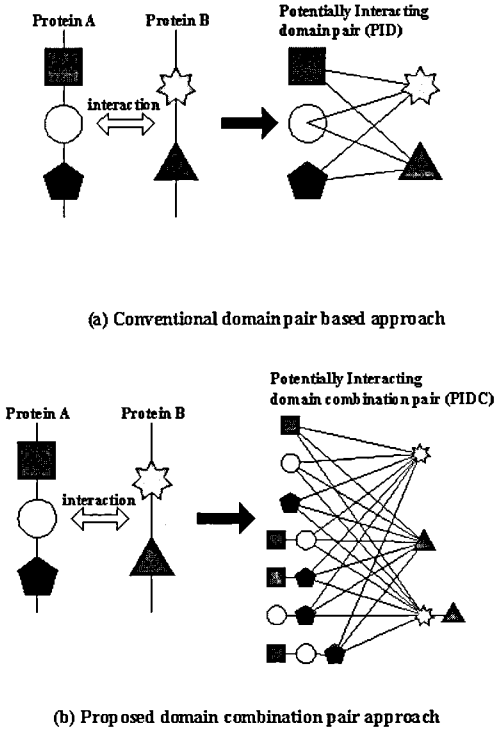


그림 1. (a)도메인 쌍에 기반한 기존의 예측 접근방법, (b)도메인 조합에 기반한 새로운 예측 접근방법

도메인 조합 기반 단백질 상호작용 예측 기법

도메인 조합 기반 단백질 상호작용 예측 기법에서는 상호작용이 있는 단백질 쌍 집합과 상호작용이 없는 것으로 가정된 단백질 쌍 집합에 대해서 각각 *dc-pair* 의 출현 빈도를 측정하여 출현 확률 배열(Appearance Probability Matrix or AP matrix) 에 저장한다. 그리고 이 배열을 토대로 단백질-단백질 상호작용 확률 예측 모델을 구축한다. 도메인 조합 기반 상호작용 예측 기법은 도메인 쌍에 대한 정보를 *dc-pair* 정보 안에 포함하고 있으므로, 종래의 도메인 쌍에 기반한 방법에 비교할 때 더 포괄적이다. 또한, 상호작용 가능성에 대한 확률 값을 제시함으로써 좀 더 실질적인 정보를 생물학자에게 제공하는 것이 가능하다.

그림 2 는 도메인 조합 기반 단백질 상호작용 예측 기법의 전체 구조를 보여준다. 첫 번째 과정은 예측을 준비하는 과정이며 두 번째 과정에서는 예측을 수행하는 서비스 과정이다. 예측 준비 과정은 다시 세 개의 단계를 포함한다. 첫 번째 단계에서는 상호작용이 있는 것으로 알려진 단백질 쌍 집합과 상호작용이 없는 것으로 추정되는 단백질 도메인 쌍 집합으로부터 각각 도메인 조합 정보와 그 출현 빈도를 추출한다. 이 정보들은 출현 확률 배열(Appearance Probability Matrix;AP matrix)라고 불리는 배열 구조에 저장된다. 두 번째 단계에서는 AP matrix 를 기반으로 단백질-단백질 상호작용 예측 확률식(Primary Interaction Probability or PIP)을 정의한다. 이 확률식은 미 정의된 상수 k 를 포함하게 되며 이 상수는 maximum likelihood estimation 적용을 통하여 결정한다. 마지막 세 번째 단계에서는 상호작용이 있는 것으로 알려진 단백질 쌍 집합과 상호작용이 없는 것으로 추정되는 단백질 도메인 쌍 집합의 확률 값 분포를 얻게 된다. 두 번째 과정에서는 첫 번째 과정에서 얻어진 분포에 기초하여, 단백질-단백질 상호작용을 예측하는 또 다른 확률식이 정의되며, 이 확률식을 이용하여 단백질-단백질 상호작용을 예측하는 최종 확률을 계산한다. 각 단계의 세부사항은

참고문헌 [Han03][Han04]에 자세히 기술되어 있다.

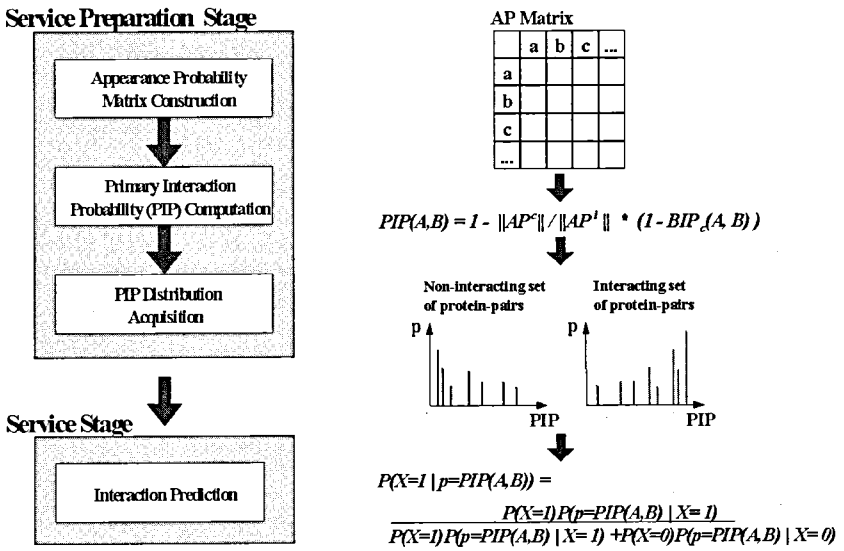


그림 2 예측 시스템의 전체 구조

예측 정확도 검증

도메인 조합 기반 단백질 상호작용 예측 기법의 유효성은 효모(yeast)에서 상호작용이 있는 것으로 알려진 단백질 쌍 집합과 상호작용이 없는 것으로 추정되는 단백질 쌍 집합을 대상으로 검증하였다. DIP 데이터 베이스[3,12]의 상호작용이 있는 것으로 알려진 단백질 쌍 집합의 80%를 학습 집단으로 사용했을 때, 제안된 예측 시스템은 매우 높은 sensitivity(84%)와 specificity(75%)를 보여 주어 제안된 예측 시스템의 유용성이 확인되었다.

PreSPI 기능 및 구성

앞 장에서 소개한 도메인 조합 기반 단백질 상호작용 예측 기법이 널리 사용되기 위해서는 생물학자들이 손쉽게 접근하여 사용할 수 있는 서비스 시스템의 구현이 요청된다. 본 장에서는 이와 같은 서비스 시스템 PreSPI(Prediction System for Protein Interaction)의 기능 및 설계에 관하여 기술한다.

PreSPI 기능

PreSPI 가 제공하는 기능은 크게 핵심 기능 및 부가 기능 그리고 일반 서비스 등으로 분류된다. 핵심기능은 본 논문에서 소개한 방식을 서비스 형태로 만들어 제공하는 기능으로 입력 단백질 쌍에 대한 상호작용 예측, PIP 분포 표시, 복수의 단백질 쌍에 대한 카테고리 결정 및 상호작용 가능성 서열 결정 기능 등을 포함한다. 부가 기능은 핵심 기능으로부터 파생되는 기능으로 AP 배열 상에서 높은 값을 갖는 dc-pair 검색 및 본 논문에서 소개한 기법을 이용하여 예측한 단백질 상호작용 데이터에 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

분류	Function Name	Function Description
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 상호작용 가능성 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 [] 에서 제시한 방식으로 분류한 뒤 그 서열을 결정한다.
	상호작용 단백질 검색	이 기능은 주어진 하나의 단백질과 상호작용할 가능성이 있는 단백질 들을 찾고 이것들을 그 서열에 의하여 상호작용 확률과 함께 리스트 형태로 출력한다.
부가 기능 (Subsidiary Function)	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 배열 상에 나타나는 dc-pair 중 큰 값을 갖는 300 개의 dc-pair 리스트를 보여주고 특정 dc-pair 에 대해서는 AP 배열 상의 해당 값을 찾아준다. 또한 특정 도메인 조합을 포함하면서 특정 값 이상을 갖는 dc-pair 에 대해서도 리스트 형태로 출력하는 것이 가능하다. 생물학자는 이 것을 통해 단백질 상호작용에 영향을 미치는 주요 도메인 또는 도메인 조합 쌍에 대한 정보를 얻을 수 있다.
	단백질 상호작용 네트워크 구성	이 기능은 PreSPI 에 의해 상호작용이 있는 것으로 예측된 단백질 쌍을 기반으로 구성된 단백질 상호작용 네트워크를 보여준다. 사용자가 입력하는 확률 이상의 상호작용 가능성이 있는 단백질 쌍만을 대상으로 네크워크를 구성하는 것도 가능하며 주어진 쌍(two or more) 이상의 상호작용이 있는 것으로 추정되는 단백질만을 대상으로 네크워크를 구성하는 것도 가능하다.
	예측 시스템 정확도 표시	이 기능은 도메인 조합에 기반한 단백질 상호 작용 예측 기법의 정확도를 Yeast 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_ID 를 상호 변환한 결과를 출력한다.

표 1 PreSPI 의 서비스 기능

도메인 정보 검색 기능 및 DIP_ID, SWISSPROT_ID, PIRID 등의 Accession_ID 를 상호 변환해 주는 기능 등을 포함한다. 핵심 기능 및 부가 기능이 본 논문에서 소개한 기법을 구현하는 PreSPI 만이 제공할 수 있는 독특한 기능이다. 반면에 일반 서비스 기능은 본 논문에서 소개한 기법과는 직접적인 연관이 없이 제공될 수 있다는 점에서 핵심 및 부가 기능과는 구분된다. 표 1 은 PreSPI 가 제공하는 기능을 위의 분류에 따라서 정리한 내용이다.

PreSPI 구성

일반적으로 계산을 통한 단백질 상호작용 예측 시스템은 다양한 포맷으로 지속적으로 경신되는 분산된 데이터를 기반으로 대량의 계산을 통하여 예측을 시도한다. 따라서 많은 경우 예측에 많은 시간이 소요되는 것이 보통이다. PreSPI 도 예측을 위해서는 종별로 약 10 억개 이상의 요소를 갖는 AP 배열을 생성하는 것이 필요하고 다시 학습에 사용된 단백질 쌍을 입력으로 한 PIP 분포를 얻는 것이 필요하다. 이와 같이 서비스를 위한 준비가 완료된 뒤에는 전 절에서 소개한 다양한 기능을 제공할 수 있지만 서비스에 소요되는 시간은 제공하는 기능에 따라서는 수초에서 수십 시간이 소요되기도 한다.

또한 PreSPI 가 제공하여 주는 서비스는 그것이 곧바로 생물학자가 요구하는 최종적인 답을 제공하기 보다는 그러한 정보를 이용하여 다양한 시도를 할 수 있도록 하는 단초를 제공하여 주는 측면이 강하다. 따라서 다른 응용 프로그램이나 외부의 시스템이 예측 시스템이 제공하여주는 서비스에 손쉽게 접근할 수 있는 수단을 제공하여 주는 것이 필요하다.

PreSPI 는 위와 같은 요구 사항을 반영하여 시스템의 성능과 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 서비스화함으로써 외부의 응용 프로그램이나 시스템이 손쉽게 접근할 수 있도록 하였다. 시스템의 확장성을 위해서는 시스템을 데이터 모듈과 서비스 모듈로 명확히 구분하고 새롭게 갱신되는 데이터를 주기적으로 시스템에 반영할 수 있도록 구성하였다. 다음은 이러한 목표를 달성하기 위하여 설계된 PreSPI 의 구성도를 보여주고 있다.

PreSPI 는 크게 서비스 준비를 위해서 관련 데이터베이스를 생성하는 데이터 모듈과 준비된 데이터베이스에 기반하여 사용자로부터 서비스 요청을 받아서 서비스하는 서비스 모듈로 구성되어 있다. 데이터 모듈은 다시 인터넷 상에 산재해 있는 각종 단백질 및 도메인 관련 데이터를 모아서 단백질-도메인 사전을 구축하는 단백질-도메인 사전 구축 모듈, 상호작용이 있는 단백질 쌍 집합과 임의의 단백질 쌍 집합으로부터 AP 배열을 생성하여 데이터베이스에 저장하는 AP 배열 생성 모듈. 그리고 PIP 함수를 각 AP 배열의 원소에 적용하여 PIP 값 분포를 얻어 데이터베이스에 저장하는 PIP 값 분포 생성 모듈을 포함한다. 이와 같은 세 가지 데이터베이스가 데이터 모듈에 의해서 생성이 완료되면 PreSPI 는 사용자로부터

각종 서비스 요구를 받아서 처리할 초기 준비가 1차 완료된 상태로 볼 수 있다.

서비스 모듈은 데이터 모듈에 의해서 생성된 세 개의 데이터베이스를 기반으로 표에서 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, 그리고 사용자 서비스 요청을 받은 후 해석하여 그것에 해당하는 서비스 루틴을 찾고 구동하는 연결 층, 그리고 해당 서비스 루틴 들을 포함하고 있는 서비스 라이브러리 루틴 층으로 구성된다. PreSPI 는 서비스 모듈과 데이터 모듈을 분리함으로써 인터넷 상에 새로운 데이터가 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 Layer 구조의 특징을 유지시키는 것이 가능하다. 또한 인터넷 상의 새로운 데이터의 추가는 서비스 모듈에는 영향을 미치지 않고 데이터 모듈의 해당 부분만을 확장함으로써 수용 가능하다.

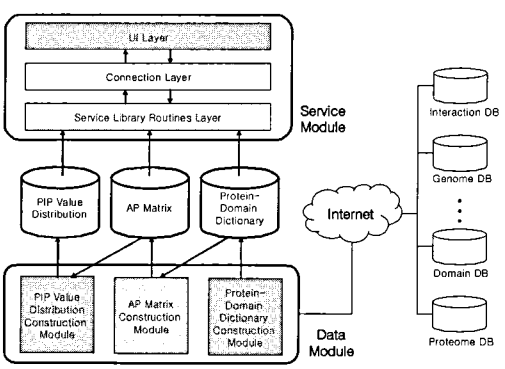


그림 3 PreSPI의 아키텍처

한편 PreSPI 는 많은 기능을 웹 서비스 API 형태로 제공함으로써 응용 프로그램이나 외부 시스템에 의하여 손쉽게 접근할 수 있는 개방성을 지원한다. 표 2 는 PreSPI 에 의하여 외부에 제공되는 API 들을 정리한 것이다. WSDL 은

『http://silver.icu.ac.kr:8080/axis/services/PreSPI?wsdl』 에서 자세히 살펴 볼 수 있다.

구현

본 장에서는 본 논문에서 구현된 PreSPI 시스템의 대표적인 사용자 인터페이스 및 사용법에 관하여 간략히 소개한다. PreSPI 시스템의 각 페이지는 페이지 Description 과 Usage Guide 형태로 구성하여 사용자가 페이지 Description 을 통하여 페이지의 의미를 파악하고 Usage Guide 를 참고하여 PreSPI 가 제공하는 기능을 사용할 수 있도록 하였다. PreSPI 시스템의 구현에 있어서 데이터 모듈 부분은 주로 Python 2.2.2 를 사용하여 코딩하였으며 MySQL 3.23.5 을 데이터베이스 관리 시스템으로 사용하였다. 사용자 인터페이스는 웹 상에서 제공하기 위하여 Java 를 사용하여 구현하였으며 Jython 을 사용하여 Java 와 Python 모듈을 연결하였다. 또한 jakarta-tomcat-4.1.24 를 웹 서버로 사용하였으며 Java 에서 MySQL 데이터베이스 접근을 위한 JDBC 드라이버로는 mysql-connector-java-3.0.8-stable 을 사용하였다.

PreSPI 웹 서비스 API 는 웹 서버로 Apache tomcat 4.1 을 사용하고 Container 로 Axis-

1_1 을 사용하여 구축하였다. PreSPI 사이트(http://silver.icu.ac.kr:8080/torajim/)를 방문하여 사용할 수 있다.

사용자는 PreSPI 서비스 웹 기능을 웹 페이지로 접근하길 원하는 일반

Name	Parameters	Description
getDomain()	In: String protein, int kind	kind 타입의 protein 에 대하여 해당 도메인을 반환한다.
	Out: String	(0: DIP, 1: SWISSPROT, 2: PIR, 3: GI)
		-1: Domain 정보를 가지고 있지 않음
checkProPair()	In: String proteinA, String proteinB	proteinA 와 proteinB 를 PreSPI 에서 계산가능 여부를 알려준다.
	Out: int	(0: Domain 정보를 가지고 있음
		1: Domain 정보를 가지고 있으며 실험적으로 상호작용 함으로 증명된 pair 임
		-1: Domain 정보를 가지고 있지 않음
transID()	In: String protein, int from, int to	from 타입의 protein 을 to type 의 ID 로 변환한다.
	Out: String	(0: DIP, 1: SWISSPROT, 2: PIR, 3: GI)
		ID 가 없을 경우 -1 리턴
getDomainCombination()	In: String protein, int kind	kind 타입의 protein 에 대하여 해당 도메인의 멱집합(DC)을 반환한다.
	Out: String	(0: DIP, 1: SWISSPROT, 2: PIR, 3: GI)
		-1: Domain 정보를 가지고 있지 않음
getDomainCombinationPairs()	In: String proteinA, String proteinB, int aKind, int bKind	aKind 타입의 proteinA 와 bKind 타입의 proteinB 에 대하여 DC-Pairs 를 반환한다.
	Out: java.util.Vector	(0: DIP, 1: SWISSPROT, 2: PIR, 3: GI)
		도메인 정보가 없을 경우 size 가 0 인 Vector 반환
getProteinsForGivenDomains()	In: String[] domains, int andOr	domain 들을 입력받아, andOr 조건으로 검색하여 포함하고 있는 proteins 을 리턴한다.
	Out: java.util.Vector	(0: or, 1:and)
getPIPValue()	In: String proteinA, String proteinB	proteinA 와 proteinB 를 입력받아 PIP value 를 리턴한다. ID 는 DIP 형태를 지원한다.
	Out: String	
getInterProb()	In: String pipValue	PIP value 를 입력받아 Interaction Probability 를 리턴한다.
	Out: String	
getInterProb()	In: String proteinA, String proteinB	proteinA 와 proteinB 를 입력받아 Interaction Probability 를 리턴한다.
	Out: String	-1: Domain 정보를 가지고 있지 않음
getIsInPIPDist()	In: String pipValue	pipValue 를 입력받아 PIP distribution 에 존재하는지 여부를 리턴한다.
	Out: boolean	
getIsInPIPDist()	In: String proteinA, String proteinB	proteinA 와 proteinB 를 입력받아 PIPdistribution 에 존재하는지 여부를 리턴한다.
	Out: boolean	

getIsProved()	In: String proteinA, String proteinB	proteinA 와 proteinB 를 입력받아 실험적으로 상호작용이 증명되었는지
	Out: boolean	여부를 리턴한다.
getInteractingPairs()	In:	PreSPI 에서 사용하고 있는 Interacting pair 의 DIP ID 와 Domain 들을
	Out: java.util.Vector	리턴한다.
getAllProteins()	In:	PreSPI 에서 사용하고 있는 모든 protein 의 DIP ID 를 리턴한다.
	Out: java.util.Vector	

표 2 PreSPI가 외부에 제공하는 API

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. 사용자는 이 분포를 대비해 봄으로써 두개의 집단이 PIP 값을 매개로 잘 분리될 수 있는지에 대하여 직관적으로 판단할 수 있다. 다음 Snapshot 은 Regular 간격을 사용하여 두개의 집단의 PIP 값 분포를 보여주고 있다. Regular 간격을 사용한 PIP 값 분포에서는 수평선 라인의 0 과 1 사이에 나타나는 서로 다른 PIP 값의 간격을 PIP 값에 관계없이 그 상대적 크기에 따라 동일한 간격으로 배열한다. 따라서 약 10000 여개의 서로 다른 PIP 값이 존재하는 경우에는 각 PIP 값의 간격은 1/10000 으로 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.

그림 4 에서 붉은 선은 상호작용이 있는 것으로 알려진 단백질 쌍 집단의 PIP 값 분포를 나타내고 파란 선은 임의로 짝지어진 단백질 쌍 집단의 PIP 값 분포를 나타낸다. 그림에서 보듯이 두개의 집단은 PIP 값을 매개로 대체로 잘 분리되고 있음을 알 수 있다. 이것은 PIP 값이 두개의 집단을 분리해 주는 분리자로서의 역할을 잘 수행하고 있음을 의미한다.

PIP 분포 시각화에서 수평축의 PIP 값을 Regular 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 를 선택하면 100%를 사용한 경우의 PIP 분포를 보여준다.

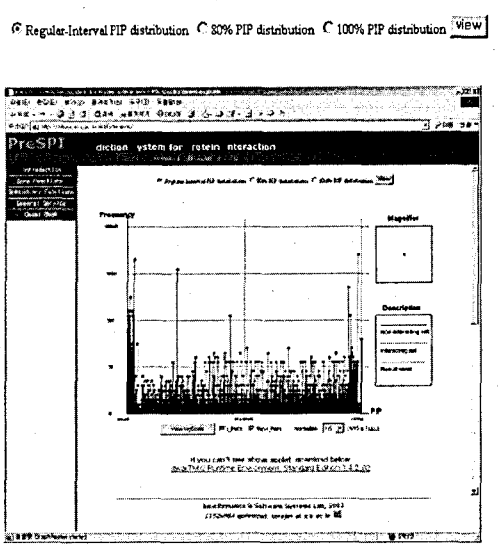


그림 4 Regular 간격을 사용한 두개의 단백질 집단의 PIP 값 분포

단일 단백질 쌍 상호작용 예측

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 정보와 PIP 값, 계산된 PIP 값이 PIP 분포 상에 존재하는지의 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_ID 등을 사용하는 것이 가능하다.

그림 5 는 단백질 6500N(GID Number)과 5307N 을 입력으로 하여 단일 단백질 쌍 상호작용 예측을 시행한 결과를 보여주고 있다. 그 결과 단백질 6500N 은 하나의 도메인(IPR001126)을 갖고 단백질 5307N 은 다섯 개의 도메인(IPR002314, IPR002320, IPR004095, IPR004154, IPR006195)을 갖는 것을 보여주고 있다. 한편 이 단백질 쌍의 PIP 값은 1.0 이고 In-pip-distribution 필드가 *true* 이어서 이 값은 PIP 분포 상에 존재하는 값임을 알 수 있다. 또한 이 단백질 쌍의 상호작용은 실험을 통하여 밝혀져 있지는 않은 상태이고 PreSPI 를 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 비교적 상세하고 유용한 정보를 모아서 보여주고 있다.

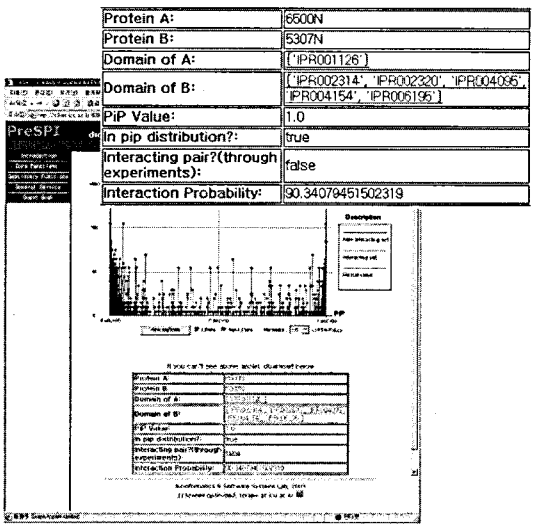


그림 5 6500N과 5307N의 상호작용 예측 결과

복수 단백질 쌍 상호작용 예측

복수 단백질 쌍 상호작용 예측 페이지에서 사용자는 복수의 단백질 쌍에 대해서 이들의 상호작용 가능성을 예측함과 동시에 어느 단백질 쌍이 더 상호작용을 일으킬 가능성이 높은지를 예측하는 기능이다. 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.

PreSPI 는 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 6 은 네 개의 단백질 쌍에 대해서 상호작용 가능성을 예측하고 그 결과를 상호작용 확률 값으로 정리하여 보여주고 있다. 마찬가지로 PreSPI 가 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.

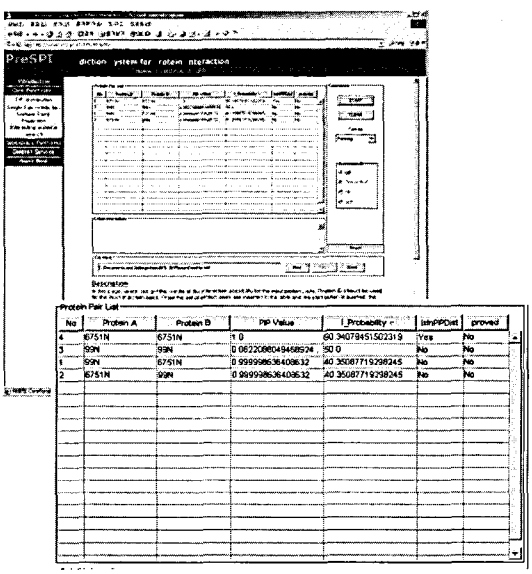
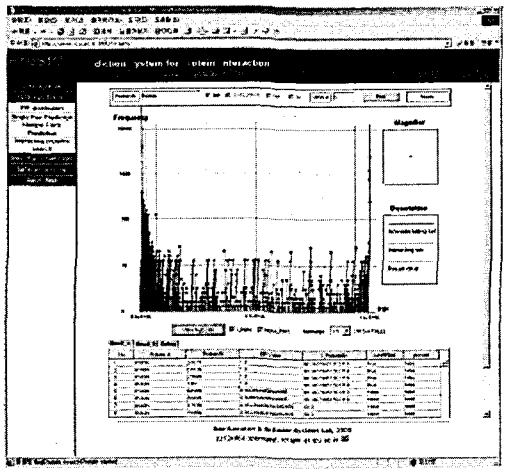


그림 6 복수 단백질의 상호작용 예측

상호작용 단백질 검색

상호작용 단백질 검색 페이지에서 사용자는 하나의 입력 단백질에 대해서 이것과 상호작용할 가능성이 있는 단백질 들을 찾고 이것들을 그 서열에 의하여 상호작용 확률과 함께 리스트 형태로 출력하는 것이 가능하다. PreSPI 는 입력으로 주어진 단백질과 나머지 단백질(Yeast 의 경우에는 대략 3 천여개)과 쌍을 형성하여 각각의 단백질 쌍에 대한 상호작용 확률을 예측하고 그 결과를 정리하여 보여준다. 이 작업은 주어진 단백질과 쌍을 이루는 모든 단백질쌍에 대해서 예측이 되어야 하는 만큼 많은 시간이 소요되는 것이 보통이다. 현재의 시스템에서는 하나의 단백질 쌍에 대한 상호작용 예측에는 약 5-8 초가 소요되는 만큼 3000 여쌍(Yeast 의 경우)에 이르는 단백질 쌍에 대해서 모두 예측하는 데에는 약 6-7 시간이 걸리게 된다. 그럼에도 불구하고 생물학자들은 PreSPI 가 제공하는 이 기능을 통해서 관심이 있는 몇 개의 단백질과 상호작용을 일으킬 가능성이 있는 단백질에 관한 유용한 정보를 제공받을 수 있다.

그림 7 은 단백질 6500N 에 대해서 이것과 상호 작용할 확률이 높으면서 PIP 값이 큰 8 개의 단백질 쌍을 보여주는 화면이다. 화면은 약 30 여분이 경과했을 때 약 2800 여 쌍의 대상 단백질 쌍 중에서 약 200 여 쌍의 단백질의 상호작용 확률이 계산되었고 그 중에서 8 개의 단백질 쌍을 보여주고 있다. 8 개를 제외한 나머지 190 여개의 단백질 쌍에 대한 예측 결과는 화면의 ' Result B' 버튼을 누르면 볼 수 있다.



Result_A | Result_B | Setting

No	Protein A	Protein B	PIP Value	I_Probability	isInPIPDist	proved
1	6500N	5307N	1.0	90.340794515023 19	true	false
2	6500N	6782N	1.0	90.340794515023 19	true	false
3	6500N	2300N	1.0	90.340794515023 19	true	false
4	6500N	14N	1.0	90.340794515023 19	true	false
5	6500N	6456N	0.9999999999999996	90.340794515023 19	false	false
6	6500N	5584N	0.9999999999999996	90.340794515023 19	false	false
7	6500N	4763N	0.06220880494 6892404	50.0	false	false
8	6500N	1656N	0.06220880494 6892404	50.0	false	false

그림 7 상호작용 가능성이 있는 단백질 검출

이 밖에도 PreSPI 에는 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

결론

본 논문에서는 도메인 조합에 기반한 단백질 상호작용 예측 서비스 시스템인 PreSPI 시스템을 성능과 확장성 그리고 개방성 달성을 목표로 설계하고 구현하였다. PreSPI 시스템을 구현한 결과, 예상대로 서비스 제공에 많은 시간이 소요됨을 확인할 수 있었으며 클러스터 시스템 상에서 시스템의 병렬화를 통하여 비교적 손쉽게 시스템의 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 PreSPI 가 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 기술을 활용하였고, 그 일환으로 PreSPI 기능을 웹 서비스 API 형태로 제공함으로써 개방성을 지원할 수 있는 것을 확인하였다.

한편 PreSPI 는 인터넷 상에 새롭게 공표되는 도메인 및 단백질 상호작용에 관한 데이터를 지속적으로 갱신하는 것이 필요하다. 이를 위해서 PreSPI 는 데이터 모듈과 서비스 모듈을 분리하여

구성하였으며 이러한 구성이 새롭게 갱신되는 데이터에 대해서 손쉽게 대처할 수 있는 구조임도 확인하였다. 본 논문에서 고안한 PreSPI 구조는 비록 도메인 조합 기반 단백질 상호 작용 예측 기법의 구현을 위하여 설계되었지만 또 다른 단백질 상호 작용 예측 기법을 구현하는 경우에도 참조 모델로 활용될 수 있을 것으로 기대한다.

PreSPI 가 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 Yeast 단백질을 중심으로 제공되는 서비스를 다른 종으로 확장하는 것이 필요하고 사용자가 단백질과 관련된 풍부하고 다양한 형태의 정보를 손쉽게 접근할 수 있는 기능이 지속적으로 보강되어야 한다.

참고문헌

[1] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni and F. Servant, The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29, 37-40, 2001.

[2] 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.

[3] I. Xenarios and D. Eisenberg, Protein interaction databases. *Curr. Opinion in Biotechnology*, 12, 334-339, 2001.

[4] E. Sprinzak and H. Margalit, Correlated sequence-signatures as markers of protein-protein inter- action. *J. Mol. Biol.*, 311, 681-692, 2001.

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

[6] 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.

[7] 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.

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

[9] 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.

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

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

[12] 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.

[13] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247, 536-540, 1995.

[14] F. M. G. Pearl, D. Lee, J. E. Bray, I. Sillitoe, A. E. Todd, A. P. Harrison, J. M. Thornton and C. A. Orengo, Assigning genomic sequences to CATH. *Nucleic Acids Research*, 28, 277-282, 2000.

[15] L. Holm, and C. Sander, The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.*, 24, 206-210, 1996.

[16] J. Park, M. Lappe and S. A. Teichmann, Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.*, 307, 929-938, 2001.

[17] K. Han, B. Park, H. Kim, H, J. Kim and J. Park, Protein Interactions in the Whole Human Genome, *Genome Informatics*, 13, 318-319, 2002.

[18] B. H. Ju, B. Park, J. H. Park, and K. Han, Visualization and Analysis of Protein Interactions, *Bioinformatics*, 19, 317-318, 2003.

[19] W. K. Kim, J. Park, J. K. Suh, Large Scale Statistical Prediction of Protein-Protein Interaction by Potentially Interacting Domain (PID) Pair, Genome Informatics, No. 13, 2002.

[20] N. Goffard, V. Garcia, F. Iragne, A. Groppi and A. de Daruvar, IPPRED: Server for Proteins Interactions Inference. *Bioinformatics*, 19, 903-904, 2003.

[Han03] D. Han, H. Kim, J. Seo, and W. Jang, A Domain Combination Based Probabilistic Framework for Protein-Protein Interaction Prediction, Genome Informatics, No. 14, 250-259, 2003

[Han04] D. Han, H. Kim, W. Jang, S.Lee, Domain Combination Based Protein-Protein Interaction Possibility Ranking Method, Proceedings of 4-th IEEE Symposium on Bioinformatics and Bioengineering, 434-441, 2004.