

# 연관성 규칙에 기반한 보존된 단백질 도메인 조합의 식별

(Identification of Conserved Protein Domain Combination based on Association Rule)

정석훈<sup>\*</sup> 장우혁<sup>\*</sup>

(Suk-Hoon Jung) (Woo-Hyuk Jang)

한동수<sup>††</sup>

(Dong-Soo Han)

**요약** 도메인은 단백질의 진화와 삼차구조 및 분자 기능의 기본 단위체이다. 단백질은 한 개 이상의 도메인들로 구성되며, 단백질의 기능 또한 각 도메인이 가진 기능의 집합으로 구현된다. 단백질은 특정 기능을 담당하기 위해 진화되어 왔으므로, 도메인 또한 단백질 내에서 기능을 위한 특정 조합 패턴, 즉 보존도메인 조합을 가진다. 본 논문은 각 도메인 조합의 진화상 보존 정도를 측정할 수 있는 연관성 규칙 기반 계산 기법을 제안한다. 제안된 기법은 기존 기법에서 주로 고려되었던 도메인 조합의 빈도뿐 아니라, 조합 내 소속 도메인간의 상호 의존도를 측정하여 주어진 조합의 보존 정도를 산출한다. 이를 기반으로 *S.cerevisiae*의 단백질을 대상으로 보존 도메인 조합을 추출하였으며, Gene Ontology를 이용하여 그 생물학적 의미를 분석하였다. 그 결과 제안된 기법으로 추출된 보존 도메인 조합은 기존의 것에 비해 조합 내 기능의 유사도가 높았으며, 따라

서 제안된 기법이 생물학적 기능의 협업 위해 보존된 도메인 조합의 추출에 우수하다 할 것이다. 또한 *S.cerevisiae* 단백질체에는 서로 의존도가 높고 자주 나타나는 보존 도메인 조합이 존재하며, 그러한 조합들은 molecular function의 협업과 관련 있음을 밝혀냈다.

**키워드 :** 단백질 도메인, 도메인 조합, 보존 도메인 조합, 연관성 규칙

**Abstract** Protein domain is the conserved unit of compact tree-dimensional structure and evolution, which carries specific function. Domains may appear in patterns in proteins, since they have been conserved through the evolution for functional formation of proteins. In this paper, we propose a formulated method for conservation analysis of domain combination based on association rule. Proposed method measures mutual dependency of domains in a combination, as well as co-occurrence frequency of them, which is conventionally used. Based on the method, we extracted conserve domain combinations in *S.cerevisiae* proteins and analyzed their functions based on Gene Ontology. From the results, we drew conclusions that domains in *S.cerevisiae* proteins form patterns whose members are highly affiliated to one another, and that extracted patterns tend to be associated with molecular function. Moreover, the results testified to proposed method superior to conventional ones for identifying domain combinations conserved for functional cooperation.

**Key words :** protein domain, domain combination, conserved domain combination, association rule

## 1. 서 론

모든 단백질은 하나 또는 여러 개의 도메인으로 이루어져 있다[1]. 단백질 도메인은 진화를 통해 보존되어온 삼차원의 구조체로서, 특정 분자 기능을 가지며, 단백질은 진화의 과정에서 도메인이라는 단위체의 조합 및 재조합으로 세포 기능을 위해 진화하여 왔다[2]. 때문에 단백질 도메인 분석은 새로운 발견된 단백질의 기능을 이해하기 위한 첫 번째 단계로 간주되어 왔다.

도메인이 분자 기능을 위한 기초 단위체이지만 환경 조건에 영향을 많이 받는 폴리펩타이드의 특성상 주변 도메인과 연계하여 그 기능을 발현함은 당연하다 할 것이다. 즉, 한 단백질 내의 도메인들은 직접적으로 연계하여 단백질의 기능을 발현하거나, 또는 간접적인 물리화학적 도움을 줄 것으로 생각된다. 이러한 동일 단백질 내 도메인간 기능 협업을 위한 상호 연계 작용은 특정 유전자에 대한 몇몇 실험에서 밝혀졌으나[3], 일반적인 단일 단백질 내 도메인 상호 연계 메커니즘은 충분히 연구되지 않았다.

단백질이 특정 기능을 위해 발전하여 왔음을 고려할

\* 이 논문은 교육과학기술부, 한국산업기술재단, 한국과학재단의 특장기초 연구지원사업(학제기초)(No. R01-2008-000-20765-0) 및 지역혁신인력양성 사업으로 수행된 연구결과임

† 이 논문은 제35회 추계학술대회에서 '연관성 규칙에 기반한 보존된 단백질 도메인 조합의 추출'의 제목으로 발표된 논문을 확장한 것임

†† 학생회원 : 한국과학기술원 정보통신공학과  
sh.jung@kaist.ac.kr  
torajim@kaist.ac.kr

\*\* 종신회원 : 한국과학기술원 전산학과 교수  
dshan@kaist.ac.kr

논문접수 : 2008년 12월 19일  
심사완료 : 2009년 3월 1일

Copyright©2009 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지 : 컴퓨팅의 실제 및 레터 제15권 제5호(2009.5)

때에, 하위 기능 구조체인 도메인은 그 기능을 함께 수행하거나 영향을 주는 도메인과 함께 단백질을 이루어 왔을 것이다. 이러한 대전제 하에, 단백질의 발전에서 단지 서열 및 단일 도메인만이 보존되는 것이 아니라 기능을 위한 도메인 조합 또한 보존되어 왔을 가능성이 크다 할 수 있을 것이다.

실제로 도메인 조합 보존의 관점에서 계산, 통계적 기법을 사용한 몇몇의 연구가 이루어져 왔으며, 제한된 조건에서의 도메인 조합 보존 현상이 밝혀져 왔다. 쌍 도메인 조합에 대한 통계학적 연구가 그 대표적 예이다 [1]. 이들 연구에서는 대부분의 도메인이 적은 수의 선호하는 이웃 도메인을 가짐을 밝혀냈으며, 몇몇 경우에 대해 도메인 조합과 생물학적 기능과의 상관도가 높음을 밝혀냈다. 하지만 이들 연구에서는 제한조건으로 인해, 전체적 관점에서의 도메인 조합 기작을 밝혀냈다고는 말할 수 없다. 이들은 도메인 조합의 출현 빈도와 조합 내 두 도메인의 단 방향 의존성을 기반으로 도메인 조합을 분석하였으며, 따라서 도메인 조합을 쌍 도메인 조합으로 제한시킬 수밖에 없었다. 이들은 주로 기존의 보존 서열 추출 방법과 마찬가지로 도메인 조합의 출현 빈도를 기반으로 보존 정도를 측정하였다. 이는 각각 도메인이 분자기능의 특이성 혹은 다양성에 따라 출연빈도가 결정될 수 있기에 조합의 의미나 도메인의 협업 가능성을 내포할 수는 없다. 예를 들어 키나아제 도메인의 경우 세포 내 신호전달 경로에서 광범위하게 필요한 기능을 포함하므로 여러 단백질에서 나타나며, 따라서 키나아제를 포함하는 조합 또한 높은 출현 빈도를 가진다. 이러한 도메인 조합의 높은 빈도는 해당 기능의 다양성 혹은 유용성을 나타낼 뿐, 타 조합에 비해 조합 내 기능협업의 의미가 크다고는 말할 수 없다.

또한 기존의 연구는 도메인 조합의 생물학적 의미 또는 기능에 대한 분석이 충분하지 않았다. 단일 단백질 내 도메인들의 기능 연계를 밝혀내려면, 이러한 제한조건을 극복한 방법론과 체계적 연구가 필요할 것이다.

보존된 도메인 조합을 알아내고, 이의 기능적 특성을 분석하기 위해서는 각각의 도메인 조합을 보존 정도를 평가할 수 있는 방법론의 정립이 필요하다. 본 논문에서는 보존 도메인 조합을 밝혀내기 위해, 도메인 조합의 출현 빈도를 나타내는 *support*와 함께, 연관성 규칙 기법의 하나인 *all-confidence*를 사용할 것을 제안한다. *all-confidence*는 도메인 조합내의 도메인 간 상호 의존성을 수치화 할 수 있으며[4], 이는 이전 연구들에서 단방향 의존성을 측정한 것과는 차별화 된다. 따라서 제안된 기법은 기존의 제한 조건을 극복할 수 있다.

본 연구에서는 *S.cerevisiae* 단백질을 대상으로 제안된 방법론을 이용하여 보존 정도가 높은 도메인 조합을

도메인 패턴으로 정의하고, 각 도메인의 Gene Ontology(GO) term[5]을 사용하여 도메인 패턴에 대한 생물학적 의미를 분석하였다. 그 결과 *S.cerevisiae* 단백질에는 보전 정도가 높은 도메인 조합이 존재하며, 이는 GO term의 molecular function과 상관관계가 높음을 밝혀냈다. 또한 제안된 방법론을 이용하여 추출된 보전 도메인 조합은 기존의 출연빈도만을 보는 것보다 기능 웅집도와의 상관관계가 큼을 알 수 있었다. 따라서 제안된 방법론이 구성원간 협업하는 보전도메인조합을 추출하는데 기존의 방법론보다 우수하다 할 수 있으며, 또한 *S.cerevisiae* 단백질의 도메인 조합이 우연히 이루어 진 것이 아니라 분자적 기능을 수행하기 위해 특정 도메인들이 함께 진화, 보전되어왔다고 결론 내릴 수 있다.

## 2. 보존된 도메인 조합 추출

### 2.1 연관성 규칙

도메인의 동시 출연 빈도와 조합 내 도메인간의 상호 의존도는 해당 조합의 진화상 보전 정도를 나타내는 기준이 된다. 본 연구진은 보존 도메인 조합을 추출하기 위하여 데이터마이닝에서 널리 사용되는 연관성 규칙의 개념 중 *support*(식 (1))와 *all-confidence*(식 (2))를 차용하여 출연빈도 및 상호의존도를 측정하였다.

$$supp(dc) = \frac{|\{p \mid p \in \wedge dc \subset p\}|}{|P|} \quad (1)$$

식 (1)에서 *dc*는 하나의 도메인 조합을 나타내며, *P*는 한 생물 종의 전체 단백질 집합, *p*는 *P*에 소속된 하나의 단백질을 나타낸다. 즉 식 (1)의 *support*는 전체 단백질 중 해당 도메인 조합을 가지는 단백질의 비율을 나타낸다.

*all-confidence*는 연관성 규칙을 이용한 기법에서 일 반적으로 사용되는 *confidence*를 용용한 개념이다. *confidence*란 어떠한 조합 내의 두 원소가 함께 나타날 신뢰도를 나타내는 척도로써, 두 원소 *X*, *Y*의 *confidence*는 *X* => *Y*로 표현되며, *X*가 출현할 때 *Y* 또한 출현할 확률을 말한다.

*all-confidence*는 이를 확장한 개념으로 하나의 집합 내의 모든 부분집합 및 원소들이 가질 수 있는 모든 *confidence*의 최소값을 뜻한다. 도메인 조합 추출의 경우, *confidence*를 적용시키면, 실행되는 조건, 즉 기준 도메인이 존재하므로 조합 내 도메인 간의 상호 의존성을 측정할 수는 없다. 이에 반해 *all-confidence*는 기준 도메인 없이 상호의존성을 나타내므로 보존 도메인 조합 추출에 알맞은 척도이다. 도메인 조합 (*d1*, *d2*, *d3*)의 *all-confidence*가 1일 경우 그 조합은 항상 함께 나타난다고 볼 수 있으며, 따라서 단백질의 기능 구성을 위해 서로 꼭 필요하다고도 볼 수 있다. *all-confidence*

의 계산식은 식 (2)와 같다.

$$all-conf(dc) =$$

$$\frac{|\{p \mid p \in P \wedge dc \subseteq p\}|}{MAX\{i \mid \forall l(l \in PowerSet(dc) \wedge l \neq \emptyset \wedge l \neq dc \wedge i = |\{p \mid p \in P \wedge l \subseteq p\}|\})} \quad (2)$$

## 2.2 도메인 패턴 추출

본 논문에서는 단백질에서 나타나는 도메인 조합 중 높은 보존 정도를 가지는 조합을 도메인 패턴이라 정의하였다. 즉 도메인 패턴은 일정 수준 이상의 보존 정도를 가지는 조합을 말하며, 미리 설정된 최소 support와 all-confidence을 이용하여 추출할 수 있다.

도메인 패턴 추출에 사용된 한가지 다른 개념은 *maximal property*이다. 어떠한 도메인 조합이 가지는 부모집합들은 부모집합의 support와 all-confidence값과 같은 값, 혹은 초과 값을 가질 수 있다. 초과 값을 가지는 경우, 부분집합은 부모집합보다 더 높은 보존 정도를 가지지만, 같은 support와 all-confidence 값을 가질 경우에는 그렇지 않으므로, 부모집합의 일부로서의 의미만을 가진다. 따라서 도메인 패턴은 그 조합을 포함하고, 같은 보존 정도를 가지는 부모집합이 존재하지 않아야 하며, 본 논문에서는 이러한 특성을 *maximal property*로 정의하였다.

도메인 조합은 level-wise pattern tree 생성 알고리즘으로 쉽게 구할 수 있으며, 이를 바탕으로 생성된 조합을 미리 정의된 최소 support와 all-confidence 및 maximal property를 이용하여 도메인 패턴을 추출한다.

## 3. 도메인 조합의 생물학적 기능 분석

본 연구에서는 보존된 도메인 조합의 도메인들이 기능적 협업을 위해 보존되었다는 가설을 바탕으로 조합에 소속된 도메인에 대한 Gene ontology[5] 주석(GO term)을 분석하여 조합 내 도메인들의 기능 응집도, 즉 기능의 유사도를 계산하였다. Gene ontology는 단백질 RNA, 도메인과 같은 유전자 생성물에 대한 주석을 cellular component, biological process, molecular function, 세 분류로 설명하여 도메인 조합의 생물학적 의미를 더욱 체계적으로 설명할 수 있으며, 따라서 기능 분석 또한 세 분류 각각 따로 실시하여야 한다. 두 GO term의 기능적 유사성을 측정하기 위해 FuSSiMeg 평선[6]을 사용하였으며, 계산된 수치를 바탕으로 조합 내의 도메인 전체에 대한 기능의 유사 정도를 나타내는 식 Internal Function Similarity(IFS)를 고안하였다. FuSSiMeg는 Jiang and Conrath's semantic similarity measure를 사용하여 information content와 각 GO term의 분산을 고려한 온톨로지 상의 거리를 계산하는 방법으로, 두 GO term의 유사성을 측정하는 데에 널리 사용된다. IFS는 조합 내 도메인들이 가질 수 있

는 모든 도메인 쌍의 기능 유사도의 평균값으로, 도메인 조합이 생물 기능에 대해 협업하는가를 나타내는 척도로써, 주어진 조합 dc의 IFS는 식 (3)으로 정의된다. 식에서  $d$ 는 조합 dc에 소속된 하나의 도메인이며, FussiMedg() 함수는 두 도메인의 GO 상 거리를 계산하는 함수이다.

$$IFS(dc) =$$

$$\frac{\sum\{s \mid \forall d_i d_j (d_i \in dc \wedge d_j \in dc \wedge i < j \wedge s = FussiMedg(d_i, d_j))\}}{\binom{|dc| \cdot (|dc|-1)}{2}} \quad (3)$$

## 4. *S.cerevisiae* 단백질의 보존 도메인 조합 및 그의 생물학적 의미 분석

### 4.1 IFS 분포

본 연구에서는 UniProt Knowledgebase[7]에 보고된 7449개의 *s.cerevisiae* 단백질을 대상으로 제안된 방법론을 적용하여 각각 도메인 조합의 보존 정도를 측정하였다. 7449개의 *s.cerevisiae* 단백질은 총 1535개 도메인들의 조합으로 이루어지며, 단백질에서 한번 이상 발견되는 도메인 조합은 총 340459개로 밝혀졌다. 이렇게 찾아진 각 조합들의 support와 all-confidence를 계산하였으며, 이중 1758개의 maximal property를 가지는 조합을 도메인 패턴의 후보군으로 간주하여 기능 응집도, 즉 IFS를 계산하였다. 그림 1은 GO term 중 molecular function에 대하여, 추출된 도메인 패턴 후보군의 support에 따른 IFS 분포를, 그림 2는 all-confidence에 따른 IFS의 분포를 나타낸다.

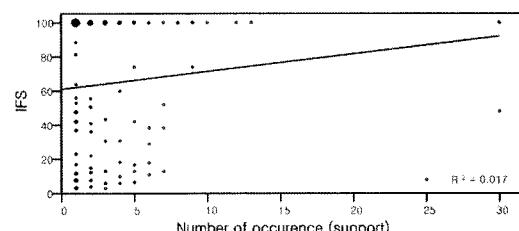


그림 1 support에 따른 molecular function의 IFS분포

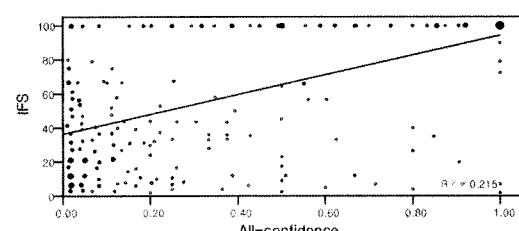


그림 2 all-confidence에 따른 molecular function의 IFS분포

그림 2 그래프의 각 점의 크기는 도메인 조합의 개수를 나타낸다. 그 분포에서 보듯이 각 조합의 *all-confidence*가 커질수록 IFS 또한 높아지는 양상을 볼 수 있다. 그림 1의 그래프에서도 *support*가 커질수록 평균 IFS 또한 높아지는 양상을 볼 수 있지만 *all-confidence*의 경우 보다는 그 상관관계가 작다고 말할 수 있다. GO term의 다른 생물학 기능 분류인 biological process와 cellular component의 경우에도 molecular function의 경우와 유사한 경향이 나타나나, 그 상관관계는 뚜렷하지 못하다. 표 1은 계계적인 분석을 위하여 실험 결과의 피어슨 상관관계를 측정한 표이다. 상관관계 분석 결과 모든 GO term 분류에서 IFS와 *all-confidence*, 그리고 IFS와 *support*는 어느 정도 상관관계를 보인다. 하지만 cellular component에 경우에는 *p*값이 유효수준 0.01을 넘지 못함으로 통계적으로 상관관계가 인정되지 않는다. 특이한 사항은 *support*보다 *all-confidence*가 IFS와 상관관계가 높다는 것이다. 즉 기존에 보존된 도메인 조합을 찾기 위해 사용되던 기법인 동시 출현 빈도(*support*)보다 새로이 제안된 *all-confidence*가 그 효용성이 크다 말할 수 있다.

표 1 피어슨 상관관계 분석

IFS	<i>all-confidence</i>		<i>support</i>		candidates used
	correlation	<i>p</i>	correlation	<i>p</i>	
<i>IFS<sub>mol</sub></i>	0.464	0.000	0.205	0.000	389
<i>IFS<sub>bio</sub></i>	0.268	0.000	0.129	0.058	217
<i>IFS<sub>cell</sub></i>	0.267	0.076	0.180	0.237	45

*p*: significant value

*mol*: molecular function go term category

*bio*: biological process go term category

*cel*: cellular component go term category

#### 4.2 *S.cerevisiae* 단백질의 도메인 패턴

높은 보존 정도를 지닌 도메인 조합, 즉 도메인 패턴을 추출하기 위해서는 기준이 되는 최소 *support* 및 *all-confidence*의 설정이 필요하다. 본 연구에서는 각 값에 따른 조합의 분포를 고려하여, 최소 *support*를 0.00057로, 최소 *all-confidence*를 0.3으로 정하였으며, 그 결과 1,758개의 후보 중 579개의 도메인 패턴이 추출되었다. 이를 바탕으로 추출된 도메인 패턴의 생물학적 의미를 보기 위해, 패턴과 패턴으로 승격되지 못한 일반 도메인 조합의 IFS를 비교하였다.

그림 3은 패턴과 일반 도메인 조합의 IFS 분포를 상자 그림으로 나타내고 있다. GO term 분류 중 molecular function에 대하여, 각 그룹은 잘 구분된 IFS 값을 가지는 것으로 보인다. 패턴 그룹의 IFS 중간 값은 100 인데 비하여 일반 도메인 조합의 중간 값은 패턴 그

룹의 4사분위수와 같은 50에 근사한 값을 가진다. Biological process 분류에 대해서는 패턴 및 일반 조합이 크게 다른 양상을 갖지 않는다. 이는 단백질 조합의 보존 현상이 biological process와는 크게 상관이 없음을 나타낸다. Cellular component의 경우는 molecular function과 다르게 패턴 및 일반 조합이 모두 IFS 100을 가지는 양상을 보인다. 이는 단백질 및 도메인의 생물학적 의미에 부합한다고 볼 수 있다. 단백질 내 도메인들은 다양한 분자기능(molecular function)을 가지고 이의 직/간접적 협업으로 최종적인 단백질 기능을 구성하지만, 하나의 단백질에 소속되기 때문에 세포 내의 위치, 즉 cellular component는 같을 수밖에 없는 것이다.

표 2는 그림 3 분포의 통계적 분석을 위해 T-test를 실시한 결과로, 그림 3의 분석과 동일한 양상을 가진다. Molecular function의 경우에만 *p*가 유효수준 0.01 이상의 값을 가지므로, 패턴과 일반 조합이 통계적으로 다른 그룹이라고 말할 수 있다. 따라서, 본 연구에서 추출한 도메인 패턴은 그 도메인 원소들 사이에 기능적 유사성이 일반 조합에 비하여 크다고 할 수 있다. 이러한 결과를 바탕으로, 보존된 도메인 조합은 분자기능과 같은 소단위 생물학적 기능의 협업과 관련 있음을 알 수 있다.

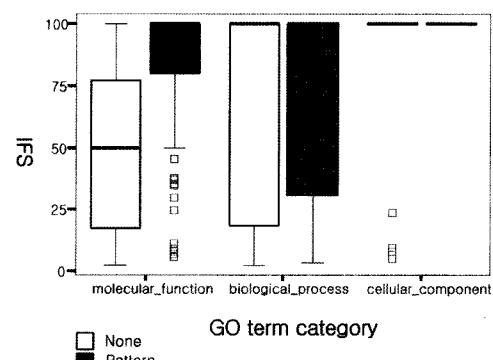


그림 3 패턴과 일반 도메인 조합의 IFS 분포 비교

표 2 T-test 결과

	#Case	Avg.	SD	t	p	
molecular function	<i>pattern</i>	82	81.5	30.9	8.30	0.000
	<i>none</i>	399	49.7	34.5		
biological process	<i>pattern</i>	60	72.1	38	1.36	0.175
	<i>none</i>	217	64.5	39.6		
cellular component	<i>pattern</i>	11	100	0	2.33	0.024
	<i>none</i>	45	89.9	28		

Avg.: Average of functional similarity.

SD: Standard deviation

t: t-Value

*p*: significant value

## 5. 결 론

본 논문은 하나의 단백질체에서 보존된 도메인 조합의 추출을 위한 체계적인 기법으로, *support*로 표현되는 도메인의 동시 출현 빈도 및 *all-confidence*로 표현되는 조합 내 도메인간의 상호 의존도를 기반한 기법을 제안하였다. 제안된 기법은 동시 출현 빈도만을 사용하는 기준의 기법과는 달리 조합 내의 상호 의존성 또한 고려하므로, 기준의 방법론 보다 더욱 보존 도메인 조합의 생물학적 의미에 부합한다 할 수 있다. 제안된 기법을 *s.cerevisiae* 단백질에 적용하고 각 조합의 기능을 분석한 결과, 조합 내 도메인간 기능 유사도는 *support*보다 *all-confidence*와 더욱 상관관계가 큼을 관찰하였다. 이는 도메인간 상호의존도를 고려하는 본 기법이 동시 출현 빈도만을 고려하는 기준 기법에 비해 기능 협업을 위해 보존된 도메인 조합을 추출하는 데에 우수함을 뜻한다.

제안된 기법을 사용하여 *s.cerevisiae* 단백질에서 597개의 보존 정도가 높은 도메인 조합, 즉 도메인 패턴을 추출하였으며, 이를 바탕으로 보존도메인조합의 생물학적 특성을 분석하였다. 기능 분석에는 GO term을 이용하여, 조합 내 도메인들이 유사한 기능을 가지고 협업하는 정도, 즉 IFS(inner functional similarity)를 고안하여 측정하였으며, 그 결과 추출된 도메인 패턴은 GO 분류의 molecular function과 깊은 관련이 있음을 밝혔다. 이는 도메인 패턴이 우연히 생성된 것이 아닌, 진화상에서 molecular function의 협업을 위해 결성되고 보존된 하나의 팀임을 뜻한다 하겠다.

본 연구 결과, 보존 도메인 조합 추출을 위해서는 조합의 출현 빈도뿐 아니라 조합 내 도메인의 상호 의존도를 고려하는 것이 타당하다 할 것이다. 또한 단백질의 분자기능을 밝힐 때에는 기준과 같이 도메인을 각각 따로 살펴보는 것 보다는 도메인 조합, 특히 보존된 조합의 기능 협업을 살펴보는 것이 중요하다 할 수 있을 것이다. 본 기법은 도메인 조합의 보존 정도를 수치화 하므로, 대량 계산을 이용한 단백질 기능 예측 기법에 활용될 수 있으며, 잘 정제된 도메인 패턴은 단백질 기능 예측 및 도메인 간 협업 프로세스를 밝히는 데에 중요한 단서로 작용할 것으로 기대된다.

향후 본연구진은 제안된 기법을 바탕으로 기능 분석의 고도화를 통하여 도메인간 협업에 대한 심화 연구를 계획하고 있다. 본 연구에서는 도메인들이 유사한 기능을 가지고 직접적으로 협업하는 경우를 중심으로 분석하였으나, 몇몇 유전자에 대한 기능 협업 연구는 다른 간접적 협업의 형태를 보고하고 있다. 따라서 도메인 조합의 기능적 의미를 밝히기 위해서는 좀더 세밀한 협업 모델이 필요할 것이다.

## 참 고 문 헌

- [1] Apic G., Gough J. and Teichmann S. "Domain combinations in archaeal, eubacterial and eukaryotic proteomes," *J. Mol. Biol.*, Vol.310, pp. 311–325, 2001.
- [2] Jacob, F., Evolution and Tinkering, *Sci.*, Vol. 196 pp. 1161–1166, 1977.
- [3] Achila D, Banci L, Bertini I, Bunce J, Ciofi-Baffoni S, HuffmanDL, "Structure of human Wilson protein domains 5 and 6 and theirinterplay with domain 4 and the copper chaperone HAH1 in copperuptake," *Proc. Natl. Acad. Sci. U S A*, Vol. 103(15): pp. 5729–5734, 2006.
- [4] E. R. Omiecinski, "Alternative interest measures for mining associations in databases," Vol.15, No.1, pp. 57–69, 2003.
- [5] Consortium, T. G. O., "Gene ontology: tool for the unification of biology," *Nature Genet.*, 25, pp. 25–29, 2000.
- [6] F. Couto, M. Silva, and P. Coutinho. "Implementation of a functional semantic similarity measure between gene products," *Department of Informatics*, pp. 3–29, 2003.
- [7] <http://au.expasy.org/sprot/>