Regular paper | Journal of the KIECS. pp. 1397-1404, vol. 13, no. 6, Dec. 31 2018, t. 92, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2018.13.6.1397

인간 단백질 분석을 위한 빅 데이타 기반 RMF 방법

김은미* · 정종철* · 이배호**

A Big Data Based Random Motif Frequency Method for Analyzing Human Proteins

Eun-Mi Kim* · Jong-Cheol Jeong* · Bae-Ho Lee**

요 약

입체적 단백질 구조를 이용한 단백질의 분석은 3차원 데이타를 생성하기 위한 기술적인 어려움과 요구되는 높은 비용으로 인해 크게 발전하지 못하였다. 모티프(motif)는 단백질이나 유전자 염기서열의 단편(segment) 정보로 정의된다. 단순성 때문에 모티프는 다양한 분야에서 활발하고 폭넓게 응용되고 있다. 그러나 모티프 자체에 대한 포괄적인 이해와 연구는 미미하다. 이 논문이 가지는 중요성은 인공지능 기법을 활용하여 인간 단백질을 분석하는 방법으로 3가지 측면에서 찾아볼 수 있다.

(1) 현재 단백질 데이타 뱅크 (PDB)에 저장된 모든 인간의 단백질 구조를, 이에 상응하는 효소위원회 (EC)의 데이타베이스와 단백질의 구조적 특성에 따른 분류 데이타베이스 (SCOP)를 연동하여, 단백질이 가지는 고유의 특성을 모티프를 응용한 새로운 방법으로 컴퓨터를 이용하여, 분석한 최초의 종합적이고 심층적인 인간 단백질의 분석법이다. (2) 본 연구는 모티프에 의해 생성된 새로운 단백질의 특성을 계층적 클러스터링을 이용하여 단백질이 가지는 고유한 특징을 패턴 분석법과 통계 그리고 단백질 기능 분석의 세 가지 범주로 단백질의 특성을 분석한다. (3) 임의로 생성된 모티프가 단백질 내에서 가지는 빈도에 대해 빅 데이타를 활용하여 모티프의 길이를 다양화시킴과 동시에 접촉 염기와 단백질의 기능을 다각도로 분석할 수 있는 임의 모티프 빈도 (RMF)를 이용한 단백질 분석 방법론을 제안한다.

ABSTRACT

Due to the technical difficulties and high cost for obtaining 3-dimensional structure data, sequence-based approaches in proteins have not been widely acknowledged. A motif can be defined as any segments in protein or gene sequences. With this simplicity, motifs have been actively and widely used in various areas. However, the motif itself has not been studied comprehensively. The value of this study can be categorized in three fields in order to analyze the human proteins using artificial intelligence method: (1) Based on our best knowledge, this research is the first comprehensive motif analysis by analyzing motifs with all human proteins in Protein Data Bank (PDB) associated with the database of Enzyme Commission (EC) number and Structural Classification of Proteins (SCOP).

(2) We deeply analyze the motif in three different categories: pattern, statistical, and functional analysis of clusters. (3) At the last and most importantly, we proposed random motif frequency(RMF) matric that can efficiently distinct the characteristics of proteins by identifying interface residues from non-interface residues and clustering protein functions based on big data while varying the size of random motif.

키워드

Big Data Pattern Analysis, Sequence-based Method, Protein Analysis Random Motif Frequency Method
빅 데이터 패턴 분석, 염기 서열 기반 방법, 단백질 특성 분석, 임의 모티브 빈도 방법

* 전남대학교 전자컴퓨터공학과
(koreaeunmi@gmail.com korjcjeong@gmail.com)
** 교신저자 : 전남대학교 전자컴퓨터공학부
· 접 수 일 : 2018. 10. 05
· 수정완료일 : 2018. 11. 09
· 게재확정일 : 2018. 12. 15
· Received : Oct. 05, 2018, Revised : Nov. 09, 2018, Accepted : Dec. 15, 2018
· Corresponding Author : Bae-Ho Lee
Dept. of Electronics &Computer Engineering, Chonnam National University,
Email : bhlee@chonnam.ac.kr

Ⅰ. 서 론

단백질이 가지는 고유 기능과 단백질간의 상호 작용이 질병과 밀접한 관계를 가지고 있다는 것은 일반적으로 알려진 사실이다. 이러한 단백질의 특성을 분석하는 방법은 크게 네 가지로 분류할 수 있다. 우선, in-vitro 방식은 생활 반응이 있는 유기체를 제외한 통제된 환경에서 수행하는 방법으로 단순화한 조건에서 빠른 정화과정을 허용하는 Tandem Affinity Purification(: TAP)방법이 대표적이다[1-3]. 다음으로는 임상 실험을 포함한 in-vivo가 있으며, yeast two-hybrid 방법을 이용하여 이전보다 비용 절감 효과를 가져왔다. 너무 단순화된 조건과 현실의 조건 절대 보존해야 하는 in-vitro와 in-vivo의 단점을 보완하기 위해 이들의 중간단계라 할 수 있는 in-situ 방법이 사용된다[4-5]. 대표적인 방법으로는 형광 물질을 이용하여 유기체의 물리적인 위치를 식별하는 Fluorescence in situ hybridization (:FISH) 방법이다[5].

이러한 방법들은 많은 비용을 요구함으로 대용량의 데이타를 이용하여 데이터마이닝 분석 방법으로는 적합하지 않다. 이러한 단점을 극복하기 위해 최근에는 컴퓨터를 이용하여 분석하는 in-silico방법이 사용되고 있다[6]. 특히 단백질을 분석하는 방법으로는 크게 염기 서열에 기반한 방법(sequence-based method)과 3D 단백질 구조에 기반한 방법(structure-based method)으로 나눈다[7-9]. 기술의 발달로 많은 양의 3차원 단백질의 구조가 Protein Data Bank(:PDB)와 같은 공용 데이타베이스에 저장되어 쉽게 접근할 수 있지만, 기술 한계와 3D 단백질 구조 데이타를 생성하기 위한 높은 시간적 물리적인 비용으로 인해 3D 데이타의 수는 아직 염기서열 데이타의 수와 현저한 차이를 나타낸다[8-9]. 하지만 3D 데이타 정보의 양과 질은 염기서열에 비해 훨씬 우위에 있다.

따라서 염기서열 데이타로부터 3차원 단백질 구조 데이타에 견줄 수 있는 특징 추출하는 데이터마이닝 방법은 인공지능 기술 향상과 함께 주목받고 있다[10]. 이러한 염기서열을 이용한 특징 추출의 대표적인 방법으로는 모티프로 불리는 단백질 또는 유전자 염기 서열의 단편적인 정보를 이용하여 단백질을 분석하는 방법으로, 다양한 분야에서 활발히 응용되고 있다. 따라서 본 논문에서는 임의 모티프 검색을 통해 고유한 패턴 분석법과 통계, 그리고 단백질 기능 분석의 세 가지 범주로 단백질의 특성을 분석한다.

Ⅱ. 방법론

논문에 사용된 데이타는 PDB, EC, 그리고 SCOP 데이타베이스를 다음과 같은 방법으로 연동하며 생성하였으며 전체적인 빅 데이타의 가공방법과 분석은 다음과 같다.

2.1 PDB와 외부 데이타베이스의 연동 방법

1. PDB와 다른 데이타베이스를 연동하기 위해서 관련 파일들을 EMBL-EBI 홈페이지[1]에서 다운로드한다.

2. PDB에 존재하는 전체 인간 단백질 데이타를 얻기 위해 생물분류학 (Taxonomy) 파일에서 인간 분류학 번호 9606과 연계된 PDB 고유번호를 추출하고 이에 상응하는 PDB번호를 가진 단백질이 EC와 SCOP 데이타 양쪽 모두에 존재 할 경우 연구 데이타에 첨부하여 관련 파일을 PDB에서 다운로드한다. 주의할 점은 최근 업데이트로 인해 PDB에서 다운로드 하는 파일들의 확장자가 'cif'로 변경되었다. 이 논문에서 사용된 BioPython이 지원하는 양식은 'pdb'이므로 프로그램을 통해 'cif'를 'pdb'양식으로 다시 컴파일 하는 과정을 가졌다. 이 과정에서 4udf, 4v6m, 4v6x, 4v98, 5lzw, 5lzx, 5lzy는 그 크기와 복잡도로 인해 'pdb'로 변환 할 수 없었다.

2.2 접촉 염기의 식별법과 임의 모티프 생성

1. PDB에서 다운로드된 단백질 구조는 BioPython[2]과 PDB 체인 정보를 이용하여 접촉 염기를 특징한다. 이때 접촉되는 염기 지정 방법은 두 체인이 접촉할 때 상호 근접하는 염기의 최대 거리가 4.5Å 이하일 때 이를 접촉 염기로 정의하였고, 이에 선택되지 않은 염기는 비접촉 염기로 정의된다.

2. 임의 모티프 (Random Motif: RM)는 일반적인 20가지 염기 서열을 임의의 길이 내에서 조합하는 방법으로 생성이 된다. 예를 들면, 3가지 염기 서열

1) http://www.ebi.ac.uk/pdbe/docs/sifts
2) https://biopython.org/

'ABC'만 존재한다는 가정을 할 때 부분적인 중복('AAA', 'BBB', 'CCC', 'AA', 'BB', 'CC') 을 허용함으로써'AAA', 'BBB', 'CCC', 'AAB', 'AAC', 'ABB', 'ABC', 'ACC', 'BBC', 'BCC'.을 생성할 수 있다.

$$Motif = \left\{ \bigcup_{k} AA_{\binom{n}{k}},\ \alpha \right\} \quad (1)$$

$$\alpha = \left\{ \bigcup_{i=1, j=1}^{n} [AA_{i,j}]_{k=2}^{3} \right\}$$

$$where\ \ i = j,\ \ k = 1 \cdots 4$$

식 (1)에서 k는 RM의 길이를, n은 염기 (AA)의 전체 갯수를 나타내며, n = 20으로 일반적인 20가지 염기 모두를 사용하였다. α는 부분적인 염기의 중복으로 생성하는 전체 조합을 나타낸다.

2.3 염기 서열 기반의 임의 모티프 빈도 측정법 (Sequence-based Random Motif Frequency: SRMF)

SRMF은 접촉 염기가 가질 수 있는 최대 허용치를 RM의 길이 내에서 검색하는 방법으로 RM과 동일한 패턴이 존재하는 횟수를 특정된 단백질의 전체 염기 서열 내에서 측정한 방법이다. 단백질의 길이와 PDB가 갖는 체인의 수에 대한 편중을 최소화하기 위해 식 (2)를 이용하여 표준화하였다.

$$RMF = \frac{M_f}{L \cdot \binom{C}{2}} \quad (2)$$

식 (2)에서 Mf는 MR이 전체 단백질 염기 서열 내에서 매칭 횟수, L은 전체 염기의 길이, C는 특정 단백질의 체인 갯수를 나타낸다.

2.4 위치 기반의 임의 염기 서열 모티프 빈도 측정법 (Position-based Random Motif Frequency: PRMF)

PRMF는 SRMF가 가지는 복잡도를 낮추어 검색을 빠르고 효율적으로 할 수 있는 발견법적 방법(heuristic method)으로 접촉 염기는 공간적으로 근접한다는 사실에 기반한 모티프 빈도 측정법이다. 동일한 가정으로 PRMF와 SRMF 모두 모티프가 갖는 서열의 특성은 검색에 반영되지 않는다. 다시 말해 'ABC'와 'CBA'는 동일한 모티프로 간주한다. 이러한 가정으로 인해, PRMF는 SRMF와 달리 각각의 접촉 염기가 가질 수 있는 모들의 수를 검색하는 대신, 접촉 염기들이 서로 붙어 있는 가상의 염기 서열을 기반으로 모티프 빈도를 측정하는 방법이다. 이러한 방법으로 본 논문에서는 빅 데이타베이스로서 SCOP 데이타베이스 4,180개 그리고 EC 데이타베이스 3,796개, 전제 7,976개의 단백질복합구조를 생성하였다.

III. 실 험

3.1 패턴분석을 통한 단백질 특성분석법

이 실험에서는 EC와 SCOP으로부터 얻은 단백질 구조로 부터 접촉염기와 비접촉 염기의 패턴을 비교 분석한다.

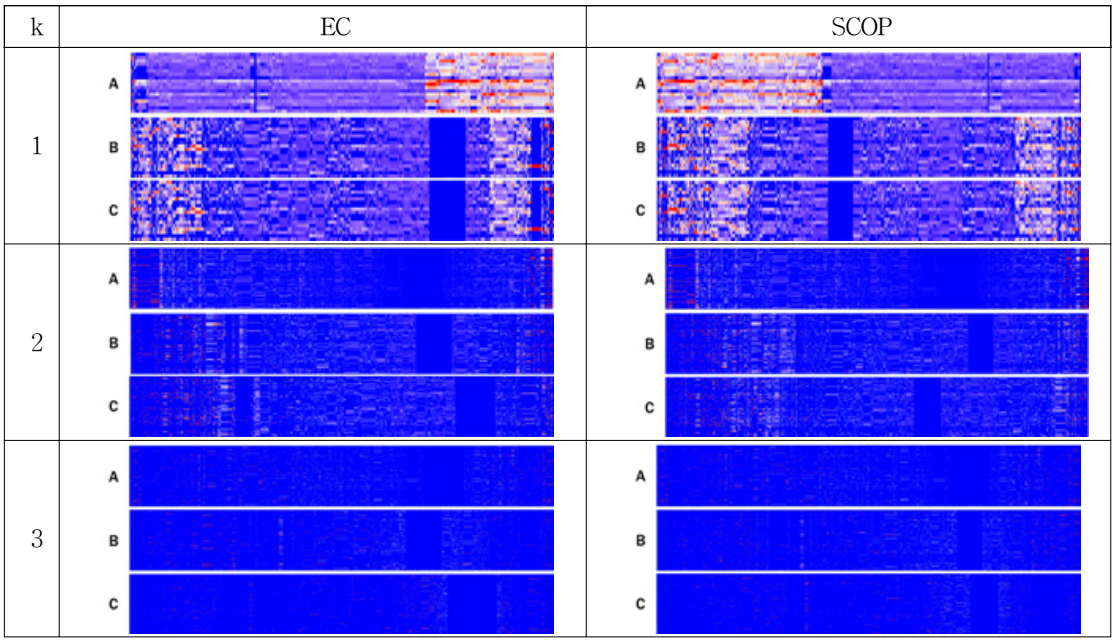

그림 1. RM 히트맵 비교(Random Motif Heatmap Comparison).
(A) 서열 기반 비접촉 임의염기서열모티브 측정으로부터의 히트맵 ,
(B) 서열 기반 접촉 임의염기서열모티브 측정으로부터의 히트맵
(C) 위치 기반 접촉 임의염기서열모티브 측정으로부터의 히트맵

이를 비교 분석하기 위해 유클리디안 거리측정을 통해 RMF 벡터를 클러스터링하고 이를 히트맵을 통해 도식화하였다. 적용된 유클리디안 거리는 식 (3)에 의해 정의된다.

$$Distance = \sqrt{\sum_{i=1}^{\binom{20}{k}+[\alpha]} (x_i - y_i)^2} \qquad (3)$$

식 (3)에서 k는 모티프의 길이를 나타내며, x와 y는 특정 단백질로부터 생성된 RMF 벡터를 나타낸다. α는 식 (1)과 동일하게 정의된다.

그림 1 에서 보듯이, EC와 SCOP는 k = 1 에서 같은 클러스터 패턴을 보여 주며, 비접촉 염기들은 접촉 염기보다 더 느슨하게 클러스터링되는 경향을 보이고 있다. 여기서 SRMF와 PRMF가 k = 1 에서 동일하다는 점이다. 이러한 결과는 k = 1 일 때 두 방법 모두 개별 염기 빈도를 계산하기 때문이다.

저해상도 이미지로 인해 구분하기가 어렵지만, k를 증가시킴으로써 (즉, k = 2 및 3), 클러스터를 더 작은 그룹으로 좁혀 클러스터 수를 증가시키는 경향이 나타낸다. 즉, 접촉 염기들은 비접촉 염기들에 비해 패턴의 특수성이 높은 경향을 보인다. 이 실험의 결과는 기존의 단백질 분석 방법론들이 모티프에 의존하는지를 부분적으로 설명할 수 있지만, RMF에 의해 생성되는 클러스터의 특성에 관해서는 설명할 수가 없다.

3.2. 통계학적 분석법

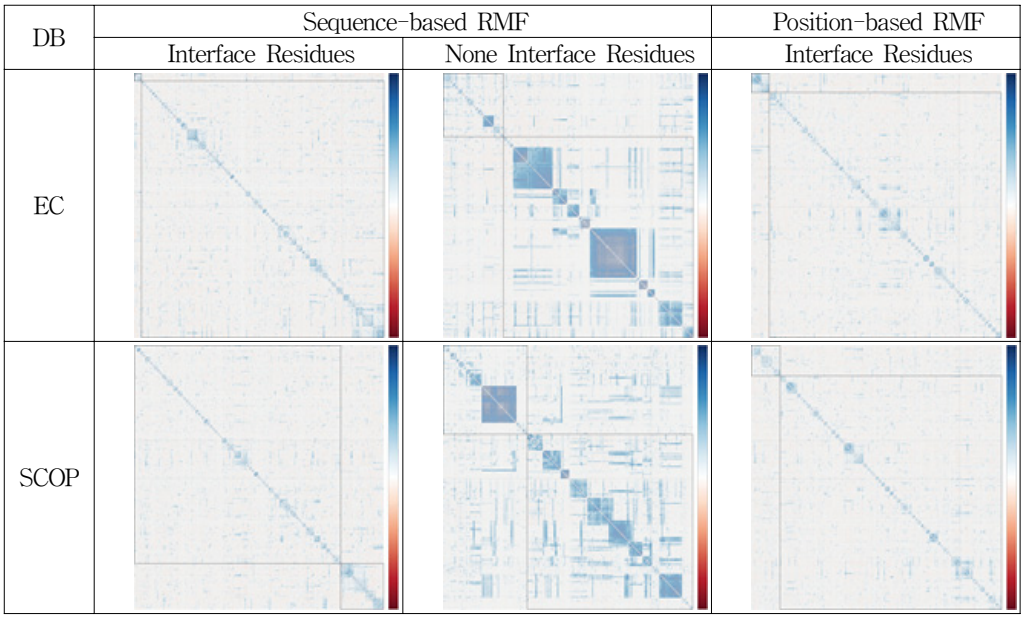

그림 2. RMF사이의 상관관계(Correlation between Random Motif Frequencies (RMF)).

클러스터의 통계적 유의성 및 경향을 확인하기 위해 RMF 벡터 간의 상관관계를 계산하였다. 그림 2는 모티프 길이가 2 (k = 2)인 개별 RMF 벡터 간의 상관관계를 색상 강도 그래프로 표시한 것이다. 두 번째 열에 나타난 비접촉 염기(None Interface Residues)는 EC와 SCOP 데이타 모두에서 RMF 상관관계가 매우 강하거나 상반되는 상관관계를 관측할 수 있다. 이는 개별 RMF에서 검색된 정보가 중복되어 샘플의 속성을 효율적으로 특성화하지 못할 수 있음을 의미한다.

반면, 접촉 염기에 대한 결과는 상관관계가 희박하게 분포되어 있음을 보여 주고 있으며, 이는 개별 RMF가 접촉 염기의 독립적인 특성을 관측하고 있음을 보여준다. 세 번째 열인 PRMF 또한 접촉 염기와 비접촉 염기의 뚜렷한 차이를 보여준다. 여기에서 주목할 사항은 위치 기반 또는 시퀀스 기반에서 파생된 RMF는 매우 유사한 패턴을 가진다는 점이다. 이는 접촉 염기들의 염기 서열과 공간적 위치 상관관계에 있음을 보여준다.

또한 이러한 실험결과는 RMF의 계산시간을 획기적으로 단축할 수 있는 발견적 교수법의 이론적 타당성을 보여준다. 이 실험의 모티프가 접촉 염기의 특징을 추출하는 원리와 성향에 대한 이해에 도움을 주지만, 모티프에 의해 특징되어진 정보의 종류에 관해서는 다음 섹션에서 다룬다.

3.3 단백질 기능의 분석

모티프에 의해 수집된 정보를 보다 잘 이해하기 위해 서로 다른 길이의 모티프를 가진 RMF 벡터로부터 생성된 클러스터를 분석한다. 그림 3 에서 좌측 상단의 그림은 EC 데이타에서 모티프의 길이가 2인 (k = 2) RMF로부터 형성된 클러스터 중 가수분해에 의해 생성된 펩티다아제를 나타내는 EC 번호 '3.4' 인 그룹을 나타낸다.

이 그룹 내에는 다른 EC 번호가 있지만 k = 3 으로 모티프의 길이가 길수록 보다 다양한 EC 코드가 클러스터 내에 포함됨을 EC와 SCOP 데이타 양측 모두에서 관측할 수 있다. SCOP 데이타의 경우 k = 2 에서 SCOP 번호 'b.17' 트립신 유사 세린 프로테아제가 주축으로 이루어진 그룹이 k = 3 으로 변화

하면서 EC 데이타보다 더욱 다양한 그룹을 포함하게 된다. 즉, RMF 기반 클러스터가 SCOP 데이타보다는 EC 데이타의 특성을 잘 표현한다.

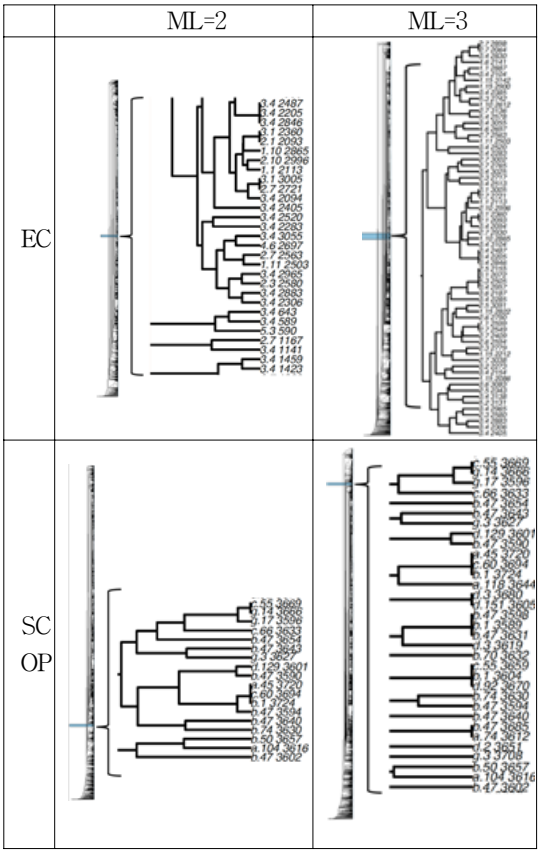

그림 3. EC와 SCOP 데이타베이스로부터 식별된 클러스터의 비교.

이는 단백질 구조에 기반한 SCOP보다는 화학 반응상태에 기반한 EC 데이타의 특성을 잘 반영하여 특정 단백질의 구조보다는 단백질의 생화학적인 역할을 보다 잘 특징화하는 경향이 있음을 보인다. RMF로부터 생성되는 클러스터들을 이해하기 위해 Gene Set Enrichment Analysis (GSEA)를 이용하여 분석하였다. GSEA는 단백질이 아닌 유전자 목록을 분석하기 위해 설계되었기 때문에 각 PDB ID는 UniProt ID에 매핑 된 다음 다시 유전자 이름으로 매핑하여 GSEA에 적용하였다.

모티프의 길이가 2 (k = 2)인 SRMF 로부터 생성된 EC 클러스터에서 통계학적으로 매우 의미가 깊은 (p-value=1.38e-10) 유전자 그룹 (GSTP1, SOD1, F7, GSTA1, WARS, HMOX1, HNMT, SULT1E)을 발견하였으며 이들은 모두 GO: 1901564 에 속한 유전체들이다.

또한, 모티프의 길이가 3 (k = 3)인 PRMF 로부터 생성된 EC 클러스터에서 통계학적으로 매우 중요한 (p-value=1.36e-12) 유전자 그룹 (F10, F2, MMP3, MMP10, F11)을 발견하였으며 이들은 모두 GO: 0017171에 속한 유전체들이다.

다른 예로 모티프의 길이가 3 (k = 3)인 SRMF 로부터 생성된 SCOP 클러스터에서 통계학적으로 매우 신빙성있는 (p-value=3.81e-10) 유전자 그룹 (GSTP1, GSTM1, GSTM2, GSTA1)을 발견하였다.

Ⅳ. 결 론

이 논문에서는 빅 데이터 기반의 심층적 분석이 전무한 모티프를 패턴, 통계, 단백질 기능 분석에 인공지능 기법을 통해 모티프의 특성에 관해 심도있게 살펴보았으며, RMF기반의 새로운 단백질 분석법을 제안하였다.

이 논문에서는 RMF를 이용한 단백질의 분석에 초점을 맞추었지만 제안된 빅 데이터기반 분석법은 차세대 유전체 분석법의 정보를 활용 통합함으로써 효과적인 치료제의 개발과 치료법의 효능을 예측 할 수 있는 광범위한 영역에 적용될 수 있는 잠재적 활용도가 매우 높다. 특히 PRMF는 SRMF 기반 모티프 검색에서의 계산 복잡도를 현저히 줄여 대용량의 단백질 데이타의 처리 능력을 향상하였다.

References

[1] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. Kumar, “Protein-Protein Interaction Detection: Methods and Analysis,” *Int. J. of Proteomics* , vol. 2014, Feb. 2014, pp. 147648.

[2] S. Xing, N. Wallmeroth, K. W. Berendzen, and C. Grefen, “Techniques for the Analysis of Protein-Protein Interactions in Vivo,” *Plant*

Physicology, vol. 171, issue 2, 2016, pp. 727-58.
[3] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Seraphin, "The Tandem Affinity Purification (TAP) Method : A General Procedure of Protein Complex Purification," *Methods*, vol. 24, issue 3, July 2001, pp. 218-229.
[4] A. Bruckner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner, "Yeast Two-Hybrid, a Powerful Tool for Systems Biology," *Int. J. Mol. Sci.*, vol. 10, issue 6, June 2009, pp. 2763-2788.
[5] M. Werner, L. Wilkens, M. Aubele, M. Nolte, H. Zitzelsberger, and P. Komminoth, "Interphase cytogenetics in pathology: principles, methods, and applications of fluorescence in situ hybridization (FISH)," *Histochem. Cell Biol.*, vol. 108, issue 4-5, 1997, pp. 381-90.
[6] X. W. Chen and J. C. Jeong, "Sequence-based prediction of protein interaction sites with an integrative method," *Bioinformatics*, vol. 25, issue 5, Mar. 2009, pp. 585-591.
[7] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein-protein interaction using a deep-learning algorithm," *BMC Bioinformatics*, vol. 18, issue 1, May 2017, pp. 277.
[8] 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.*, vol. 28, issue 1, Jan. 2000, pp. 235-42.
[9] J. Jeong, "A New Methodology For Identifying Interface Residues Involved In Binding Protein Complexes," Master's Thesis, *University of Kentucky*, 2011.
[10] H. Ceong and C. Park, "Enzyme Metabolite Analysis Using Data Mining," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 10, Oct. 2016, pp. 969-982.

저자 소개

김은미(Eun-Mi Kim)

2001년 전남대학교 공학대학 전자컴퓨터공학 전공(공학사)
2004년 전남대학교 대학원 전자컴퓨터공학과 졸업(공학석사)
2004년 ~ 현재 전남대학교 대학원 전자컴퓨터공학과 박사과정 및 수료

※ 관심분야 : Machine Learning, Bioinformatics 인공지능

정종철(Jong-Cheol Jeong)

2000년 전남대학교 공학대학 컴퓨터공학과 졸업(공학사)
2002년 전남대학교 공학대학원 컴퓨터공학과 졸업(공학석사)
2011년 University of Kansas, Dept. of Computer Science(공학석사)
2013년 University of Kansas, Dept. of Bioinformatics(공학박사)
2017년~현재 University of Kentucky, Assistant Professor in the division of biomedical informatics in college of medicine

※ 관심분야 : Bioinformatics, 인공지능

이배호(Bae-Ho Lee)

1978년 한양대학교 전자공학과졸업(공학사)
1980년 한국과학기술원 대학원 전기및전자공학과 졸업(공학석사)
1993년 University of Missouri, Columbia 졸업(공학박사)
1993년 ~ 현재 전남대학교 전자컴퓨터공학부 교수

※ 관심분야 : 인공지능, Machine Learning, 영상처리, 컴퓨터비전 등