18
논문번호 09-03-03

단백질 구조 및 기능 분석을 위한 FEATURE 시스템 개선

Deciphering FEATURE for Novel Protein Data Analysis and Functional Annotation

유 승 학*, 윤 성 로*★
Seunghak Yu*, Sungroh Yoon*★

Abstract

FEATURE is a computational method to recognize functional and structural sites for automatic protein function prediction. By profiling physicochemical properties around residues, FEATURE can characterize and predict functional and structural sites in 3D protein structures in a high-throughput manner. Despite its effectiveness, it has been challenging to apply FEATURE to novel protein data due to limited customization support. To address this problem, we thoroughly analyze the internal modules of FEATURE and propose a methodology to customize FEATURE so that it can be used for new protein data for automatic functional annotations.

요 약

FEATURE는 단백질 내에서 특정 기능이나 구조를 가지고 있는 site의 미세환경분포를 이용하여 다른 단백질 내에서 이와 유사한 미세환경을 가지고 있는 부분을 찾아 그 분분이 site일 확률을 수치적으로 제시해 줌으로써 사용자로 하여금 site의 존재 유무와 그 위치를 판단하는데 기준을 제공해주는 유용한 툴이다. 하지만 기존의 FEATURE에서 사용된 데이터 이외의 새로운 단백질 구조 데이터를 FEATURE에 적용하기 위해서는 FEATURE 내부의 module을 입력 데이터 구조에 맞게 수정해야 한다. 그러나 FEATURE 내부의 module 구조를 수정하는 방식이 직관적이지 않기 때문에 많은 연구자들이 FEATURE를 원활하게 사용하지 못하였다. 따라서 본 논문에서는 FEATURE의 내부 구조를 분석하고 FEATURE를 새로운 단백질 데이터에 적용하기 위한 방법을 제시한다.

Key words : FEATURE, protein structure, microenvironment, computational biology, biophysical properties

Ⅰ. 서론

구조 유전체학의 발달로 인해 수많은 단백질 3차 구조의 정보들이 데이터베이스화 되고 있다 [1]. 이를 통해 단백질 1차 구조(서열) 혹은 단백질 2차 구조만 가지고는 분석할 수 없었던 단백질의 구조적 기능에 대한 연구가 가능하게 되었다. 따라서 분자생물학의 중요 목표 중 하나인 단백질의 기능을 이해하는 것에 좀 더 가까이 갈 수 있게 된 것이다 [2]. 하지만 단백질의 기능을 개별적으로 연구하는 기존의 방식으로는 해결할 수 없는 속도로 단백질 구조 정보가 범람하게 되었다 [3].

이에 따라 단백질의 3차원 구조를 자동적으로 분석하는 도구에 대한 수요가 증가하였고 여러 가지 도구가 개발되기 시작하였다 [3]. Wallace가 개발한 PRO CAT은 기하학적으로 residue를 분석하여 catalytic site를 찾아낸다. 이 방식으로는 새로운 단백질 내에서 caltalytic의 존재유무를 아는데 유용하게 사용할 수 있다 [4-5]. Fetrow와 Skolnick이 개발한 Fuzzy Functional Forms (FFF)는 residue 간의 거리를 계산해 냄으로서 가장 근접한 residue 쌍을 찾아내어 단백질이 꼬이는 구조를 알아낼 수 있다 [6-7]. FEA

* 高麗大學校 電氣電子電波工學部
(School of Electrical Engineering, Korea University)
★ 교신저자 (Corresponding author)
接受日:2009年 9月 4日, 修正完了日: 2009年 9月 23日

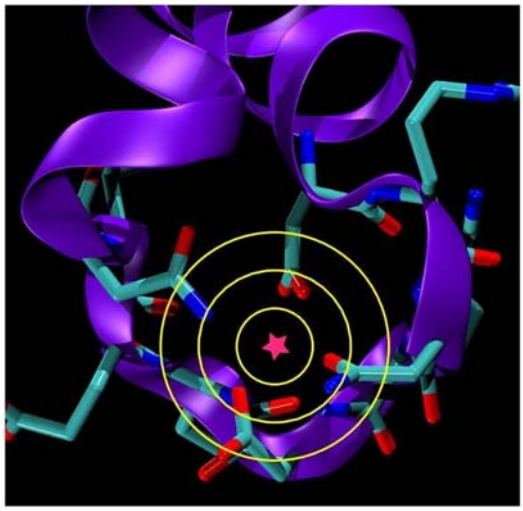

Fig. 1. Example of site [8]
그림 1. site의 예시 [8]

TURE[2]는 이러한 도구 중의 하나로 단백질 내의 특정 기능이나 구조를 가지고 있는 site를 찾아 단백질 구조의 분석을 돕는다.

특히 FEATURE는 다른 방식과 다르게 특정 단백질의 확정된 특성들을 미리 알 필요가 없고 심지어 그것들을 자동적으로 찾아내는 기능을 하기 때문에 성능 면에서 가장 월등하다고 볼 수 있다 [2]. 따라서 본 논문에서는 FEATURE의 중점을 두어 단백질 구조분석과 FEATURE를 새로운 단백질 데이터에 적용하는 방법에 대해 알아 볼 것이다.

먼저 FEATURE는 그림 1에서 알 수 있듯이 site를 중심으로 하는 동심구를 사용자가 원하는 수만큼 생성하여 각 구체 내에 들어있는 물리화학적 특성요소의 분포를 수치화 한 뒤 이를 통해 site의 미세환경분포를 구성한다 [3,11].

이렇게 구성된 site의 미세환경분포와 유사한 분포를 가지고 있는 부분을 단백질 구조 데이터베이스 내에서 찾아냄으로써 site의 존재 유무를 확률로 나타내어 사용자로 하여금 판단의 기준을 제공해 주는 유용한 도구라 할 수 있다.

하지만 FEATURE는 개선되어야 할 부분들이 있다. 먼저 FEATURE가 supervised machine learning에 의존한다는 것이다. 이는 site를 검색해 내는 기능을 기존에 알고 있는 site로 제한하게 되어 새로운 site를 발견 할 수 없게 된다. 또한 FEATURE가 site의 model을 만들 때 non-site의 영향을 직접적으로 받기 때문에 non-site 선택에 주의를 기울일 필요가 있다. 마지막으로 입력으로 들어가는 데이터의 차원이 고정되어 있기 때문에 사용자가 원하는 신규 데이터를 적용하기에는 어려움이 따른다.

따라서 본 논문에서는 FEATURE 시스템의 내부구조를 분석한 뒤 FEATURE를 구성하고 있는 핵심 module의 설명과 더불어 FEATURE의 module을 수정하는 방법을 제시하여 다른 연구자들이 FEATURE를 편리하게 이용할 수 있게 하였다.

II. 본론

1. FEATURE 개요

FEATURE는 supervised machine learning 알고리즘을 사용하여 단백질 내에서 특정 구조 혹은 기능을 지닌 site의 미세환경분포를 추출한 뒤 다른 단백질 내에서 이와 유사한 미세환경분포를 가진 부분을 찾아내는 기능을 하는 시스템이다.

먼저 사용자가 원하는 구조 혹은 기능의 site를 결정하고 이와 대조군으로 non-site를 선택한다. 이때 non-site란 site의 미세환경분포에서 어떤 특징이 site를 결정하는데 주요한 역할을 하는지 알아내기 위해 사용되는 것으로 랜덤으로 선택할수록 편향되지 않은 site model을 만들 수 있다 [9].

site와 non-site가 선택되면 Wilcoxon ranksum[2]을 통하여 site의 미세환경분포 model을 만들게 된다 [10]. 이러한 model이 가지고 있는 데이터의 범주는 크게 세 가지로 나눌 수 있는데 site가 non-site에 비해 수치가 큰 부분, site와 non-site의 수치가 비슷한 부분, site가 non-site보다 수치가 적은 부분이다.

이렇게 만들어진 model을 바탕으로 특정 단백질 내에서 알고 싶은 부분을 선택한 뒤 이에 대한 주변환경의 물리화학적 분포를 얻고 Bayesian scoring을 사용하여 site 판단 기준을 제시한다 [2]. 다시 말해서 site의 model이 가지고 있는 정보를 바탕으로 query sample의 데이터가 site에 가까운지 non-site의 가까운지를 확률적으로 계산하여 정수로 바꾼 뒤 이들의 합으로 score를 매기는 것을 말한다.

2. FEATURE 내부구조

FEATURE의 내부구조는 그림 2에서 볼 수 있듯이 크게 3가지로 나눌 수 있다. 먼저 site, non-site, query sample 주변의 미세환경분포를 추출하는 featurize module이 있고 site와 non-stie의 미세환경분포를 이용하여 site의 model을 만드는 buildmodel module이 있다. 마지막으로 이렇게 만들어진 model을

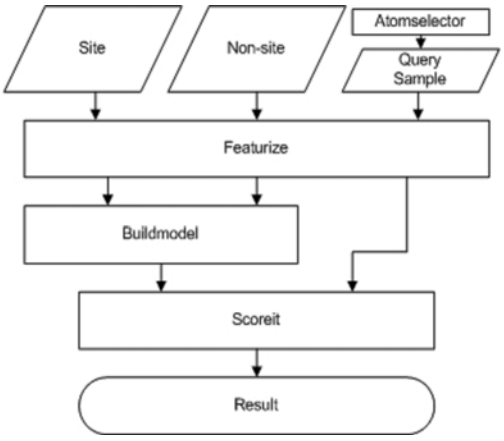

Fig. 2. Flowchart of FEATURE
그림 2. FEATURE의 순서도

이용하여 query sample의 미세환경분포와 비교분석을 통해 query sample의 site 가능성을 제시해 주는 scoreit module이 있다. 그 밖에 atomseletor module이 있어서 query sample을 선택하는데 있어 site와 동일한 residue를 가지는 sample을 선택하여 계산량을 줄여주는 역할을 한다. 각각의 내부구조를 좀 더 자세히 살펴보면 다음과 같다.

가. Featurize

단백질 데이터 뱅크(PDB)[2]의 ID를 입력으로 하여 그 부분의 미세환경분포를 추출해 내는 함수이다. 이를 사용하여 site와 non-site 그리고 query sample의 미세환경분포를 얻어 낼 수 있다. 이때 사용되는 property는 80 x 6 으로 6개의 동심구에서 각각 관찰된 80개의 물리화학적 특성들로 이루어져 있다. 이때 6개라는 동심구의 숫자는 featurize 내부의 파라미터를 바꾸어 조절할 수 있지만 80개의 특성은 고정적이기 때문에 이를 원하는 값으로 바꾸기 위해서는 사용자가 직접 본 module을 수정해야 한다. 이를 통해 단백질 내의 특정 부분의 물리화학적 정보가 수치화 되어 얻어지게 된다.

나. Buildmodel

앞서 Featurize를 사용하여 얻어낸 site와 non-site의 미세환경분포를 이용하여 site의 model을 만들어 낸다. 이러한 model은 앞서 설명한 것과 같이 site와 non-site의 수치들을 비교하여 그림 3과 같은 결과를 생성하게 된다. 그림에서 초록색으로 표시된 부분은 site가 non-site에 비해 수치가 높은 것이고 흰색은

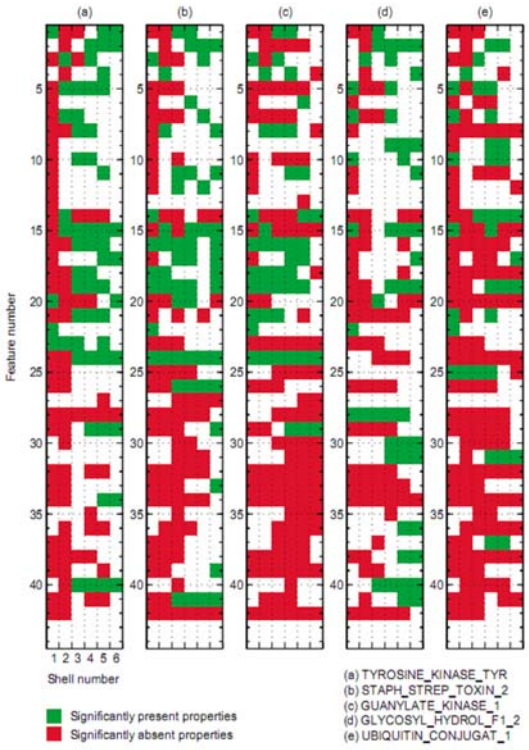

Fig. 3. Fingerprints of over- or underrepresented features. [1]
그림 3. 통계적으로 유의한 특성 분포 [1]

site와 non-site가 비슷한 수치를 가지는 것이다. 마지막으로 붉은색은 non-site가 site 보다 높은 수치를 지니는 것을 말한다.

따라서 랜덤으로 선택된 non-site를 가정 했을 때 query sample의 데이터가 site model의 초록색 부분과 근사한 부분이 많을수록 site에 가까운 것을 알 수 있다.

다. Atomselector

만들어진 site의 model을 바탕으로 단백질을 분석하기 위해서는 site와 같은 residue와 atom을 지닌 query sample을 얻어내는 과정이 필요하다. 이를 통해 residue와 atom이 다른 부분과 비교하는 과정을 생략하여 시간과 계산량을 단축시킬 수 있기 때문이다.

atomselector라는 module을 이용하여 원하는 residue와 atom을 입력하면 단백질 내에서 해당되는 부분의 PDB ID를 출력해주게 된다. 이를 featurize module을 이용하여 sample의 미세환경분포를 얻어내야 앞서 만든 site의 model과 비교 가능하다.

Table. 1. Description of the 44 FEATURE properties [1]
표. 1. 본 연구에 사용된 44개 특성의 목록 [1]

라. Scoreit

Site의 model과 query sample의 미세환경분포가 준비되었다면 scoreit 함수를 사용하여 sample이 얼마나 site에 가까운지 확률을 계산하게 된다.

이때 Bayesian scoring을 사용하여 site일 확률을 score로 환산해 사용자로 하여금 판단의 기준을 제공하는 것이다. site와 비슷할수록 큰 양의 정수를 가지며 site와 다를수록 큰 음의 정수를 가지게 된다 [2].

이렇게 나온 score를 토대로 사용자는 단백질 내의 어느 부분이 site에 가까운지를 결정하는데 도움을 받을 수 있다.

3. 신규 데이터를 위한 내부구조 개선

가. PropertyConstants

FEATURE의 입력은 앞서 말한 대로 80 x 6 차원의 데이터로 되어있다. 이는 따라서 원하는 데이터가 그 규격이 아니라면 FEATURE는 동작하지 않게 된다. 이를 가능하게 하기 위해서는 몇 가지의 함수를 수정하여야 하는데 가장 우선시 되는 것은 Property Constants 안의 enumeration이다. enumeration은 80개의 특성으로 이루어져 있는데 이를 원하는 개수의 특성으로 수정하여야 한다. 우리는 80개의 특성을 44개로 축소시킨 44 x 6 차원의 데이터를 사용하였다. 44개의 특성은 표 1번과 같다.

나. PropertyList

PropertyConstants와 PropertyList는 밀접한 관계가 있기 때문에 PropertyList 안의 구조체 역시 Property Constants안의 enumeration과 그 명칭과 순서가 동일하도록 수정해 주어야 한다. 다시 말해 구조체가 표 1에 명시된 특성을 다룰 수 있도록 표의 정보에 따라 명칭과 순서를 변경해 주어야 한다는 것이다. 이와 같이 PropertyConstants에 이어 PropertyList를 수정하면 원하는 데이터 구조를 입력할 수 있게 된다.

다. SAtomProperties

PropertyConstants와 PropertyList가 입력으로 받는 데이터 구조를 결정한다면 SAtomProperties에서는 실제 구조체로 값들을 입력받는 역할을 한다. 따라서 앞서 수정한 PropertyConstants와 PropertyList와 동일하게 SAtomProperties를 변경해 주어야 데이터의 값들이 FEATURE의 module 안으로 들어가게 된다.

III 결론

FEATURE를 사용하면 특정 기능을 하거나 구조를 지니는 site를 자동적으로 찾아낼 수 있기 때문에 단백질의 3차원 구조를 분석하는데 많은 도움을 받을 수 있다. 하지만 FEATURE에 입력으로 들어가는 데이터의 차원이 고정되어 있기 때문에 사용자의 신규 데이터를 가지고 이용하기에는 어려움이 따른다.

따라서 이를 개선하기 위한 노력을 하였고 기존의 FEATURE에서 사용하는 고정된 규격의 데이터 입력 구조를 원하는 데이터 구조로 변경 할 수 있는 방법을 찾아내었다. FEATURE 안의 PropertyConstants, PropertyList 그리고 SAtomProperties module을 수정함으로써 원하는 데이터 구조로 입력할 수 있게 되었다. 이 module들을 수정함으로써 FEATURE를 더욱 자유로이 이용할 수 있게 되었고 앞으로 남은 FEATURE의 문제점들을 개선해 나갈 수 있는 발판을 마련하였다. 차후에는 이를 토대로 44 x 6 차원의 데이터를 이용한 FEATURE의 정확도 향상을 연구할 계획이다.

참고문헌

[1] Sungroh Yoon, Jessica C. Ebert, Eui-Young Chung, Giovanni De Micheli and Russ B. Altman " Clustering protein environments for function prediction: finding PROSITE motifs in 3D" *BMC(BioMedCentral) Bioinformatics 8(Suppl 4):S10*, 2007.

[2] Liping Wei and Russ B. Altman, "Recognizing complex, asymmetric functional sites in protein structures using a bayesian scoring function," *Journal of Bioinformatics and Computational Biology* Vol. 1, pp. 119-138, 2003.

[3] Steven C. Bagley, Liping Wei, Carol Cheon, and Russ B. Altman "Characterizing oriented protein structural sites using biochemical properties" *Proc Int Conf Intell Syst Mol Biol. 3,* pp. 12-20, 1995.

[4] Wallace, A.C., N.Borkakoti, and J.M. Thornton, "TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites" *Protein Sci.* 6, 11(1997), pp. 2308-2323 , 1997.

[5] Wallace, A.C., R.A. Laskowski, and J.M. Thornton, "Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases" *Protein Sci.* 5, 6(1996), pp. 1001-1013, 1996.

[6] Fetrow, J.S. and J. Skolnick, "Method for prediction of protein function from sequence using the sequence-to structure-to function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases" J Mol Biol. 281, 5(1998) pp. 949-968, 1998.

[7] Fetrow, J.S., A. Godzik, and J. Skolnick, "Functional analysis of the Escherichia coli genome using the sequence- to structure-to-funtion paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity" J Mol Biol. 282, 4(1998), pp. 703-711, 1998.

[8] Inbal Haplperin, Dariya S Glazer, Shirely Wu and Russ B Altman "The FEATURE framework for protein function annotation: modelling new functions, improving performance, and extending to novel applications" *BMC Genomics 9(Suppl 2):52*, 2008.

[9] M.P. Liang, D.L. Brutlag, R.B. Altman, "Automated construction of structural motifs for predicting functional sites on protein structures," *The Pac Symp Biocomput.* pp. 204-215, 2003.

[10] Liping Wei, Russ B. Altman, Jeffrey T. Chang " Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences" *Pac Symp Biocomput.* pp. 465-76, 1997.

[11] Liping Wei and Russ B. Altman "Recognizing protein binding sites using statistical descriptions of their 3D environments" *Pac Symp Biocomput.* pp. 497-508, 1998.

저 자 소 개

유 승 학 (학생회원)

2008년 : 고려대학교 전기전자전파공학부 졸업 (공학사)

2008년 9월~현재: 고려대학교 대학원 전자전기공학과 (석사과정)

<주관심분야>

윤 성 로 (정회원)
1996년 : 서울대학교 전기공학부 졸업 (공학사)
2002년 : Stanford University, MS in Electrical Engineering (공학석사)
2006년 : Stanford University, PhD in Electrical Engineering (공학박사)
2007년 9월~현재: 고려대학교 전기전자전파공학부 교수
<주관심분야>