Function Prediction of Gene products by Term based Probabilistic Model

단어 기반의 확률 모델을 이용한 단백질 기능 예측

Daewon Park[1], Hyukchul Kwon[2]

[1] Department of Computer Science, Pusan National University, Pusan, Korea
[2] Division of Computer Science and Engineering, Pusan National University, Pusan, Korea

E-mail: bluepepe@pusan.ac.kr

Abstract

유전 연구를 통해 밝혀지고 있는 단백질은 각각의 기능적 특성을 가지고 서로 영향을 주고 받으며 상호 작용한다. 단백질의 기능적 특성은 생물체에서는 단백질이 나타내는 기능으로 단백질 이름은 이들 단백질의 기능을 정확히 나타낼 수 있도록 붙여진다. 기능적 특성에 의해 명명된 단백질은 단백질을 구성하는 단어도 단백질과 유사한 기능 특성을 가질 가능성이 높다. 이는 텍스트 기반의 연구에서 단어가 가지는 중요성에서 비롯된다. 본 논문에서는 단백질을 구성하는 단어들을 단백질의 기능적 특성으로 분류하고, 이 기능 분포에 의해서 단백질의 기능을 역으로 예측하고 판단하고자 하였다.

Introduction

인간 지놈 프로젝트를 비롯한 생물 정보학 분야의 활발한 연구활동은 많은 유전자 데이터 및 유전 정보의 축적을 가져왔으며 현재도 지속적으로 데이터의 축적이 이루어지고 있다. 많은 유전 정보의 축적은 생물 정보학 분야의 연구에 긍정적인 영향을 주고 있으나, 아울러 많은 데이터의 축적으로 필요한 정보 획득에 많은 시간과 노력을 필요로 하게 되었다. 이들 많은 데이터로부터 필요한 정보를 쉽고 빠르게 획득하기 위해 노력은 데이터베이스화되어 축적된 데이터로부터의 정보 검색, 추출에 관한 연구로 이어지고 있다.

유전자 분석 과정에서 새롭게 발견된 데이터는 기존 연구로 그 기능 및 특성이 이미 밝혀진 유전자 데이터와의 염기서열 분석 등과 같은 비교 분석을 통해, 유전자 기능을 밝혀내고, 기능에 적합한 이름으로 명명된다. 단백질 이름은 단백질의 기능적 특성 및 생물학적 특성을 나타내는 하나 이상의 단어로 이루어진다. 단백질 이름은 단백질의 기능 특성을 나타내므로 단백질을 구성하는 단어도 단백질의 기능을 나타내는 중요한 요소라 할 수 있다.

본 논문에서는 단백질 이름을 구성하는 단어를 이용하여 단백질의 기능을 기초로 하는 기능적인 확률 분포를 구성하고,

단어들의 기능 분포에 의해 단백질의 기능을 예측하고자 하였다. Molecular function으로 분류된 단백질 데이터로부터 단어를 추출하고, 추출한 단어들을 molecular function에 의해 재 분류하였다. 분류된 단어들은 molecular function class에서의 출현 빈도 정보를 가지게 된다. 각 단어들이 가지는 molecular function에 대한 분포 확률로 단백질의 기능을 예측한다.

Systems and Methods

본 논문에서 제시하는 단어의 기능적 특성 분포에 따른 단백질의 기능 예측 분류는 두 가지 단계로 구성된다. 단백질의 기능 분류에 기초하여 단백질을 구성하는 단어들의 기능적 분포를 구성하는 학습단계와 단어의 기능 분포로 단백질의 기능을 예측하여 분류하는 단계로 구성된다.

Molecular Function Classes

단백질의 기능을 분류하기 위한 분류 기준으로 GO 분류를 이용하였다. GO분류는 molecular function, biological process, cellular component로 구성되며, molecular function은 activities를 기술한다. GO molecular function 분류는 최상위 28개의 그룹으로 구성되며, 각각의 그룹은 하위 그룹을 포함하는 계층구조를 이루고 있다.

GO molecular function은 하위 계층구조에 같은 activities를 포함하기도 한다. 이는 하나의 activity가 여러 상위 기능과 연관되어 구성된 경우이다. 예를 들어, 'ATP-binding cassette (ABC) transporter activity'는 binding, catalytic activity, transporter activity 그룹에 모두 포함된다. 이와 같이, 여러 그룹에 중복되어 포함하는 activity는 새로운 기능 그룹으로 구분하였다.

본 논문에서 적용하는 molecular function classes는 GO molecular function 분류의 상위 28개 그룹과 상위 그룹에 중복되어 포함된 하위 기능을 새 그룹으로 구분하여 구성한다. 즉, 기능적으로 독립된 56개의 그룹으로 단백질을 분류한다.

Terms' Functional Distribution

유전자 염기서열 분석, 패턴 분석 과정을 통해 새롭게 발견된 단백질은 단백질이 가지는 생물학적 특성, 기능적 특성에 의해 이름이 명명된다. 단백질 이름은 단백질의 기능을 파악할 수 있는 요소가 될 수 있다. Information retrieval, information extraction과 같은 텍스트를 기반으로 하는 연구에서 단어는 중요한 분석 요소이므로, 단백질 이름을 구성하는 단어 또한 단백질 분석의 중요 요소로 인식될 수 있다. 단백질의 기능 예측과 분류는 단어의 기능적 분포에 의해 이루어진다.

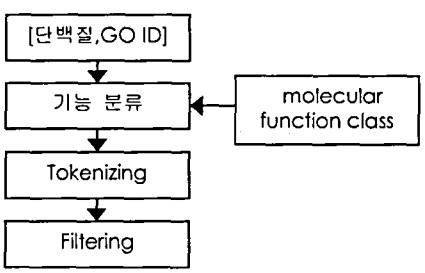

Figure 1. Process of term extraction

단어의 기능적 분포 특성은 단백질 이름에서의 중요 단어 요소 추출에 의해 파악될 수 있다. 단어의 기능적 특성은 단어를 추출한 단백질에 영향을 받으므로 molecular function classes로 분류된 단백질 이름 데이터로부터 단어를 추출하고, 이를

단백질의 기능 분류에 따라 분류한다. GO molecular function으로 분류된 단백질은 GO ID mapping table을 이용하여 56개 molecular function class로 다시 분류한다. Molecular function classes로 분류된 단백질 데이터에서 단백질의 기능적 특성을 나타낼 수 있는 단어를 추출한다.

단어의 추출은 단백질 이름으로부터의 단어의 분리와 분리된 단어의 filtering 과정을 통해 단백질의 기능 특성을 반영할 수 있는 단어를 구별하게 된다. 단백질 이름 표현 형태를 분석하여 단어의 구분자를 추출하고 이를 이용하여 단어를 분리하였다.

단백질 이름은 하나 이상의 단어들로 구성되며, 여러 형태의 단어들을 포함한다. 즉, 단백질 이름을 구성하는 단어 모두가 중요 분석 요소가 되지는 않는다. 단백질 이름은 다수의 명사, 동사와 함께 관사, 접속사와 같은 문법적 기능을 하는 단어를 포함하기도 한다. 문법적 기능을 하는 관사, 접속사 등의 단어는 단백질 기능 분류에 noise로 작용할 수 있으므로 단어의 기능 분포 구성에서 제외한다. 단백질 이름에 포함된 단어 중 의미적으로 중요하지 않은 단어는 filtering 규칙과 filtering 단어 list를 이용하여 제거한다.

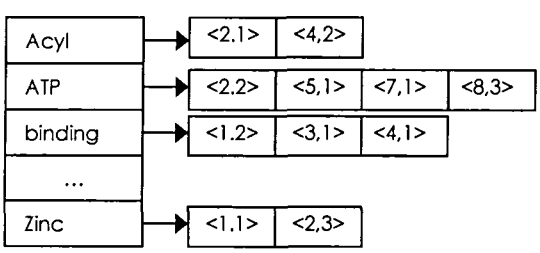

Figure 2. Structure of term distribution list

단백질 이름으로부터 추출한 단어는 단백질의 기능에 의해 molecular function classes로 분류된다. 단백질 이름은 하나 또는 그 이상의 단어로 구성되고 각 단어는 단백질의 기능적 특성을 나타낼 수 있는 요소이므로 단어의 기능적 특성은 단백질의 기능적 특성을 반영하게 된다.

단백질 이름에서 추출한 단어는 단백질의 기능과 같은 기능으로 분류되어 molecular function classes에 의한 단어의 분류결과는 molecular function classes에 대한 각 단어의 출현 빈도로 나타나게 된다. 즉, 단어는 단백질과 같은 기능 특성과 molecular function class 내에서의 빈도 정보를 가진다. Molecular function classes에 대한 단어의 출현 빈도는 각 단어의 기능적 특성에 대한 확률로 나타나게 된다.

Function Prediction

단백질의 기능 예측 및 분류는 단백질 이름을 구성하는 단어의 기능적 확률에 의해 이루어진다. 단백질 이름에서 추출한 단어가 가지는 molecular function class에서의 출현빈도와 단어를 포함하는 function class의 수를 이용하여 function class에서의 출현 확률을 계산한다. 각 단어의 class에 대한 확률은 단백질의 function class에 대한 확률로 귀착될 수 있다.

$$Pr(\text{Ci} \mid \text{protein}) = Pr(\text{Ci} \mid \langle \text{term1}, \text{term2}, \ldots, \text{term k} \rangle)$$

$$= \Sigma_{(i=1,k)} Pr(\text{Ci} \mid \text{term i})$$

$$Pr(\text{Ci} \mid \text{term k}) = tf_k / \Sigma_{(i=1,N)}(tf_{ki}) * idf_k$$

tf : term frequency in the molecular function class
idf : number of function class include term t
Ci : molecular function class i (i <= N)
N :: number of molecular function class

학습 데이터를 통해 구성된 단어들의

molecular function classes에 대한 분포 데이터로부터 각 단어가 가지는 molecular function과 molecular function class에서의 출현빈도 값을 얻는다. Function class에서의 출현빈도는 단어의 기능적 특성을 상대적인 확률로 계산되어진다.

특정 molecular function class에서 많이 나타난 단어는 다른 기능에 비해 현재의 function으로 이용될 확률이 높음을 나타낸다. 단백질은 하나 이상의 단어들로 구성되므로 단백질 이름을 구성하는 각 단어의 기능적 확률이 높게 나타나는 방향으로 단백질의 기능이 예측되어진다.

$$\text{argmax}_i P(C_i \mid \text{protein}) > \theta$$

θ : threshold

단백질의 기능은 특정 한 단어에 의해 결정되지 않고 단백질 이름에서 추출한 전체 단어들에 의해 결정된다. 각 단어들이 molecular function classes에 고르게 분포한다면 단백질의 기능을 예측하기가 어려워진다. 각 단어의 특징적인 분포와 그에 따라 확률로 단백질 기능이 결정된다.

Results

본 논문에서는 기능 특성에 의해 명명된 단백질의 이름에 나타나는 단어들이 단백질의 기능적 특성의 반영과 단어의 기능 분포 확률에 의해 단백질 기능 예측 정확도 산출을 중심으로 실험하였다.

Protein dataset

단백질 데이터는 GO(Gene Ontology) 사이트에서 제공하는 non-redundant human proteome set의 단백질에 대한 GO assignments 데이터를 이용하였다. 단백질 데이터에서 단백질 이름과 단백질의 기능 분류를 나타내는 GO ID를 추출하여 사용하였다. 단백질 이름은 full name을 대상으로 하였으며, symbol과 synonym은 제외하였다.

Experiments

단백질 데이터는 단어의 기능 분류 확률 구성을 위한 학습 데이터 set과 단백질 기능 예측 테스트를 위한 데이터 set으로 나누어 구성하였다.

Rate (%)	Data Number			Accuracy (%)		
	All	One word	Two word	All	One word	Two word
60	8261	657	1588	84.82	77.78	79.85
70	9685	733	1768	86.08	80.35	83.09
80	11114	751	1933	88.01	83.75	86.86
90	12551	841	2075	89.52	89.42	89.78
100	14040	905	2175	91.05	93.26	93.79

Table 1.Results of prediction

위 테이블은 단백질을 구성하는 단어의 기능 분포 확률에 의한 기능 예측 분류 결과를 나타낸다.

검색결과는 테스트 데이터로 나눈 단백질을 단어의 기능분포에 따라 분류 예측한 결과의 분류 정확도이다. 주어진 단백질의 기능과 단어의 기능 확률로 예측한 결과와 비교한 결과이다. Rate는 단어의 기능 분포를 구성하는 이용된 단백질 이름의 비율을 말한다. 60%의 단백질 데이터로 학습데이터를 사용했을 때는 84.82%의 정확도를 보였으며 90%에서는 89.52%의 결과를 보였다. 전체 데이터를 모두 학습 데이터로 사용하여 단어의 기능 분포를 구성하여 분류 실험한 결과 91.05%의 정확도를 보였다.

Discussion

실험을 통해, 단백질 이름을 구성하는 단어는 단백질의 특성을 나타내는 중요한 요소로 작용함을 알 수 있다. 특히, 텍스트에 나타나는 단백질 이름은 단백질의 기능 및 생물 특성을 표현할 수 있도록 그 이름이 명명됨으로, 단백질의 이름을 구성하는 단어는 단백질의 특성과 밀접한 관계가 있음을 알 수 있다.

단백질의 기능이 단백질 이름에 포함된 단어에 의해 예측이 가능한 수준이므로 텍스트에서의 단백질간의 상호 관련성 연구에 도움을 줄 수 있을 것으로 본다. 단백질간의 상호 관련성은 단백질이 가지는 기능에 많은 영향을 받을 것으로 기대되며, 단어에 의한 단백질 기능 예측으로 단백질간의 상호 관련성을 추출이 용이할 것으로 보인다.

Future Works

본 논문에서의 실험은 molecular function으로 분류된 단백질 데이터를 대상으로 단어의 기능 분포를 확률 모델로 구성하여 단백질의 기능을 예측하고자 시도하였다. Molecular function으로 분류된 단백질 데이터뿐만 아니라 텍스트에 나타내는 단백질 이름을 대상으로 단어에 의한 기능 예측 실험이 필요할 것으로 보인다. 그리고 단백질에 포함된 단어의 추출을 언어처리 방법을 적용하지 않고 간단한 분리, 제거 과정으로 추출하였으나 단백질 이름에 쓰이는 단어의 정확한 추출을 위해서는 언어처리 방법의 적용이 필요할 것으로 보인다.

References

[1] Frequent Term-Based Text Clustering, Florian Beil, Martin Ester, Xiaowei Xu, SIGKDD 2002

[2] Large-scale prediction of Saccharomyces cerevisiae gene funtion using overlapping transcriptional clusters., Lani F. Wu, Timothy R. Hughes, Armaity P. Davierwaa, Mark D. Robinson, Roland Stoughton, Steven J. Altschuler, Nature Genetics, volume 31, july 2002

[3] Prediction of human protein function according to Gene Ontology categories., Jensen LJ, Gupta R, Staerfeldt HH, Brunak S., Bioinormatics., March 2003.

[4] Predicting protein function from protein/protein interaction data : a probabilistic approach., Letovsky S, Kasif S., Bioinformatics. July 2003.

[5] The Lexical Properties of the Gene Ontology(GO)., Alexa T. McCray, Allen C. Browne, Oliver Bodenreider, Proceedings of the AMIA 2002 Annual Symposium, 2002

[6] Gene Ontology Consortium (http://www.geneontology.org)

[7] Annotating protein function through lexical analysis.,, Rajesh Nair, Burkhard Rost, AI Magazine, 2003.

[8] A Probabiistic Model for Identifying Protein Names and their Name Boundaries., Kazuhiro Seki, Javed Mostafa, Proceedings of the Computational Systems Bioinformatics, 2003.

[9] Selecting Text Features for Gene Name Classification : from Documents to Terms,, Goran Nenadic, Simon Rice, Irena Spasic, Sophia Ananiadou, Benjamin Stapley,

Proceeding of ACL, 2003

[10] Functional Discrimination of Gene Expression Patterns in Terms of the Gene Ontology, Liviu Badea, Pac Symp Biocomput, 2003

[11] Inferring sub-cellular localization through automated lexical analysis, Rajesh Nair, Burkhard Rost, Bioinformatics vol. 18, 2002

[12] Construction of Database for Protein Classifications, Jiro Araki, Genome Informatics 12, 2001

[13] Detecting Protein Function and Protein-Protein Interactions from Genome Sequences, Edward M. Macrotte, Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, David Eisenberg, SCIENCE, 1999