ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

단백질 의약품 특성정보필드 유용성 평가☆

Evaluation of Usefulness of the Protein Drug Feature Information Filed

변 재 희[1] 최 유 주[2] 이 주 환[2] 서 정 근[2*]
Jaehee Byeon Yoo-Joo Choi Ju-Hwan Lee Jung-Keun Suh

요 약

단백질 의약품 산업이 성장함에 따라 단백질 의약품 개발 시 단백질 정보는 필수적으로 인식 되고 있다. 단백질 정보 서비스를 제공하는 대표적인 바이오데이터 센터로는 미국의 NCBI, PDB, 유럽의 EMBL, 일본의 DDBJ 등이 있으며, 각 센터별로 특화된 단백질 정보들이 제공되고 있다. 사용자가 원하는 단백질 정보를 얻기 위해서는 독립된 단백질 정보를 검색하고 이를 통합하고 분석해야하며, 보다 편리하게 접근할 수 있도록 대표적인 데이터 센터 혹은 소규모 프로젝트 별로 바이오데이터에 대한 다양한 웹서비스가 연구 개발되고 있다. 단백질 의약품 정보 서비스에 대한 필요성이 높아지면서 캐나다의 DrugBank, 미국의 GDSC 등에서 의약품 정보와 단백질 데이터를 통합하여 서비스하고 있다. 하지만 사용자가 요구하는 다양한 단백질 정보를 반영하지 못하는 실정이다. 국내의 경우 바이오인포매틱스 인프라가 부족하고, 단백질 의약품 정보 서비스 또한 의약품의 기본 정보와 유통정보만을 제공하는 것에 한정되어 있다. 본 연구에서는 기존 서비스의 한계를 벗어난 한국형 단백질 의약품 전용 서비스 설계를 위한 사전 연구로 기존 대표적 데이터베이스에서는 적용하고 있지 않은 단백질 특성정보필드들을 제시하고 이에 대한 유용성 평가를 전문 종사자를 대상으로 진행하여 기존 바이오데이터베이스의 단백질 정보 필드와 비교하였다. 그 결과 본 연구에서 제시한 단백질 특성정보필드들이 단백질 의약품 정보 서비스에 요구되는 유용한 데이터 필드임을 검증하였다.

☞ 주제어 : 바이오인포매틱스, 단백질 의약품, 특성정보

ABSTRACT

As the protein drug industry is growing, protein informations are indispensable for the protein drug development. NCBI and PDB in the U.S., the EMBL in Europe and the DDBJ in Japan are the representative centers for bio information and each center provides specific data for protein information. To obtain specific protein information, users are to be collect them from the service sites of each center and then combine or analyze for their purpose. To facilitate the accessibility to bio data, various R&D activities are running for development of diverse web services relevant to bio data in major data centers or small-scale projects. With the recognition of protein information as pivotal for the protein drug development, DrugBank in Canada, GDSC in the U.S. start to provide integrated informations between drugs and proteins. However, those service does not meet users' demands due to lack of diversity. In Korea, infra structures for bioinformatics are limited and the current services for protein drug information are providing only basic information of the drug including distribution data. This is a pilot study to construct a specialized service for protein drug information in Korean style breaking through the limitations of current services. This study proposed new fields for protein characterization information which had not been provided by current services and evaluated their effectiveness and usability by comparing them to the existing fields with expert survey. As a result, the newly proposed fields for protein characterization have been proven to be useful data fields for the service of protein drug information.

☞ keyword : Bioinformatics, Protein Drug, Characterization

1. 서 론

1980년대 Human Genome 프로젝트가 시작되면서 생화학적 실험이 대규모로 이뤄졌고 수많은 데이터를 분석하여 처리하고 서비스하기 위한 바이오인포매틱스(Bioinformatics) 분야가 발전하였다[1-3].

바이오인포매틱스는 IT 기술을 기반으로 분산된 바이오 데이터 및 문헌 정보를 통합하여 관리, 분석, 처리하여

[1] Dept. of Stereoscopic Media, Korean German Institute of Technology, Seoul, 157-930, Korea.
[2] Dept. of Newmedia, Korean German Institute of Technology Seoul, 157-930, Korea.
* Corresponding author (jksuh@kgit.ac.kr)
[Received 07 February 2014, Reviewed 20 February 2014, Accepted 14 May 2014]
☆ 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2014년도 산업계 맞춤형 인력양성 지원사업의 연구결과로 수행되었음

서비스하는 융합기술이다. 바이오인포매틱스는 신약발굴, 시스템 생물학 등 여러 분야에 응용될 수 있어 다양한 분야에서 응용 지원 기술 등의 요구가 증가하여 서비스가 급증하고 있다[4]. 국외에서는 바이오인포매틱스에 대한 다양한 정보를 얻을 수 있도록 바이오데이터베이스 센터를 운영하고 있으며, 미국의 NCBI, PDB, 유럽의 EMBL, 일본의 DDBJ 등이 대표적이다. 각 센터 모두 웹 브라우저를 기반으로 유전자와 연관된 서열정보, 문헌정보 등을 포괄적으로 서비스하고 있으며, 특화된 서비스를 별도로 제공하고 있다[1].

하지만 바이오데이터베이스는 분산 운영되고 있어 원하는 정보를 검색하기 위해서는 여러 독립된 데이터베이스를 검색하고, 수작업으로 통합해야 하는 문제점이 있다. 특히 통합된 데이터에서 의미 있는 지식을 추출하기 까지 많은 노력이 들어 비효율적이고, 이용하는 서비스마다 정보 격차가 크다는 문제점이 있다[2,3,5]. 이러한 문제를 해결하기 위해 최근 바이오 데이터베이스에 대한 통합 검색 시스템을 개발하기 위한 다양한 연구가 진행 중이다[2,4,6].

특정 단백질의 작용기전을 활용하여 특정 질병에 대해 치료용으로 개발된 단백질 의약품의 경우 최근 글로벌 시장이 성장함에 따라 국내에서도 이 분야의 산업이 빠르게 확대되고 있다. 단백질 의약품의 개발에 있어서 단백질 정보는 필수적이다. 따라서 기존 약물의 효과를 극대화시키거나 약물의 위험 요소를 중점적으로 관리하기 위해 바이오데이터베이스를 이용한 단백질 및 의약품 통합 서비스 연구가 점차 증가하고 있다.

국외에서는 의약품 검색에 대한 요구가 증가하고 수많은 바이오 데이터가 공개되면서 각 의약품 별 단백질 데이터를 제공하는 연구와 표준화가 진행되고 있다. 특히 in silico 신약개발 기술을 중심으로 생화학, IT 등의 다양한 학문적 이론을 융합하여 발전하고 있다[3,7,8].

하지만 국내에서는 바이오인포매틱스 인프라가 약하고 이 또한 여러 기관에 독립적으로 분산 운용되고 있어 단백질 의약품에 대한 바이오인포매틱스 활용은 매우 미미한 실정이다. 약물 표적 단백질을 예측하기 위한 연구가 진행되고 있지만 사용자가 원하는 정보를 얻기 쉽지 않고 최근 정보를 반영한 단백질 의약품 전용 정보 서비스는 부재한 상태이다[9,10].

따라서 국내 바이오산업의 확대에 발맞추어 단백질의 약품에 대한 정보화 및 서비스 시스템 구축이 필요하며, 이를 위해 국외의 대표적 바이오인포매틱스 서비스를 비교 분석하여 국내 실정에 맞는 서비스를 제공하는 방안이 요구된다.

본 논문에서는 효율적인 단백질 의약품 전용 서비스를 위한 사전 연구로 기존 대표적 데이터베이스에서는 적용하고 있지 않은 단백질 의약품 특성 정보 필드들을 제시하였다. 제시한 정보의 유용성을 검증하기 위하여 기존 데이터베이스 서비스에서 제공하는 필드의 중요도를 평가하였으며, 제시한 단백질 의약품 특성 정보 필드의 중요도와 비교 분석하여 그 유용성을 검증하였다. 검증된 결과를 바탕으로 본 논문에서 제시한 단백질 의약품 특성 정보 필드가 사용자 입장에서 유용한 데이터 필드임을 입증하였다.

2. 관련연구

2.1 단백질 의약품 정보 서비스

단백질 의약품의 가장 대표적 서비스로는 DrugBank가 있으며, 암 세포에 대한 항암제 정보를 제공하는 서비스로는 GDSC가 대표적이다. 국내에서는 의약품관리종합정보센터에서 국내에 유통되는 의약품 정보를 서비스하고 있다.

DrugBank와 GDSC는 의약품의 적용기전을 찾을 수 있도록 약에 포함된 단백질 데이터베이스 서비스를 제공하고 있으나, 다양한 바이오인포매틱스의 최근 정보를 통합하여 반영하기 어렵다는 한계점이 있다. 의약품관리종합정보센터는 약물의 기본적인 정보 및 유통과 관련된 정보만 제공하고 있어 약물의 단백질, 화학 정보는 확인할 수 없다는 한계점이 있다.

2.1.1 DrugBank

DrugBank는 포괄적인 약물 표적 정보인 단백질과 추가 약물 정보인 화학 물질의 데이터를 결합한 바이오인포매틱스/화학 정보에 대한 데이터베이스 서비스다. DrugBank의 데이터베이스에는 3,200개 이상의 실험 약물과 FDA에서 승인한 저분자, 바이오 의약품 800개 이상, 4,100개 이상의 약물 항목이 있다. 또한 14,000개 이상의 단백질 또는 약물 표적 정보는 각 의약품 항목에 맵핑된다. 각 의약품 항목은 80개 이상의 데이터 필드로 구성되어 있으며, 단백질과 관련된 약물 표적 정보와 추가 약물 정보인 화학 물질 데이터로 이루어져 있다.

DrugBank 데이터베이스의 특징은 대표적 단백질 데이터베이스인 NCBI, EMBL, PDB의 데이터를 하이퍼링크와 다양한 구조의 애플릿을 통해 보여준다는 것이다. 더

불어 단백질 의약품과 관련된 관계형 쿼리 검색을 지원하여 단백질 의약품에 대한 광범위한 데이터서비스가 가능하다.

DrugBank는 in silico 환경에서 약물 표적을 발견하고, 약물을 설계하는 것, 마약 심사, 약물 대사 예측 등 단백질 의약품 정보와 관련된 다양한 잠재적 서비스 제공을 목표로 하고 있다[3].

2.1.2 GDSC(Genomics of drug sensitivity in cancer)

GDSC는 암 세포에 대한 항암제의 약물 반응 감도 및 게놈 데이터베이스를 제공하는 바이오인포매틱스 서비스로 누구든 상관없이 자유롭게 데이터 이용이 가능하다. GDSC는 현재 약 700개의 암 세포주와 138개의 항암제의 반응을 기술하고 75,000건의 항암제의 약물 반응 감도에 대한 데이터를 포함하고 있다. GDSC에서 제공하는 데이터를 통해 암 유전자의 증폭, 결실 조직 유형 및 체세포 돌연변이의 정보를 분석할 수 있으며, 이는 게놈 데이터 세트에 통합되어 제공된다. GDSC의 데이터 분석은 특정 항암제 및 암 유전자의 쿼리를 기반으로 웹 포털을 통해 제공되며, 원하는 데이터는 완전히 다운로드하여 사용할 수있다.

GDSC는 암 치료의 효과를 증대 시키기 위해 서비스되고 있으며, 대규모의 암과 관련된 전반적인 약물 정보를 서비스 한다[8].

2.1.3 의약품관리종합정보센터

의약품관리종합정보센터는 의약품의 생산, 사용에 관한 현황정보를 관리하고 물류 흐름을 파악하기 위해 2007년 정부에서 설립한 건강보험심사평가원 산하 기관이다. 의약품관리종합정보센터에서는 의약품의 정보 수집, 조사, 분석 서비스를 제공하고 의약품유통정보 데이터베이스를 구축, 운영하고 있다. 구축된 정보를 기반으로 의약품의 유통실태에 대한 현황을 파악하고 유통 선진화와 관련된 연구를 수행한다. [11].

2.2 바이오데이터베이스 서비스

대표적인 바이오데이터베이스 센터로는 미국의 NCBI, PDB, 유럽의 EBI, 일본의 DDBJ가 있다. 각 센터들은 방대한 양의 정보시스템을 구축하여, 바이오데이터를 활용한 응용 서비스 부문에 큰 영향을 미치고 있다. 또한 다른 센터의 데이터들을 링크를 통해 유기적으로 공유하기도 한다. 하지만 상호간의 데이터베이스 갱신 등 상호작용은 하지 않는다. 따라서 원하는 정보를 얻기 위해서는 각 데이터베이스 특징에 맞게 정보를 검색하고 사용자는 이를 통합하여 분석해야하는 단점이 있다. 특히 세 데이터베이스 모두 단백질의 특성 분석을 확인하기 위해서 관련 문헌 정보를 수집한 후 분석해야하는 단점이 있다.

2.2.1 NCBI

NCBI(The National Center for Biotechnology Information)는 1988년 분자 생물학을 위한 정보 시스템 개발을 위해 설립된 센터이다. NCBI는 웹 사이트를 통해 GenBank, 생물학적 데이터 분석 및 검색 리소스, 단백질 등의 데이터를 제공하고 있다. 더불어 일본의 DDBJ, 유럽의 EMBL 등 바이오와 관련된 데이터를 유기적으로 수집, 분석한다.

NCBI는 바이오인포매틱스와 관련된 다양한 분석을 위해 Protein, BLAST, PubMed 등 데이터를 체계적으로 카테고리화 하여 서비스하고 있으며, NCBI에서 제공하는 모든 정보는 FTP나 Entrez를 통해 열람할 수 있다. 최근에는 BLAST, Entrez 등을 통해 방대한 양의 바이오데이터에 대해 전문화된 데이터 검색이 가능하도록 연구하고 있다[12].

2.2.2 EMBL

EMBL(European Molecular Biology Laboratory)은 분자생물학 연구를 위해 1974년 유럽의 국가들 중심으로 공공의 목적을 위해 설립한 연구 기관으로, 총 5개의 분과로 운영된다. 그 중에서도 EBI(European Bioinformatics Institute)는 바이오인포매틱스 연구 및 교육을 위한 데이터베이스 서비스의 허브 역할을 하며, 공공 데이터베이스 서비스로서 모든 프로젝트 데이터를 공개한다. EBI는 웹을 기반으로 최신 분자 데이터베이스를 주제별로 카테고리화 하여 서비스하고 있으며, 데이터베이스는 회원국으로 등록된 각 국가들의 프로젝트와 NCBI 등 국외의 공공 데이터베이스의 정보를 수집하고 분석하여 서비스한다[13].

2.2.3 PDB

PDB(Protein Data Bank)는 RCSB(the Research Collaboratory for Structural Bioinformatics)의 산하 기관으로 생물학과 관련된 연구 및 교육을 위해 단백질 구조를 수집하고, PDB 아카이브의 3D 고분자 데이터를 사용하여 단백질 시

각화 도구와 자원을 개발하여 웹 서비스한다. PDB는 미국의 PDB, 유럽의 PDBe, 일본의 PDBj와 BioMagResBank 등 전세계 회원 단체의 데이터를 수집하여 가공하고 유통한다. 최근에는 화학 성분에 특화된 도메인 기반의 구조적인 정렬이 가능한 간단한 검색 시스템에 대해 연구하고 있으며, iPad 등 모바일 서비스를 통해 데이터베이스에 대한 사용자 접근성을 높였다[14].

2.3 단백질 특성정보

단백질 의약품의 경우 대단히 복잡한 구조를 가지고 있다. 따라서 구조를 규명하기 위해서는 단백질 특성정보의 확보가 필수적이다. 단백질 의약품 개발 과정에서는 필수적으로 확인(Identity), 순도(Purity), 함량(Quantity), 물리화학적 특성(Physicochemical properties) 및 안정성(Stability)에 대한 단백질 특성정보를 제공해야 한다[15-17]. 이와 관련된 연구로써 확인, 순도, 함량, 물리화학적 특성 및 안정성에 대한 분석 항목으로 Identity, HPLC, Spectroscopy, Glycan의 분류방법을 제시한다.

2.3.1 Identity

Identity는 단백질 의약품에 대한 단백질 구조 및 조성 정보를 제공하며, 상세 필드는 다음과 같다[18].

(표 1)단백질 의약품에 대한 Identity 필드
(Table 1) Identity field for Protein Drug

필 드	정 의
Total Mass	단백질 의약품의 분자량 정보
N-terminal Sequence	단백질 의약품의 N-terminal 아미노산 서열 정보
C-terminal Sequence	단백질 의약품의 C-terminal 서열 정보
Amino Acid Composition	단밸질 의약품의 아미노산 함량 정보
Extinction Coefficient	단백질 의약품의 흡광 계수 정보
Peptide Mapping	단백질 의약품의 펩타이드 맵핑 정보
Disulfide Bond	단백질 의약품의 아미노산 서열 정보

2.3.2 HPLC

HPLC는 단백질 의약품에 대한 단백질 함량 및 변형체 정보를 제공하며, 상세 필드는 다음과 같다[18].

(표 2)단백질 의약품에 대한 HPLC 필드
(Table 2) HPLC field for Protein Drug

필 드	정 의
RP-HPLC	RP-HPLC 분석 정보
SEC-HPLC	SEC-HPLC 분석 정보
IEX-HPLC	IEX-HPLC 분석 정보
Hydrophobic HPLC	Hydrophobic HPLC 분석 정보
HILIC-HPLC	Hydrophobic interaction HPLC분석 정보

2.3.3 Spectroscopy

Spectroscopy는 단백질 의약품에 대한 단백질의 2차, 3차 구조 정보를 제공하며, 상세 필드는 다음과 같다[18].

(표 3)단백질 의약품에 대한 Spectroscopy 필드
(Table 3) Spectroscopy field for Protein Drug

필 드	정 의
UV	UV spectroscopy 분석 정보
Circular Dichroism(CD)	CD spectroscopy 분석 정보
Fluorescence	Fluorescence spectroscopy 분석 정보
FT-IR	FT-IR 분석 정보
DSC	differential scanning calorimetry 분석 정보

2.3.4 Glycan

Glycan은 단백질 의약품에서 단백질의 당화에 대한 구조 및 조성 정보를 제공하며, 상세 필드는 다음과 같다[18].

(표 4)단백질 의약품에 대한 Glycan 필드
(Table 4) Glycan field for Protein Drug

필 드	정 의
Monosacchaid	단백질 성분당 함량 분석 정보
Sialic Acid	시알란 함량 분석 정보
Charged N-glycan Profiling	Charged N-glycan 함량 및 구조 분석 정보
Antennary N-glycan Profiling	Antennary N-glycan 함량 및 구조 분석 정보
O-glycan Profiling	O-glycan 함량 및 구조 분석 정보
N-glycosylation Site	N-glycosylation 위치 및 구조 분석 정보
O-glycosylation Site	O-glycosylation 위치 및 구조 분석 정보
Glycan Site Profiling	glycosylation 위치에서의 N-glycan 함량 및 구조 분석 정보

3. 단백질 특성정보필드의 유용성 평가 실험 설계

3.1 연구 목표와 가설 설정

본 논문에서는 단백질 의약품 전용 서비스를 위한 사전 연구로 바이오 전문가를 대상으로 단백질 기본정보, 서열정보, 특성정보필드의 중요도를 조사하였다. 세부 연구 가설은 다음과 같다.

1. 단백질 기본정보필드 별 중요도에는 차이가 있을 것이다.
2. 단백질 서열정보필드 별 중요도에는 차이가 있을 것이다.
3. 단백질 특성정보필드 별 중요도에는 차이가 있을 것이다.

3.2 연구대상 및 자료 수집

본 논문에서는 단백질 의약품에 대한 단백질 기본정보, 서열정보, 특성정보필드 별 중요도 조사를 위해 해당 분야에 종사하고 있는 전문가 95명을 대상으로 설문을 진행하였다. 해당 분야 종사자 모두 단백질 정보 제공 사이트와 의약품 검색 사이트를 이용하는 이용자들이며, 수집된 설문지 중 불성실한 응답 10부를 제외한 85부를 설문 분석에 사용하였다.

3.3 조사방법 및 체계

설문 항목 내용은 크게 4가지로 개인정보, 단백질 기본정보필드, 단백질 서열정보필드, 단백질 특성정보필드이다. 개인정보는 명목척도로 하여 피험자의 인구통계학적 정보를 수집하였다. 단백질 기본, 서열, 특성정보필드의 항목은 리커트 5점 척도로 하여 필드 별 중요도를 수집하였다.

단백질 의약품 정보 서비스를 위한 사전연구에서 대표적 단백질 정보 검색 서비스의 이용현황을 분석 한 결과 설문 응답자 85명 중 72명인 84.7%가 NCBI를 이용하였다. 이에 단백질 기본정보, 서열정보필드에 대한 설문항목은 NCBI의 단백질 데이터 필드를 기준으로 구성하였다[19].

단백질 특성정보필드의 설문항목은 단백질 의약품 개발 시 반드시 필요한 정보 중 허가기관에서 제시하고 있는 가이드라인에 기반하여 [15-17] 단백질 특성정보를 Identity, HPLC, Spectroscopy, Glycan의 필드로 구성하였다.

설문 항목의 내용은 다음과 같다.

(표 5) 설문 내용
(Table 5) Contents of Survey

설문영역		설문항목
개인정보		학력, 업계 종사기간, 자주 이용하는 단백질 검색 사이트, 단백질 검색 사이트 방문 횟수, 자주 이용하는 의약품 검색 사이트, 의약품 검색 사이트 이용 목적
NCBI의 단백질 기본정보		LOCUS, DEFINITION, ACCESSION, VERSION, DBSOURCE, KEYWORD, SOURCE, REFERENCE, COMMENT
NCBI의 단백질 서열정보		FEATURES, FEATURE/source, FEATURE/Region, FEATURE/Site, FEATURE/Protein, FEATURE/CDS, ORIGIN
단백질 특성 정보	Identity	Total Mass, N-terminal Sequence, C-terminal Sequence, Amino Acid Composition, Extinction Coefficient, Peptide Mapping, Disulfide Bond
	HPLC	RP-HPLC, SEC-HPLC, IEX-HPLC, Hydrophobic HPLC, HILIC-HPLC
	Spectro-scopy	UV, Circular Dichroism(CD), Fluorescence, FT-IR, DSC
	Glycan	Monosacchaid, Sialic Acid, Charged N-glycan Profiling, Antennary N-glycan Profiling, O-glycan Profiling, N-glycosylation Site, O-glycosylation Site, Glycan Site Profiling

단백질에 대한 기본정보, 서열정보, 특성정보필드 별 중요도의 차이가 있는지 검증하기 위하여 각 단백질 정보 필드를 독립변수로 하여 One-Way Repeated Measures ANOVA 분석을 하였다. Mauchly의 구형성 검정을 만족할 경우 Within-Subjects Effect의 Sphericity Assumed을, 만족하지 못할 경우 Greenhouse-Geisser 값을 참조하였다.

4. 단백질 의약품 정보 서비스를 위한 설문 결과 분석

4.1 피험자 분석

설문에 응답한 피험자의 개인정보의 분포를 보면 85명 중 48명이 석사졸업으로 56%를 차지하였고, 업계 종사기간은 2년 이상인 피험자가 55명으로 64.7%를 차지하였다.

(표 6) 피험자 분석
(Table 6) Subject of Survey

구분		빈도	백분율(%)
학력	대학재학	3	4
	대학교 졸	14	16
	대학원 졸(석사)	48	56
	대학원 졸(박사)	20	24
업계 종사 기간	2년 미만	30	35
	2년 이상~5년 미만	21	25
	5년 이상~10년 미만	13	15
	10년 이상	21	25

단백질 의약품과 관련하여 자주 이용하는 단백질 검색 사이트로는 72명이 NCBI를 이용하는 것으로 84.7%를 차지하였다. 단백질 의약품과 관련하여 자주 이용하는 의약품 검색 사이트는 41명이 DrugBank를 이용하는 것으로 48.2%를 차지하였다.

이에 본 연구에서는 대다수의 응답자가 사용하는 NCBI의 단백질 필드를 바탕으로 단백질 의약품 검색 시 단백질 기본정보와 단백질 서열정보필드에 대한 중요도를 분석하였다. 더불어 단백질 의약품 개발 과정에서 반드시 필요하기에 본 연구에서 제시한 단백질 특성정보인 Identity, HPLC, Spectroscopy, Glycan에 대한 필드 별 중요도를 분석하였다.

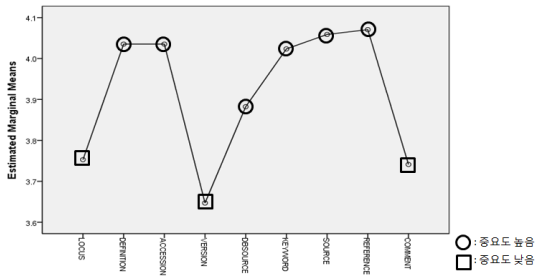
4.2 단백질 기본정보필드 중요도 분석

단백질 기본정보에 대한 필드 별 중요도를 분석하기 위해 Mauchly 구형성 검증을 실시한 결과 (Mauchly's W=.331, df=35, p<.05) 구형성 가정을 만족하지 못하여, Within-subjects effect의 Greenhouse-Geisser 보정을 사용하여 분석하였다. 그 결과 단백질 기본정보필드 별 중요도에 유의한 차이가 있었음으로(F(8,672)=6.566, p<.05), 가설 1이 채택되었다.

(표 7) 단백질 기본정보필드 중요도 분석 결과
(Table 7) Analysis Result of Field Importance for General Information of Protein

Source	Mauchly's test of sphericiy (p-value)	Within subjects effect (df)	Within subject effects (mean square)	Within subjects effect (F-value)	Within subjects effect (p-value)
단백질 기본정보 필드	.000	8	2.284	6.566	.000
Error (단백질 기본정보 필드)		672	.348		

단백질 기본정보필드에 대해 Pairwise Comparisons을 통해 필드 간 중요도의 우선순위를 분석하였다. 그 결과 DEFINOTION, VERSION, DBSOURCE, KEYWORD, SOURCE, REFERENCE의 중요도가 높음을 확인하였다.



(그림 1)단백질 기본정보의 추정된 주변평균
(Figure 1) Estimated Marginal Means of Protein Basic Information

4.3 단백질 서열정보필드 중요도 분석

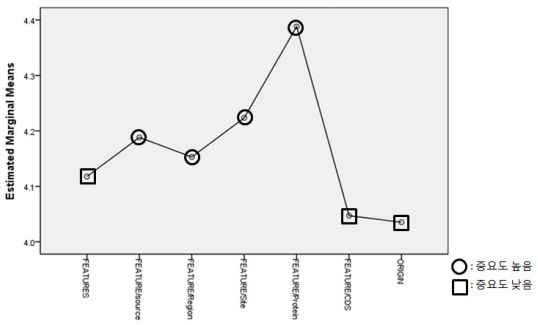
단백질 서열정보에 대한 필드 별 중요도를 분석하기 위해 Mauchly 구형성 검증을 실시한 결과 (Mauchly's W=.553, df=20, p<.05) 구형성 가정을 만족하지 못하여, Within-subjects effect의 Greenhouse-Geisser 보정을 사용하여 분석하였다. 그 결과 단백질 서열정보필드 별 유의한 차이가 있었음으로(F(6,504)=5.539, p<.05), 가설 2가 채택되었다.

(표 8) 단백질 서열정보필드 중요도 분석 결과
(Table 8) Analysis Result of Field Importance for Sequence Information of Protein

Source	Mauchly's test of sphericiy (p-value)	Within subjects effect (df)	Within subject effects (mean square)	Within subjects effect (F-value)	Within subjects effect (p-value)
단백질 서열정보 필드	.000	6	1.231	5.539	.000
Error (단백질 서열정보 필드)		504	.222		

단백질 서열정보필드에 대해 Pairwise Comparisons을 통해 필드 간 중요도의 우선순위를 분석하였다. 그 결과

FEATURE/Source, FEATURE/Region, FEATURE/Site, FEATURE/Protein의 중요도가 높음을 확인하였다.



(그림 2) 단백질 서열정보의 추정된 주변평균
(Figure 2) Estimated Marginal Means of Protein Sequence Information

4.4 단백질 특성정보필드 중요도 분석

단백질 특성정보에 대한 필드 별 중요도를 분석하기 위해 단백질 의약품 개발 시 반드시 필요한 단백질의 Identity, HPLC, Spectroscopy, Glycan에 대한 필드 별 중요도를 분석하였다.

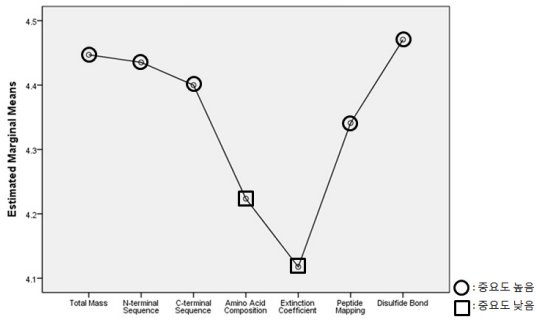
4.4.1 Identity 필드 중요도 분석 결과

단백질 특성정보 중 Identity에 대한 필드 별 중요도를 분석하기 위해 Mauchly 구형성 검증을 실시한 결과 (Mauchly's W=.280, df=20, p<.05) 구형성 가정을 만족하지 못하여, Within-subjects effect의 Greenhouse-Geisser 보정을 사용하여 분석하였다. 그 결과 Identity 필드 별 유의한 차이가 있었다(F(6,504)=7.646, p<.05).

(표 9) Identity 필드 중요도 분석 결과
(Table 9) Analysis Result of Field Importance for Identity

Source	Mauchly's test of sphericiy (p-value)	Within subjects effect (df)	Within subject effects (mean square)	Within subjects effect (F-value)	Within subjects effect (p-value)
Identity	.000	6	1.470	7.646	.000
Error (Identity)		504	.192		

Identity 필드에 대해 Pairwise Comparisons을 통해 필드 간 중요도의 우선순위를 분석하였다. 그 결과 Total Mass, N-terminal Sequence, C-terminal Sequence, Amino Acid Composition, Extinction Coefficient, Peptide Mapping, Disulfide Bond의 중요도가 높음을 확인하였다.



(그림 3) Identity의 추정된 주변평균
(Figure 3) Estimated Marginal Means of Identity

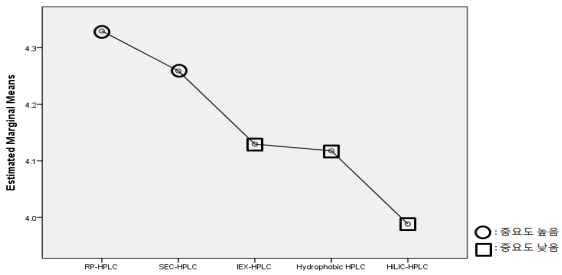
4.4.2 HPLC 필드 중요도 분석 결과

단백질 의약품 특성 정보 중 HPLC에 대한 필드 별 중요도를 분석하기 위해 Mauchly 구형성 검증을 실시한 결과 (Mauchly's W=.604, df=9, p<.05) 구형성 가정을 만족하지 못하여, Within-subjects effect의 Greenhouse-Geisser 보정을 사용하여 분석하였다. 그 결과 HPLC 필드 별 유의한 차이가 있었다(F(4,336)=6.429, p<.05).

(표 10) HPLC 필드 중요도 분석 결과
(Table 10) Analysis Result of Field Importance for HPLC

Source	Mauchly's test of sphericiy (p-value)	Within subjects effect (df)	Within subject effects (mean square)	Within subjects effect (F-value)	Within subjects effect (p-value)
HPLC	.000	4	1.500	6.429	.000
Error (HPLC)		336	.233		

HPLC 필드에 대해 Pairwise Comparisons을 통해 필드 간 중요도의 우선순위를 분석하였다. 그 결과 RP-HPLC, SEC-HPLC의 중요도가 높음을 확인하였다.



(그림 4) HPLC의 추정된 주변평균
(Figure 4) Estimated Marginal Means of HPLC

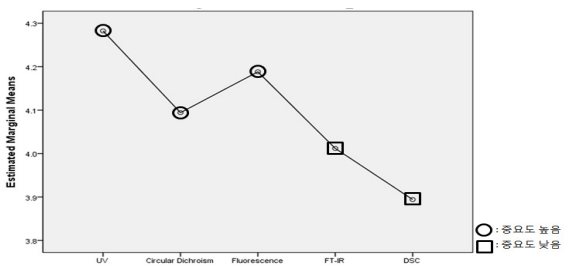
4.4.3 Spectroscopy 필드 중요도 분석 결과

단백질 의약품 특성 정보 중 Spectroscopy에 대한 필드 별 중요도를 분석하기 위해 Mauchly 구형성 검증을 실시한 결과 (Mauchly's W=.741, df=9, p<.05) 구형성 가정을 만족하지 못하여, Within-subjects effect의 Greenhouse-Geisser 보정을 사용하여 분석하였다. 그 결과 Spectroscopy 필드 별 유의한 차이가 있었다(F(4,336)=8.900, p<.05).

(표 11) Spectroscopy 필드 중요도 분석 결과
(Table 11) Analysis Result of Field Importance for Spectroscopy

Source	Mauchly's test of sphericiy (p-value)	Within subjects effect (df)	Within subject effects (mean square)	Within subjects effect (F-value)	Within subjects effect (p-value)
Spectro-scopy	.000	4	1.935	8.900	.000
Error (Spectro-scopy)		336	.217		

Spectroscopy 필드에 대해 Pairwise Comparisons을 통해 필드 간 중요도의 우선순위를 분석하였다. 그 결과 UV, Circular Dichroism, Fluorescence의 중요도가 높음을 확인하였다.



(그림 5) Spectroscopy의 추정된 주변평균
(Figure 5) Estimated Marginal Means of Spectroscopy

4.4.4 Glycan 필드 중요도 분석 결과

단백질 의약품 특성 정보 중 Glycan에 대한 필드 별 중요도를 분석하기 위해 Mauchly 구형성 검증을 실시한 결과 (Mauchly's W=.122, df=27, p<.05) 구형성 가정을 만족하지 못하여, Within-subjects effect의 Greenhouse-Geisser 보정을 사용하여 분석하였다. 그 결과 Glycan의 필드 별 유의한 차이가 없었다(F(7,588)=.783, p>.05).

(표 12) Glycan 필드 중요도 분석 결과
(Table 12) Analysis Result of Field Importance for Glycan

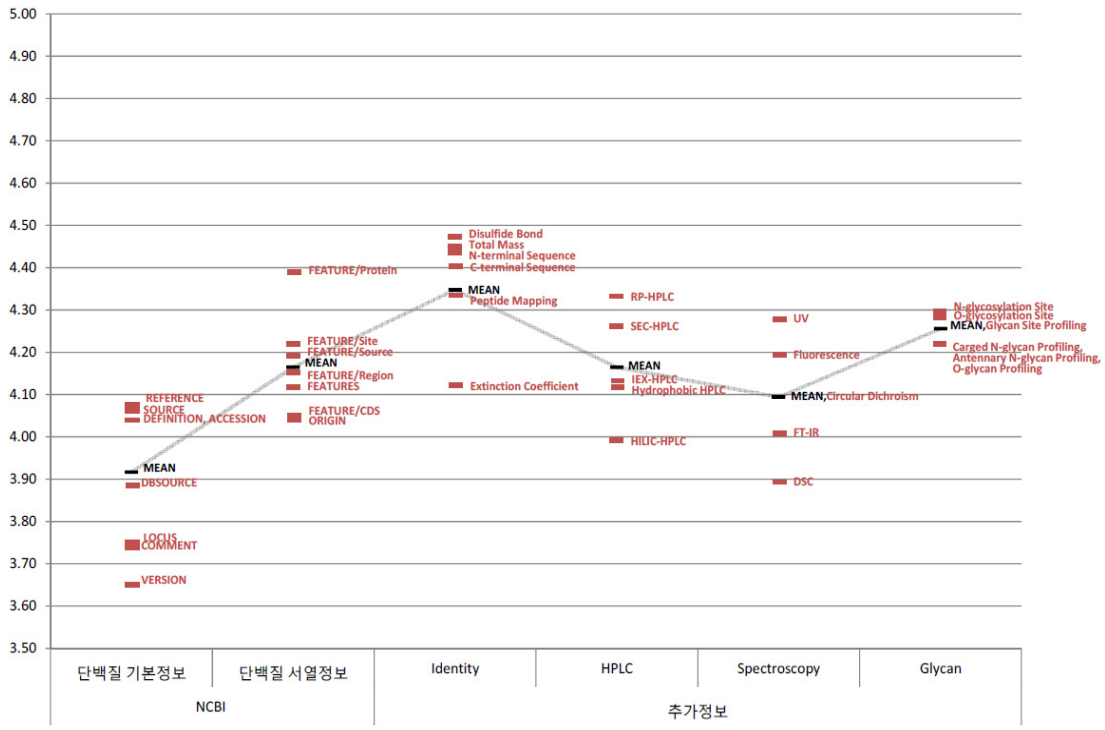
Source	Mauchly's test of sphericiy (p-value)	Within subjects effect (df)	Within subject effects (mean square)	Within subjects effect (F-value)	Within subjects effect (p-value)
Spectro-scopy	.122	7	.100	.783	.601
Error (Spectro-scopy)		588	.128		

4.4 결 과

단백질 의약품 정보 서비스를 위한 단백질 기본정보, 서열정보, 특성정보필드에 대한 설문 분석 결과 NCBI에서 제공하는 단백질 기본정보, 서열정보필드와 본 연구에서 제시한 단백질 특성정보필드 사이에서 Glycan을 제외한 모든 필드의 중요도에는 유의미한 차이가 있음을 확인하였다. Glycan은 필드 별 중요도에는 차이가 없었지만 필드의 중요도가 모두 4.2를 넘어 다른 필드들과 비교했을 때 중요도가 높아 Glycan에 대한 모든 필드는 유용한 정보라 판단 할 수 있음으로 단백질 특성정보필드 별 중요도에는 유의한 차이가 있어 가설 3이 채택되었다.

단백질 특성정보의 중요도 평균치는 Identity는 4.34, HPLC는 4.16, Spectroscopy는 4.09, Glycan 4.26으로 단백질 기본정보필드의 중요도 평균치인 3.91 보다 높았다. 단백질 서열정보필드의 중요도 평균치는 4.16으로 Identity와 Glycan, HPLC의 중요도가 서열정보필드의 중요도 평균치보다 높았으며, Spectroscopy 또한 중요도 평균치 차이가 0.07로 큰 차이가 나지 않았다.

따라서 기존에 제공되던 단백질 기본정보, 서열정보필드와 단백질 특성정보인 Identity, HPLC, Spectroscopy, Glycan의 필드를 비교했을 때 단백질 특성정보에 대한 요구도가 높다는 것을 확인할 수 있었으며, 단백질 의약품 서비스를 위한 단백질 특성정보필드들이 유용한 데이터 필드임을 입증할 수 있었다.



(그림 6) 단백질 의약품 정보 필드의 추정된 주변 평균

(Figure 6) Estimated Marginal Means of Field for Drug Information of Protein

5. 결 론

바이오인포매틱스의 응용분야가 증가함에 따라 방대한 양의 데이터베이스에서 원하는 정보를 쉽게 통합하고 분석하는 서비스에 대한 요구가 증가하고 있다. 기존의 바이오데이터베이스 서비스는 표준이 서로 달라 사용자가 의미있는 정보를 도출하기 위해서는 정보 검색 결과를 통합하고 분석해야 한다는 한계점이 있다.

특히 바이오데이터의 단백질 정보는 신약개발 또는 기존 의약품의 효과를 극대화하는데 많은 영향을 끼쳐 중요한 정보로 인식되고 있다. 국외에서는 이미 의약품과 단백질 정보를 통합한 DrugBank 등의 서비스가 제공되고 있다. 하지만 국내에서는 바이오데이터베이스 연구 및 교육, 인프라가 취약하다는 단점이 있다. 더불어 의약품에 대한 정보 또한 기본정보, 유통정보에만 국한되어 있다.

단백질 의약품에 대한 바이오데이터베이스 분야의 경쟁력 확보를 위해서는 기존의 바이오데이터를 통합하고 이를 국내 실정에 맞게 서비스 할 수 있는 독창적 데이터베이스에 대한 연구가 필요하다.

본 연구에서는 단백질 의약품 개발 과정 및 가이드라인을 평가하여 단백질 의약품에 필수적으로 제공되는 단백질 특성정보를 Identity, HPLC, Spectroscopy, Glycan의 분류방법에 따라 제시하였으며 이를 해당 분야 종사자를 대상으로 유용성을 평가하여 단백질 의약품에 대한 단백질 의약품 정보 필드를 도출하였다. 그 결과 해당 분야 종사자가 주로 이용하는 NCBI에서 제공하는 단백질 기본정보, 서열정보에 대한 필드와 본 연구에서 제시한 단백질 특성정보의 필드 사이에서 유의미한 차이가 있음을 검증하였다. 이를 통해 해당 분야 종사자들은 Identity, HPLC, Spectroscopy, Glycan의 분류방법에 따라 제시한 필드에 대한 요구도가 높다는 것을 확인하였다.

본 논문의 결과는 단백질 의약품 전용 서비스를 위한 사전 연구로써 사용자들이 본 연구에서 제시한 단백질 특성정보에 대한 서비스의 요구 사항이 높다는 것을 확인한 바 이에 대한 서비스를 제공할 필요가 있음을 확인하였다. 이러한 서비스의 제공을 통해 관련 산업의 경쟁력을

확보하고, 신약 개발 및 기존 약물 연구 및 개발에 소요되는 시간과 비용을 축소할 수 있을 것으로 기대된다. 더불어 신규 서비스 모델 발굴 및 관련 분야 기반으로 확대되어 활용될 수 있을 것이다.

참 고 문 헌(Reference)

[1] H. Y. Jung, S. J. Park, and S. H. Park, "Bioinformatics Technology," Electronics and Telecommunications Trends, Vol. 20, No. 5, 2005, pp. 93-104.

[2] S. J. Lee, H. S. Yong, "Development of Integrated Retrieval System of the Biology Sequence Database Using 11 Service," The KIPS Transactions:PartD, Vol. 11D, No. 4, 2004, pp. 755-764.

[3] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," Nucleic Acids Research, Vol. 34, 2006, pp, D668-D672.

[4] B. Y. Ahn, J. M. Han, and S. C. Hong, "Design of Integrated Retrieval System for Bioinformatics," 2007 KoCon Autumn Comprehensive Academy Conference, Vol. 5, No. 1, 2007, pp. 11-14.

[5] D. H. Im, "Trends of Bioinformatics for Establishing a Science-Technology-Oriented Society in the 21st Century," Graduate School of Sejong University, Master's thesis, 2004.

[6] Y. H. Choi, "Implementation of WSBAT: 11 Services based Biodata Analysis Tool," Graduate School of Sejong University, Master's thesis, 2004.

[7] B. A. Aksoy, J. Gao, G. Dresdner, W. Wang, A. Root, X. Jing, E. Cerami, and C. Sander, "PiHelper: An Open Source Framework for Drug-Target and Antibody-Target Data," Bioinformatics, 2013.

[8] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, "Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," Nucleic Acids Research, Vol. 41, 2013, pp. D955 – D961.

[9] J. S. Sohn, D. H. Kim, and I. J. Chung, "Representation of drug informations and their relations using ontology," Proceeding of Korea Computer Congress , Vol. 37, No. 1(C), 2010, pp. 317-322.

[10] H. S. Jung, B. R. Hyun, S. H. Jung, W. H. Jang, and D. S. Han, "Drug Target Protein Prediction using SVM," KISS Korea Computer Congress, Vol. 34, No. 2(B), 2007, pp. 17-21.

[11] Korea Pharmaceutical Information Service, http://www.kpis.or.kr/

[12] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," Nucleic Acids Research, Vol. 41, 2013, pp. D8 – D20.

[13] EMBL, http://www.embl.org/

[14] P. W. Rose, C. Bi, W. F, Bluhm, C. H. Christie, D. Dimitropoulos, S. Dutta, R. K, Green, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, A. G. Ramos, J. D. Westbrook, J. Young, C. Zardecki, H. M. Berman, and P. E. Bourne, "The RCSB Protein Data Bank: new resources for research and education," Nucleic Acids Research, Vol. 41, 2013, pp. D475 – D482.

[15] Guideline on Characterisation and Specifications for Therapeutic Monoclonal Antibody Products. 2011. Korea Food and Drug Administration.

[16] ICH Q6B "Test Procedures and Acceptance Criteria for Biotechnological/Biological Products" (CPMP/ICH/365/96). 1999. European Medicines Agency.

[17] Guideline on Development, Production, Characterization and Specifications for Monoclonal Antibodies and Related Products. 2009. European Medicines Agency.

[18] D. K. Han, "A study on the media of bio-information based on 3D modeloing for protein tertiary structure," Korean German Institute of Technology, Master's thesis, 2012.

[19] J. H. Byeon, Y. J. Choi, J. H. Lee, and J. K. Suh, "The Present Situation Analysis on Services of Protein Drug Information," Proceedings of the Korea Information Processing Society Conference, 2014.

◐ 저 자 소 개 ◑

변 재 희 (Jaehee Byeon)
2010년 덕성여자대학교 컴퓨터시스템학과(공학사)
2012년 호서대학교 벤처전문대학원 IT응용기술학과 졸업(공학석사)
2013~현재 한독미디어대학원대학교 입체영상미디어학과 석사과정중
관심분야 : 바이오인포매틱스, 컴퓨터그래픽스, etc.
E-mail : bjaeh9188@gmail.com

최 유 주 (Yoo-Joo Choi)
1989년 이화여자대학교 전자계산학과(이학사)
1991년 이화여자대학교 일반대학원 전자계산학과(이학석사)
2005년 이화여자대학교 과학기술대학원 컴퓨터공학학과(공학박사)
1991년~1993년 한국컴퓨터주식회사 기술연구소 주임연구원
1994년~1999년 포스데이타주식회사 기술연구소 주임연구원
2005년~2010년 서울벤처대학원대학교 컴퓨터응용기술학과 조교수
2010년~현재 한독미디어대학원대학교 뉴미디어학부 부교수
관심분야 : 컴퓨터그래픽스, 가상현실, HCI, 컴퓨터비전, etc.
E-mail : yjchoi@kgit.ac.kr

이 주 환 (Ju-Hwan Lee)
1997년 경상대학교 심리학과(심리학학사)
2003년 연세대학교 심리학과(심리학석사)
2007년 연세대학교 심리학과(심리학박사)
2000년~2007년 연세대학교 인지과학연구소 연구원/전문연구원
2007년~2009년 영국 옥스퍼드대학교 실험심리학과 박사후연구원
2009년~2010년 성균관대학교 인터랙션사이언스학과 연구교수
2010년~현재 한독미디어대학원대학교 뉴미디어학부 조교수
관심분야 : HCI, Multisensory, Interaction, UX, etc.
E-mail : jhlee@kgit.ac.kr

서 정 근 (Jung-Keun Suh)
1987년 서울대학교 식물학과(이학사)
1989년 서울대학교 대학원 식물학과(이학석사)
1996년 University of Texas at Austin Department of Biochemistry(이학박사)
2000년~2007년 LG생명과학 기술연구원 부장
2007년~2009년 한독산학협동단지 BT연구센터 센터장
2009년~현재 한독미디어대학원대학교 뉴미디어학부 교수
관심분야 : 데이터베이스, 바이오인포매틱스, etc.
E-mail : jksuh@kgit.ac.kr