빅데이터 분석을 위한 임상 및 바이오 정보 통합 데이터베이스의 설계

Design of an Integrated Database of Clinical and Bio Information for Big Data Analysis

임종태[1], 류은경[1], 김기연[1], 김천중[1], 윤수용[1], 박선용[1], 노연우[1], 육미선[1], 정지원[1], 최기태[1], 유석종[2], 유재수[1]*
[1]충북대학교 정보통신공학과
[2]한국과학기술정보연구원

Jongtae Lim[1], Eunkyung Ryu[1], Kiyeon Kim[1], Cheonjung Kim[1], Sooyong Yoon[1], Sunyong Park[1], Yeonwoo Noh[1], Miseon Yuk[1], Jiwon Jeong[1], Kitae Choi[1], Seokjong Yu[2], Jaesoo Yoo[1]*
[1]Chungbuk National University,
[2]Korea Institute of Science and Technology Information

요약

생명과학분야에서는 생명현상을 이해하기 위해 신호 전달 네트워크에 대한 연구가 진행되고 있다. 하지만 신호전달 네트워크와 임상 정보를 결합하여 질병관점에서 신호 전달 네트워크를 통합하고 결합하는 관점의 연구가 부족하다. 따라서 본 논문에서는 빅데이터 기술을 활용하여 임상 및 신호전달 정보를 연계 분석할 수 있는 시스템을 구축하고자 빅데이터 분석을 위한 임상 및 바이오 정보 통합 데이터베이스를 설계한다. 설계한 임상 및 바이오 정보 통합 데이터베이스는 빅데이터 분석 기술을 적용한 확장 분석 기법 및 통합 분석 시스템 개발에 활용할 수 있다.

Ⅰ. 서론

최근 미국, 일본과 같은 선진국의 주도하에 생물학적 반응 및 질병 분석 연구가 활발하게 진행되고 있다. 연구자들은 생물학적 반응 및 질병 분석을 통해 세포의 동력학적 행동을 시뮬레이션하고 질병의 발병기작과 같은 생명현상을 이해하여 이를 신약 개발, 맞춤의료 구현, IT 융합 의학 산업 창출 등에 활용하고 있다. 바이오 정보 수집 기술이 발전하고, 바이오 정보를 제공하는 많은 데이터베이스 시스템이 출현하면서 연구자들은 보다 높은 수준의 연구를 수행할 수 있게 되었다. 하지만 한편으로는 바이오 정보들이 너무 빠르고 방대하게 생산되어 이를 수용하여 연구를 하는 것이 힘들어졌다. 따라서 이렇게 빠르고 방대하게 생산되는 바이오 정보를 빅데이터 처리 기법을 이용하여 분석하는 방법이 중요하게 연구되기 시작했다[1].

기존 생명과학분야에서는 생명현상을 이해하기 위해 신호 전달 네트워크에 대한 연구가 진행되고 있다. 하지만 빅데이터 분석 기술을 통해 보다 고도화된 연구를 진행할 수 있음에도 불구하고 이를 잘 활용하지 못하는 것이 현실이다. 특히 신호전달 네트워크와 임상 정보를 결합하여 질병관점에서 신호 전달 네트워크를 통합하고 결합하는 관점의 연구가 부족하다. 질병의 발병기작과 같이 생명현상을 이해하기 위해서는 질병관점에서 생명현상을 분석하는 연구가 필요하며, 이를 지원하기 위한 질병과 관련된 임상 및 바이오 정보의 통합된 데이터베이스가 필요하다.

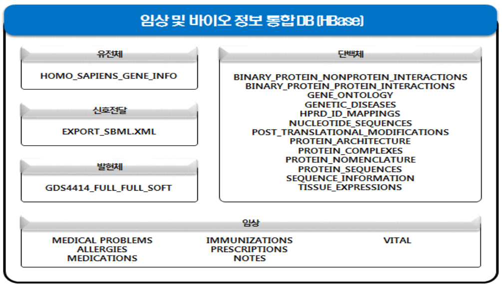
본 논문에서는 빅데이터 분석을 위한 임상 및 바이오 정보 통합 데이터베이스를 설계한다. 설계한 데이터베이스를 이용하면 기존 신호전달 네트워크와 질병과 관련된 임상정보까지 폭 넓은 정보를 통합하여 보다 질병관점에서 생명현상을 분석 할 수 있다. 또한 설계한 임상 및 바이오 정보 통합 데이터베이스는 빅데이터 분석 기술 중 하둡(Hadoop) 기반의 HBase를 기반으로 구축한다.

Ⅱ. 임상 및 바이오 정보 통합 DB 설계

1. 전체 시스템 구조

그림 1은 임상 및 바이오 정보 통합 데이터베이스를 나타낸 것이다. 임상 및 바이오 정보 통합 데이터베이스는 유전체, 단백체, 신호전달, 발현체, 임상 정보로 구성된다. 각 정보는 대표적인 바이오 데이터베이스인 HPRD[2], NCBI[3], KEGG[4], OpenEMR[5] 등으로부터 수집한 데이터의 분석을 통해 데이터베이스로 설계된다. 이렇게 수집한 데이터는 하둡 및 하둡 에코시스템과 연동하여 돌아갈 수 있도록 HBase에 설계된 데이터베이스에 저장된다. 데이터는 바이오, 임상, 질병, 임상 및 바이오 연계 정보로 분류되어 저장된다.

* 교신저자 : yjs@chungbuk.ac.kr
본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학 IT연구센터육성 지원사업(NIPA-2014-H0301-14-1022), 그리고 한국과학기술정보연구원의「고성능 컴퓨팅 기반 빅데이터 기술 개발 (K-14-L06_C04-S01)」사업으로부터 지원받아 수행된 연구임.



▶▶ 그림 1. 임상 및 바이오 정보 통합 데이터베이스

2. 세부 스키마 설계

그림 2는 유전체 데이터베이스 스키마를 나타낸다. 유전체 정보는 총 15개의 컬럼으로 구성되어 있다. 본 논문에서는 유전체 컬럼을 4개의 컬럼 패밀리로 구분했다.

Homo_sapiens.gene_info.gz

Gene_info						Location		Desc						Version
tax_id	GeneID	Symbol	LocusTag	Synonyms	dbXrefs	Chromosome	map_location	Description	type_of_gene	Symbol_from_nomenclature_authority	Full_name_from_nomenclature_authority	Nomenclature_status	Other_designations	Modification_date
9606	1	A1BG	-	A1B|ABG|GAB|HYST2477	HGNC:5|MIM:138670|Ensembl:ENSG00000121410|HPRD:00726|	Vega:OTTHUMG00000183507	19	19q13.4 alpha-1-B glycoprotein	protein-coding	A1BG	alpha-1-B glycoprotein	O	HEL-S-163pA|alpha-1B-glycoprotein|epididymis secretory sperm binding protein Li 163pA	20140706

▶▶ 그림 2. 유전체 데이터베이스 스키마

그림 3은 단백체 데이터베이스 스키마를 나타낸다. 단백체 정보는 총 13개의 항목이 존재한다. 그 중에 대표적인 항목은 Binary Protein Protein Interaction과 Genetic Diseases이다. 단백체 정보는 총 85개의 컬럼으로 구성되어 있다. 본 논문에서는 단백체 컬럼을 총 26개의 컬럼 패밀리로 구분했다.

9. BINARY_PROTEIN_PROTEIN_INTERACTIONS.txt

Hprd_info		Interactor					
interactor_1_geneSymbol	interactor_1_hprd_id	interactor_1_refseq_id	interactor_2_geneSymbol	interactor_2_hprd_id	interactor_2_refseq_id	experiment_type	reference_id
ITGA7	2761	NP_001138468.1	CHRNA1	7	NP_001034612.1	in vivo	10910772

2. GENETIC_DISEASES.txt

Hprd_info			Disease	
hprd_id	geneSymbol	refseq_id	disease_name	reference_id
9	CHRNG	NP_005190.4	Escobar syndrome	1.682.652.016.826.530

▶▶ 그림 3. 단백체 데이터베이스 스키마

그림 4는 발현체 데이터베이스 스키마를 나타낸다. 발현체 데이터베이스는 S. Wang이 제안한 스키마[6]을 사용했다. 발현체 정보는 총 8개의 컬럼으로 구성되어 있다. 본 논문에서는 발현체 컬럼을 4개의 컬럼 패밀리로 구분했다.

GDS4414_full_full_soft

info	Raw		Log		Zscore	
Assay ID	Gene1+Probeset1	Gene2+Probeset2	Gene1+Probeset1	Gene2+Probeset2	Gene1+Probeset1	Gene2+Probeset2

▶▶ 그림 4. 발현체 데이터베이스 스키마

그림 5는 신호전달 데이터베이스 스키마를 나타낸다. 신호전달 정보는 총 8개의 컬럼으로 구성되어 있다. 본 논문에서는 신호전달 컬럼을 4개의 컬럼 패밀리로 구분했다.

export_sbml.xml

info	compartment	species			reactions		
name	compartment_name	name	compartment	initialAmount	name	reactants	products
	X0:X1:X3					X0:X1:X3	X0:X1:X3

▶▶ 그림 5. 신호전달 데이터베이스 스키마

그림 6은 임상 데이터베이스 스키마를 나타낸다. 임상 정보는 총 64개의 컬럼으로 구성되어 있다. 본 논문에서는 임상 컬럼을 7개의 컬럼 패밀리로 구분했다.

Who							
Name	DOB	S.S.	Marital Status	User Defined	External ID	Sex	License/ID
Phil Belford	1972-02-09	333222333	Single		1	Male	

Contact												
Address	City	State	Postal Code	Country	Mother's Name	Guardial's Name	Emergency Contact	Emergency Phone	Home Phone	Work Phone	Mobile Phone	Contact Email
6666 String Street	Longview	Florida	44433	USA	Gardner		Wilma	222-333-4444	333-444-2222	555-444-3333	222-444-2222	heya@invalid.email.com

Choice												
Provider	Referring Provider	Pharmacy	HIPAA Notice Received	Allow Voice Message	Leave Message With	Allow Mail Message	Allow SMS	Allow Email	Allow Immunization Registry	Allow Immunization Info Sharing	Allow Health Information	Allow Patient Portal

▶▶ 그림 6. 임상 데이터베이스 스키마

Ⅲ. 결론

본 논문에서는 빅데이터 분석을 위한 임상 및 바이오 정보 통합 데이터베이스를 제안했다. 제안하는 데이터베이스는 하둡 및 하둡 에코시스템과 연동하여 돌아갈 수 있도록 HBase에 저장된다. 제안하는 데이터베이스는 빅데이터 분석 기술을 적용한 확장 분석 기법 및 통합 분석 시스템 개발에 활용할 수 있다. 향후 연구로는 제안하는 데이터베이스를 Mahout이나 R과 같은 빅데이터 분석 도구와 연동하여 질병 중심의 임상 및 바이오 정보 연관 분석을 수행하는 빅데이터 분석 시스템을 구현할 것이다.

▮참 고 문 헌▮

[1] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, S.. Y. Rhee, "Big data: The Future of Biocuration", Nature, Vol.455, pp.47-50, 2008.

[2] http://www.hprd.org/

[3] http://www.ncbi.nlm.nih.gov/

[4] http://www.genome.jp/kegg/

[5] http://www.open-emr.org/

[6] S. Wang, I. Pandis, I. Emam, F. Guitton and Y. Guo, "High Dimensional Biological Data Retrieval Optimisation with NoSQL Technology", Proc. of the International Symposium on Bioinformatics Research and Applications, 2013.