Korean J. Malacol. 30(4): 429-431 2014
http://dx.doi.org/10.9710/kjm.2014.30.4.429

연체동물 전용 BLAST 서버 업데이트 (Version II)

강세원[1], 황희주[1], 박소영[1], 왕태훈[1], 박은비[1], 이태희[2], 황의욱[3], 이준상[4], 박홍석[5], 한연수[6], 임채은[7], 김순옥[7], 이용석[1]

[1]순천향대학교 자연과학대학 생명시스템학과, [2]순천향대학교 공과대학 건축학과
[3]경북대학교 사범대학 생물교육학과, [4]강원대학교 환경연구소, [5](주)지앤시바이오
[6]전남대학교 농업생명과학대학 식물생명공학부, [7]국립생물자원관 유용자원분석과

Mollusks Sequence Database: Version II

Se Won kang[1], Hee Ju Hwang[1], So Young Park[1], Tae Hun Wang[1], Eun Bi Park[1], Tae Hee Lee[2], Ui Wook Hwang[3], Jun-Sang Lee[4], Hong Seog Park[5], Yeon Soo Han[6], Chae Eun Lim[7], Soonok Kim[7] and Yong Seok Lee[1]

[1]Department of Life Science, Soonchunhyang University, Asan, Chungnam, 336-745 Korea.
[2]Department of Architecture, Soonchunhyang University, Asan, Chungnam, 336-745 Korea.
[3]Department of Biology Education, Teacher's College, Kyoungpook National University, Daegu 702-701, Korea.
[4]Institute of Environmental Research, Kangwon National University, Chuncheon, Gangwon 220-701, Korea.
[5]Research Institute of GnC BIO Co., LTD., Daejeon 305-150, Korea.
[6]Division of Plant Biotechnology, College of Agriculture and Life Sciences, Chonnam National University, Gwangju 500-757, Korea.
[7]Biological and Genetic Resources Assessment Division, National Institute of Biological Resources, Inchon 404-708, Korea.

ABSTRACT

Since we reported a BLAST server for the mollusk in 2004, no work has reported the usability or modification of the server. To improve its usability, the BLAST server for the mollusk has been updated as version II (http://www.malacol.or.kr/blast) in the present study. The database was constructed by using the Intel server Platform ZSS130 dual Xeon 3.20 GHz CPU and Linux CentOS system and with NCBI WebBLAST package. We downloaded the mollusk nucleotide, amino acid, EST, GSS and mitochondrial genome sequences which can be opened through NCBI web BLAST and used them to build up the database. The updated database consists of 520,977 nucleotide sequences, 229,857 amino acid sequences, 586,498 EST sequences, 23,112 GSS and 565 mitochondrial genome sequences. Total database size is 1.2 GB. Furthermore, we have added repeat sequences, *Escherichia coli* sequences and vector sequences to facilitate data validation. The newly updated BLAST server for the mollusk will be useful for many malacological researchers as it will save time to identify and study various molluscan genes.

Key words: Mollusks, Sequence, BLAST

Received: December 20, 2014; Revised: December 24, 2014; Accepted: December 27, 2014
Co-corresponding author : Soonok Kim
Tel: +82 (32) 590-7422 e-mail: sokim90@me.go.kr
Co-corresponding author : Yong Seok Lee
Tel: +82 (41) 530-3040 e-mail: yslee@sch.ac.kr
1225-3480/24556

서 론

거듭되어진 생물학의 발달로 인하여 생물학 데이터는 계속해서 대량화되고 있다. 대량화 되어지는 정보들은 종의 정보는 물론 유전체 정보, 유전자 정보, 아미노산 서열 정보 등을 포함하고 있다. 이러한 정보들 중에서 우리가 원하는 정보를 찾을 수 있게 도와주는 방법 중 가장 보편적으로 사용되어 지고 있는 것은 상동성 검사이다. 이를 수행하는 프로그램으로서는 NCBI (National Center for Biotechnology information)

Table 1. Ongoing genome projects related with mollusks (https://gold.jgi-psf.org/)

Class	Sequencing Centers	Nation
Gastropoda (16)	Biodiversity and Climate Research Centre (1)	Germany (1)
	Broad Institute (1)	USA (9)
	DOE Joint Genome Institute (6)	
	Rutgers University (1)	
	Washington University in St. Louis (1)	
	University of Tartu, Institute of Molecular and Cell Biology (1)	Estonia (1)
	University of Nottingham (1)	UK (1)
	Beijing Genomics Institute (1)	China (2)
	Hong Kong Baptist University (1)	
	Queensland University of Technology (1)	Australia (1)
Bivalvia (16)	Institute of Clinical Molecular Biology, Kiel (1)	Germany (1)
	DOE Joint Genome Institute (3)	USA (7)
	Marine Biological Laboratory (2)	
	Ocean Genome Legacy (1)	
	University of New Hampshire (1)	
	Marine Genomics Unit, Okinawa Institute of Science and Technology (2)	Japan (2)
	Beijing Genomics Institute (1)	China (3)
	Ocean University of China (2)	
	Genoscope (1)	France (1)
	Deakin University (1)	Australia (1)
Cephalopoda (3)	Baylor College of Medicine (1)	USA (2)
	Rice University (1)	
	Yellow Sea Fisheries Research Institute (1)	China (1)

에서 제공하는 BLAST (Basic Local Alignment Search Tool) 가 주로 사용되어 지고 있다 (Altschul *et al*., 1990; McGinnis and Madden, 2004). 하지만 NCBI 에서 제공하는 데이터베이스는 몇몇 모델생물에만 국한되어 있기 때문에 연체동물 연구에는 적합하지 않은 실정이다. 이에 본 연구팀은 연체동물을 연구하기에 적합한 연체동물 전용 블라스트 서버를 2004년도에 구축한 바 있다 (Lee *et al*., 2004). 하지만 NGS (Next Generation Sequencer) 등 자동염기서열분석기의 거듭된 진보를 통하여 약 10년간 데이터의 수치는 약 1900% 증가하였다. 이에 발맞추어 기 구축된 연체동물 전용 블라스트 서버를 업데이트하였다. 또한 이전에 연체동물 전용 블라스트 서버는 본 연구팀에서 보유중인 서버에 구축되어 있어서 연구자들이 쉽게 접근이 힘든 단점이 있었다. 이러한 단점을 극복하기 위하여 한국패류학회 홈페이지와 연동이 되게 하여 더 많은 연체동물 연구자들이 이용할 수 있게 하였다 (http://www.malacol.or.kr/blast).

재료 및 방법

1. 서버 구축

사용된 서버는 Intel Server Platform ZSS130에 dual Xeon 3.20 GHz cpu 시스템이며, 운영체제 (operation system) 는 Linux CentOS release 3.9 를 사용하였다. 운영체제 설치 후 Apache 웹서버의 설정에서 일반 사용자가 cgi (common gate interface) 를 사용할 수 있도록 환경설정을 한 후 WebBLAST 패키지를 설치하였다.

2. 데이터베이스 구축

NCBI에 등록되어 있는 연체동물과 관련된 유전자서열 정보, 아미노산서열 정보, 미토콘드리아서열 정보, EST 서열 정보, GSS (genome survey sequecne) 서열 정보를 taxonomy browser와 연계하여 모두 다운 받은 후 multifasta 형태의 정보를 만든 후 BLAST용 데이터베이스를 구축하였다. 또한 부가적으로 반복서열, *Escherichia coli* 서

Table 2. Status of Mollusks Sequence Database: Version II

Database	2004 (Ver. I)		2014 (Ver. II)		Rate of increase	
	Sequences	Total Letters	Sequences	Total Letters	Sequences	Total Letters
Nucleotide	64,851	38,014,072	520,977	739,594,084	803%	1,946%
Amino Acid	14,923	2,999,586	229,857	69,685,049	1,540%	2,323%
Mitochondrial Genome	17	272,450	565	9,690,106	3,324%	3,557%
Express Sequence Tags	-	-	586,498	380,518,782	-	-
Genome Survey Sequence	-	-	23,112	11,486,149	-	-

열, 벡터서열을 데이터베이스화하여 데이터 검증에 용이하도록 하였다.

3. 웹 인터페이스 구축

모든 데이터베이스를 독립적으로 검색이 가능하도록 구성하였으며, query 및 데이터베이스가 허용하는 한 5가지 BLAST 프로그램이 모두 수행 가능하도록 하였다. 또한 Multi DB 메뉴를 만들어 라이브러리 검증 등을 할 때 용이하도록 하였다. 또한 국내외의 모든 연구자를 고려하여 영문으로 인터페이스를 구성하였으며 photoshop 7.0 및 Illustrator CS6를 활용하여 직관성과 단순성을 위주로 디자인 하였다.

결과 및 고찰

전 세계적으로 약 56,020 개의 genome project가 진행 중이거나 완성되었다 (2014년 8월 18일 기준). 이는 2004년의 1,238 개의 project에 비해 약 45배 증가한 수치이다. 연체동물로 국한하여 살펴보면 2004년 4개에서 8배 이상 증가된 35개의 genome project가 현재 연구진행 중이다 (Tabls 1) . 진행 중인 35개의 genome project 는 복족류와 이매패류가 각각 16개, 두족류가 3개를 차지하고 있다. 미국에서 가장 많은 18개의 project가 진행 중이며 중국이 5개의 project가 진행 중이다.

NCBI에 등록된 유전자 서열들을 확인 한 결과 nucleotide 서열의 수는 약 800%, amino acid 서열은 약 1500% 증가하였다. Mitochondrial genome 서열의 경우 2004년에 17개에서 2014년 565개로 약 3300% 증가함을 확인할 수 있었는데 이는 앞서 언급한 바와 같이 자동염기서열분석기의 발전으로 인한 것임을 알 수 있다. NCBI에 등록된 565개의 mitochondrial genome 서열은 이매패류에서 257개, 복족류에서 193개, 두족류에서 101 개 등으로 이루어져 있다. 이러한 서열들을 활용하여 베타분류 연구자들이 매우 유용하게 사용되어 질 수 있을 것으로 생각된다. 추후 업데이트는 Version1 에 있었던 프라이머 제작기능 및 기타 사용자들의 요구를 반영할 예정이다.

요 약

본 연구를 통하여 연체동물 전용 BLAST 서버 (Version II) 가 웹주소 http://www.malacol.or.kr/blast 에 구축되었다. 연체동물을 대상으로 하는 연구에 있어 필요한 정보를 매우 빠르게 얻을 수 있었다. 본 시스템을 사용하여 앞으로도 많은 연구가 진행되어지길 바라며, 아울러 많은 연체동물 연구자들에게 많은 도움이 되리라고 사료된다.

사 사

이 연구는 국립생물자원관에서 추진하는 "자생 생물자원의 유전자 다양성 연구" 사업에서 지원하는 "주요 동물자원의 유전자(체) 다양성 연구(NIBR201403102)" 과제로 수행되었습니다.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**: 403-410.

McGinnis, S. and Madden, T. L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, **32**: W20-W25

Lee, Y. S., Jo, Y. H., Kim, D. S., Kim, D. W., Kim, M. Y., Choi, S. H., Yon, J. O., Byun, I. S., Kang, B. R., Jeong, K. H. and Park, H. S. (2004) Construction of BLAST Server for Mollusks. *Korean journal of malacology*, **20**:165-169.