Korean J. Malacol. 35(1): 73-75 2019
https://doi.org/10.9710/kjm.2019.35.1.73

연체동물 대량염기서열 분석을 위한 PANM 데이터베이스 버전 3 업데이트

강세원[1], 박소영[2], 황희주[3], 정종민[3], 상민규[3], 민혜린[3], 박지은[3], 조항철[3], Bharat Bhusan Patnaik[4,5], 이용석[3]

[1]한국생명공학연구원 생물자원센터, [2]국립낙동강생물자원관 다양성연구팀,
[3]순천향대학교 자연과학대학 생명시스템학과, [4]전남대학교 농업생명과학대학 식물생명공학부,
[5]Trident School of Biotech Sciences, Trident Academy of Creative Technology

PANM DB ver 3.0 : An update of the bioinformatics database for annotation of large datasets from sequencing of species under Protostomia clade

Se Won Kang[1], So Young Park[2], Hee Ju Hwang[3], Jong Min Chung[3], Min Kyu Sang[3], Hye Rin Min[3], Jie Eun Park[3], Hang Chul Cho[3], Bharat Bhusan Patnaik[4,5] and Yong Seok Lee[3]

[1]Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Jeongeup, Jeonbuk 56212, Korea
[2]Biodiversity Research Team, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongbuk 37242, Korea
[3]Department of Life Science and Biotechnology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam 31538, Korea
[4]Division of Plant Biotechnology, Institute of Environmentally-Friendly Agriculture (IEFA), College of Agriculture and Life Sciences, Chonnam National University, Gwangju-61186, Korea
[5]School of Biotech Sciences, Trident Academy of Creative Technology (TACT), Chandrasekharpur, Bhubaneswar, Odisha-751024, India

ABSTRACT

The PANM Database (Protostome DB) was constructed for the first in the year 2015. The aim of the construction was to provide a public platform for researchers to efficiently annotate the next-generation sequencing (NGS) data arising out of genome sequencing projects of species belonging to Arthropoda, Mollusca, and Nematoda (Protostome group). The web-based interface has paramount importance in improving the process of sequencing data with a greater number of noteworthy hits. In the subsequent year, we released an update to the database (PANM-DB v1.0) that contained twice the number of protein sequences (7,571,246) as compared to the original database and the inclusion of Cephalopod sequences. We now release the third database update (PANM-DB v3) with the incorporation of 11,615,243 protein sequences. The size of PANM-DB v3 is smaller, only 6% of NCBI nr database, and significantly reduces the NGS data annotation time. Like the previous versions of the database, PANM-DB v3 is freely downloadable from http://panm.sch.ac.kr/ for local BLAST analysis.

key words: PANM-DB, Arthropoda, Nematoda, Mollusks, NGS

서 론

최근 유전자 및 유전체 분석에서 차세대 염기서열 분석 방법인 NGS (Next Generation Sequencing) 기술이 주로 사용되면서 무제한적인 서열 데이터의 확보가 쉬워짐에 따라 (Metzker, 2010) 축적되어지는 NCBI SRA (Sequence Read Archive) NGS 데이터의 양이 최근 10년간 약 만 배 이상 매우 빠르게 증가하고 있다 (https://www.ncbi.nlm.nih.gov/ sra/). 이에 따라 대량의 서열 정보를 NCBI nr (All non- redundant) 데이터베이

Received: March 20, 2019; Revised: March 27, 2019; Accepted: March 31, 2019
Corresponding author: Yong Seok Lee
Tel: +82 (41) 530-3040, e-mail: yslee@sch.ac.kr
1225-3480/24725

스를 이용하여 BLAST (Basic Local Alignment Search Tool) 분석을 진행할 경우 긴 시간이 소요되는 단점을 극복하고 annotation 결과의 정확성을 높이는 것은 NGS 기술발달과 맞물려 생물정보학적 분석에 있어서 가장 큰 문제로 대두되고 있는 실정이다 (Altschul *et al*., 1990; McGinnis and Madden, 2004).

2015년 연체동물 NGS 데이터의 분석시간 단축과 annotation 결과의 정확성을 높이기 위해 PANM 데이터베이스를 구축한 이후 2016년 9월 버전 2로 업데이트 하였다 (Kang *et al*., 2015, 2016). 그 이후 NCBI 의 아미노산 서열 정보들이 대량으로 축적되었으며, 버전 2이 구축 이후 NCBI 의 아미노산 정보를 확인한 결과 정보의 양이 약 150% 정도 상승한 것을 확인하였다. 이에 본 연구를 통해 PANM 데이터베이스의 버전을 업데이트하고 최신의 유전정보를 제공하여 annotation 결과의 질적인 향상을 제공하고자 하였다.

재료 및 방법

1. 데이터베이스 업데이트

2016년 9월에 웹인터페이스 구축 및 업데이트가 진행된 PANM 데이터베이스 버전 2를 버전 3으로 업데이트하기 위하여 2016년 9월부터 2019년 2월까지의 NCBI에 등록된 Arthropoda, Nematoda, Mollusca 의 모든 유전자의 아미노산 서열 정보를 확인 후 다운로드 하였다. 확보한 유전자의 아미노산 서열 정보는 PANM 데이터베이스 버전 2에 추가하여 NCBI에서 제공하는 BLAST가 가능할 수 있도록 formatdb 프로그램을 사용하여 데이터베이스화 하였다.

2. 서버 구축 및 웹 인터페이스 개선

웹상에서 BLAST를 서비스하기 위하여 Xeon E3-1220 CPU를 장착한 PowerEdge(TM) R230 Rack Mount Server에 Linux CentOS 7.6 운영체제를 이용하여 웹서버를 구축하였다. 구축된 서버에 NCBI에서 제공하는 WebBLAST 패키지를 설치하고 PANM 데이터베이스 버전 3을 이용한 BLAST가 수행 가능하도록 인터페이스를 설정하였다. 또한 이전 버전과 마찬가지로 많은 연구자들이 사용할 수 있게 PANM 데이터베이스 버전 3의 다운로드 링크를 탑재하였다.

결과 및 고찰

PANM 데이터베이스의 업데이트를 위해서 NCBI에서 Arthropoda, Nematoda, Mollusca 분류별로 2016년 9월부터 2019년 2월까지 등록된 4,040,997개의 유전자의 아미노산 서열을 다운로드하여 버전 2에 추가하여 총 11,612,243개의 서열, 4,813,465,883개의 아미노산으로 이루어진 PANM 데이터베이스 버전 3을 구축 완료하였다. 버전 3의 데이터를 확인해보면 이전 버전과 마찬가지로 Arthropoda의 유전자 아미노산 서열이 9,313,972개로 가장 많았으며, Nematoda와 Mollusca가 각각 1,585,378개와 712,893개를 차지하는 것으로 확인되었다. 버전 1과 비교해보면 유전자 개수는 약 287%

Table 1. Status of updated amino acid sequences in PANM DB version 3

		PANM-DB	Arthropoda	Nematoda	Mollusca
Total sequences	Ver. 1	4,051,323	3,111,849	652,125	287,349
	Ver. 2	7,571,246	6,178,888	964,027	428,331
	Ver. 3	11,612,243	9,313,972	1,585,378	712,893
	Rate of Increase (1 → 3)	287%	299%	243%	248%
	Rate of Increase (2 → 3)	153%	151%	164%	166%
Total letters	Ver. 1	1,522,609,669	1,196,730,565	226,168,391	99,710,713
	Ver. 2	3,114,590,190	2,590,040,078	366,349,624	158,200,488
	Ver. 3	4,813,465,883	3,939,129,978	569,496,601	304,239,304
	Rate of Increase (1 → 3)	316%	329%	252%	305%
	Rate of Increase (2 → 3)	155%	152%	155%	192%

가 증가하였고, 아미노산 개수는 316%가 증가한 것을 확인하였다. 업데이트된 버전 3의 PANM 데이터베이스를 검증하기 위하여 본 연구팀이 보유한 연체동물 전사체 분석에서 얻은 1,000개의 유전자 서열을 샘플로 하여 BLASTX 분석을 진행한 결과 분석시간은 조금 증가하였지만 annotation 결과 개수가 전체의 37%에서 43%로 약 6% 가량 상승하는 것으로 확인되었다.

NCBI에 등록된 Arthropoda, Nematoda, Mollusca 유전자 아미노산 서열들을 확인한 결과 1999년 최초로 유전자가 등록되었으며, PANM 데이터베이스 버전 1이 발표된 2015년 6월 까지 16년 동안 NCBI에 등록된 유전자 수치가 버전 3이 업데이트된 4년 정도의 시간 동안 약 3배정도 증가하는 것을 확인하였다. 이는 서두에 밝혔듯이 NGS 라는 차세대염기서열 분석기의 가파른 발전과 유전자원에 대한 접근 및 그 이용으로부터 발생하는 이익의 공정하고 공평한 공유에 관한 나고야의정서 시행에 따른 전 세계 생물자원들의 유전체 및 전사체 관련연구가 기하급수적으로 증가하고 있음을 의미한다.

PANM 데이터베이스 버전 2에서는 단독서버가 아니라서 많은 사용자로 인한 과부하 문제로 BLAST 웹서비스를 중단하고 DB 데이터의 다운로드만 제공하여 일반 연구자들이 사용하기 어렵다는 단점이 있었다. 하지만 PANM 데이터베이스 버전 3로 업데이트가 되면서 단독 서버로 웹서비스가 가능한 홈페이지가 구축되었으며, 이를 통하여 웹상에서 바로 PANM 데이터베이스를 활용한 BLAST가 가능하게 되었다. 뿐만 아니라 이전 버전과 마찬가지로 로컬 서버에서도 사용할 수 있도록 PANM 데이터베이스를 다운로드할 수 있도록 하였다.

요 약

버전 3으로 업데이트된 PANM 데이터베이스는 이전 버전과 비교하여 약 1.5배 정도 유전자의 아미노산 서열정보를 추가적으로 확보하였다. 현재 일반적인 annotation 에 주로 사용되고 있는 NCBI nr 데이터베이스에 비하여 그 크기가 6%에 지나지 않으며, 이를 통하여 NGS를 통해 얻어진 대량의 유전자 서열들에 대한 annotation 시간을 NCBI nr 에 비해 더욱 줄여줄 수 있게 되었다. 이전 버전과 마찬가지로 PANM 데이터베이스 버전 3 역시 모든 연구자들이 사용할 수 있도록 오픈하여 다운로드가 가능하게 하였다. 앞으로 Arthropoda, Nematoda, Mollusca의 유전자나 유전체 또는 전사체를 연구하는 연구자들에게 본 연구를 통해 구축된 PANM 데이터베이스가 유용하게 이용될 것으로 사료된다.

사 사

본 논문은 환경부의 재원으로 국립낙동강생물자원관의 지원(NNIBR201901110) 및 순천향대학교 학술연구비의 지원을 받아 수행하였습니다.

REFERENCE

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology,* **215**: 403-410.

Kang, S.W., Hwang, H.J., Park, S.Y., Wang, T.H., Park, E.B., Lee, T.H., Hwang, U.W., Lee, J.-S., Park, H.S., Han, Y.S., Lim, C.E., Kim, S., and Lee, Y.S. (2014) Mollusks Sequence Database: Version II. *The Korean Journal of Malacology,* **30**: 429-431.

Kang, S.W., Park, S.Y., Patnaik, B.B., Hwang, H.J., Kim, C., Kim, S., Lee, J.S., Han, Y.S., and Lee, Y.S. (2015) Construction of PANM Database (Protostome DB) for rapid annotation of NGS data in Mollusks. *The Korean Journal of Malacology,* **31**: 243-247.

Kang, S.W., Park, S.Y., Patnaik, B.B., Hwang, H.J., Chung, J.M, Song, D.K., Park, Y.-S., Lee, J.S., Han, Y.S., Park, H.S., and Lee, Y.S. (2016) The Protostome database (PANM-DB): Version 2.0 release with updated sequences. *The Korean Journal of Malacology,* **32**: 185-188.

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. *The Korean journal of malacology,* **20**: 165-169.

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**: W20-25.

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics,* **11**: 31-46.