Korean J. Malacol. 38(3): 125-128 2022
https://doi.org/10.9710/kjm.2022.38.3.125

[단보, Short communication]

PANM DB ver 5.0 : An update of the PANM database for invertebrate NGS data analysis

Jun Yang Jeong[1,2], Jie Eun Park[1,2], Dae Kwon Song[1,2], Chan-Eui Hong[1,2], Yong Tae Kim[1,2], Hyeonjun Shin[1,2], Ziwei Liu[1,2], Min Kyu Sang[3], Hongray Howrelia Patnaik[2], Bharat Bhusan Patmaik[2,4], Se Won Kang[5], So Young Park[6], Jun Sang Lee[2],Yeon Soo Han[7], Hong Seog Park[8] and Yong Seok Lee[1,2,3]

[1]Department of Biology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam 31538, Korea,
[2]Korea Native Animal Resources Utilization Convergence Research Institute
[3]Research Support Center(Core-Facility) for Bio-Bigdata Analysis and Utilization of Biological Resources
[4]PG Department of Biosciences and Biotechnology, Fakir Mohan University, Balasore, Odisha, 756089, India
[5]Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Jeongeup, Jeonbuk 56212, Korea,
[6]Biodiversity Research Team, Animal & Plant Research Department, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongbuk, 37242, Korea
[7]Department of Applied Biology, Institute of Environmentally-Friendly Agriculture (IEFA), College of Agriculture and Life Sciences, Chonnam National University, Gwangju 61186, Korea
[8]Research Institute, GnC BIO Co., LTD., 621-6 Banseok-dong, Yuseong-gu, Daejeon, 34069, Korea

ABSTRACT

The PANM database (Protostome DB) is a public database platform established in 2015 for the efficient annotation of next-generation sequencing (NGS) data arising from genome sequencing projects involving the Invertebrates (include Arthropoda, Mollusca, and Nematoda). The database has been updated on a routine basis to further increase the accuracy and speed of NGS data annotation. In PANM DB version 5.0 release a total of 21,276,123 protein sequences belonging to the Protostomes have been made available (registered sequences of NCBI). The PANM DB version 5.0 contains about 4% of the total NCBI-nr data shortening the time for overall annotation of large-scale NGS data of the Protostomes. PANM DB version 5.0 can be downloaded for free from http://panm.sch.ac.kr/ for annotation based on local BLAST analysis.

Keywords: PANM-DB, NGS Data, BLAST, Annotation

서 론

NGS 기술의 발전으로 대량의 유전체 및 전사체 서열 데이터를 더 낮은 생산 비용과 더 적은 시간으로 확보가 가능해지면서 NCBI의 Sequence Read Archive (SRA) 의 NGS 데이터의 양이 빠르게 증가하고 있다 (https://www.ncbi.nlm.nih.gov/genbank/statistics). 최근에는 NGS 2세대 서열분석기 즉 Long read sequencer (PacBio SMRT, Oxford Nanopore 등) 의 활용성이 증대되어 Whole Genome Shotgun (WGS) 데이터의 길이 (Read length) 및 데이터 양이 급증하고 있어, Gene prediction 및 annotation의 중요성이 더욱 강조 되어지고 있다.

과거에는 NGS를 통해 얻은 데이터의 annotation을 위해 NCBI에서 제공하는 BLAST와 NCBI nr 데이터 베이스를 이용하는 것이 일반적이었다. 그러나 NGS의 발전에 의한 서열데이터의 급격한 증가로 인해 NCBI nr 데이터베이스를 이용하여 BLAST 분석을 진행할 경우 매우 오랜 시간이 걸린다는 점이 대두되었다 (Altschul *et al*., 1990, McGinnis and Madden, 2004). 본 연구진은 이러한 문제를 극복하고 annotation 결과의 정확성을 높이기 위해 2004년 연체동물 전용 BLAST 서버를 구축하였다 (Lee *et al*., 2004). 하지만 곤충을 포함한 무척추동물 관련 유전정보의 분석을 수행하기에는 적합하지 않아 동일한 인터페이스를 활용하여 2015년

Received: September 16, 2022; Revised: September 21, 2022; Accepted: September 28, 2022
Corresponding author: Yong Seok Lee
Tel: +82 (41) 530-3040, e-mail: yslee@sch.ac.kr
1225-3480/24821

PANM (Arthropoda, Nematoda, Mollusca) 데이터베이스 version 1.0을 구축하였다 (Kang *et al*., 2015). 데이터의 지속적인 증가로 인해 PANM DB는 2016년 9월 version 2.0으로 업데이트를 진행하였으며, 아미노산 서열을 추가하여 2019년 3월 version 3.0으로 업데이트 하였다 (Kang *et al*., 2016, Kang *et al*., 2019). 이후 2021년 2월 NCBI에 2019년 3월 이후 추가적으로 업데이트 된 Arthropoda, Nematoda, Mollusca의 모든 아미노산 서열을 PANM 데이터베이스에 추가하여 version 4.0으로 업데이트를 진행 (Sang *et al*., 2021) 하였다. PANM database의 annotation 속도가 매우 빨라서 해외 연구자들도 많이 활용하고 있다 (Capt *et al*., 2018).

Version 4.0으로 업데이트 이후 2022년 9월 현재까지 NCBI에 등록되어진 Arthropoda, Nematoda, Mollusca 유전자의 아미노산 서열의 수가 약 133.47% 증가한 것을 확인하였다. 이에 annotation 정확도 등 결과의 질적 향상을 위하여 최신 유전정보를 추가하여 PANM 데이터베이스를 업데이트 하였다.

재료 및 방법

1. 데이터베이스 업데이트

PANM database version 5 업데이트를 위하여 2021년 3월부터 2022년 9월까지 NCBI에 등록된 Arthropoda, Nematoda, Mollusca의 모든 amino acid sequence를 다운로드하였다. 확보한 유전자의 amino acid sequence를 PANM database version 4.0에 추가하여 BLAST가 가능하도록 formatdb 프로그램으로 데이터베이스화를 진행하였다.

2. 웹 인터페이스 개선

기존 PANM database version 4.0이 구축된 Web server에 PANM database version 5.0을 새로 적용하여 BLAST 수행이 가능하도록 인터페이스를 수정하였으며, PANM database version 4.0과 같이 많은 연구자들이 활용할 수 있도록 PANM database version 5.0을 탑재하여 다운로드가 가능하도록 하였다.

결과 및 고찰

PANM 데이터베이스 업데이트를 위하여 NCBI에서 2021년 3월부터 2022년 9월까지 등록된 Arthropoda, Nematoda, Mollusca 유전자의 amino acid sequence 총 5,386,132개를 다운로드하였으며, version 4.0에 추가하여 총 21,276,123개의 서열, 11,470,671,405개의 아미노산으로 이루어진 PANM database version 5.0을 구축하였다. Version 5.0의 데이터는 이전버전과 같이 Arthropoda, Nematoda, Mollusca 3개 phylum의 amino acid 서열로 구성하였으며, 각 phylum의 아미노산 서열은 Arthropoda이 전체의 약 83.46%를 차지하는 17,384,642개였으며, Nematoda는 약 8.16%인 2,014,594개, Mollusca는 약 8.38%인 1,876,887개로 확인되었다. 이는 이전 버전과 비교하였을 때 유전자의 수는 약 133%가 증가하였고, 아미노산의 수는 약 158%가 증가한 것으로 나타났다.

Mollusca Bivalvia의 분석을 위해 NCBI browser를 통하여 50,826개의 Nucleotide sequence data와 52,814개의 amino acid sequence data를 다운받아 Bivalvia 전용 database를 구축하였으며, 국내로 수입되는 Bivalvia의 분류학적 연구를 위하여 국내에 수입되는 Bivalvia와 국외에서 직접 채집한 Bivalvia를 대상으로 COI 유전자를 sequencing한 서열정보를 데이터베이스화 하였다 (Chung *et al*., 2019).

또한 복족류의 분석을 위하여 NCBI와 BOLD에서 각각 98,070개와 124,802개의 COI 서열을 다운로드하여 복족류 전용 COI 데이터베이스를 구축하고, 국내로 수입된 복족류와 중국, 미얀마, 영국, 일본, 호주, 베트남 등 6개국에서 확보한 복족류를 대상으로 화보 및 분석한 COI 유전자의 서열정보를 데이터베이스화 하였다 (Sang *et al*., 2021).

NCBI에 등록된 Arthropoda, Nematoda, Mollusca의 amino acid 서열들을 확인해본 결과 1999년 최초로 등록되기 시작하였으며, PANM database version 1.0이 발표된 2015년 6월까지 4,051,323개의 amino acid 서열이 등록되었다. 이후 매년 평균적으로 약 200만개의 Amino acid 서열이 등록되고 있다. 이는 차세대 염기서열 분석기의 빠른 발전과 나고야 의정서 시행에 따른 생물자원의 유전체 및 전사체 연구가 증가하고 있으며, 연구 목적 및 방법에 따라 다양화되고 있음을 의미한다.

요 약

업데이트된 PANM database version 5.0은 이전 version과 비교하여 약 1.3배의 amino acid 서열들이 추가되었다. 일반적으로 annotation에 사용하는 NCBI-nr database와 비교하면 약 4%의 서열로 구축되어 있으나, Arthropoda, Nematoda, Mollusca에 해당하는 유전자의 amino acid sequence를 이용하여 구축하였기 때문에 3개의 phylum에 해당하는 종의 유전자, 유전체 또는 전사체 sequence에 대한 annotation 시간을 줄이고 정확도를 향상시킬 수 있었다. 또한 PANM database version5는 이전 version과 같이 모든 연구자들이 사용할 수 있도록 오픈하여 다운로드가 가능하게 하였으며, 앞으로 무척추동물을 연구하는 연구자들에게 유용하

Table 1. Status of updated amino acid sequences in PANM DB version 5

		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
	Ver. 4	15,889,991	12,943,654	1,858,096	1,088,241
	Ver. 5	21,276,123	17,384,642	2,014,594	1,876,887
	Rate of Increase (1 → 5)	525%	558%	308%	653%
	Rate of Increase (2 → 5)	281%	280%	312%	438%
	Rate of Increase (3 → 5)	183%	186%	127%	263%
	Rate of Increase (4 → 5)	133%	134%	108%	172%
Total Letters	Ver. 1	1,552,609,669	1,196,730,565	226,168,391	99,710,713
	Ver. 2	3,114,590,190	2,590,040,078	266,349,624	158,200,488
	Ver. 3	4,813,465,883	3,939,129,978	569,496,601	304,239,304
	Ver. 4	7,245,891,909	6,032,532,729	697,011,463	516,347,717
	Ver. 5	11,470,671,405	9,573,324,156	935,627,103	961,720,146
	Rate of Increase (1 → 5)	739%	800%	414%	965%
	Rate of Increase (2 → 5)	368%	370%	351%	608%
	Rate of Increase (3 → 5)	238%	243%	164%	316%
	Rate of Increase (4 → 5)	158%	159%	134%	186%

게 이용될 것으로 판단된다.

사 사

본 연구는 교육부에서 지원하는 지역대학 우수과학자(한국연구재단, NRF-2017R1D1A3B03034971), 중점연구소(NRF-2021R1A6A1A03039503) 및 국가연구시설장비진흥센터(2022R1A6C101B794), 순천향대학교 학술연구비의 지원을 받아 수행하였습니다.

REFERENCES

C. Capt, S. Renaut, F. Ghiselli, L. Milani, N. A. Johnson, B. E. Sietman, D. T. Stewart, and S. Breton. (2018) Deciphering the Link between Doubly Uniparental Inheritance of mtDNA and Sex Determination in Bivalves: Clues from Comparative Transcriptomics. *Genome Biology and Evolution,* **10**(2): 577-590.

J. M. Chung, H. R. Min, M. K. Sang, J. E. Park, H. C. Cho, S. W. Kang, S. Y. Park, B. B. Patnaik, Y. S. Lee, K. A. Lutaenko, E. H. Shin, J.-S. Lee, Y. S. Han, H. J. Hwang, and Y. S. Lee. (2019) Molecular phylogenetic study of bivalvia from four countries (China, Japan, Russia and Myanmar) using 3 types. *The Korean Journal of Malacology,* **35**(2): 137-148.

M. K. Sang, J. E. Park, D. K. Song, J. Y. Jeong, C.-E. Hong, Y. T. Kim, H. J. Hwang, S. W. Kang, S. Y. Park, J. S. Lee, Y. S. Han, H. S. Park, and Y. S. Lee. (2021) PANM DB ver 4.0 An update of the bioinformatics database for annotation of large datasets from sequencing of species under Invertebrates. *The Korean Journal of Malacology,* **37**(1): 33-36.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. (1990) Basic local alignment search tool. *Journal of Molecular Biology,* **215**(3): 403-410.

S. McGinnis, and T. L. Madden. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research,* **32**: W20-25.

S. W. Kang, S. Y. Park, H. J. Hwang, J. M. Chung, M. K. Sang, H. R. Min, J. E. Park, H. C. Cho, B. B. Patnaik, and Y. S. Lee. (2019) PANM DB ver 3.0 An update of the bioinformatics database for

annotation of large datasets from sequencing of species under Protostomia clade. *The Korean Journal of Malacology,* **35**(1): 73-75.

S. W. Kang, S. Y. Park, B. B. Patnaik, H. J. Hwang, J. M. Chung, D. K. Song, Y.-S. Park, J. S. Lee, Y. S. Han, H. S. Park, and Y. S. Lee. (2016) The Protostome database (PANM-DB): Version 2.0 release with updated sequences. *The Korean Journal of Malacology,* **32**(3): 185-188.

S. W. Kang, S. Y. Park, B. B. Patnaik, H. J. Hwang, C. Kim, S. Kim, J. S. Lee, Y. S. Han, and Y. S. Lee. (2015) Construction of PANM Database (Protostome DB) for rapid annotation of NGS data in Mollusks. *The Korean Journal of Malacology,* **31**(3): 243-247.

Y.-S. Lee, Y.-H. Jo, D.-S. Kim, D.-W. Kim, M.-Y. Kim, S.-H. Choi, J.-O. Yon, I.-S. Byun, B.-R. Kang, K.-H. Jeong, and H.-S. Park. (2004) Construction of BLAST Server for Mollusk. *The Korean Journal of Malacology,* **20**(2): 165-169.