Korean J. Malacol. 37(1): 33-36 2021
https://doi.org/10.9710/kjm.2021.37.1.33

무척추동물의 NGS 데이터 분석용 PANM 데이터베이스 업데이트 (Version Ⅳ)

상민규[1], 박지은[1], 송대권[1], 정준양[1], 홍찬의[1], 김용태[1], 황희주[2], 강세원[3], 박소영[4], 이준상[2], 한연수[5], 박홍석[6], 이용석[1]

[1]순천향대학교 자연과학대학 생명과학과, [2]순천향대학교 자연과학대학 기초과학연구소, [3]한국생명공학연구원 생물자원센터, [4]국립낙동강생물자원관 전략기획실 혁신성과부, [5]전남대학교 농업생명과학대학 식물생명공학부, [6]㈜지앤시바이오

PANM DB ver 4.0 : An update of the bioinformatics database for annotation of large datasets from sequencing of species under Invertebrates

Min Kyu Sang[1], Jie Eun Park[1], Dae Kwon Song[1], Jun Yang Jeong[1], Chan-Eui Hong[1], Yong Tae Kim[1], Hee Ju Hwang[2], Se Won Kang[3], So Young Park[4], Jun Sang Lee[2], Yeon Soo Han[5], Hong Seog Park[6] and Yong Seok Lee[1]

[1]Department of Biology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam 31538, Korea,
[2]Institute for Basic Sciences, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam, 31538, Korea,
[3]Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Jeongeup, Jeonbuk 56212, Korea,
[4]Performance Management Division, Strategic Planning Department, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongbuk, 37242, Korea,
[5]Department of Applied Biology, Institute of Environmentally-Friendly Agriculture (IEFA), College of Agriculture and Life Sciences, Chonnam National University, Gwangju 61186, Korea
[6]Research Institute, GnC BIO Co., LTD., 621-6 Banseok-dong, Yuseong-gu, Daejeon, 34069, Korea

ABSTRACT

The PANM database (Protostome DB) was first established in 2015 as a public database platform aimed at the efficient annotation of next-generation sequencing (NGS) data of species belonging to Mollusks. It was updated to version 2 in 2016 and version 3 established in 2019, with a total of 11,615,243 protein sequences. The PANM DB version 4 was updated by integrating the protein sequences of Arthropoda, Nematoda, and Mollusca registered in NCBI from the update to PANM DB v3 until February 2021. The PANM DB v4 accounts for only about 4% of the NCBI-nr database but significantly reduces the Invertebrates' NGS data annotation time.

Keywords: PANM-DB, Arthropoda, Nematoda, Mollusks, NGS

서 론

최근 유전자, 유전체 및 전사체의 분석에서 저렴한 비용으로 대량의 염기서열 데이터의 확보가 용이한 차세대 염기서열 분석 방법인 NGS (Next Generation Sequecing) 기술이 주로 사용되면서, 연구의 목적에 따라 다양한 방식의 NGS 기술을 이용하고 있다 (Metzker, 2010; Goodwin *et al*., 2016). 이에 따라 NCBI SRA (Sequence Read Archive) 와 NCBI GenBank의 NGS 데이터 양이 매우 빠르게 증가하고 있다 (https://www.ncbi.nlm.nih.gov/sra). 이러한 상황으로 인해 대량의 염기서열 분석 정보를 NCBI nr (All non-redundant) 데이터베이스를 이용해 BLAST (Basic Local

Received: march 19, 2021; Revised: march 26, 2021; Accepted: march 31, 2021
Corresponding author: Yong Seok Lee
Tel: +82 (41) 530-3040, e-mail: yslee@sch.ac.kr
1225-3480/24786
This is an Open Access Article distributed under the terms of the Creative Commons Attribution Non-Commercial License with permits unrestricted non-commercial use, distribution, and reproducibility in any medium, provided the original work is properly cited.

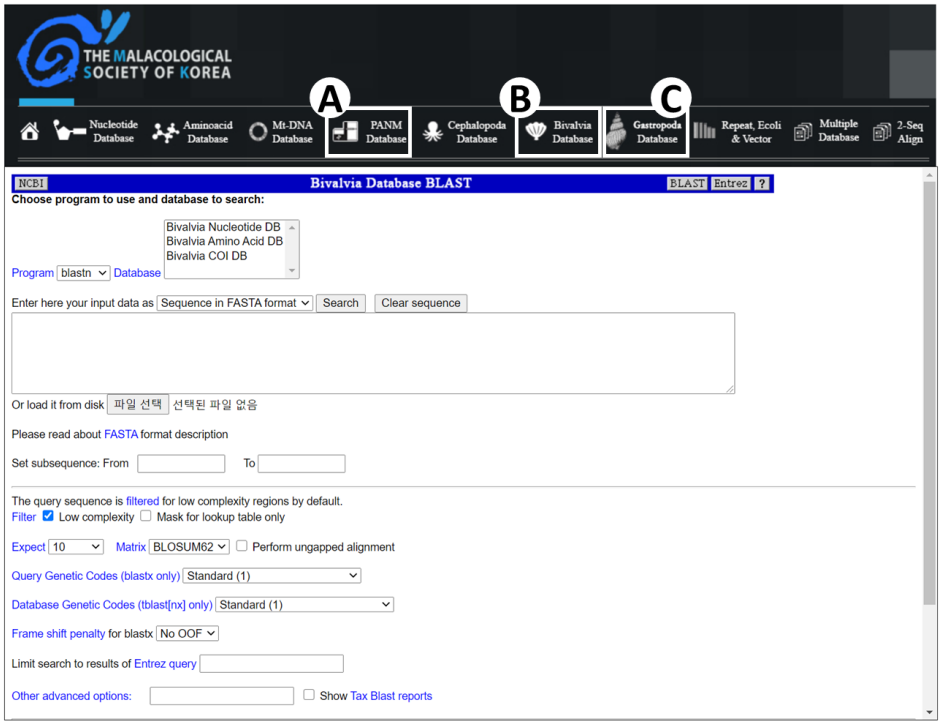

Fig. 1. A web interface of PANM-DB version Ⅳ (https://www.panm.sch.ac.kr). (**A**) PANM database, (**B**) Bivalvia database, (**C**) Gastropoda database.

Alignment Search Tools) 분석을 진행할 경우 시간이 매우 오래 걸리는 단점이 부각되고 있으며, 이러한 단점을 극복하고 annotation 결과의 정확성을 높이는 것은 생물정보 분석에 있어 가장 큰 문제로 대두되고 있다 (Altschul *et al*., 1990; McGinnis and Madden, 2004).

지난 2015년, 본 연구진은 연체동물 NGS 데이터의 annotation 결과 정확도 향상과 분석에 소요되는 시간을 단축하기 위하여 무척추동물 전용 데이터베이스를 구축하였다 (Kang *et al*., 2015). 이후 데이터 증가 추세에 따라 2016년 9월 버전 2로 업데이트하였으며, 아미노산 서열을 추가, 2019년 3월 버전 3로 업데이트 하였다 (Kang *et al*., 2016; Kang *et al*., 2019). 마지막 업데이트 이후 2021년 2월 현재까지의 NCBI 확인 결과, 절지동물문, 선형동물문, 연체동물문의 유전자 아미노산 서열 수가 약 140% 증가되었다. 이에 annotation 결과의 질적 향상을 위하여 최신의 유전정보를 추가해 PANM 데이터베이스를 업데이트 하였다.

재료 및 방법

1. 데이터베이스 업데이트

PANM 데이터베이스 버전 3의 업데이트를 위하여 절지동물, 선형동물, 연체동물의 2019년 3월부터 2021년 2월까지 NCBI에 등록된 모든 유전자 아미노산 서열을 확인하여 다운로드하였다. 확보한 유전자의 아미노산 서열을 PANM 데이터베이스 버전 3에 추가하였으며, BLAST가 가능하도록 formatdb 프로그램을 이용하여 데이터베이스화 하였다.

2. 웹 인터페이스 개선

기존 PANM 데이터베이스 버전 3가 구축된 웹서버에 새로 적용한 PANM 데이터베이스 버전 4를 이용하여 BLAST 수행이 가능하도록 인터페이스를 설정하였으며, PANM 데이터베이스 버전 3와 같이 많은 연구자들이 이용할 수 있도록 PANM 데이터베이스 버전 4를 다운로드 할 수 있도록 탑재하였다.

결과 및 고찰

PANM 데이터베이스의 업데이트를 진행하기 위하여 NCBI에서 2019년 3월부터 2021년 2월까지 등록된 Arthropoda, Nematoda, Mollusca 유전자의 아미노산 서열 총 4,277,748개를 다운로드하였으며, 버전 3에 추가하여 총 15,889,991개의 서열, 7,245,891,909개의 아미노산으로 이루어진 PANM 데이터베이스 버전 4를 구축하였다 (Fig. 1-A). 버전 4의 데이터는 이전버전과 같이 절지동물, 선형동물, 연체

Table 1. Amino acid sequence status updated in PANM DB version 4

		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
	Rate of Increase (1 → 4)	392%	416%	285%	379%
	Rate of Increase (2 → 4)	210%	209%	193%	254%
	Rate of Increase (3 → 4)	137%	139%	117%	153%
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	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
	Rate of Increase (1 → 4)	476%	504%	308%	518%
	Rate of Increase (2 → 4)	233%	233%	262%	326%
	Rate of Increase (3 → 4)	151%	153%	122%	170%

동물의 3개 문 (phylum) 의 유전자 아미노산 서열로 구성하였으며, 각 문 (phylum) 의 아미노산 서열은 절지동물이 전체의 약 81.5%를 차지하는 12,943,654개였으며, 선형동물은 약 11.7%인 1,858,096개, 연체동물은 약 6.8%인 1,088,241개로 확인되었다. 이전 버전과 비교해보면, 유전자의 수는 약 137%가 증가하였고, 아미노산의 수는 약 151%가 증가한 것으로 확인되었다 (Table 1).

연체동물문 이매패류 분석을 위해 NCBI taxonomy browser를 통하여 50,286개의 뉴클레오타이드 서열 정보와 52,814개의 아미노산 서열 정보를 다운받아 이매패류 전용 데이터베이스를 구축하였으며, 국내로 수입되는 이매패류의 분류학적 연구를 위하여 국내에 수입되는 이매패류와 국외에서 직접 채집한 이매패류를 대상으로 COI 유전자를 시퀀싱한 서열정보를 데이터베이스화 하였다 (Fig. 1-B) (Chung *et al.*, 2019).

또한 복족류의 분석을 위하여 NCBI와 BOLD (Barcode of Life Data) 에서 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 확보·분석한 COI 유전자의 서열정보를 데이터베이스화 하였다 (Fig. 1-C).

NCBI에 등록된 절지동물문, 선형동물문, 연체동물문의 유전자 아미노산 서열들을 확인해본 결과 1999년 최초로 등록되기 시작하였으며, PANM 데이터베이스 버전 1이 발표된 2015년 6월까지 4,051,323개의 유전자 아미노산 서열이 등록되었다. 이후 매년 평균적으로 약 200만개의 유전자 아미노산 서열이 꾸준히 등록되었으며, 최근에는 더욱 긴 서열의 유전자 아미노산 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 데이터베이스 버전 4는 이전 버전에 비해 약 1.3배의 유전자 아미노산 서열이 추가되었다. 일반적으

로 annotation에 사용하는 NCBI-nr 데이터베이스와 비교하면 약 4%의 서열로 구되어 있으나, 절지동물문, 선형동물문, 연체동물문에 해당하는 유전자의 아미노산 서열을 이용하여 구축하였기에 3개 문 (phylum) 에 해당하는 종의 유전자, 유전체 또는 전사체 서열에 대한 annotation 시간을 줄이고 정확도를 향상시킬 수 있었다. 또한 PANM 데이터베이스 버전 4는 이전 버전과 같이 모든 연구자들이 사용할 수 있도록 오픈하여 다운로드가 가능하게 하였으며, 앞으로 무척추동물을 연구하는 연구자들에게 유용하게 이용될 것으로 판단된다.

사 사

본 논문은 교육부 (한국연구재단, NRF-2017R1D1A3B06034971) 및 순천향대학교 학술연구비의 지원을 받아 수행하였습니다.

REFERENCES

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

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

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**: 333-351.

Kang, S.W., Park, S.Y., Hwang, H.J., Chung, J.M., Sang, M.K., Min, H.R., Park, J.E., Cho, H.C., Patnaik, B.B., and Lee, Y.S. (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**: 73-75.

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.

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.

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

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Gene*t., **11**: 31-46.