Korean J. Malacol. 32(3): 185-188 2016
http://dx.doi.org/10.9710/kjm.2016.32.3.185

연체동물 NGS 데이터 분석을 위한 PANM 데이터베이스 업데이트 (Version II)

강세원[1], 박소영[2], Bharat Bhusan Patnaik[1,3], 황희주[1], 정종민[1], 송대권[1], 박영수[4], 이준상[5], 한연수[6], 박홍석[7], 이용석[1]

[1]순천향대학교 자연과학대학 생명시스템학과, [2]국립낙동강생물자원관 다양성보전·변화연구부, [3]Trident School of Biotech Sciences, Trident Academy of Creative Technology, [4]순천향대학교 의과대학 간호학과, [5]강원대학교 환경연구소, [6]전남대학교 농업생명과학대학 식물생명공학부, [7](주)지앤시바이오

The Protostome database (PANM-DB): Version 2.0 release with updated sequences

Se Won Kang[1], So Young Park[2], Bharat Bhusan Patnaik[1,3], Hee Ju Hwang[1], Jong Min Chung[1], Dae Kwon Song[1], Young-Su Park[4], Jun Sang Lee[5], Yeon Soo Han[6], Hong Seog Park[7] and Yong Seok Lee[1]

[1]Department of Life Science and Biotechnology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam 31538, Korea
[2]Biodiversity Conservation & Change Research Division, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongbuk 37242, Korea
[3]Trident School of Biotech Sciences, Trident Academy of Creative Technology (TACT), Bhubaneswar 751024, Odisha, India
[4]Department of Nursing, College of Medicine, Soonchunhyang University, Asan, Chungnam 31538, Korea
[5]Institute of Environmental Research, Kangwon National University, Chuncheon, Gangwon 24341, Korea
[6]College of Agriculture and Life Science, Chonnam National University, Gwangju 61186, Korea
[7]Research Institute, GnC BIO Co., LTD. Daejeon 34069, Korea

ABSTRACT

PANM-DB (version 1.0) was constructed as a web-based interface for the analysis and annotation of Next-Generation Sequencing (NGS) data of Mollusca, Arthropoda, and Nematoda. The database collected the sequences of Protostomes (Mollusca, Arthropoda, and Nematoda) from the NCBI Taxonomy Browser, and the same were compiled in a multi-FASTA format and stored using the formatdb program. This improved the processing of the RNA-seq sequences in terms of speed and hit percentage. PANM-DB has been successfully used for the transcriptome annotation of butterfly, land snail, and other commercial mollusca. We have improved the database by updating the same with new sequences and version 2.0 contains a total of 7,571,246 protein sequences (two times more as compared to version 1.0). Furthermore, the updated version contains the Cephalopoda database. The constructed web interface is available that independently analyses following these updates that is an improvement of the mollusks BLAST server. The updated version of PANM-DB will be helpful for the analysis of the NGS based sequencing data of non-model species, especially Mollusca, Arthropoda, Nematoda. (http://malacol.or.kr/blast/PANM.html)

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

Received: September 24, 2016; Revised: September 27, 2016; Accepted: September 30, 2016
Corresponding author : Yong Seok Lee
Tel: +82 (41) 530-3040,e-mail: yslee@sch.ac.kr
1225-3480/24626

서 론

유전체의 분석에서 가장 기본이 되는 것은 염기서열을 알아내는 시퀀싱 과정이다. 과거에는 sanger 시퀀싱 방식이 주로 사용되어 졌지만 최근에는 대량의 염기서열을 얻을 수 있는 NGS (Next Generation Sequencing) 방식이 주로 사용되고 있다 (Sanger *et al.*, 1977; Metzker, 2010). 이러한 기술을 통하여 유전체 및 전사체 서열 정보의 생산 비용과 시간이

현저히 단축되어 지고 있고, 이에 따라서 축적되어지는 NGS 데이터의 양이 가파른 상승곡선을 그리며 많아지고 있는 추세이다.

NGS를 통해 얻은 데이터의 annotation 을 위해서는 NCBI 에서 제공하는 BLAST (Basic Local Alignment Search Tool) 와 NCBI nr (All non-redundant) 데이터베이스를 이용하는 것이 매우 일반적이었다 (Altschul *et al*., 1990; McGinnis and Madden, 2004). 그러나 NGS를 통하여 염기서열 데이터의 양이 많아지면서 NCBI nr 데이터베이스를 이용하여 분석을 진행하면 시간이 매우 오래 걸린다는 단점이 지적되고 있다. 이러한 단점을 극복하기 위하여 본 연구진은 기존에 사용되고 있던 연체동물 전용 BLAST 서버 (Lee *et al*., 2004; Kang *et al*., 2014) 에 추가로 선구동물에서 많은 비중을 차지하는 절지동물, 선형동물의 데이터를 엮은 NGS 전용 PANM (Protostome DB) 데이터베이스를 구축하여 사용 중이었다 (Kang *et al*., 2015).

전 세계적으로 NGS 기법의 등장이후 annotation에 사용되는 NCBI에 등록되는 서열 역시 기하급수적으로 증가하고 있기 때문에 최신의 분석 결과를 위해서는 데이터베이스의 주기적인 업데이트가 필요하다. 또한 현재의 웹 인터페이스에서는 PANM 데이터베이스 전용 페이지는 없는 상태이기 때문에 많은 연구자들이 더 쉽게 이용할 수 있도록 웹 인터페이스의 개선도 이루어져야 할 필요가 있다. 이에 따라 본 연구는 PANM 데이터베이스를 최신의 상태로 업데이트를 진행하는 것과 웹 인터페이스의 개선이 진행되었다. 추가적으로 연체동물에 속하는 두족류의 분류에 대한 연구들이 늘어나고 있는 상황에 대비하여 두족류 전용 데이터베이스 구축이 진행되었다.

재료 및 방법

1. 데이터베이스 업데이트 및 구축

PANM 데이터베이스 버전 I 의 구축 이후에서부터 2016년 8월 31일까지 NCBI 등록되어 있는 연체동물, 절지동물, 선형동물의 아미노산 서열정보를 taxonomy browser를 통해 모두 다운로드하였다. 다운로드한 데이터는 기존에 PANM 데이터베이스 버전 I 과 multiFASTA 형태로 결합한 후 BLAST 에서 제공하는 formatdb 프로그램을 사용하여 BLAST가 가능하도록 데이터베이스화 하였다. 또한 두족류 전용 데이터베이스를 구축하기 위하여 taxonomy browser를 통해 두족류 데이터를 다운로드하여 데이터베이스화 하였으며, 분류학적인 연구에 사용이 더 용이하도록 국내외 두족류들에 대하여 COI, 16S 유전자 서열들을 따로 모아서 데이터베이스화 하였다.

2. 웹 인터페이스 개선

기존 연체동물 전용 BLAST 인터페이스에서 아미노산 데이터베이스 페이지에 삽입되어 있던 PANM 데이터베이스를 새로운 페이지로서 독립시켜 PANM 데이터베이스 쉽게 사용할 수 있게 하였다. 또한 인용에 필요한 정보 및 PANM 데이터베이스 업데이트 소식을 추가하였으며, 버전 I 과 마찬가지로 연구자의 독립된 서버에서 바로 사용이 가능하도록 PANM 데이터베이스를 압축하여 웹에서 다운로드가 가능하도록 하였다.

결과 및 고찰

NCBI의 taxonomy browser를 통하여 연체동물, 절지동물, 선형동물 유전자의 아미노산 서열정보를 다운로드한 뒤 PANM 데이터베이스의 업데이트를 진행하여 총 7,571,246 개의 유전자 서열이 포함되어 있는 PANM 데이터베이스 버전 II 를 구축하였다. 자세히 살펴보면 절지동물 유전자 서열이 6,178,888 개로 가장 많은 부분을 차지하고 있었고 선형동물이 964,027 개, 연체동물이 428,331 개로 이루어져 있는 것을 확인할 수 있었다. 아미노산 총 개수를 살펴보면 전체 3,114,590,190 개의 아미노산으로 이루어져 있었다. PANM 데이터베이스 버전 I 과 비교한 결과 유전자 서열의 개수는 약 187% 증가하였고, 아미노산 총 개수는 205% 증가하였음을 확인할 수 있었다. PANM 데이터베이스 버전 I 의 경우 NCBI의 등장 이후에서부터 2015년 6월까지의 정보를 포함하고 있었는데 약 1년여 만에 그간의 데이터의 약 2배에 이르는 증가를 보인 것은 역시 NGS를 통한 서열분석 등으로 전 세계에서 유전체와 관련된 연구가 많이 진행되는 것을 대변하는 결

Table 1. Status of the available amino acid sequences in PANM-DB Version II

	Version I		Version II		Rate of Increase	
Database	Total Seq.	Total letters	Total Seq.	Total letters	Total Seq.	Total letters
PANM-DB	4,051,323	1,522,609,669	7,571,246	3,114,590,190	187%	205%
- Arthropoda	3,111,849	1,196,730,565	6,178,888	2,590,040,078	199%	216%
- Nematoda	652,125	226,168,391	964,027	366,349,624	148%	162%
- Mollusca	287,349	99,710,713	428,331	158,200,488	149%	159%

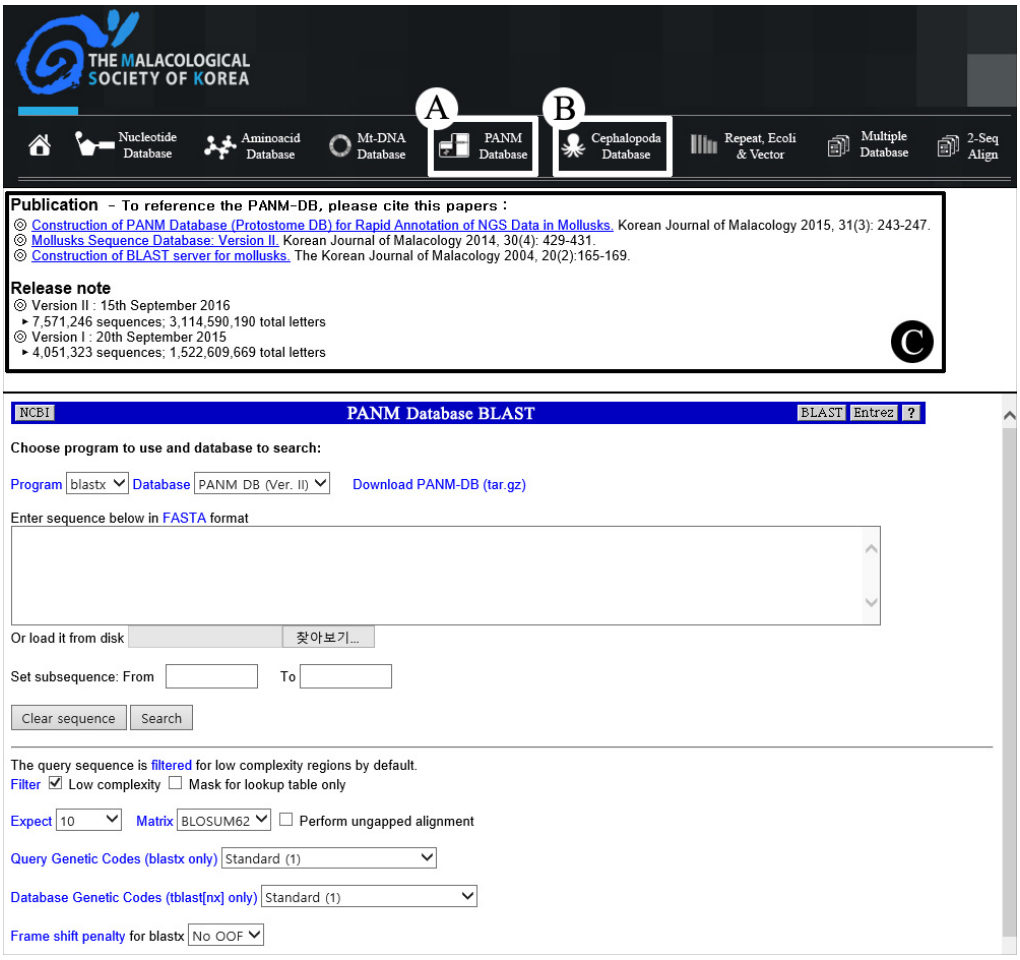

Fig. 1. A web interface for PANM-DB Version II. Screenshot from MSK (the Malacological Society of Korea, http://malacol.or.kr/blast/PANM.html). (**A**) Users can click PANM-DB (**B**) Users can click Cephalopoda-DB (**C**) Information of publication & release note.

과이기도 하다.

연체동물 중 두족류의 분석을 위하여 taxonomy browser를 통하여 49,693 개의 뉴클레오타이드 정보와 73,346 개의 아미노산 정보를 다운로드하여 두족류 전용 데이터베이스를 구축하였다. 또한 국내에 수입되는 두족류들의 분류학적인 연구를 위하여 국외에서 직접 채집한 두족류 및 수입된 두족류들을 대상으로 COI 및 16S 유전자를 직접 시퀀싱한 서열정보를 데이터베이스화하였다 (Hwang *et al*., 2016).

PANM 데이터베이스 버전 I 의 경우 웹 인터페이스가 연체동물전용 BLAST 서버에서 아미노산데이터베이스 페이지에 삽입되어져 있었다. 하지만 버전 II 로 업데이트를 하면서 새로운 메뉴바를 생성하여 PANM 데이터베이스 이용이 더욱 편리해졌다. 또한 PANM 데이터베이스와 관련된 논문 정보와 버전 정보 등의 제공으로 인하여 신뢰도를 높였으며, PANM 데이터베이스를 직접 다운로드 할 수 있도록 하여 관련 연구자들이 개별 서버에서 더욱 더 빠른 분석이 가능하도록 하였다.

요 약

본 연구를 통하여 업데이트된 PANM 데이터베이스 버전 II는 버전 I 에 비해 많은 양의 정보가 추가되었다. 하지만 여전히 NCBI nr 데이터베이스에 비해 적은 양으로서, NGS 분석에 있어 많은 시간을 절약하게 해줄 수 있다. 또한 웹 인터페이스의 개선으로 인하여 직관성 및 신뢰성을 더욱 더 확보할 수 있었다. 개별적인 서버를 운용하여 NGS 데이터를 분석하는 연구자들을 위해 PANM 데이터베이스의 다운로드가 가능하도록 하였고 이로 인해 NGS 데이터 분석 시간이 줄어들 수 있을 것이다. 앞으로 꾸준한 PANM 데이터베이스 업데이트를 통하여 연체동물을 연구하는 연구자들은 물론 절지동물, 선형동물을 연구하는 연구자들에게도 많은 도움이 될 것으로 생각

되며, 추가적으로 구축된 두족류 전용 데이터베이스 역시 두족류를 연구하는 연구자들에게 매우 유용하리라 사료되어진다.

사 사

본 논문은 정부(환경부)의 재원으로 국립생물자원관의 지원(NIBR201603205) 및 순천향대학교 학술연구비 지원으로 수행되었습니다.

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.

Hwang, H.J., Kang, S.W., Park, S.Y., Chung, J.M., Song, D.K., Park, H., Park, H.S., Han, Y.S., Lee, J.-S., and Lee, Y.S. (2016) Classification and Phylogenetic Studies of Cephalopods from four countries of South-East Asia. *The Korean Journal of Malacology,* **32**: 55-62.

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.

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

Sanger, F., Nicklen, S., and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* **74**: 5463-5467.