Korean J. Malacol. 26(2): 185-188, 2010

웹기반 전복류 (*Haliotis*) SNP 데이터베이스 구축

정지은, 이재봉, 강세원, 백문기, 한연수[1], 최태진[2], 강정하[3], 이용석

인제대학교 의과대학 기생충학교실, [1]전남대학교 농업생명과학대학 식물생명공학부
[2]부경대학교 미생물학과, [3]국립수산과학연구원 생명공학과

Construction of web-based Database for *Haliotis* SNP

Ji-Eun Jeong, Jae-Bong Lee, Se-Won Kang, Moon-Ki Baek, Yeon-Soo Han[1], Tae-Jin Choi[2], Jung-Ha Kang[3] and Yong-Seok Lee

Department of Parasitology, College of Medicine and Frontier Inje Research for Science and Technology, Inje University, Busan, 614-735, Korea
[1]Department of Agricultural Biology, Chonnam National University, Gwangju 500-757, Korea
[2]Department of Microbiology, Pukyong National University, Busan, Korea
[3]Biotechnology research division, NFRDI, Busan, Korea

ABSTRACT

The Web-based the genus *Haliotis* SNP database was constructed on the basis of Intel Server Platform ZSS130 dual Xeon 3.2 GHz cpu and Linux-based (Cent OS) operating system. *Haliotis* related sequences (2,830 nucleotide sequences, 9,102 EST sequences) were downloaded through NCBI taxonomy browser. In order to eliminate vector sequences, we conducted vector masking step using cross match software with vector sequence database. In addition, poly-A tails were removed using Trimmest software from EMBOSS package. The processed sequences were clustered and assembled by TGICL package (TIGR tools) equipped with CAP3 software. A web-based interface (*Haliotis* SNP Database, http://www.haliotis.or.kr) was developed to enable optimal use of the clustered assemblies. The Clustering Res. menu shows the contig sequences from the clustering, the alignment results and sequences from each cluster. And also we can compare any sequences with *Haliotis* related sequences in BLAST menu. The search menu is equipped with its own search engine so that it is possible to search all of the information in the database using the name of a gene, accession number and/or species name. Taken together, the Web-based SNP database for *Haliotis* will be valuable to develop SNPs of *Haliotis* in the future.

Key Word : *Haliotis*, SNP database, Web-interface

서 론

전복 (全鰒) 은 전복과 (Haliotidae), 전복속 (*Haliotis*) 에 속하는 연체동물의 총칭으로, univalve (single-shelled) 해양 복족류이다. 우리나라에서 제일 많이 발견되는 대표 종으로 둥근전복 (*Haliotis discus discus*) 이 알려져 있다 (Lee & Min 2002). 예로부터 전복은 식용으로 뿐만 아니라 약으로써도 많은 효험을 나타내 귀하게 여겨져 왔다. 특히 국내산 전복과 5종은 실명 (blindness) 을 완화하는 치료약으로 사용되고 있다 (정평림 *et al*. 2000).

한국, 중국, 일본 등지에서 순종간의 교배에 의한 방법인 selective breeding program의 성공으로 양식 산업이 발전하였으며 최근에는 분자육종에 의한 품종개발 관련 연구들이 수행되어지고 있다. 하지만 전복의 경우, 아직 밝혀진 유전체 또는 유전자 서열이 매우 적어 분자육종에 필요한 SNP 마커 등의 발굴이 매우 어려운 실정이다. 몇몇 종의 전복에 관한 SNP 분석의 보고가 있지만 (QI *et al*. 2008; Qi *et al*. 2009), 본 연구에서는 현재까지 알려진 *Haliotis* 의 모든 염기서열을 정리하여 SNP 발굴이 용이하도록 웹데이터베이스를 구축하였다.

Received May 11, 2010; Revised June 17, 2010; Accepted June 25, 2010
Co-corresponding author: Yong Seok Lee & Jung Ha Kang
Tel: +82-51-890-6462 e-mail: yslee@inje.ac.kr
Tel: +82-51-720-2462 e-mail: kjh0124@nfrdi.go.kr
1225-3480/24351

재료 및 방법

1. 서버구축 및 환경설정

사용된 서버는 Intel Server Platform ZSS130 (Samsung) 에 Xeon 3.2 GHz cpu 시스템을 사용하였으며, 운영체제 (operating system) 는 Cent OS를 사용하였다. 운영체제 설치 후 Apache, PHP, Mysql 연동 시스템을 구축하였으며, 서버의 설정에서 웹 접속 사용자가 cgi (common gate interface) 를 사용할 수 있도록 환경설정을 한 후 Web BLAST 패키지를 설치하였다.

2. 전복 관련 서열의 확보 및 분석

NCBI taxonomy browser 를 통해 *Haliotis* 에 속하는 모든 서열정보 (Core Nucleotide sequences, EST sequences, Amino Acid sequences 및 Mitochondrial genome sequences) 들을 다운로드 하였다 (Table 1). 보다 정확한 SNP의 발굴을 위하여 다운로드 한 염기서열 및 EST 서열을 대상으로 크로스매치 및 벡터데이터베이스를 이용한 벡터마스킹 작업을 수행하였으며, EMBOSS package의 Trimest 프로그램을 활용하여 잔재하고 있는 poly-A 서열들을 제거하였다 (Rice *et al*. 2000). 정리 되어진 서열들은 TGICL package 를 이용하여 30 bp 이상, 94% 동일한 서열의 경우 같은 클러스터로 분류하고, 클러스터에 들어간 서열들은 cap3 를 이용하여 assembly 하였다 (Huang & Madan 1999; Pertea *et al*. 2003). 본 작업을 통해 만들어진 contigs 들은 모두 연체동물전용 BLAST DB 및 nr 데이터베이스에 BLASTx 를 통해 annotation 작업을 수행하였다 (Altschul *et al*. 1990).

3. 웹 인터페이스 구축

Blast 메뉴에서는 핵산, 아미노산, EST 3개의 소메뉴로 구성하였다. *Haliotis* 속에서 현재까지 밝혀진 핵산 서열, 아미노산 서열, EST 서열을 대상으로 각각 BLAST 가 가능하도록 구성하였으며, query 및 데이터베이스가 허용하는 한 blastp, blastn, blastx, tblastn, tblastx 모두 수행 가능하도록 하였다. Vector, *E. coli,* Repeat 서열을 따로 검색 할 수 있도록 하였으며, multi DB 메뉴를 만들어 라이브러리 확인 (insert size 측정) 등을 할 때 용이하도록 하였다. 또한 perl script를 기반으로 한 검색엔진을 설치하여 Annotation 이 되어진 *Haliotis* EST 서열정보를 종 이름, 유전자 이름 및 NCBI accession number 등을 query로 하여 찾을 수 있도록 하였다. 그리고 primer3 등 실험 시 부가적으로 필요한 웹용 프로그램을 설치하여 연구자들이 편리하게 이용하도록 하였다 (Fig1).

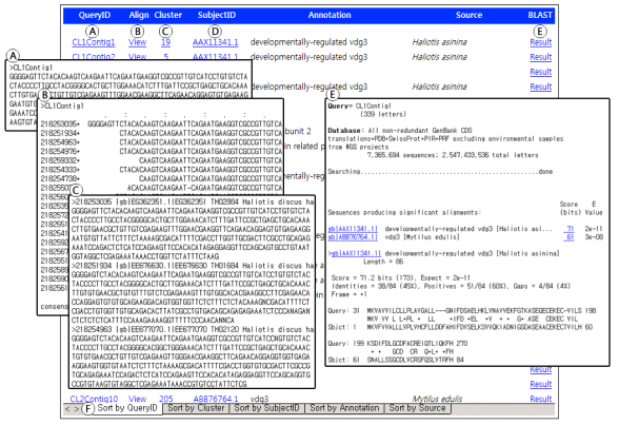

Fig. 1. The annotation results by clustering and assembly of nucleotide and EST sequences downloaded through NCBI taxonomy browser. (**A**) Contig sequence, (**B**) Alignment results, (**C**) Sequences contained in the contig, (**D**) Subject ID (NCBI protein database link), (**E**) Blast results (against NCBI nr), (**F**) Tab menu for Result view by sorting.

결과 및 고찰

1. 염기서열 분석

로컬서버에 다운로드 되어진 *Haliotis*의 염기서열은 총 11,932 개 이었는데 그 중 가장 많은 서열은 *Haliotis discus* 의 EST 서열로 총 7,260 개 이었다. 11,932 개의 서열을 대상으로 vector sequence database 와 cross_match 프로그램을 사용하여 vector sequence를 제거한 결과 587 개의 벡터 서열들이 제거 되었으며, poly-A를 가지고 있는 서열들은 모두 3,580 개 이었다. 두 과정을 모두 거친 후, 100 bp 이하의 서열은 총 148 개로 본 분석에서 제외하였으며, 그리하여 최종 데이터로 사용된 서열은 11,197 개의 서열이었다.

정리된 11,197 개의 서열은 BLAST 및 cap3 소프트웨어를 엔진으로 한 TGICL package (TIGR tools)를 이용하여 clustering 및 assembly 를 수행한 결과 1,317 개의 클러스터에서 1,415 개의 contig 와 3,962 개의 singleton 이 생성되었다.

생성되어진 1,415 개의 contig를 연체동물 전용 데이터베이스 (Lee *et al*. 2004) 에 BLASTx 로 검색한 결과 1,116 개의 유효한 결과를 얻을 수 있었으며 *Haliotis* NT database 에 BLASTn 으로 검색한 결과 1,326 개의 유효한 결과를 얻을 수 있었다 (Table 1).

2. 데이터베이스 구축

Home menu 에서는 전반적인 data processing 방법을 도식화하여 보여주고 있으며, clustering res. 메뉴에서는 clustering 결과에 의해 만들어진 contig 의 서열, align 결

Table 1. Summary of the nucleotides and EST assembly

Information about *Haliotis* sequences downloaded from NCBI	Number of *Haliotis* sequences	11,932
	Haliotis asinine	1,760
	Haliotis discus	7,260
	Haliotis diversicolor	12
	Haliotis midae (perlemoen abalone)	70
	Others (*Haliotis* spp.)	142
After editing the sequence data	Number of vector sequences	587
	Poly-A tail removed	3,580
	Sequences below 100 bp	148
Total result of clustering and assembly	Number of sequences for assembly	11,197
	Number of clusters	1,317
	Number of contigs	1,415
	Significant BLAST hit against Mollusks DB AA	1,116
	Significant BLAST hit against *Haliotis* DB NT	1,326
	Number of Singleton	3,962

과, cluster 를 이루는 각 read 의 서열을 웹상에서 볼 수 있도록 하였다. contig 서열을 NCBI nr database 에 BLASTx 로 homology test 한 결과 중 best hit 에 해당하는 annotation 및 source (species) 를 바로 볼 수 있으며 전체 BLAST 결과를 볼 수 있는 링크를 구성하였다. 또한 결과는 Query ID, Align results, Cluster 를 이루는 각 read 의 서열, Annotation, Source (종별) 등으로 sorting 된 결과를 보여 줄 수 있도록 하였다 (Fig. 1).

Blast 메뉴에서는 핵산, 아미노산, EST 3개의 소메뉴로 구성하였다. 현재까지 밝혀진 핵산 서열, 아미노산 서열, EST 서열 및 미토콘드리아 게놈 (Mitochondrial Genome) 등과 이번에 분석한 *Haliotis* EST 를 포함하여 총 7종류의 데이터베이스를 대상으로 BLAST 가 가능하도록 구축하였다. BLAST 는 query 및 데이터베이스의 관계가 허용하는 한 blastp, blastn, blastx, tblastn, tblastx 프로그램의 수행이 모두 가능하도록 구축하였다. 또한 Vector, *E. coli*, Repeat 서열을 따로 검색 할 수 있도록 구축하여, multi DB 메뉴를 만들어 라이브러리 확인 (삽입체 크기 측정) 등을 할 때 용이하도록 하였다 (Fig. 2).

Search 메뉴에서는 자체적으로 구축된 CGI 와 perl script 를 이용한 검색엔진을 설치하여 유전자 이름, accession number, 종 이름 등을 query 로 하여 데이터베이

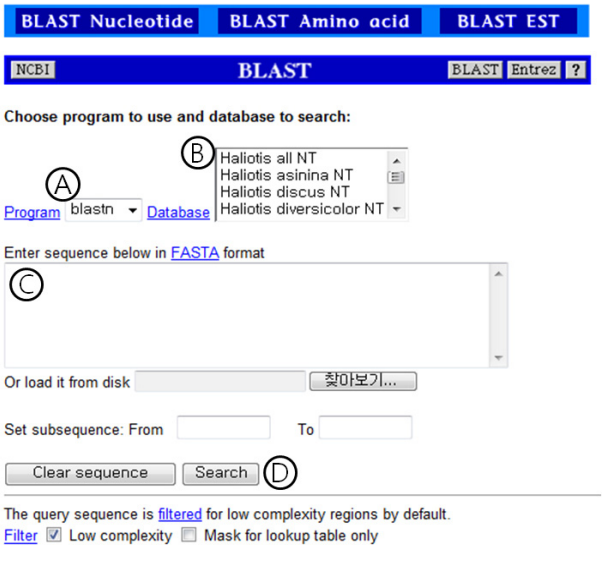

Fig. 2. The screen of searching the BLAST through constructed database. (**A**) Selection of blast program, (**B**) Selection of blast database, (**C**) Insertion of query sequences,(**D**) Start to search.

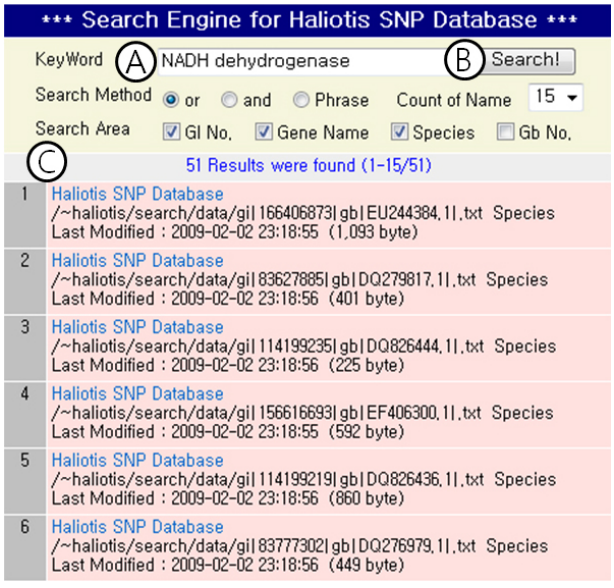

Fig. 3. Installation of search engine within constructed *Haliotis* database. (**A**) Insertion of keywords, (**B**) Start to search, (**C**) Results of searching.

스 내부의 모든 정보를 검색할 수 있도록 구축하였다 (Fig. 3). 또한 웹페이지 자체에 Primer3 엔진을 탑재하여 필요한 서열의 시발체를 데이터베이스 내부에서 제작 할 수 있도록 하였다. 2-sequence 메뉴에서는 BLAST engine 을 이용하여 두 개의 sequence 의 alignment 를 생성하는 local alignment 도 구축하였다.

요 약

- 본 웹 데이터베이스 서버의 구축을 통해 *Haliotis* 속간의 염기서열과 일치하는 서열을 자체 BLAST 를 통해 매우 빠른 속도로 추출 할 수 있었다.
- Repeat elements, *E. coli*, vector 등의 서열들과 동시에 BLAST를 시행할 수 있어 cDNA 또는 genomic DNA 라이브러리를 구축할 때 라이브러리의 오염, 삽입체의 길이 등의 상태를 쉽게 확인 할 수 있었다.
- Clustering Res. 인터페이스를 통해 SNPs 발굴이 용이하게 되었으며 자체 구축된 primer3 를 통해 실험용 시발체를 제작할 수 있게 되었다 (Evans *et al*. 2001).
- 이러한 SNP 데이터베이스 구축은 SNP 발굴 작업을 극대화 시킬 수 있어 차후 수행될 *Haliotis* 관련 분자육종 관련연구에 많은 도움이 될 것으로 기대된다.

사 사

본 연구는 국립수산과학원 (RP-2010-BT-020) 지원에 의해 수행되었습니다.

REFERENCES

Altschul S.F., Gish W., Miller W., Meyers E.W. & Lipman D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology,* **215**: 403-10.

Huang X. & Madan A. (1999) CAP3: A DNA sequence assembly program. *Genome Res,* **9**: 868-77.

Lee J.-S. & Min D.-K. (2002) A Catalogue of Molluscan Fauna in Korea. *Korean Journal of Malacology,* **18**: 93-217.

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. & Park H.-S. (2004) Construction of BLAST Server for Mollusks. *Korean Journal of Malacology,* **20**: 165-9.

Pertea G., Huang X., Liang F., Antonescu V., Sultana R., Karamycheva S., Lee Y., White J., Cheung F., Parvizi B., Tsai J. & Quackenbush J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics,* **19**: 651-2.

QI H.-G., LIU X. & ZHANG G.-F. (2008) Characterization of 12 single nucleotide polymorphisms (SNPs) in Pacific abalone, *Haliotis discus hannai*. *Molecular Ecology Resources,* **8**: 974-6.

Qi H., Liu X., Wang S. & Zhang G. (2009) Development of gene-associated intronic TR markers for the Pacific abalone *Haliotis discus hannai*. *Anim Genet,* **40**: 575.

Rice P., Longden I. & Bleasby A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics,* **16**: 276-7.

정평림, 박갑만, 정영헌, 용태순, 임경일 & 소진탁 (2000) 한국의 약용패류. *한국패류학회,* **16**: 55-60.